Big Data Privacy Preservation - A Survey

Craig T. Chosney
CSc 84030 Big Data Analytics
CUNY Graduate Center
New York City, USA
cchosney@cuny.gradcenter.edu

As individuals engage with social networks, the "internet of things," and outsourced cloud computing, they produce an exponentially growing amount of "big data" concerning themselves. There are many types of privacy attacks that can be leveraged on datasets to reveal sensitive information to adversarial actors. Because the big data being constantly generated is so telling, there is a great need for data privacy preservation in order to avoid such privacy attacks and the subsequent key privacy threats of surveillance, disclosure, discrimination, personal embracement and abuse that these privacy attacks permit.

Key methods of big data preservation include perturbation and anonymization. This survey will discuss randomization, data distribution, and differential privacy as methods of data perturbation and the classic data anonymization algorithms of k-anonymity, l-diversity, and t-closeness. Such privacy preservation methods consume resources, however, and as always when choosing methods of privacy preservation there are trade-offs between privacy and utility. We will then discuss evaluation metrics for these privacy preservation methods and some future research directions including new frameworks for the k-anonymity algorithm that result in effective big data privacy protection with less information loss and computing time and the "Data lake" concept for privacy preservation utilizing machine learning techniques. The structure of the survey is as follows:

1. ***Introduction to big data and privacy***
2. ***Data privacy attacks and threats***
3. ***Methods of data perturbation: randomization, data distribution, and differential privacy***
4. ***Methods of data anonymization: k-anonymity, l-diversity, and t-closeness***
5. ***Evaluation Metrics: data utility, robustness, complexity, and efficiency***
6. ***Future Research Directions***
7. ***Conclusion***

Keywords—Big Data, Privacy Protection, Data Perturbation, Data Anonymization

I. INTRODUCTION TO BIG DATA

As our lives become increasingly entangled with technology, the amount of "Big Data" being created is increasing exponentially. All of our internet connected devices, including our personal computers, tablets, smartphones, and household appliances are generating constant streams of data about things such as our shopping habits, location, the people we are around or with whom we interact, and personal metrics like health and medical history amongst others. This "Big Data" that we are constantly creating can be defined as "a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis" [2]. Big Data is further defined by five characteristic attributes: volume, variety, velocity, veracity, and value. These attributes are known colloquially as the "5 V's." Simply put, volume is the amount of data generated and processed, variety refers to the type and quality of knowledge that allows processing of the data, velocity refers to the rate at which data is collected and processed, veracity is the term that refers to the quality of the information collected, and value refers to the data's utility and significance [8].

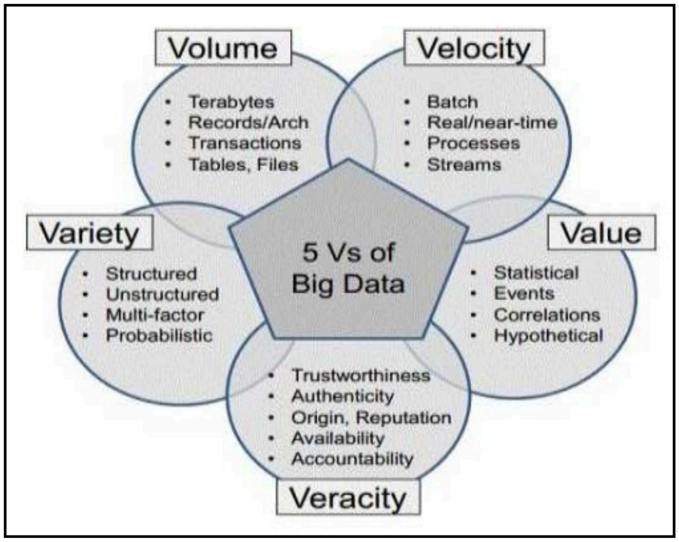


Figure 1. The 5 V's of Big Data [2]

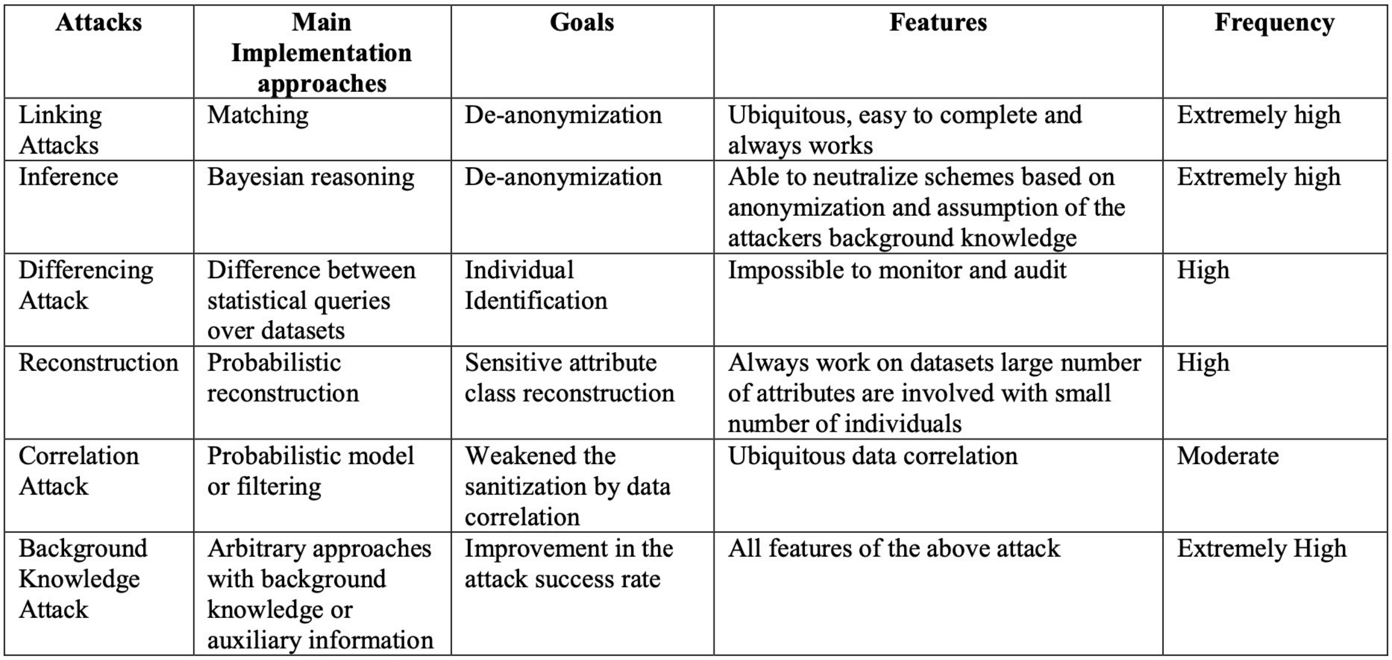
The "value" of this data is of specific concern to us because it is this value from which big data derives its importance, and in turn, this value can create immeasurable benefits for society. However, when dealing with such large amounts of data there is always potential for abuse or malfeasance. This potential for abuse is why data privacy is so important. It is now important to distinguish between data security and data privacy. The two are closely related, but data privacy is of particular concern to this survey. Data security is built upon the pillars of confidentiality, integrity, and availability. These pillars concern the safe sharing, exchange, storage, and handling of sensitive data [2]. Wang, et al. [11] write "Specifically, security is the practice of preventing unauthorized and malicious access, use, disruption and modification of information." In contrast, these authors construct privacy as relating to the idea that information about individuals and groups that is inherently special or sensitive should not be advertised to others; thus, privacy allows these individuals and groups to "express themselves selectively" [11]. If, as individuals, we are constantly leaving a trail of data "breadcrumbs" about ourselves, data security relates to say, those breadcrumbs being dropped in an effective path without being eaten by birds, being washed away by a storm, or being taken and replaced by pebbles that look like breadcrumbs. Privacy relates to only the breadcrumb trails we consent to being followed as being the ones that are followed.

Data itself is neutral, but the constituents with access to data can have either benevolent or malicious intentions. We can consider three main actors in privacy related scenarios involving data. The first actor is the "data owner," who owns the data. The second is the "data holder" who collects that data from the data owners, and the third is the "data consumer" who carries out data analytics. Any of these three main actors (owner, holder, or consumer) can be either trusted, semi-trusted, or untrusted [9]. Privacy protection can be implemented by any of the main actors. For example, if a data owner is untrusted, a trusted data holder can implement privacy protection methods or vice versa. Conventionally, however, the greatest onus for implementing privacy protection methods is upon the data holder. The data holder is the social networking application, website, mobile app, e-commerce site, bank, hospital, or business that holds data after it has been generated and collected. The primary responsibility of ensuring the privacy of the user's data falls to these data holders [7].

II. Data Privacy Attacks and Threats

A. *Privacy Attacks*

Malicious actors or adversaries can leverage attacks on insecure data in attempts to breach this data and obtain sensitive information. Six types of such privacy attacks are delineated: linking attacks, inference attacks, differencing attacks, reconstruction attacks, correlation attacks, and background knowledge attacks. The first, linking attacks, involve an adversary matching anonymized data in one dataset with non-anonymized data in a different dataset. Once the datasets are "linked," this external data is combined with the anonymized data to identify individuals in that dataset, thus causing a privacy breach. This type of attack is ubiquitous and easy to implement in this era of big data, rich data collection, and substantial data mining. The second type of attack, inference, involves the adversary using techniques such as association rules and Bayesian reasoning to access some public information and potentially infer sensitive information of real value from this information with high confidence. Thus, the adversary is inferring a fact that should be preserved at a high privacy level from information pieced together at a lesser privacy level. The third attack type, differencing, involves an adversary determining sensitive information by using the differences between statistical queries or aggregate queries over datasets. For example, an adversary could query a cancer database, "How many people in the database have cancer?" and "How many people, not named Alice, in the database have cancer?" The adversary can then determine the cancer status of Alice by comparing the difference between the two queries. The fourth type of privacy attack is a reconstruction attack. In this type of attack an adversary reconstructs a probabilistic version of the original dataset from the adversary's observations and auxiliary information, thus determining if an individual has a high probability of being in an attribute-group. The fifth type of attack, correlation, is when an adversary with background knowledge or auxiliary/correlational information violates privacy by using this information to obtain privacy information. The correlation attack type is enabled by real-world data often exhibiting strong correlations. For example,



Attacks	Main Implementation approaches	Goals	Features	Frequency
Linking Attacks	Matching	De-anonymization	Ubiquitous, easy to complete and always works	Extremely high
Inference	Bayesian reasoning	De-anonymization	Able to neutralize schemes based on anonymization and assumption of the attackers background knowledge	Extremely high
Differencing Attack	Difference between statistical queries over datasets	Individual Identification	Impossible to monitor and audit	High
Reconstruction	Probabilistic reconstruction	Sensitive attribute class reconstruction	Always work on datasets large number of attributes are involved with small number of individuals	High
Correlation Attack	Probabilistic model or filtering	Weakened the sanitization by data correlation	Ubiquitous data correlation	Moderate
Background Knowledge Attack	Arbitrary approaches with background knowledge or auxiliary information	Improvement in the attack success rate	All features of the above attack	Extremely High

Figure 2. Data Attack Summary Table [2]

more information than expected may be disclosed by interactive relationships in social networking data and spatiotemporal correlations in location trajectories. The sixth and final type of data privacy attack is the background knowledge attack. This is an all encompassing type of attack that has features of all the prior discussed attack types: linking, inference, differencing, reconstruction, and correlation. All of these attack types more or less involve an adversary's background knowledge, so they can all be qualified as this type of attack—one in which an adversary leverages background knowledge, or rather an association between private information and other information—to improve the attacker's success rate [11]. Figure 2 gives further summary of these attack's types and features.

B. *Privacy Threats*

When sensitive information is obtained by adversarial actors, there are four main key threats that these malicious actors can leverage against individuals and groups using the personal data of these individuals and groups. The threats to data privacy are delineated as surveillance, disclosure, discrimination, and personal embracement and abuse. The first, surveillance, involves continuous and sustained monitoring of an individual or group's online activity. The second, disclosure, is when sensitive information that has been released from a first party to a second party with consent is then released from the second party to a third party without consent of the first party. The third, discrimination, describes the bias or inequality which can happen when some private information about an individual or group is disclosed. The fourth and final major key threat to data privacy, personal embracement and abuse, is another result of sensitive information being disclosed to bad actors [7]. Wang, et al. [11] delineate three scenarios regarding data privacy infringements that help us to elucidate the key threats to data privacy. In the first scenario, a user "Alice" utilizes a location-based service (LBS) to find the restaurants, subway stations, service businesses, etc. nearest to her. Alice's smartphone communicates her position to the LBS provider, and her smartphone provides her with information about such nearby points of interest. Additionally, however, the LBS collects and curates Alice's current and historical positions, allowing it to predict Alice's next location through location-prediction mechanisms, in what it perceives as an extension of its customer service. This location communication between Alice and the LBS provider impacts Alice's privacy and may be perceived as untrustworthy. This location data may subsequently be used to infer Alice's home, place of work, personal points of interest, hobbies, and social relationships, with this inference exemplifying multiple threats to her privacy, primarily surveillance. In the second scenario, a "smart meter," much like the five million Consolidated Edison has installed in the New York City area to read electric meters remotely, allows a customer "Amy" to enjoy superior power services by instantaneously communicating her electric consumption and its cost to the electric company. However, a potential adversarial third party could use the smart meter data to infer the appliances Amy owns and her daily routine by interpreting her household power usage. This invasion of privacy primarily exemplifies the threat of disclosure, in which the data communicated consentingly between Amy and her electric company is now communicated without consent to a third party. The third and final data privacy infringement scenario involves the mining of a communication network to acquire plentiful and accurate information about entities using the network. This mining involves the leaking of sensitive information such as the frequent calls, emails, and messages between "Ada" and "Bob," revealing an intimate relationship between the two. Also leaked are records of the communication between "Tom" and his cancer doctor that provide hints about his health. Furthermore, the leak of data in a network of familial hereditary diseases may result in an adversary inferring the potential diseases faced by "Jack" on the basis of the hereditary diseases of his other relatives. Thus, the data mining and subsequent leak of this communication network represent the threats of discrimination and personal embracement and abuse for Ada, Bob, Tom, and Jack, who have had their sensitive data compromised. It is hoped that by elaborating upon data privacy attack types and subsequent threats to the privacy of individuals and groups we have established the challenges facing privacy preservation and established the need for the privacy preservation methods to be discussed in the next section.

III. Data Perturbation

Now that we have established a need for privacy preservation methods for big data, a discussion of such methods will occur. This survey will discuss the classic and most commonly used methods utilized for data privacy preservation although myriad methods exist. We will first discuss three methods of data perturbation: randomization, data distribution, and differential privacy. Perturbation refers simply to a deviation of a system or process, so these three perturbation techniques are methods of "disturbing" or deviating data.

A. *Randomization*

This first technique of data perturbation, randomization, involves distorting data by adding noise or noisy signals to the original data. This distortion is generally done by probability distribution, multiplying disturbance, random projections of random rotation techniques, or random rotating techniques [8]. This noise is combined into a dataset in order to hide the actual values of the individual records. To protect the individual value of the records the added noise in the data is significantly large. Although the data now contains noise, it is still possible to obtain the data distributions from the random records within this "noisy" data, and the "noisy" data is sufficient for various data mining tasks. The aggregate behavior of the data distribution can be reconstructed by subtracting the noise from the "noisy" data. Consider A is our plain data and B is our added noise. The addition of A and B creates new distribution C, representing our "noisy" data. Given that B is known, we can estimate the distribution by subtracting B from C. $C = A + B$; then $A = C — B$ [6]. Benefits of the randomization technique for privacy preservation are that this technique does not require knowledge of other records in the data. It is a technique that can be applied during data collection and the time of pre-processing, and there is no anonymization overhead for this technique. A detriment to randomization is that applying this technique to large datasets is not possible due to time complexity and data utility concerns [7].

B. Data Distribution

As the second technique of data perturbation, data distribution involves distributing data across many sites, either horizontally or vertically. By decentralizing the data either horizontally or vertically over different sites, this technique protects data by making it less vulnerable to a single attack or privacy breach [7]. The difference between horizontal and vertical distribution is that in a horizontal distribution individual records are distributed across multiple entities and in a vertical distribution attributes of the records are distributed across multiple entities. A benefit to this perturbation technique is that it can improve site performance when regular transactions involving certain views of data occur while still maintaining data availability and privacy. A detriment to this technique is that when data is distributed across sites under the custodianship of different parties, the security of its system depends upon the trust between its participants [6]. The following figure (3) demonstrates the schema of horizontal and vertical data distribution. As we see in the pictorial representation of horizontal data distribution the records of our two individuals are distributed across multiple sites—the sites being represented pictorially by a horizontally oriented box or rectangle. When data is distributed vertically, attributes of the records, in this case four per actor, are distributed across multiple entities, again represented by a vertically oriented box or rectangle.

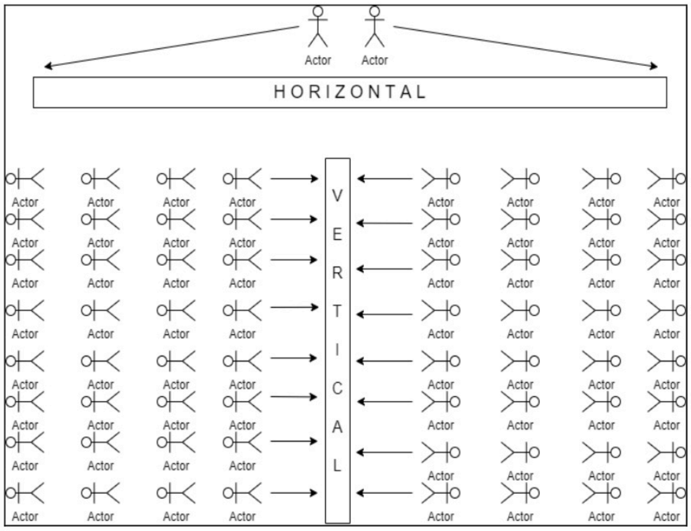


Figure 3. Horizontal and Vertical Distribution [8]

C. Differential Privacy

Differential privacy is a final data perturbation model proposed by Dwork in 2006 with the aim of masking the differences in computation results of a function *f* on neighboring datasets which differ on at most one data item [9]. This method perturbs the raw records of individuals at random in order to make the raw record unidentifiable to an adversary regardless of that adversary's background knowledge. It is important to note that differential privacy is not an algorithm or a mechanism; rather, it is a definition, and it can be implemented utilizing varying algorithms and mechanisms [11]. It will benefit us to present one conventional definition of differential privacy elaborated by Biswas, Fole, Khare, and Agrawal [1]: Suppose D_1 and D_2 are adjacent datasets, A is a privacy mechanism and if A satisfies ε-differential privacy for any output $A(D_1) \rightarrow R$, $S \in R$, then

$$Pr[A(D_1) \in S] \leq e^{\epsilon}[A(D_2) \in S]$$

(1)

Let D_1 and D_2 be adjacent datasets; the global sensitivity, the measure of change in other records due to modification of another record can be defined as follows:

$$GS = max_{D_1,D_2} ||f(D_1) - f(D_2)||_1$$

(2)

where f is the query function, ||.||1 is the L1-norm. The differential privacy can be realized by adding noise 'e' to the output.

$$A(D) = f(D) + e$$

(3)

Because it involves the perturbation of data by the addition of noise, differential privacy can be considered a more specific method of data randomization. It is afforded its own section in this survey due to it being a well researched and published privacy preservation method. The included figure (4) delineating many different differential privacy mechanisms is indicative of the work that has been done to develop this perturbation model. This figure evidences varying implementations, levels of privacy, computational overhead, data utilities, and time complexities for differing mechanisms of the method.

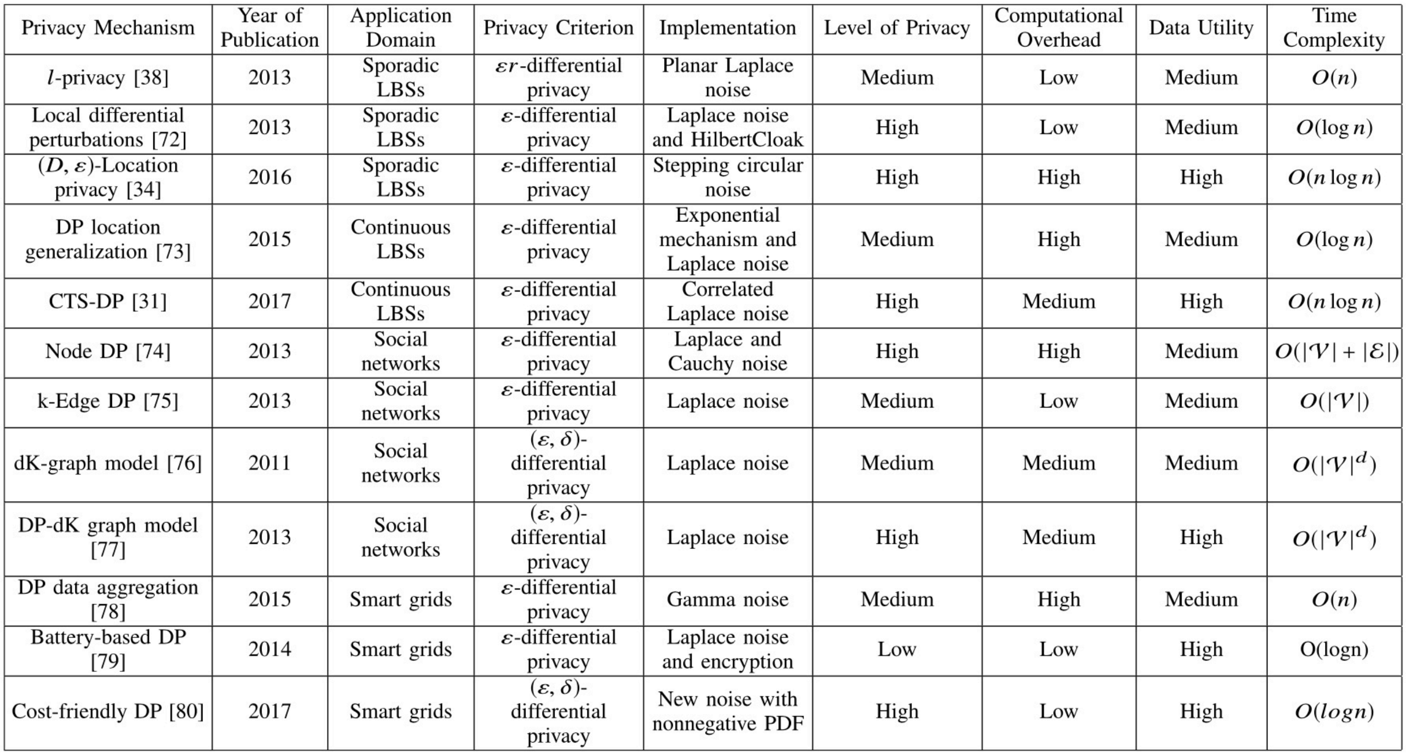
There are two key advantages to differential privacy. The first advantage is that differential privacy defines the maximum background knowledge. A potential adversary knows all the information about individuals except one sensitive record, thus still guaranteeing the privacy of the individuals. The second advantage is that due to being built upon a probability statistical model, differential privacy is able to quantitatively analyze the risk of privacy disclosure [11]. The main disadvantage of differential privacy is its lack of uniformity in implementation mechanisms. This lack of uniformity results in varying strengths and weaknesses for each mechanism.

IV. Data Anonymization

What follows now is a discussion of four techniques for the data anonymization method of data privacy preservation. The classic algorithmic techniques for data anonymization are k-anonymity, l-diversity, and t-closeness. Anonymization techniques work by replacing the values for an attribute with less particular yet consistent values, thereby "anonymizing" the data [2].

A. K-Anonymity

Introduced in 1998 by Samarati and Sweeney, the k-anonymity model was developed to deter the indirect



Privacy Mechanism	Year of Publication	Application Domain	Privacy Criterion	Implementation	Level of Privacy	Computational Overhead	Data Utility	Time Complexity
l-privacy [38]	2013	Sporadic LBSs	εr-differential privacy	Planar Laplace noise	Medium	Low	Medium	$O(n)$
Local differential perturbations [72]	2013	Sporadic LBSs	ε-differential privacy	Laplace noise and HilbertCloak	High	Low	Medium	$O(\log n)$
(D, ε)-Location privacy [34]	2016	Sporadic LBSs	ε-differential privacy	Stepping circular noise	High	High	High	$O(n \log n)$
DP location generalization [73]	2015	Continuous LBSs	ε-differential privacy	Exponential mechanism and Laplace noise	Medium	High	Medium	$O(\log n)$
CTS-DP [31]	2017	Continuous LBSs	ε-differential privacy	Correlated Laplace noise	High	Medium	High	$O(n \log n)$
Node DP [74]	2013	Social networks	ε-differential privacy	Laplace and Cauchy noise	High	High	Medium	$O(\|\mathcal{V}\| + \|\mathcal{E}\|)$
k-Edge DP [75]	2013	Social networks	ε-differential privacy	Laplace noise	Medium	Low	Medium	$O(\|\mathcal{V}\|)$
dK-graph model [76]	2011	Social networks	(ε, δ)-differential privacy	Laplace noise	Medium	Medium	Medium	$O(\|\mathcal{V}\|^d)$
DP-dK graph model [77]	2013	Social networks	(ε, δ)-differential privacy	Laplace noise	High	Medium	High	$O(\|\mathcal{V}\|^d)$
DP data aggregation [78]	2015	Smart grids	ε-differential privacy	Gamma noise	Medium	High	Medium	$O(n)$
Battery-based DP [79]	2014	Smart grids	ε-differential privacy	Laplace noise and encryption	Low	Low	High	O(logn)
Cost-friendly DP [80]	2017	Smart grids	(ε, δ)-differential privacy	New noise with nonnegative PDF	High	Low	High	$O(logn)$

Figure 4. Summary of Differential Privacy Mechanisms [11]

identification of records from public databases. The name "k-anonymity" comes from the quality of a table of data to be called "k-anonymous" if every record in the table is similar to at least $k-1$ other records with respect to every set of quasi-identifiers [6]. The initial algorithm divides all the attributes of a dataset into four categories: (1) personally identifiable information, (2) quasi-identifier (QID), (3) sensitive attribute, and (4) non-sensitive attribute. The k-anonymity algorithm checks that if one record in the dataset has some value QID then at least $k-1$ other records also have the same QID values. In other words, at least k records in the data set must have the same QID value, and the produced table is called k-anonymous with a probability of $1/k$ to identify any individual record [3]. Thus, sensitive information is hidden into $k-1$ dummies with the same QID, making it difficult for adversaries to identify the actual record [9]. This method successfully protects agains linkage attacks, but can still leave data susceptible to background knowledge attacks [11].

B. L-Diversity

To resolve the k-anonymity model's susceptibility to certain attacks, l-diversity was proposed by Machanavajjhala, et al. as an extension of the k-anonymity model used to secure individual identity from disclosure. The l-diversity algorithm builds upon k-anonymity by requiring each group of QIDs that contains at least one representative and distinct sensitive attributes that have roughly equal proportion to protect against background knowledge attacks. When $k=1$, l-diversity automatically satisfies k-anonymity [11]. Thus, there must be l well represented values for the sensitive attributes in each equivalence class [7]. A problem with l-diversity arises when it is found that different values may belong to the same category, only guaranteeing the diversity of sensitive attribute values and allowing attribute linkage attacks when the overall distribution of a sensitive attribute is skewed [6].

C. T-Closeness

To prevent the problem of attribute linkage attacks when the overall distribution of a special attribute is skewed, Li, Li, and Venkatasubramanian proposed the t-closeness model. T-closeness requires the distribution of a special attribute in any QID group to be close to the distribution attribute in the overall dataset [9]. By this method, the distance between the two distributions is smaller than a threshold t. The t-closeness algorithm uses the earth mover's distance (EMD) function to measure the closeness between distributions of sensitive values and requires closeness to be within t [6]. A drawback to t-closeness is that because it requires the distribution of special attribute values to be the same in all QID groups, its enforcement greatly degrades the data utility [9]. The progressive development of k-anonymity, l-diversity, and t-closeness demonstrates that no single approach can guarantee complete privacy; however, the weaknesses of one approach can be partially or properly remedied by another [2].

V. Evaluation Metrics

Now that some classic methods of data privacy preservation have been surveyed, it is important to be able to evaluate these methods in order to compare and contrast their efficacy. The four concepts of data utility, robustness, complexity, and efficiency are used as metrics for this evaluation.

A. *Data Utility*

Our first concept, data utility, refers to the extent that the original informational content of the data can be preserved when applying a privacy preservation method to the data. Data utility is a very important concept because data utility and data privacy have an inverse relationship. This relationship dictates that as data utility increases, or that as more of our original data's content is preserved by our privacy preservation method, data privacy decreases [1]. Figure 5 illustrates this relationship by calculating data privacy as the number of attributes that have been anonymized and by using Information Gain to calculate data utility. Interestingly in this relationship between utility and privacy, privacy is an individual concept and utility is an aggregate concept. A good privacy preservation method protects the privacy for each individual while maintaining the maximum utility of the data without compromising the underlying privacy constraints [6]. Data Utility is the most important metric of evaluation that computer scientists and engineers use to develop privacy preservation methods, as without adequate data utility we would diminish the informational content of our data to a point where our data would lose its value.

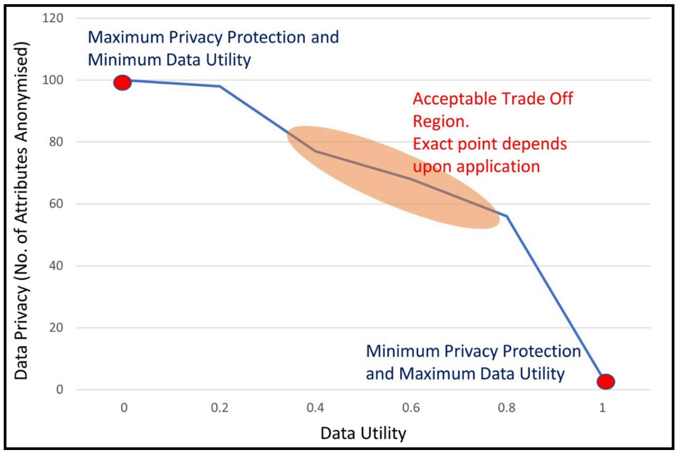


Figure 5. Data Utility vs. Data Privacy [1]

B. *Robustness*

Secondly, the robustness of a privacy preservation method refers to the degree of protection it provides against attacks. A robust privacy preservation method is strong against different types of attacks and linkages. The robustness of any proposed privacy preservation method in practice depends on different reasons [6].

C. *Complexity*

Complexity is an important third dimension used to evaluate any privacy preservation method, and it refers to the computational or algorithmic complexity of a privacy preservation method. An optimal privacy preservation has time and space complexity in $O(n^2)$, where n is the number of data records. A good complexity allows a privacy preservation method to execute with good performance in terms of all the resources implied by the method [6].

D. *Efficiency*

The final concept, efficiency, refers to the computational efficiency of executing a privacy preservation method. Due to the tremendous amount of big data being produced, efficiency is an important challenge for our privacy preservation methods. The three main factors for measuring the efficiency of a privacy method are CPU time, space requirements (related to memory usage and the required storage capacity), and communication requirements. Ensuring efficiency for our methods allows the quick processing of streaming and near-real-time data and the accurate analysis of structured and unstructured data [6]. To clarify the distinction between efficiency and complexity, we distinguish that efficiency relates to computational hardware and complexity relates to software.

VI. Future Research Directions

Although current privacy preservation methods are feasible and effective, it is always possible to improve such methods in terms of our evaluation metrics of data utility, robustness, complexity, and efficiency. For such improvement, immense research effort is still required. A portion of this research effort is committed to improving the k-anonymity algorithm to overcome its shortcomings. Various "improved" k-anonymity methods have been proposed, such as scalable k-anonymization SKA [3], improved scalable k-anonymization (ImSKA) [4], and (a,k)-anonymity [10]. The first approach, SKA, outperforms existing approaches in terms of information loss and running time [3]. The next, ImSKA, extends SKA for the "variety" aspect of data, working well with unstructured big data [4]. The third approach, (a,k)-anonymity improves the efficiency of computing and reduces computing time by using the distributed MapReduce to classify and group data for massive datasets [10]. Although these three exemplary algorithmic extensions of k-anonymity are developed and published, they are presented as "future research directions" to show that computer scientists and engineers are always pushing privacy preservation algorithms forward in terms of our evaluation metrics. Rather, the "future research direction" is one of moving algorithms towards greater data utility, robustness, complexity, and efficiency.

The utilization of machine learning techniques also opens a new frontier for developing new and more appropriate solutions to privacy problems. One novel privacy preservation model utilizing machine learning is suggested by Rao, Krishna, and Kumar [7] based on the "Data lake" concept. "Data lake" is a repository to hold data from diverse sources in their raw format. The proposed model, which uses advanced computing techniques to ensure very fast processing while implementing its privacy preserving algorithms is shown in figure 6.

This figure shows that in this novel privacy preservation model data will be ingested from a variety of sources using Apache Flume and SQOOP. Following the ingestion of data into the "Data lake," an intelligent algorithm based on machine learning will be applied to identify sensitive attributes dynamically. This intelligent algorithm will be rigorously trained with existing datasets with known sensitive attributes leading to deep learning techniques and improving accuracy. Within the "Data lake" the data remains in its native form, either structured or unstructured. When it is time for processing, it can be transformed into HIVE tables, and a

Hadoop MapReduce job using machine learning can be executed on the data to classify the sensitive attributes. The data can then be vertically distributed to separate the sensitive attributes from the rest of the data, and tokenization can be applied to map the vertically distributed data. The data without any sensitive attributes can be published for data analytics [7]. Although currently theoretical, the “Data lake” model may greatly improve data utility and efficiency while ensuring privacy for any sensitive attributes of a dataset.

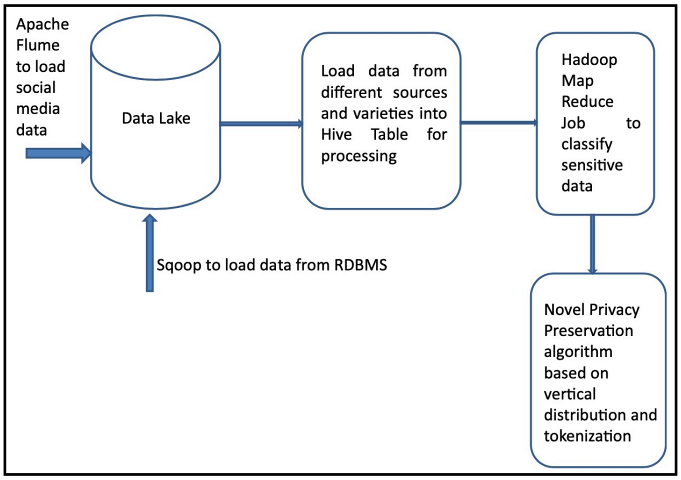

Apache Flume to load social media data
Data Lake
Load data from different sources and varieties into Hive Table for processing
Hadoop Map Reduce Job to classify sensitive data
Sqoop to load data from RDBMS
Novel Privacy Preservation algorithm based on vertical distribution and tokenization

Figure 6. “Data lake” Novel Privacy Preservation Model [7]

VII. Conclusion

Through this survey of data privacy preservation methods we have been introduced to big data and privacy, delineated types of data privacy attacks and threats, established the classic methods of data perturbation and anonymization, conceptualized the metrics used to evaluate these methods, and discussed future directions for the research of privacy preservation and its methods. This field of research is massive, and much more on the topic can be written and discussed. It is my hope that this survey provides a clear and focused introduction to this vastly important area of research and study.

References

1. Biswas, Sreemoyee, Anuja Fole, Nilay Khare, and Pragati Agrawal. “Enhancing Correlated Big Data Privacy Using Differential Privacy and Machine Learning.” Journal of Big Data 10, No. 1 (2023): 30–23.
2. Devi, A. Sangeerani, and A. Chinnasamy. “Privacy Preservation of Sensitive Data in Big Data Analytics - A Survey.” In 2021 10th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON), 01–05. IEEE (2021).
3. Mehta, Brijesh B, and Udai Pratap Rao. “Privacy Preserving Big Data Publishing: a Scalable k-Anonymization Approach Using MapReduce.” IET Software 11, No. 5 (2017): 271–276.
4. Mehta, Brijesh, Udai Pratap Rao, Ruchika Gupta, and Mauro Conti. “Towards Privacy Preserving Unstructured Big Data Publishing.” Journal of Intelligent & Fuzzy Systems 36, No. 4 (2019): 3471–3482.
5. N, Gayathri Devi, and Manikandan K. “Improved Perturbation Technique Privacy-Preserving Rotation-Based Condensation Algorithm for Privacy Preserving in Big Data Stream Using Internet of Things.” Transactions on Emerging Telecommunications Technologies 31, No. 12 (2020).
6. Pramanik, M. Ileas, Raymond Y. K. Lau, Md Sakir Hossain, Md Mizanur Rahoman, Sumon Kumar Debnath, Md Golam Rashed, and Md Zasim Uddin. “Privacy Preserving Big Data Analytics: A Critical Analysis of State-of-the-art.” Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery 11, No. 1 (2021).
7. Ram Mohan Rao, P., Murali Krishna, S. & Siva Kumar, “Privacy Preservation Techniques in Big Data Analytics: A Survey." Journal of Big Data 5, 33 (2018).
8. Takle, Lalita. “A Survey on Data Privacy Threats and Preservation Techniques.” International Journal of Advanced Research in Computer Science 11, No. 2 (2020): 57–63.
9. Tran, Hong-Yen, and Jiankun Hu. “Privacy-Preserving Big Data Analytics - A Comprehensive Survey.” Journal of Parallel and Distributed Computing 134 (2019): 207–218.
10. Wang, Jie, Hongtao Li, Feng Guo, Wenyin Zhang, and Yifeng Cui. “D2D Big Data Privacy-Preserving Framework Based on (a, K)-Anonymity Model.” Mathematical Problems in Engineering 2019 (2019): 1–11.
11. Wang, Tao, Zhigao Zheng, Mubashir Husain Rehmani, Shihong Yao, and Zheng Huo. “Privacy Preservation in Big Data From the Communication Perspective - A Survey.” IEEE Communications Surveys and Tutorials 21, No. 1 (2019): 753–778.