Information Theoretic-based Privacy Risk Evaluation for Data Anonymization

Anonymous Author(s)*

ABSTRACT

Data anonymization aims to enable data publishing without compromising the individuals' privacy. The re-identification and sensitive information inference risks of a dataset are important factors in the decision-making process for the techniques and the parameters of the anonymization process. If correctly assessed, measuring the reidentification and inference risks can help optimize the balance between protection and utility of the dataset, as too aggressive anonymization can render the data useless, while publishing data with a high risk of de-anonymization is troublesome. In this paper a new information theoretic-based privacy metric (ITPR) for assessing both the re-identification risk and sensitive information inference risk of datasets is proposed. We compare the proposed metric with existing information theoretic metrics and their ability to assess risk for various cases of dataset characteristics. We showed that ITPR is the only metric that can effectively quantify both re-identification and sensitive information inference risks. We provide several experiements to illustrate the effectiveness of ITPR.

CCS CONCEPTS

• **Computer systems organization → Embedded systems**.

KEYWORDS

Privacy-preserving Data Publishing, Re-identification Risk, Inference Risk, Anonymity Metric, Information Theory.

ACM Reference Format:

Anonymous Author(s). 2019. Information Theoretic-based Privacy Risk Evaluation for Data Anonymization. In *Proceedings of ACSAC '19*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Digital services today rely on the availability and processing of often sensitive data, while data analytics brings about important benefits both for the service providers and the individuals. However, the protection of sensitive data during this extensive processing has become a growing concern for Internet users, as the widespread use of IoT services, mobile devices and location services, lead to constant monitoring and vast amounts of sensitive data being gathered and stored.

There is a trade-off between data protection, data availability and data utility that needs to be tackled, in order to achieve services which ensure privacy protection and produce usable results at the same time. This need becomes more imminent today, both due to the individuals' sensitization to data protection matters, which can lead to their lack of cooperation if they do not trust the service, and the data protection legislation being put in action, which holds the data handler responsible for any data breaches [14].

Another important issue concerns privacy-preserving data publishing [7]. Anonymization techniques are being used to sanitize data sets prior to data publishing, so that similar data processing results are produced, while preserving data privacy. However, oftentimes the data characteristics are such that re-identification is easier than expected. De-anonymization techniques have advanced significantly and the abundance of external auxiliary information concerning individuals can lead to the de-anonymization of seemingly safe datasets [12].

In this paper, we study and compare existing information theoretic-based privacy metrics to show that, in several cases, they cannot correctly assess both the re-identification risk and the sensitive information inference risk. We then propose, to the best of our knowledge, the first information theoretic-based privacy metric that can correctly assess both the re-identification and the sensitive information inference risks and perform several experiments to show its effectiveness.

Outline. The paper is organized as follows: In Section 2, we describe the problem statement and the motivation for this work. In Section 3 we present the background regarding anonymisation risk metrics, as well as the related work. The proposed information theoretic-based privacy risk metric (ITPR) is presented in Section 4 and in Section 5 we describe the produced experimental results, also compared to the related work. The conclusions of this work are presented in Section 6.

2 PROBLEM STATEMENT

Privacy-preserving data publishing [7] enables the utilization of the data collected by organizations and service providers, without compromising the dataset participants' privacy. The goal is to release microdata from the dataset for processing, while protecting private information. To achieve that, the initial dataset is anonymized. Metrics such as k-anonymity [18], l-diversity [11], and t-closeness [9] have been proposed to assess the quality of the anonymization process, as well as the disclosure risk, setting thresholds for the anonymized dataset characteristics to be considered safe.

It is common that the data publisher does not know beforehand the entities that will access the released data and the operations that will be performed on the released data. Therefore, the data publisher needs a method to assess the characteristics of the dataset and the resulting disclosure risk, in order to make informed choices on what to include in the released dataset.

ACSAC '19, December 9–13, 2019, Puerto Rico, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Tailoring the anonymization techniques and parameter setting to the dataset characteristics is important, so that the maximum possible data utility is preserved, while the data remains protected. If the anonymization process affects the data more than needed, then data utility diminishes, while lighter anonymization can lead to unintended data disclosure. Therefore, the challenge in the anonymization process is achieving a balance between protection and utility [10].

While anonymized data utility measurement depends heavily on the type of analysis to be performed over the data, the measurement of the privacy protection level depends mainly on two factors: (1) the re-identification risk that measures the risk for any individual to be identified in the anonymized dataset, and (2) the inference risk that measures the risk for any sensitive information in the dataset to be linked to a specific individual. The level of privacy protection that can be ensured depends considerably on the distribution of the values of the different attributes that compose the dataset to be anonymized. So, depending on the considered dataset, we often end up with balancing the values of the re-identification and sensitive information inference risks to find an acceptable trade-off between data utility and data privacy. However, in the literature, these risks are often measured using different metrics e.g., k-anonymity [18] for the re-identification risk and l-diversity [11] and t-closeness [9] for the inference risk. The different risk measurement methods and in some case behaviors of the existing metrics often make combining the value of two risks to find the best balance between minimizing re-identification and sensitive information inference risks hard.

In this paper, we aim to provide a new information theoretic-based privacy metric that is able to assess both the re-identification and the sensitive information inference risks. We show that information theoretic-based risk metrics proposed in the literature, are mainly average values. As a result, they do not assess effectively the contribution of a single record to the risk value [2]. As individual risk values can fluctuate greatly (as we will illustrate in Section 3.2), average values are not suitable to represent both the re-identification and the sensitive information inference risks.

3 BACKGROUND AND RELATED WORK

In this section we present the relevant background regarding information theoretic anonymity metrics, as well as related work on re-identification risk metrics. The typical setting of an anonymization process involves data, contained in one or more tables. Each table row represents a data record and the columns of the table are the record attributes. These attributes are categorized into three main categories:

- Identifiers, which directly identify an individual, such as name, social security number, etc.
- Quasi-identifiers (key attributes), which can be used in combination to identify an individual, such as age, gender, occupation, postal code, etc.
- Sensitive attributes, which contain sensitive information concerning an individual, such as health data, financial data, etc.

In order to protect individuals from the disclosure of their sensitive data, anonymization techniques can be employed, such as data generalization, suppression, and perturbation, as well as noise addition techniques.

De-anonymization attacks on the released data can lead to both identity and attribute disclosure. In the case of identity disclosure, an individual is directly linked to a dataset record and the sensitive information it contains. In the case of attribute disclosure, the individual is associated with an attribute value, but not a specific record.

Anonymity is defined as the state of being not identifiable within a set of subjects, called the anonymity set [13]. Statistical disclosure control (SDC) methods propose minimum requirements for each attribute in the dataset. To conform with k-anonymity [18], it is required that all quasi-identifier groups contain at least k records. As the value of the quasi-identifier can be the same in the whole group, k-anonymity does not protect against homogeneity attacks [11]. For example, let us consider a 4-anonymity table composed of two attributes: Age and Disease. If we suppose that all individuals having the value '4*' for Age in the considered table are suffering from an 'HIV', then, to perform an homogeneity attack, an adversary only needs to know that an individual present in table is between 40 and 49 years old to know his/her disease. To address this issue l-diversity [11] has been proposed, as it requires each group to contain at least l distinct values for each quasi-identifier. Fulfilling l-diversity still fails to protect against skewness attacks [9], which allow sensitive information disclosure when their distribution in the quasi-identifier group is significantly different from the corresponding distribution over the entire dataset. To deal with this issue t-closeness [9] requires that the distance between the distributions of sensitive attributes in the quasi-identifier group and the whole dataset remains under t. However, these methods, although they provide an objective way of assessing and enforcing privacy in the datasets, do not constitute a uniform risk-assessment metric.

3.1 The limitations of k-anonymity, l-diversity, and t-closeness models

As described in the previous section, k-anonymity was proposed to mitigate identity disclosure, l-diversity was proposed to mitigate homogeneity attacks and t-closeness was proposed to prevent skewness attacks. However, when we analyze carefully the three models, we realize that they are not useful for computing the effective inference risk (disclosure risk) of sensitive attributes' information.

To illustrate, let us take the example of the two anonymized datasets in Tables 1 and 2. The anonymized dataset in Table 1b satisfies 3-anonymity, 3-diversity, and 0-closeness, while the anonymized dataset in Table 2b satisfies 5-anonymity, 3-diversity, and 0-closeness. So, if we limit our analysis to the computed three values for k-anonymity, l-diversity, and t-closeness in the two cases, the overall level of ensured privacy is better in the second dataset since $k_2 > k_1$ (i.e., the re-identification risk is lower in the second anonymized dataset than the first one) , $l_1 = l_2$, and $t_1 = t_2$. However, if we look carefully at the distribution of the values of the attribute Disease in the two anonymized datasets, we realize that if an adversary knows that an individual is in Group 1, 2 or 3, he has bigger probability (0,60) for inferring the individual's disease in the second anonymized dataset than in the first one (0,33). This example

(a) Initial dataset

#	Zip Code	Age	Disease
1	35510	21	Asthma
2	35591	42	HIV
3	35593	47	Asthma
4	35210	38	Diabetes
5	35273	32	HIV
6	35517	20	Diabetes
7	35599	49	Diabetes
8	35262	33	Asthma
9	35511	26	HIV

(b) Anonymized dataset

#	Zip Code	Age	Disease
Group 1	3551*	2*	Asthma
			Diabetes
			HIV
Group 2	3559*	4*	HIV
			Asthma
			Diabetes
Group 3	352*	3*	Diabetes
			HIV
			Asthma

Table 1: First anonymized dataset ($k_1 = 3, l_1 = 3,$ and $t_1 = 0$)

(a) Initial dataset

#	Zip Code	Age	Disease
1	35510	21	Asthma
2	35591	42	HIV
3	35593	47	Asthma
4	35210	38	Diabetes
5	35273	32	HIV
6	35517	20	Diabetes
7	35599	49	Diabetes
8	35262	33	Asthma
9	35511	26	HIV
10	35212	39	Diabetes
11	35281	32	Diabetes
12	35596	41	Diabetes
13	35592	46	Diabetes
14	35515	23	Diabetes
15	35511	26	Diabetes

(b) Anonymized dataset

#	Zip Code	Age	Disease
Group 1	3551*	2*	Asthma
			Diabetes
			Diabetes
			Diabetes
			HIV
Group 2	3559*	4*	HIV
			Asthma
			Diabetes
			Diabetes
			Diabetes
Group 3	352*	3*	Diabetes
			HIV
			Asthma
			Diabetes
			Diabetes

Table 2: Second anonymized dataset ($k_2 = 5, l_2 = 3,$ and $t_2 = 0$)

shows that the combination of the k-anonymity, l-diversity, and t-closeness models does not measure the effective disclosure risk but instead the accomplishment of the anonymization process.

To evaluate the performance of an anonymization method and to be able to compare the effectiveness amongst different methods, we need to define a common evaluation framework that can measure effectively both the re-identification risk and the sensitive attributes inference risk.

3.2 Information theoretic risk metrics

Information theory can be applied to the data protection context to evaluate the amount of information carried by a set of data and the possibility that disclosing this data leads to identity or attribute leakage. Information theoretic risk metrics provide the ability to be applied to different anonymity systems [5]. Information can be represented as a variable that can contain different values and an information theoretic risk metric aims at measuring the amount of information leaked from a dataset.

Entropy is a key concept of information theory [15] that quantifies the uncertainty of a random variable. Uncertainty enhances privacy, as it hinders an adversary from effectively estimating attribute values [19]. In the following paragraphs we provide definitions for the key concepts in entropy-based anonymity metrics. We consider X and Y be two random variables, corresponding to two attributes in a dataset.

Similar to uncertainty, information theory can be used to produce metrics that quantify information loss or gain for an adversary.

Entropy. The entropy of a discrete random variable X is:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{1}$$

where $p(x)$ is the probability of occurrence for value $x \in X$.

Conditional Entropy. The conditional entropy of a discrete random variable X, given a discrete random variable Y is:

$$\begin{aligned} H(X|Y) &= -\sum_{y} p(y) \sum_{x} p(x|y) \log p(x|y) \\ &= -\sum_{x,y} p(x,y) \log p(x|y) \end{aligned} \tag{2}$$

where $p(x|y)$ is the conditional probability of occurrence for value $x \in X$, given the occurrence of $y \in Y$. Conditional entropy expresses how much information is needed to describe X, knowing the value of Y. The maximum of conditional entropy $H(X|Y)$ is the entropy $H(X)$ [19]. Therefore, normalised conditional entropy is computed by the following fraction:

$$\frac{H(X|Y)}{H(X)} \tag{3}$$

Joint Entropy. The joint entropy of two discrete random variables X and Y is:

$$H(X,Y) = -\sum_{x \in X} p(x,y) \log p(x,y) \tag{4}$$

where $p(x,y)$ is the joint probability of occurrence for the value pair (x,y).

3.3 Related Work

In this section we present the related work on information theoretic-based privacy risk metrics. After presenting the metrics and their characteristics, we present some example cases which show that these metrics are unable to assess correctly the re-identification and inference risks of a dataset in certain cases.

3.3.1 Discrimination Rate. Discrimination Rate (DR) is an attribute-centric privacy metric, which aims to measure the degree to which attributes are able to refine an anonymity set and to measure their identification capability [16, 17]. For this purpose, attributes are represented as discrete random variables. Considering two discrete random variables X and Y, DR is used to measure the identification capacity of attribute Y over the set of X.

$$DR_X(Y) = 1 - \frac{H(X|Y)}{H(X)} \tag{5}$$

DR is bounded on [0,1], where 1 means that an identifier reduces the anonymity set to a single individual.

3.3.2 Mutual Information. Mutual Information (MI) has been proposed as a metric for the disclosure risk and the utility of a dataset [3, 4, 8]. The mutual information of two discrete random variables X and Y represents the average amount of knowledge about X gained by knowing Y, or alternatively the amount of shared information between X and Y. Therefore, mutual information is an information gain metric. Intuitively, if X and Y are independent, their mutual information is equal to zero. Mutual Information is computed as the difference between entropy $H(X)$ and conditional entropy $H(X|Y)$:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= H(X) + H(Y) + H(X,Y) \end{aligned} \tag{6}$$

3.3.3 Conditional Privacy. Conditional privacy (CP) is a privacy metric proposed in [1] for quantifying the fraction of privacy of a sensitive attribute X which is lost by revealing another attribute Y. Conditional privacy can be seen as a mutual information normalization and is formalized as following:

$$Priv_{CP} = 1 - 2^{-I(X;Y)} \tag{7}$$

where $I(X;Y)$ represents the mutual information of X and Y.

3.3.4 Maximum Information Leakage. Maximum Information Leakage (MIL) is a modification of the mutual information metric to consider only a single instance of a quasi-identifier attribute Y. It measures the maximum amount of information about a sensitive attribute X that can be learned by observing a single instance of Y [6].

$$Priv_{MIL} = \max_{y \in Y} I(X; Y = y) \tag{8}$$

where $I(X;Y)$ represents the mutual information of X and Y.

3.3.5 Entropy l-Diversity. Entropy l-Diversity (ELD) is proposed as an instantiation of the l-diversity principle [11]. It states that a table is entropy l-diverse for a sensitive attribute X if the following condition holds for all quasi-identifiers groups q:

$$-\sum_{x \in X} p(q,x) \cdot log(p(q,x)) \geq log(l)$$

where $p(q,x)$ is the fraction of records in q that have the value x from the attribute X. We note that ELD, as it is proposed in [11], does not allow to measure the inference risk of a sensitive attribute, but is used as an entropy-based condition that must be satisfied to achieve the l-diversity property for a sensitive attribute. We adapt ELD to allow inference risk measurement as following. Given a sensitive attribute X and the set of quasi-identifiers groups $\mathcal{Q}$, the inference risk of the attribute X can be measured as:

$$ELD(X) = 2^{\min_{q \in Q} \sum_{x \in X} p(q,x) \cdot log(p(q,x))}$$

Examples - Problem Cases. Previously mentioned information theoretic-based privacy metrics, do not always succeed in assessing the disclosure risk, since they do not effectively assess the contribution of individual records to the risk value. We present in Table 3 some examples of datasets, where the considered metrics fail to express one or both of the enhanced re-identification and inference risk of the data. As shown in Table 3, we consider dataset composed of eight records containing values of three attributes: Identifier, Age and Disease. Five cases are considered for the age attribute:

- case 1: All records contain unique attribute values
- case 2: All records contain the same attribute value
- case 3: Only one record contains a different attribute value
- case 4: Only two records contain a different attribute value
- case 5: Half of the records contain one and half of the records contain another attribute value

For the Disease attribute, we consider the first three cases we considered for the attribute Age.

For cases 1 and 3 of the attribute Age, we expect the highest value for the re-identification risk, since some values of the attribute uniquely identify an individual. For case 2, we expect the lowest value for the re-identification risk since the knowledge of the value of the attribute gives no additional information about the considered

Identifier	Cases of **Age** Attribute					Cases of **Disease** Attribute		
	case 1	case 2	case 3	case 4	case 5	case 1	case 2	case 3
#1	30	30	30	30	30	Diabetes	Diabetes	Diabetes
#2	62	30	30	30	30	Epilepsy	Diabetes	Diabetes
#3	37	30	30	30	30	Asthma	Epilepsy	Diabetes
#4	21	30	47	47	30	Allergies	Depression	Diabetes
#5	19	30	30	30	47	Depression	HIV	Diabetes
#6	47	30	30	47	47	HIV	Heart Disease	HIV
#7	71	30	30	30	47	Heart Disease	Cancer	Asthma
#8	73	30	30	30	47	Cancer	Allergies	Allergies

Table 3: Example of datasets

Privacy Risk Type	Considered Dataset	Results				
		DR	MI	CP	MIL	ELD
Re-identification risk	Identifier + case 1 of Age	1.0	3.0	0,875	3.0	1.0
	Identifier + case 2 of Age	0.0	0.0	0.0	0.0	0.125
	Identifier + case 3 of Age	0.18	0.54	0.31	3.0	1.0
	Identifier + case 4 of Age	0.27	0.81	0.43	2.75	0.5
	Identifier + case 5 of Age	0.33	1.0	0.5	2.0	0.25
Inference risk	Identifier + case 5 of Age + case 1 of Disease	0.33	1.0	0.5	2.0	0.25
	Identifier + case 5 of Age + case 2 of Disease	0.36	1.0	0.5	1.0	0.35
	Identifier + case 5 of Age + case 3 of Disease	0.35	0.54	0.31	1.0	1.0

Table 4: Information theoretic-based privacy metrics analysis results

identifier. For case 4, we expect a lower value for re-identification risk than the one assigned to case 3 and a higher value than the one assigned to case 5.

For the inference risk, we expect the risk assigned to case 1 (Identifier + case 5 of Age + case 1 of Disease) to be lower that the ones assigned to cases 2 (Identifier + case 5 of Age + case 2 of Disease) and 3 (Identifier + case 5 of Age + case 3 of Disease). Additionally, we expect the risk to be assigned to case 2 to be lower than the one assigned to case 3.

In Table 4, we examine the behaviour of the information theoretic-based privacy metrics previously presented in Sections 3.3.1, 3.3.2, 3.3.3 and 3.3.4, depending on the distribution of values in the considered attributes cases. According to the obtained results, DR, MI, and MIL correctly express the re-identification risk when case 1 and case 2 of the attribute Age are considered (rows 2 and 3 of Table 4), However, all three metrics fail to reflect the level of re-identification risks that is represented by cases 3, 4, and 5 of the attribute Age (rows 4, 5, and 6 of Table 4). Out of these previously mentioned three cases, the DR, MI, and MIL metrics output lower values to the case that represents a higher re-identification risk (case 3 of the attribute Age) and higher values to the case that represents a lower re-identification risk (case 5 of attribute Age). As for the CP metric, it correctly reflects the re-identification risk of the dataset instance in which case 2 of the attribute Age is considered, and fails to correctly reflect the re-identification risk for other cases. The HLD metric seems to correctly reflect both the re-identification and the inference risks for all considered cases. However, we show in Section 5 that HLD fails to measure correctly the inference risk caused by the difference between the distribution of the values of the sensitive attribute in the entire table and their distribution in the different quasi-identifiers' groups.

When it comes to measuring the inference risk, all considered metrics successfully reflect the re-identification risk of the dataset instance in which case 5 of the attribute Age and case 1 of the attribute Disease are considered (row 6 of Table 4), and fails to correctly reflect the re-identification risk for other cases (rows 7 and 8 of Table 4).

4 THE NEW INFORMATION THEORETIC-BASED PRIVACY RISK METRIC

To address the lack of ability of existing information theoretic based privacy metrics to (1) effectively assess the contribution of individual records of a dataset to the re-identification risk value and (2) correctly quantify the inference risk that stems from the correlation between a quasi-identifier attribute (e.g., Age) and a sensitive attribute (e.g., Disease), we propose a new information theoretic-based privacy risk metric (ITPR). The value of ITPR can effectively express, on one side, the probability of the attacker to refine the anonymity set and re-identify a dataset participant, based on the knowledge of an (quasi-identifier) attribute, on the other side, the probability of an adversary to refine the anonymity set and link an identity to a value of a sensitive attribute.

To develop the formula for the ITPR metric, we follow a similar logic as in the Discrimination Rate and Mutual Information metrics, still relying on information theory and entropy calculations. However, instead of using the average value of the attribute values'

#	Privacy Risk Type	Considered Dataset	ITPR Results
1	Re-identification risk	Identifier + case 1 of Age	1.0
2		Identifier + case 2 of Age	0.0
3		Identifier + case 3 of Age	1.0
4		Identifier + case 4 of Age	0.83
5		Identifier + case 5 of Age	0.33
6		Identifier + case 2 of Age + case 1 of Zip Code (Table 6)	0.6
7		Identifier + case 2 of Age + case 1 of Zip Code (Table 6)	0.75
8	Inference risk	Identifier + case 5 of Age + case 1 of Disease	0.33
9		Identifier + case 5 of Age + case 2 of Disease	0.45
10		Identifier + case 5 of Age + case 3 of Disease	1.0

Table 5: ITPR results

entropy, we take the maximum value of entropy amongst attribute values.

To compute the remaining identification information of attribute X, given attribute Y, we compute $H(X) - H(X|Y)$. We then divide with H(X) to normalise the computed value, resulting in the following representation:

$$\max_{y \in \Omega_Y} \left(\left\{ 1 - \frac{H(X|Y=y)}{H(X)} \right\} \right) \tag{9}$$

where Ω_Y is the sample space of the discrete random variable Y.

Using this equation, the results produced depend on the number of distinct values for Y, so for example for the case of two distinct values ($|\Omega_Y| = 2$, e.g. case 5 of attribute Age in Table 3) the produced results span between 0.5 and 1.0, in the case of three distinct values ($|\Omega_Y| = 3$), ITPR values span between 0.66 and 1.0 and so on. In order to counteract this behaviour, we introduce the number of distinct values in attribute Y as a parameter in the ITPR metric, leading to the following definition.

Definition 4.1. (Simple ITPR) Given two attributes X and Y of a dataset, the simple ITPR of attribute Y relative to attribute X, quantifies the capacity of attribute Y to refine the set of values of attribute X and is measured as following:

$$ITPR_X(Y) = \max_{y \in \Omega_Y} \left(\left\{ 1 - \frac{|\Omega_Y| * H(X|Y=y)}{H(X)} \right\} \right) \tag{10}$$

where $|\Omega_Y|$ denotes the number of different values of Y.

Using formula (10), the returned results for ITPR are normalised between values 0 and 1.0 and effectively represent, as we will illustrate late in the paper, (1) the re-identification risk of an (identifier) attribute X, given a (quasi-identifier) attribute Y, and (2) the inference risk caused by a sensitive attribute X when a (quasi-identifier) attribute Y is published.

Definition 4.1 can be generalized to define the combined ITPR which quantifies the ITPR measure related to the combination of the values of several attributes to perform re-identification or/and inference attacks.

Definition 4.2. (Combined ITPR) Given a set of attributes $X, Y_1, Y_2, \cdots, Y_n$ of a dataset $\mathcal{D}$, we denote $\mathcal{T}$ the set of $\langle Y_1, Y_2, \cdots, Y_n \rangle$ distinct tuples in $\mathcal{D}$. The combined ITPR of attributes $Y_1, Y_2, \cdots, Y_n$ relative to attribute X, quantifies the capacity of attributes $Y_1, Y_2, \cdots, Y_n$ to refine the set of values of attribute X and is computed according to the following Formula:

$$ITPR_X(Y_1, Y_2, \cdots, Y_n) = \max_{\langle y_1, y_2, \cdots, y_n \rangle \in \mathcal{T}} \left(\left\{ 1 - \frac{|\mathcal{T}| * H(X|Y_1 = y_1, Y_2 = y_2, \cdots, Y_n = y_n)}{H(X)} \right\} \right)$$

case 1	case 2
35000	35000
35000	35000
35000	35000
35510	35510
35510	35510
35510	35510
35510	35200
35510	35200

Table 6: Cases of the attribute Zip Code

Table 5 illustrates the expressiveness of the ITPR metric when using the same examples as used in Table 4.

As shown in Table 5, compared to the existing information theoretic-based metrics we analysed and reported in Table 4, ITPR correctly quantifies the re-identification risk for all the considered cases (rows 1 to 5 of Table 5). Rows 6 and 7 of Table 5 shows that ITPR can effectively measure the re-identification when two (quasi-identifier) attributes are combined. Moreover, the results show that the ITPR metric correctly measure the inference risk represented by the attribute Disease (Rows 8 to 10 of Table 5).

The behaviour of our ITPR privacy metric regarding the distribution of considered attributes values is more thoroughly tested and illustrated in the next Section.

5 EXPERIMENTAL RESULTS

In this section we provide the experimental results of the ITPR metric, compared to the Discrimination Rate, Mutual Information, Conditional Privacy, Maximum Information Leakage, and Entropy L-Diversity metrics. More specifically, we analyze and compare the behaviours of the considered metrics with the behaviour of our

ITPR metric for assessing the re-identification and sensitive information inference risks (Sections 5.1 and 5.2). Finally, we evaluate the computation effectiveness of our proposed metric (Section 5.3).

5.1 Re-identification risk assessment

To assess the behaviour of the functions of the considered metrics, we first calculated the metric values for a dataset of 10000 records, containing two distinct Y attribute values (for example $y_1 = Male$ and $y_2 = Female$). We denote by ϵ the maximum difference between the number of occurrences of the values of the attribute Y:

$$\epsilon = \max_{y_1, y_2 \in \Omega_Y} (|occ(y_1) - occ(y_2)|) \quad (11)$$

where Ω_Y denotes the set of distinct values of Y and $occ()$ is a function that returns the number of occurrences of value. Obliviously, the more the value of ϵ is bigger, the more the number of occurrences of y_1 (resp. y_2) is smaller which means that the number of individuals having y_1 (resp. y_2) will be smaller in the considered dataset, which should result in a high re-identification risk.

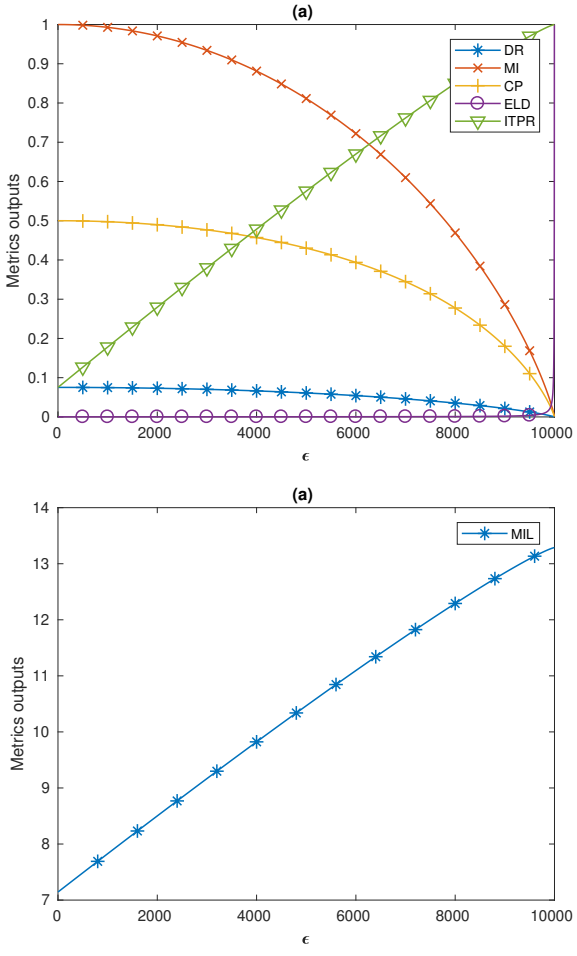

Figure 1: ITPR comparison with DR, MI, CP, ELD, and MIL for re-identification risk quantification (2 different values for Y)

The results are illustrated in Figure 1. One can observe that the value of ITPR begins from 1 for the case of $\epsilon = 9998$ (e.g., $|y_1| = 1, |y_2| = 9999$) and diminishes smoothly while the values ϵ decreases (i.e.,the sizes of the two value groups ($|Y_1|, |Y_2|$) move closer to each other), converging to a very small value ($7 * 10^{-2}$) when the value of $\epsilon = 0$ (i.e, the two group sizes are equal $|Y_1| = |Y_2| = 5000$).

In the DR case, the metric stays below 0.1 for this dataset, failing to accurately express the re-identification risk for the different cases. In the MI and CP cases, the metric output appears to increase as the number of records of each attribute value move towards being equal, failing also to express that the re-identification risk is higher when a smaller number of records contains one of the attribute values and the majority contains the other. The ELD output increases extremely slowly as the value of ϵ increases. In Figure 1(b), we observe that the MIL metric output increases as the value of ϵ increases which represents a correct behavior regarding the re-identification risk represented in the different considered cases. Unfortunately, the MIL metric suffers from two drawbacks: (1) the wide range of output values (e.g., between 7 and 13) makes the interpretation of the output of the metric difficult, and (2) as illustrated in Figure 3b, the MIL metric does not correctly express the inference risk represented by a sensitive attribute. Note that, for all studied metrics, the same behaviors can be observed when several values are considered for the attribute Y, as described in Figure 2.

As the results indicate, ITPR is able to effectively express:

- the lower existence of risk when the attribute values are distributed equally amongst the dataset records,
- the gradual enhancement of risk, as the number of records containing a certain value decreases, and
- the higher risk value when a certain value appears only in a small number of records in the dataset.

5.2 Sensitive information inference risk assessment

We compared the ability of the considered metrics to assess the inference risk represented by the publication of a sensitive attribute. For this, we consider two attributes Age and Disease in a dataset composed of 10000 records. For simplicity, we will consider the Age to be composed of a 5 different values uniformly distributed over the 10000 dataset records and the attribute Disease to be composed of instances of 10 different values. Since the inference risk depends mainly on the distribution of the different values of the considered sensitive attribute, we analyze the output of the considered metrics regarding the difference between the most used and the less used values of the attribute Disease that we denote λ.

$$\lambda = \max_{x \in \Omega_A, y_1, y2 \in \Omega_D} (|occ(y_1|Age = x) - occ(y_2|Age = x)|) \quad (12)$$

where Ω_A and Ω_D denote the set of distinct values of attributes Age and Disease respectively and $occ(y|Age = x)$ denotes the number of occurrences of y in the dataset when the value of $Age = x$. Note that the inference risk is expected to increase according to the increase of λ since the more the λ value is higher the more the number of occurrences of a specific value of a sensitive attribute is higher in an anonymity class.

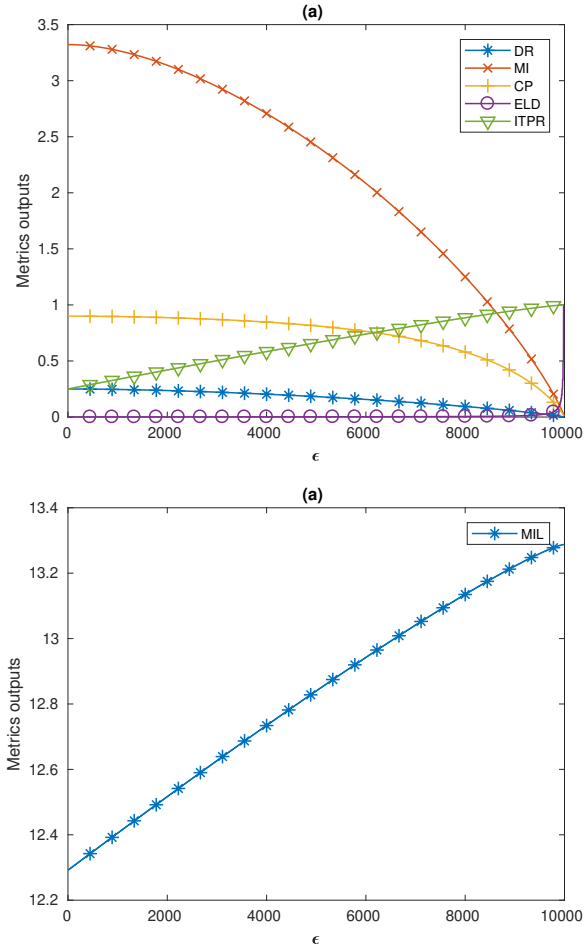

Figure 2: ITPR comparison with DR, MI, CP, ELD, and MIL for re-identification risk quantification (10 different values for Y)

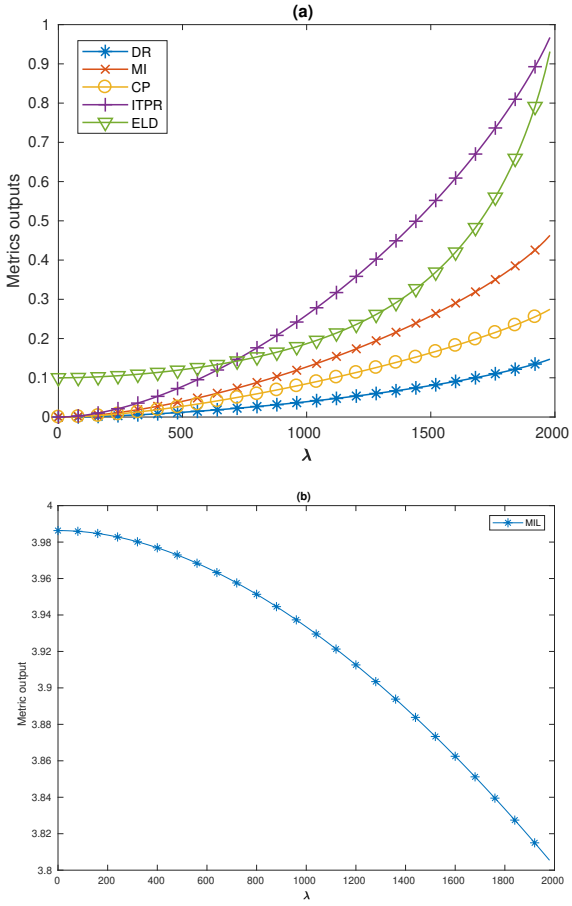

Figure 3: ITPR comparison with DR, MI, MIL, and ELD for inference risk output regarding λ

Figure 3 illustrates the obtained results. Figure 3a shows that the outputs of ITPR, DR, MI, ELD, and CP increase according to the growth of the value of λ, which is consistent with the behavior we previously expected. For the three metrics DR, MI, and CP, the particular output ranges make the interpretation of the risk difficult, since without knowing the output of the metric in the worst case it is hard to evaluate the severity of the output value. The ITPR and ELD metrics do not suffer from this limitation since their output always range between 0 and 1. As for the MIL metric, Figure 3b shows that it decreases according to the growth of the value of λ, failing to effectively assess the inference risk represented by the Disease attribute.

Li et al. in [9] showed that the inference risk does not depend only on the distribution of the values of the considered sensitive attribute in the quasi-identifiers' groups. The variation between the global distribution of the values of the sensitive attribute in the considered dataset and the local distribution of the values of the sensitive attribute in the quasi-identifiers' groups can drastically impact the inference risk. To illustrate, let us consider a dataset where the only sensitive attribute is Disease and it is composed of 10^8 records. Furthermore, suppose that each record in the dataset is associated with a different individual and that only 1000 records contain 'VIH' as a value for the attribute Disease. This means that anyone in the considered dataset has 10^{-3}% possibility of having 'VIH'. Now let us suppose that one of the quasi-identifiers' groups created by the used anonymization mechanism contains 5 records out of 100 that have 'VIH' as a value for the attribute Disease. Clearly, this presents a serious privacy risk, because anyone in the considered quasi-identifiers' group would be considered to have 5% possibility of having 'VIH', compared to the 10^{-3}% of the overall population. Thus, a correct inference risk measurement associated with a sensitive attribute X should take into consideration the variation ϑ_X between the global distribution of the values of X in the considered dataset and the local distribution of the values X in the quasi-identifiers' groups. In fact, the bigger the previous variation is, the greater the inference risk associated with the considered sensitive attribute must be. The variation ϑ_X is formalized as following:

$$\vartheta_X = \max_{q \in Q} (H(X) - H(X, q)) \quad (13)$$

where Q denotes the set of quasi-identifiers' groups, $H(X)$ denotes the entropy of X in the hole dataset, and $H(X, q)$ denotes the entropy of X in the quasi-identifier group q.

We compare the ability of the considered metrics for assessing the inference risk represented by the sensitive attribute *Disease* regarding the variation $\vartheta_{Disease}$ between the global distribution of its values in the considered dataset and the local distribution of its values in the quasi-identifiers' groups. Note that the inference risk of a sensitive attribute X is expected to increase according to the increase of the value of the variation ϑ_X. The obtained results are illustrated in Figure 4. We can observe that the ELD metric does not take into consideration the variation $\vartheta_{Disease}$ since its output is constant in the function of $\vartheta_{Disease}$. It experimentally proves that the ELD metric does not measure correctly the inference risk of the attribute Disease.

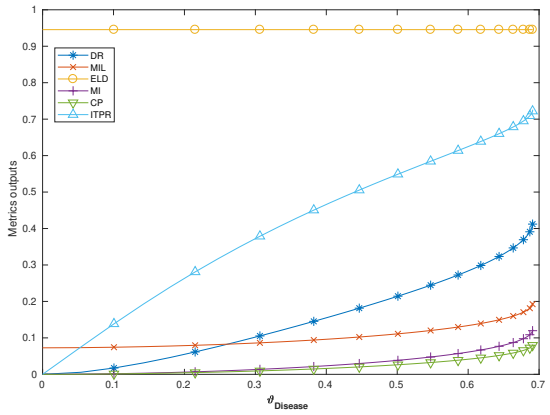

Figure 4: ITPR comparison with DR, MI, MIL, and ELD for inference risk output regarding ϑ

Moreover, we studied the behavior of the ITPR for different values of $\epsilon(0 - 1200)$ and $\vartheta(0 - 0.9)$, knowing that in this experimentation we considered a dataset with 10000 records composed of 10 different values for the attribute Age and 20 different values for the attribute Disease. The result is depicted in Figure 3.

It shows that the ITPR metric output increases smoothly while the value of ϵ or the value of ϑ increases which represents the expected behavior of a correct inference risk assessment metric. As for the MIL, MI, and CP metric, their outputs increase extremely slowly as the value of $\vartheta_{Disease}$ increases. For example, if we suppose that $\vartheta_{Disease} = 0$ represents the case we described above in which anyone in the database has 10^{-3}% possibility of having VIH, the $\vartheta_{Disease} = 0.4$ can represent the case in which anyone in a specific quasi-identifier group has 12% possibility of having VIH. However, when we examine the variations of the outputs of MIL, MI, CP between $\vartheta_{Disease} = 0$ and $\vartheta_{Disease} = 0.4$, they increase only from 0.07 to 0.09, from 0 to 0.02, and from 0 to 0.01. Finally, this experiment shows that our proposed ITPR metric has the best

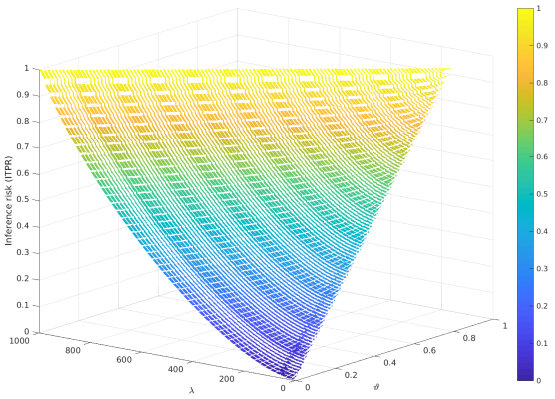

Figure 5: Inference Risk assessment using ITPR

behavior compared to considered metrics regarding the increase of the variation $\vartheta_{Disease}$.

5.3 ITPR computation effectiveness evaluation

Anonymization processes are often dealing with a large amount of data. As a result, the computation effectiveness of such a metric should be evaluated. For this, we consider a table composed of three attributes: Identifier, Age, and Disease. The attribute Age contains 120 different values while the attribute Disease contains 100 different values. We develop a Spark [20] based implementation of our ITPR metric that we evaluate on a Spark cluster of 4 nodes with 100 workers with 1 core and 1 GB per worker. Figure 6 shows the time needed for the computation of the ITPR metric regarding the number of rows in the considered table.

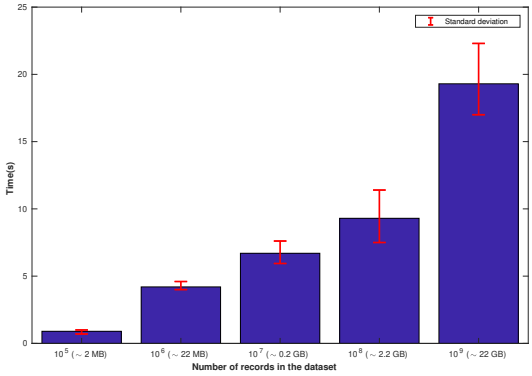

Figure 6: ITPR computation time per number of records

The previous figure illustrates the scalability and the efficiency of our ITPR implementation. Only 19 seconds are required to compute the output of the ITPR metric on a dataset of 10^9 records. Moreover, the ITPR computation time increases only by 1 order of magnitude (from 1 to 19 seconds) when increasing the size of database by 4 order of magnitudes (from 10^5 to 10^9 records).

Privacy Metrics	Re-identification risk	Inference Risk
Discrimination Rate (DR)	✗	(✓)
Mutual Information (MI)	✗	(✓)
Conditional Privacy (CP)	✗	(✓)
Maximum Information Leakage (MIL)	(✓)	✗
Entropy L-Diversity (ELD)	(✓)	✗
ITPR	✓	✓

(✓): Correct assessment but hard interpretation of the risk.

Table 7: Benefits of the ITPR compared to existing metrics

5.4 Discussion

The proposed ITPR metric can be used in both unprocessed and anonymized datasets. In the case of raw datasets, ITPR can assist the data owners with making decisions on which attributes of the dataset are more sensitive (i.e., the ones that have high inference risks) and which ones should be included in the anonymized dataset and with what anonymization parameters.

As with all information theoretic risk metrics, defining the threshold values on what constitutes low, moderate or high risk, remains an issue. The selection of such thresholds always depends on the characteristics of the dataset and the objectives of the data owner, however, in the case of ITPR, selecting thresholds for the risk values appears to be more straightforward, as the ITPR value increases gradually as the number of records containing a certain value decreases.

Another interesting use of ITPR can be the quality control and cleaning of datasets, as datasets which contain errors in their records, or contain a few outlier values will produce high values of re-identification risk, hence allowing data owners to clean errors in their datasets or decide to suppress outlier values to facilitate the successful dataset anonymization process.

6 CONCLUSIONS

In this paper the ITPR information theoretic-based metric is proposed, for assessing both the re-identification and the inference risk within datasets. This metric aims at effectively representing the contribution of individual records of a dataset to the re-identification and inference risk values. To achieve that, ITPR takes into account the maximum value of entropy amongst the dataset attribute values. To facilitate the comparison of risk values amongst different anonymization processes and between different datasets, the ITPR value is normalized and bounded between 0 and 1. The experimental results show that ITPR succeeds in expressing both the re-identification and the inference risk. The comparison with existing Information theoretic-based privacy metrics (Table 7) shows that ITPR is the sole metric that can effectively assess both the re-identification and the inference risk.

REFERENCES

[1] Dakshi Agrawal and Charu C Aggarwal. 2001. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 247–255.

[2] Michele Bezzi. 2010. An information theoretic approach for privacy metrics. *Trans. Data Privacy* 3, 3 (2010), 199–215.

[3] Michele Bezzi, Sabrina De Capitani di Vimercati, Sara Foresti, Giovanni Livraga, Pierangela Samarati, and Roberto Sassi. 2012. Modeling and preventing inferences from sensitive value distributions in data release 1. *Journal of Computer Security* 20, 4 (2012), 393–436.

[4] Thomas M Cover and Joy A Thomas. 2012. *Elements of information theory*. John Wiley & Sons.

[5] Claudia Diaz. 2006. Anonymity metrics revisited. In *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[6] Flávio du Pin Calmon and Nadia Fawaz. 2012. Privacy against statistical inference. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 1401–1408.

[7] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. 2010. Privacy-preserving Data Publishing: A Survey of Recent Developments. *ACM Comput. Surv.* 42, 4, Article 14 (June 2010), 53 pages. https://doi.org/10.1145/1749603.1749605

[8] Amaury Lendasse. 2017. Practical Estimation of Mutual Information on Non-Euclidean Spaces. In *Machine Learning and Knowledge Extraction: First IFIP TC 5, WG 8.4, 8.9, 12.9 International Cross-Domain Conference, CD-MAKE 2017, Reggio, Italy, August 29–September 1, 2017, Proceedings*, Vol. 10410. Springer, 123.

[9] N. Li, T. Li, and S. Venkatasubramanian. 2007. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *2007 IEEE 23rd International Conference on Data Engineering*. 106–115. https://doi.org/10.1109/ICDE.2007.367856

[10] Tiancheng Li and Ninghui Li. 2009. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 517–526.

[11] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. L-diversity: Privacy Beyond K-anonymity. *ACM Trans. Knowl. Discov. Data* 1, 1, Article 3 (March 2007). https://doi.org/10.1145/1217299.1217302

[12] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 111–125.

[13] Andreas Pfitzmann and Marit Köhntopp. 2001. Anonymity, unobservability, and pseudonymityâĂŤa proposal for terminology. In *Designing privacy enhancing technologies*. Springer, 1–9.

[14] General Data Protection Regulation. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. *Official Journal of the European Union (OJ)* 59, 1-88 (2016), 294.

[15] Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell system technical journal* 27, 3 (1948), 379–423.

[16] Louis Philippe Sondeck, Maryline Laurent, and Vincent Frey. 2017. Discrimination rate: an attribute-centric metric to measure privacy. *Annales des Télécommunications* 72, 11-12 (2017), 755–766. https://doi.org/10.1007/s12243-017-0581-8

[17] Louis Philippe Sondeck, Maryline Laurent, and Vincent Frey. 2017. The Semantic Discrimination Rate Metric for Privacy Measurements which Questions the Benefit of t-closeness over l-diversity. In *Proceedings of the 14th International Joint Conference on e-Business and Telecommunications (ICETE 2017) - Volume 4: SECRYPT, Madrid, Spain, July 24-26, 2017*. 285–294. https://doi.org/10.5220/0006418002850294

[18] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.

[19] Isabel Wagner and David Eckhoff. 2018. Technical privacy metrics: a systematic survey. *ACM Computing Surveys (CSUR)* 51, 3 (2018), 57.

[20] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. 2010. Spark: Cluster Computing with Working Sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud'10)*. USENIX Association, Berkeley, CA, USA, 10–10. http://dl.acm.org/citation.cfm?id=1863103.1863113