Universiteit Antwerpen

This item is the archived peer-reviewed author-version of:

The privacy issue of counterfactual explanations : explanation linkage attacks

Reference:
Goethals Sofie, Sörensen Kenneth, Martens David.- The privacy issue of counterfactual explanations : explanation linkage attacks
ACM Transactions on Intelligent Systems and Technology (TIST) - ISSN 2157-6904 - (2023), 3608482
Full text (Publisher's DOI): https://doi.org/10.1145/3608482
To cite this reference: https://hdl.handle.net/10067/1979440151162165141

uantwerpen.be
Institutional repository IRUA

The Privacy Issue of Counterfactual Explanations: Explanation Linkage Attacks

SOFIE GOETHALS, KENNETH SÖRENSEN, and DAVID MARTENS, University of Antwerp, Belgium

Black-box machine learning models are used in an increasing number of high-stakes domains, and this creates a growing need for Explainable AI (XAI). However, the use of XAI in machine learning introduces privacy risks, which currently remain largely unnoticed. Therefore, we explore the possibility of an *explanation linkage attack*, which can occur when deploying instance-based strategies to find counterfactual explanations. To counter such an attack, we propose k-anonymous counterfactual explanations and introduce *pureness* as a metric to evaluate the *validity* of these k-anonymous counterfactual explanations. Our results show that making the explanations, rather than the whole dataset, k-anonymous, is beneficial for the quality of the explanations.

CCS Concepts: • **Security and privacy** → *Social aspects of security and privacy*; **Privacy protections**; • **Computing methodologies** → *Machine learning*.

Additional Key Words and Phrases: Explainable AI, Counterfactual Explanations, Privacy, k-anonymity, Machine Learning

1 INTRODUCTION

Black-box models are used for decisions in more and more high-stakes domains such as finance, healthcare and justice, increasing the need to explain these decisions and to make sure that they are aligned with how we want the decisions to be made [21, 39]. As a result, the interest in interpretability methods for machine learning and the development of various techniques has soared [39]. At the moment, however, there is no consensus on which technique is best for which specific use case. Within the field of Explainable AI (XAI), we focus on a popular local explanation technique: counterfactual explanations [37, 58].

Counterfactual explanations, which are used to explain predictions of individual instances, are defined as the smallest change to the feature values of an instance that alters its prediction [37, 39]. *Factual instances* are the original instances that are explained and the *counterfactual instance* is the original instance with the updated values from the explanation. An example of a *factual instance*, *counterfactual instance* and *counterfactual explanation* for a credit scoring context can be seen in Figure 1. *Lisa* is the *factual instance* here, whose credit gets rejected. *Fiona*, a nearby instance in the training set whose credit was accepted, is selected as *counterfactual instance* by the algorithm and based on *Fiona*, *Lisa* receives a *counterfactual explanation* that states which features to change to receive a positive credit decision. These explanations can serve multiple objectives: they can be used for model debugging by data scientists or model experts, to justify decisions to end users or provide actionable recourse, to detect bias in the model, to increase social acceptance, to comply with GDPR, etc. [1, 36, 39].

Authors' address: Sofie Goethals, sofie.goethals@uantwerpen.be; Kenneth Sörensen; David Martens, University of Antwerp, Prinsstraat 13, Antwerp, Belgium, 2000.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2157-6904/2023/7-ART $15.00
https://doi.org/10.1145/3608482

Just Accepted

Factual instance						
Identifier	**Quasi-Identifiers**			**Private attributes**		**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
Lisa	***21***	*F*	***Brussels***	***$50K***	*Single*	*Reject*

Counterfactual explanation=
If you would be **three years older**, lived in **Antwerp** and your income would be **$10K** higher, you would have received a positive credit decision

Counterfactual instance						
Identifier	**Quasi-Identifiers**			**Private attributes**		**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
Fiona	***24***	*F*	***Antwerp***	***$60K***	*Single*	*Accept*

Fig. 1. Example of a counterfactual explanation

At the same time, there is a growing concern about the potential privacy risks of machine learning [31]. Privacy is recognized as a human right and defined by Oxford Dictionary as *"a state of being free from the attention of the public"*.[1] In a privacy attack, the goal of an adversary is to gain knowledge that was not intended to be shared [31, 47]. Different kinds of privacy attacks exist: both the training data, where the adversary tries to infer membership in a *membership inference attack* or specific attributes of an input sample in an *attribute inference attack*, as well as the model, in a *model extraction attack*, can be the target [17, 32, 47].

Unfortunately, there exists an inherent tension between explainability and privacy as the usage of Explainable AI can increase these privacy risks [1]: model explanations offer users information about how the model made a decision about their data instance. Consequently, they leak information about the model and the data instances that were used to train the model. Earlier research already shows that explanations can provide ground for membership inference attacks, where is determined whether a given instance is part of the training data, [40, 43, 46, 49] and model extraction attacks, where information about the functionality of the model is collected through query access [1, 46]. In this paper, we introduce a new kind of privacy attack based on counterfactual explanations and we call this an *explanation linkage attack*. A *linkage attack* attempts to identify anonymized individuals by combining the data with background information. An *explanation* linkage attack attempts to link the counterfactual explanation with background information to identify the counterfactual instance. We illustrate an example of an explanation linkage attack in Section 2. Unfortunately, the introduction of these attacks indicates that an attempt to make an AI system safer by making it more transparent can have the opposite effect [52]. Other researchers [6, 41, 48] also confirm the trade-off between privacy and explainability and emphasize that assessing this trade-off for minority groups is an important direction for future research [41].

Our contributions are as follows:

- We introduce a new kind of privacy attack, the *explanation linkage attack*, that can occur when using counterfactual explanations that are grounded in instances from the training set.
- As a solution for this problem, we propose k-anonymous counterfactual explanations and develop an algorithm to generate these.
- We evaluate how k-anonymizing the counterfactual explanations influences the quality of these explanations, and introduce *pureness* as a new metric to evaluate the validity of these explanations.
- We show the trade-off between *transparency*, *fairness* and *privacy* when using k-anonymous explanations: when we add more privacy constraints, the quality of the explanations and therefore the transparency decreases. This effect on

[1] https://www.oxfordlearnersdictionaries.com/definition/american_english/privacy

Just Accepted

the explanation quality is larger for minority groups, as they tend to be harder to anonymize, and this can have an impact on fairness.

2 PROBLEM STATEMENT: EXPLANATION LINKAGE ATTACKS

We introduce the privacy problem of counterfactual explanations that are grounded in instances of the training set, and illustrate this problem by using a simple toy dataset. This dataset contains individuals that are described by a set of identifiers, quasi-identifiers and private attributes [56]. Identifiers are attributes such as name, phone or social security number and need to be suppressed in any case as they often do not have predictive value and can uniquely identify a person. Quasi-identifiers are attributes such as age, zip code or gender that can hold some predictive value. They are assumed to be public information; however, even though they cannot uniquely identify a person, their combination might. It has been shown that 87% of US citizens can be re-identified by the combination of their zip code, gender and date of birth [54]. Private attributes are attributes that are not publicly known, and are meant to be kept confidential.

Let us briefly discuss the set-up of this attack: We assume that the adversary has access to the identifiers and quasi-identifiers of everyone in the the dataset. [2] In line with the literature, we look at the following two re-identification scenarios for a single individual [12, 14, 35]:

- Re-identification of a specific individual (*prosecutor re-identification scenario*): The adversary (e.g., a prosecutor) knows that a specific individual is part of the dataset, and wants to infer its private information.
- Re-identification of an arbitrary individual (*journal re-identification scenario*). The adversary (e.g., a journalist) does not care which individual is being re-identified but only wants to prove that it can be done.

If the attacker wants to execute one of the scenarios above and gets access to the private attributes of a user in the dataset, a possible avenue to achieve this is by asking for counterfactual explanations. The counterfactual explanation will never contain identifiers but if it contains a combination of quasi-identifiers that can uniquely identify a person, the attacker can deduce the person's private attributes. We name this kind of attack an *explanation linkage attack*.

Assume the following factual instance *Lisa* in Table 1:

Table 1. Factual instance *Lisa*

Identifier	**Quasi-identifiers**			**Private attributes**		**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
Lisa	*21*	*F*	*Brussels*	*$50K*	*Single*	*Reject*

Name is the identifier that is deleted from the dataset, but, as mentioned, people can often still be identified by their unique combination of quasi-identifiers. *Age*, *Gender* and *City* are the quasi-identifiers in this dataset that are assumed to be public knowledge for every adversary. A possible reasoning behind this, is that the adversary acquired access to a voter registration list as in Sweeney [54]. *Salary* and *Relationship* are private attributes that one does not want to be public information, and the target attribute in this dataset is whether the individual will be awarded credit or not. *Lisa* is predicted by the machine learning model as not creditworthy and her credit gets rejected. Logically, *Lisa* wants to know the easiest way to get her credit application accepted, so she asks for a *counterfactual explanation*, the smallest change to her feature values that result in a different prediction outcome.

In our set-up, the counterfactual algorithm looks for the instance in the training set that is nearest to *Lisa* and has a different prediction outcome (the *nearest unlike neighbor)*. The training set, with the nearest unlike neighbor highlighted, is shown in Table 2. *Fiona* has similar attribute values as *Lisa*, but is 24 years old instead of 21, lives in Antwerp instead of Brussels and earns $60K instead of $50K. When *Fiona* is used as counterfactual instance by the explanation algorithm, *Lisa* would receive the explanation: *'If you would be 3 years older, lived in Antwerp and your income was $10K higher, then you would have received the loan'*. Based on her combined knowledge of the explanation and her own attributes, *Lisa* can now deduce that *Fiona* is the counterfactual instance, as there is only one person in this dataset with this combination of quasi-identifiers (a 24-year old woman living in Antwerp). Therefore, *Lisa* can deduce the private attributes of *Fiona*, namely *Fiona*'s income and relationship status, which is undesirable.

[2]People's quasi-identifiers are often rather easy to be obtained by the public as lists like voter records are publicly available [34, 54].

Just Accepted

Table 2. Training set

Identifier	**Quasi-identifiers**			**Private attributes**		**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
Alfred	*25*	*M*	*Brussels*	*$50K*	*Single*	*Reject*
Boris	*23*	*M*	*Antwerp*	*$40K*	*Separated*	*Reject*
Casper	*34*	*M*	*Brussels*	*$30K*	*Cohabiting*	*Reject*
Derek	*47*	*M*	*Antwerp*	*$100K*	*Married*	*Accept*
Edward	*70*	*M*	*Brussels*	*$90K*	*Single*	*Accept*
Fiona	*24*	*F*	*Antwerp*	*$60K*	*Single*	*Accept*
Gina	*27*	*F*	*Antwerp*	*$80K*	*Married*	*Accept*
Hilda	*38*	*F*	*Brussels*	*$60K*	*Widowed*	*Reject*
Ingrid	*26*	*F*	*Antwerp*	*$60K*	*Single*	*Reject*
Jade	*50*	*F*	*Brussels*	*$100K*	*Married*	*Accept*

Obviously, this is just a toy example, but we envision many real-world settings where this situation could occur. For instance, when end users receive a negative decision, made by a *high-risk AI system*: these systems are defined by the EU's AI Act, which categorizes the risk of AI systems usage into four levels [15]. Among others, they include employment, educational training, law enforcement, migration and essential public services such as credit scoring. Article 13(1) states: *"High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately."* These systems are therefore obliged to provide some form of transparency and guidance to its users, which could be done by providing counterfactual explanations or any other transparency technique. Most of these settings use private attributes as input for their decisions, so it is important to make sure that the used transparency techniques do not reveal private information about other decision subjects. For example, in decisions about educational training or employment, someone's grades could be revealed, or in credit scoring, the income of other decision subjects could be disclosed.

This privacy risk occurs when the counterfactual algorithm uses instance-based strategies to find the counterfactual explanations. These counterfactuals correspond to the *nearest unlike neighbor* and are also called *native counterfactuals* [5, 25]. Other counterfactual algorithms use perturbation where synthetic counterfactuals are generated by perturbing the factual instance and labelling it with the machine learning model, without reference to known cases in the training set [25]. We focus on counterfactual algorithms that return real instances: several algorithms do this, as this substantially decreases the run time while also increasing desirable properties of the explanations such as plausibility [5]. Plausibility measures how realistic the counterfactual explanation is with respect to the data manifold, which is a desirable property[22], and Brughmans et al. [5] show that the techniques resulting in an actual instance have the best plausibility results. Furthermore, it is argued that counterfactual instances that are plausible, are more robust and therefore are less vulnerable to the uncertainty of the classification model or changes over time [2, 5, 42]. This shows that for some use cases it can be very useful to use real data points as counterfactuals instead of synthetic ones as for the latter the risk of generating implausible counterfactual explanations can be quite high [27]. Algorithms that use these *native* counterfactual explanations include NICE (without optimization setting) [5], the WIT tool with NNCE [59], FACE [44] and certain settings of CBR [25]. Perturbation-based counterfactual algorithms experience different privacy risks such as membership inference attacks: Pawelczyk et al. [43] use counterfactual distance-based attacks which leverage algorithmic recourse to determine if an instance belongs to the training data of the underlying model or not. We envisage a different scenario, where the adversary knows which instances are in the training data, but wants to gain access to its private attributes. It is worth emphasizing that some perturbation-based counterfactual algorithms could still have some vulnerability to explanation linkage attacks, although arguably less likely than native counterfactuals. Some perturbation algorithms (such as NICE with optimization settings) start from a real counterfactual instance in the dataset, and it is possible they will return the real instance without perturbations. In many cases, the instance will only be slightly perturbed, so that an ingenious adversary can still have high confidence about the private attribute values of the counterfactual instances.

Just Accepted

Table 3. Comparison between the original problem setting of k-anonymity and our problem setting.

	***k*-anonymity**	
	Dataset	**Counterfactual explanation**
Input	Dataset	Dataset Factual instance Counterfactual explanation Machine learning model
Defined over	Dataset	Counterfactual explanation
Method	Mondrian[3], Datafly[4],..	CF-K
Risk	Identifying instances in the dataset based on their combination of quasi-identifiers and inferring their private attributes	Identifying the counterfactual instance based on its combination of quasi-identifiers and inferring its private attributes
Evaluation metrics	Degree of privacy Information loss	Degree of privacy Information loss Counterfactual validity

3 PROPOSED SOLUTION

As a solution, we propose to make the counterfactual explanations *k-anonymous*. *k-anonymity* is a property that captures the protection of released data against possible re-identification by stating that the released data should be indistinguishable between k data subjects [57].

3.1 What is k-anonymity?

Before k-anonymity was introduced, data that looked anonymous was often freely shared after removing explicit identifiers such as name and address, incorrectly believing that individuals in those datasets could not be identified. Contrary to these beliefs, we have seen that people can often be identified through their unique combination of quasi-identifiers.

Consider a database that holds private information about individuals, where each individual is described by a set of identifiers, quasi-identifiers, and private attributes. k-anonymity characterises the degree of privacy, where the information for each person in the dataset cannot be distinguished from at least $k-1$ other individuals whose information was also released [55]. A group of individuals that cannot be distinguished from each other and thus have the same values of quasi-identifiers are named an *equivalence class*.

Usually k-anonymity is applied on the whole dataset: the quasi-identifiers of the data records are suppressed or generalised in such a way that one record is not distinguishable from at least $k-1$ other data records in that dataset [38]. In this way, the privacy of individuals is protected to some extent by *"hiding in the crowd"* as private data can now only be linked to a set of individuals of at least size k [20]. However, by generalising or suppressing attribute values, the data becomes less useful, so the problem studied is to make a dataset k-anonymous with minimal loss of information [20, 61]. We will measure the loss in information value with the *Normalized Certainty Penalty* (NCP) and explain this metric in Section 4.

3.2 Application to our problem

Our application differs from the original set-up of k-anonymity as it specifically aims to ensure anonymity in counterfactual explanations, rather than anonymizing the entire dataset. While the original application is suitable for situations where the entire dataset is publicly accessible. We highlight this difference in Table 3. A counterfactual instance is defined as k-anonymous if the combination of quasi-identifiers can belong to at least k individuals in the training set, and consequently, a counterfactual explanation is defined as k-anonymous if the counterfactual instance on which it is based, has a combination of quasi-identifiers that can belong to at least k individuals in the training set. We implement this by looking for close neighbours

[1]LeFevre et al. [29]
[2]Sweeney [55]

Just Accepted

Counterfactual instance						
Identifier	**Quasi-Identifiers**			**Private attributes**		**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
*	*24*	*F*	*Antwerp*	*$60K*	*Single*	*Accept*

+

Neighbor						
Identifier	**Quasi-Identifiers**			**Private attributes**		**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
*	*27*	*F*	*Antwerp*	*$80K*	*Married*	*Accept*

⇩

***K*-anonymous counterfactual instance**						
Identifier	**Quasi-Identifiers**			**Private attributes**		**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
*	*24-27*	*F*	*Antwerp*	*$60K*	*Single*	*Accept*

Fig. 2. How to generalize the counterfactual instance. As can be seen, we generalize only the values of the quasi-identifiers. The private attributes are still the same as in the original counterfactual instance as their attribute value is not public and therefore cannot be used to identify someone.

Just Accepted

of *Fiona*, that have similar values of quasi-identifiers, and that also have the desired prediction outcome. In this case, the closest neighbor to *Fiona* that has the desired prediction outcome is *Gina*, as can be seen in Table 2. Next, we generalise the quasi-identifiers of the counterfactual instance so that they can belong to both the counterfactual instance and the neighbour, resulting in a counterfactual instance that is at least 2-anonymous (see Figure 2.) However, by doing so we degrade the quality of the data as we will see in Section 4.

The k-anonymous counterfactual explanation based on the k-anonymous counterfactual instance in Figure 2 and factual instance *Lisa* (21, F, Brussels, $50K, Single) is: *'If you would be 3-6 years older, lived in Antwerp and had an income of $60K, you would have received the loan'*. This explanation is 3-anonymous because the combination of quasi-identifiers in the counterfactual instance (24-27, F, Antwerp) could point to at least three instances in the training set in Table 2, namely *Fiona*, *Gina* and *Ingrid*.

However, the fact that other instances than the ones explicitly used to generate the k-anonymous counterfactual explanation, might also fall in the range of the explanation, introduces a new issue that is specific to k-anonymous counterfactual explanations. Counterfactual explanations are defined as the smallest change to the feature values of an instance that alter its prediction outcome, but does this still hold for k-anonymous counterfactual explanations? We are no longer sure that all the value combinations in the k-anonymous counterfactual instance lead to a change in the prediction outcome and therefore we are not sure whether they are *valid* counterfactual explanations.

In this toy example, the value combination of *Ingrid* in Table 2 is also part of the k-anonymous counterfactual instance, as *Ingrid* is between 24 and 27 years old, female, single, living in Antwerp and earning $60K. However, the model predicts *Ingrid*s credit decision to be rejected. A possible reasoning behind this could be because the model has learned that for higher age groups a higher income is required to be awarded the credit (or any other pattern). Therefore, if *Lisa* would follow-up the *"advice"* in the counterfactual explanation, it is possible that she would end up in this value combination, which does not result in an altering of the prediction outcome. This is problematic as this is one of the key objectives of counterfactual explanations.

This issue leads us to a new metric: how *valid* is the k-anonymous counterfactual explanation? We discuss the evaluation metrics further in Section 4.

4 EVALUATION METRICS

We measure the quality of the explanations by using the following metrics:

- The degree of privacy is measured by k: to how many instances from the training set can this counterfactual explanation be linked?
- The validity of the counterfactual explanations is measured by the *pureness*.
- The loss in information value is measured by the *Normalized Certainty Penalty (NCP)*.

We assess how the degree of privacy influences the loss in information value and the validity of the counterfactual explanations in Figure 4.

Degree of privacy. We measure the degree of privacy by using the definition of k-anonymity. In our toy example, k is 3, as the generalised quasi-identifiers of the k-anonymous counterfactual instance could belong to three people when we look at the training set in Table 2 (*Fiona*, *Gina* and *Ingrid*). In our set-up, we will implement the degree of privacy as a minimum constraint in the algorithm.

Counterfactual validity. We define a possible value combination as a combination of attribute values that is in the range of the k-anonymous counterfactual instance. Note that we take into account all the attributes here, not only the quasi-identifiers. For a categorical attribute, we look at all the values present in the k-anonymous counterfactual instance. For a numerical attribute, we look at all the values that are in the range of the k-anonymous counterfactual instance *and* are also present in the training set. We illustrate these calculations in Table 4. The pureness of a k-anonymous counterfactual explanation can be calculated as follows:

$$\text{Pureness} = \frac{\text{\# of value combinations with desired prediction outcome}}{\text{\# of value combinations}}$$

The theoretical pureness is calculated on all the value combinations, but we will approximate this by querying the model with 100 random combinations[5] and see how many of these combinations lead to the desired prediction outcome. The pureness is the proportion of these value combinations that lead to the desired prediction outcome, which obviously should be as high as possible (preferably 100%).

Table 4. Possible value combinations and its model predictions.

Age	**Gender**	**City**	**Salary**	**Relationship status**	**Model prediction**
24	*F*	*Antwerp*	*$60K*	*Single*	*Accept*
25	*F*	*Antwerp*	*$60K*	*Single*	*Accept*
26	*F*	*Antwerp*	*$60K*	*Single*	*Reject*
27	*F*	*Antwerp*	*$60K*	*Single*	*Reject*

Table 4 shows all possible value combinations of the k-anonymous counterfactual instance, and the prediction outcome to each value combination. The goal of the counterfactual explanation was to alter the prediction outcome from *Reject* to *Accept*, so this is the desired prediction outcome. The k-anonymous counterfactual explanation in our toy example leads to the desired prediction outcome in 50% of the cases ($\frac{2}{4}$). If we sample 100 times out of the value combinations above, we expect this to approximate the theoretical pureness of 50%.

[5]We chose for 100 random value combinations instead of trying out all the possibilities as the number of combinations can quickly become very large when there is a lot of generalization. The more random value combinations we test, the more we approximate the theoretical pureness, but the longer the computation time.

Just Accepted

Loss in information value. When datasets are made k-anonymous, they tend to lose information. In general, excessive anonymization makes the data less useful because some analysis is no longer possible or the analysis provides biased and incorrect results [12].

A variety of metrics to measure information loss have been proposed, and we focus on the metrics discussed in Ghinita et al. [18], which are Normalized Certainty Penalty (NCP) [60], Discernibility metric (C_DM) [4] and the classification metric (CM) [24].

NCP penalises attributes for the way they are generalised and captures the uncertainty caused by this generalization [60]. It assigns larger penalties when attribute values are mapped to generalised values that replace many other values [33]. An advantage of this metric is that it can give different weights to different attributes, as some attributes can be more important than others for the data analysis process [60]. The NCP for each numerical (*Num*) quasi-identifier A in an equivalence class G is defined as:

$$\mathrm{NCP}_{A_{Num}}(G) = \frac{\max^G_{A_{Num}} - \min^G_{A_{Num}}}{\max_{A_{Num}} - \min_{A_{Num}}}, \tag{1}$$

where the numerator and denominator represent the range of attribute A for the equivalence class G and for the whole dataset respectively [19]. This metric thus measures which part of the total range of the numerical attribute, is present in the equivalence class. Higher values signify more generalization, and consequently, more information loss. In the case of a categorical (*Cat*) quasi-identifier A, NCP is defined as follows:

$$\mathrm{NCP}_{A_{Cat}}(G) = \begin{cases} 0, & \text{if } |A^G| = 1 \\ \frac{|A^G|}{|A|}, & \text{otherwise} \end{cases} \tag{2}$$

where $|A|$ is the number of distinct values of attribute A in the whole dataset, and $|A^G|$ is the number of distinct values of attribute A in equivalence class G [60]. So, for a categorical attribute, this metric will check which proportion of possible unique values is present in the k-anonymous counterfactual instance. The higher this number is, the more generalized this attribute will be and the more information about this attribute is lost. The NCP of equivalence class G over all quasi-identifier attributes is:

$$NCP(G) = \sum_{i=1}^{d} w_i \cdot NCP_{A_i}(G), \tag{3}$$

where d is the number of quasi-identifiers in the dataset, A_i is a (numerical or categorical) attribute with weight w_i, where $\sum_i w_i = 1$ [19]. For our experiments, we assume all attributes have an equal weight but this can easily be altered in future experiments. NCP measures the information loss for a single instance and its equivalence class. This can be aggregated to the information loss in the entire dataset [19, 60] but for our problem setting, we only need to calculate the NCP for each k-anonymous counterfactual explanation, which constitutes one equivalence class. As an illustration, we calculate the NCP of the k-anonymous counterfactual explanation (CE) in our toy example[6]:

$$NCP_{Age}(CE) = \frac{max_{CE^{Age}} - min_{CE^{Age}}}{max_{Age} - min_{Age}} = \frac{27-24}{70-23} = 0.064,$$

$$\mathrm{NCP}_{Gender}(CE) = 0 \qquad (|A^{CE}| = 1), \mathrm{NCP}_{City}(CE) = 0 \qquad (|A^{CE}| = 1),$$

$$\mathrm{NCP}(CE) = \frac{1}{3} \cdot 0.064 + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 0 = 0.021$$

For our experiments, we focus on the metrics NCP and pureness, but for completeness we also report the results with two additional metrics. The discernibility metric assigns a penalty to each tuple, based on how many tuples are indistinguishable from it after anonymizing. The idea is that it is desired to maintain discernibility between tuples as much as is allowed by a given setting of k [4]. The discernibility metric for anonymization g, and a degree of privacy k is:

$$C_{DM}(g,k) = \sum_{\forall E\ s.t. |E| \geq k} |E|^2 + \sum_{\forall E\ s.t. |E| < k} |D||E|$$

[6]See Table 2 for the range of each attribute in the training set.

Just Accepted

In this expression, E refers to the equivalence class of the tuple, and D to the dataset. Each successfully anonymized tuple (equivalence class larger than k) gets as penalty the size of the equivalence class, and each suppressed tuple (equivalence class smaller than k) gets as penalty the size of the total dataset. In our set-up, all the counterfactual explanations will be successfully anonymized so each anonymized explanation will get as penalty the size of its equivalence class. The discernibility metric for the k-anonymous counterfactual explanation in our example is 3, as this is the number of people belonging to its equivalence class (see Table 2.) This metric has been criticized because it does not take into account how much the anonymized data instances approximate the original instances [13]. NCP is a more suitable metric to measure the actual information loss incurred by anonymizing the counterfactual explanations [18, 45].

The classification metric (CM) is a class-conscious metric that attempts to create equivalence classes that consist of tuples that are uniform with respect to the class label [24].

$$CM = \frac{\sum_i^N penalty(tuple_i)}{N}$$

N is the number of anonymized tuples, which can be rows in a dataset or in our case number of the anonymized counterfactual instances. Each tuple receives a penalty of 1 if its class is different from the majority class label of its equivalence class. In the case of our toy example, the k-anonymized counterfactual instance does not receive a penalty as its label is the same as the majority class label of its equivalence class (*Accept*). This metric is related to our notion of pureness, but keep in mind that they measure different things. The classification metric looks at the instances in the equivalence class (which are Fiona, Gina and Ingrid in the case of our toy example) and their majority label. For pureness, we take all the attributes into account (so also the private attributes), and not only look at the instances present in the dataset, but at all the possible value combinations in the range of the anonymous explanation (by using sampling). This can be seen in Table 4. Pureness is therefore more suitable than the classification metric to measure how often the anonymous counterfactual explanation gives us *correct* advice.

5 MATERIALS AND METHODS

5.1 Materials

Table 5. Description of used datasets with dataset properties

Dataset	**Adult** [7]	**CMC** [8]	**German** [9]	**Heart** [10]	**Hospital** [11]	**Informs** [12]
# instances	48,842	1,473	1,000	303	8,160	5,000
# attributes	11	8	19	12	20	13
QID	*Age, Sex, Race, Relationship, Marital status*	*WifeAge, ChildrenBorn*	*Age, Foreign Personal status, Residence time, Employment, Job, Property, Housing*	*Age, Sex*	*Age Group, Race, Gender, Ethnicity, Zip Code*	*Dobmm, Dobyy, Sex, Marry, Educyear*
Sensitive attribute	*Sex*	*WifeReligion*	*Personal status*	*Sex*	*Gender*	*Race*
Target attribute	*Income*	*Contraceptive method*	*Credit decision*	*Heart disease*	*Costs*	*Income*
Uniquely identifiable (in %)	3.17	4.41	83.7	4.62	6.32	76.18
|EQ| <10 (in %)	15.39	53.78	100	79.54	37.08	100

Just Accepted

We choose the datasets described in Table 5, as they are all tabular datasets that contain various personal attributes through which individuals could be identified, and are often used in research about privacy-preserving data mining [26, 50, 51].[13] All these datasets contain private information such as financial and health data that people generally do not want to be made public. In this Table, we list general dataset description properties such as the number of instances and attributes, and the target attribute. We also mention the quasi-identifiers and sensitive attribute (on which discrimination is measured) that we used for our experiments. Additionally, we measure the privacy risk present in each dataset in two ways: 1) We measure the percentage of people that are uniquely identifiable by their combination of quasi-identifiers, and 2) We measure the percentage of instances that are not protected by k-anonymity (with k =10). This thus means that we measure the percentage of people that belong to an equivalence class with a size smaller than 10.

5.2 Methods

On every dataset, we apply the methodology as described in Figure 3. We first split the dataset in a training and test set, using a split of 60-40. We fit and tune a Random Forest model through cross-validation on the training set. The following grid is used for tuning:

$$\text{n_estimators} = [10, 50, 100, 500, 1,000, 5,000]$$
$$\text{max_leaf_nodes} = [10, 100, 500, n] \text{ with } n = \infty$$

We use the standard version (no optimization setting) of NICE [5] as counterfactual algorithm, as this will return actual instances from the training set, and fit this on the training set and the trained machine learning model. This trained machine learning model is used to make predictions on all the instances in the test set. For all the test instances[14] without the desired prediction outcome, we use NICE to generate a counterfactual explanation. We focus on the test instances without the desired prediction outcome as these are the instances that generally use counterfactual explanations to receive advice on how to change their prediction outcome. As mentioned, when using NICE without any optimization setting, the counterfactual instances are real instances from the training data so they should be anonymized. The final step is to use CF-K to make these explanations k-anonymous.

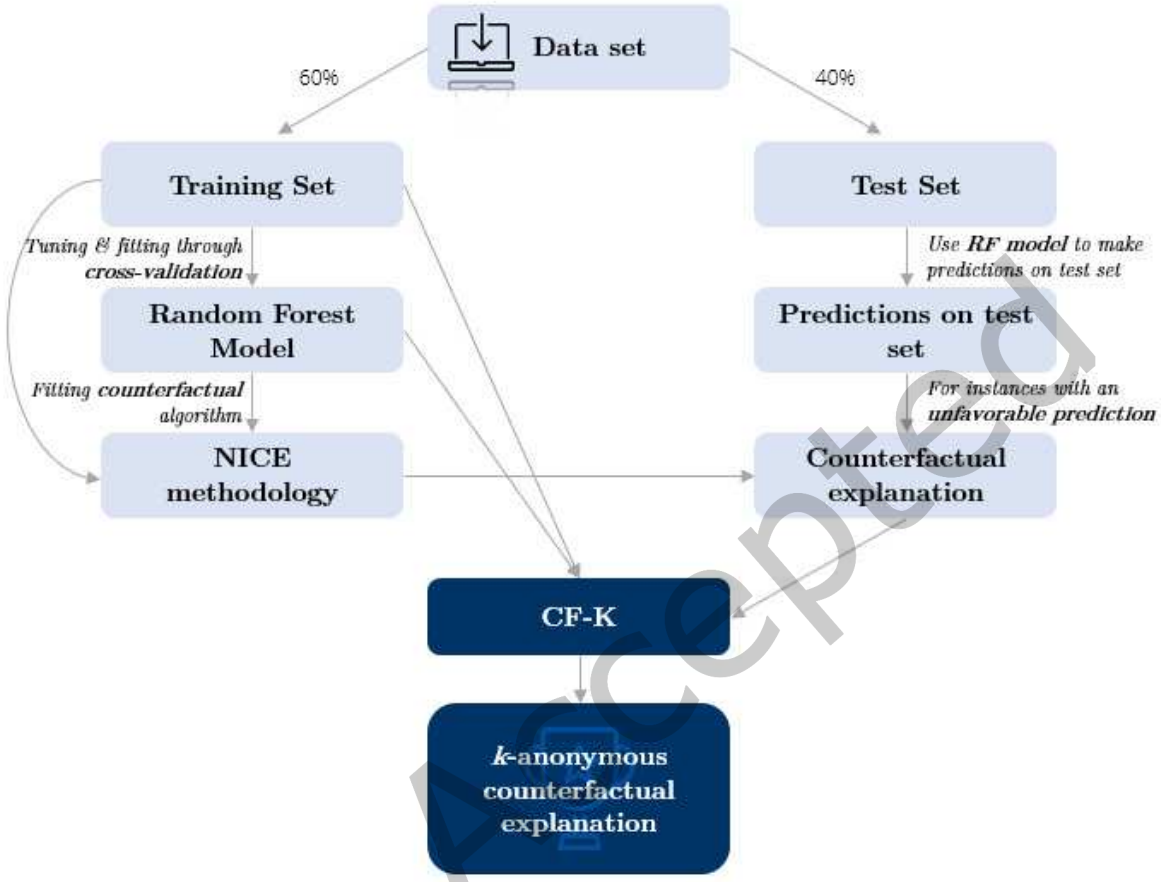
5.3 Novel algorithm CF-K

In the original application of k-anonymity, the whole dataset is made public and therefore has to be made k-anonymous. The goal is to find an optimal partition, for which both exact algorithms [28] and heuristics like genetic algorithms [24] and greedy algorithms (Mondrian [29], Datafly [55]) exist.

Our approach differs from the approaches published in the literature, as only the counterfactual explanation is made public and not the whole dataset. Therefore, we search for an equivalence class for each returned counterfactual instance separately. This changes the set-up of the problem, as making the whole dataset k-anonymous can degrade the data more than just making the counterfactual explanations k-anonymous: not every training instance is used as counterfactual explanation and unused training instances do not need to be made k-anonymous or used in the calculation for the best clustering. In the same way that local encoding achieves less information loss than global recoding [60], we hypothesize that only k-anonymizing the counterfactual instances can achieve lower information loss. We verify this claim in Section 6.2. Furthermore, specifically for our problem of k-anonymous counterfactual explanations we have to take the *counterfactual validity* of the k-anonymous explanations into account as this is essential for the goal of counterfactual explanations. We use this as additional metric in our algorithm.

[6]https://github.com/EpistasisLab/pmlb/tree/master/datasets/adult
[7]https://archive.ics.uci.edu/ml/machine-learning-databases/cmc/
[8]https://github.com/EpistasisLab/pmlb/tree/master/datasets/german
[9]https://github.com/EpistasisLab/pmlb/tree/master/datasets/heart_c
[10]https://www.opendatanetwork.com/dataset/health.data.ny.gov/82xm-y6g8
[11]https://github.com/kaylode/k-anonymity/tree/main/data/informs
[13]https://github.com/kaylode/k-anonymity
[14]We set a limit at 1,000 instances for the sake of time.

Just Accepted



Just Accepted

Fig. 3. Used methodology to generate k anonymous counterfactual explanations from a dataset.

We name our algorithm that makes counterfactual explanations k-anonymous CF-K. It is based on the metaheuristic GRASP (Greedy Randomized Adaptive Search Procedure) [16]. GRASP is a multi-start metaheuristic in which each iteration consists of two phases: construction and local search. After the two phases, the current best solution is updated. The construction phase builds a feasible solution, and the local search phase searches the neighborhood until a local optimum is found [16]. In this construction phase, GRASP combines greediness with randomness, with the purpose of escaping the myopic behavior of a purely greedy algorithm. We choose a heuristic algorithm, as it is a NP-hard problem and we are not looking for the optimal solution but for the best solution that can be found in limited computing time. Our aim is to provide a method that performs well, but we expect further optimizations to be possible in future research.

5.3.1 Algorithm description. Our algorithm starts from a counterfactual explanation that is given to one of the instances in the test set with an unfavorable prediction outcome. The counterfactual instance that this explanation is based on is an actual instance in the training set, and we want it to be unidentifiable from at least $k-1$ other instances in the training set. This is the case when at least $k-1$ other instances in the training set have the same values for the quasi-identifiers (these are the attributes that we assume to be publicly known).

Phase 1: Construct greedy randomized solution. In this phase, we construct a feasible solution. We first check for the current counterfactual instance if its values of quasi-identifiers are present for k individuals in the training set. In this case, a solution is found, the quality of the solution is calculated and the algorithm moves to the next phase. If this is not the case, we generate a list of size α by selecting the closest neighbors of the counterfactual instance in the training set with the required prediction outcome. Then, we *randomly* select a neighbour from this list and create a new *generalized instance* out of this neighbor and the counterfactual instance. The fact that we randomly select a neighbor out of this list and not just select the closest neighbour makes up the probabilistic component of GRASP. We create this *generalized instance* by generalizing the values of the quasi-identifiers so that the generalized instance includes both the values of the quasi-identifiers of the counterfactual instance as well as those of the neighbor. This happens as in Figure 2. We check again whether this generalised instance satisfies k-anonymity. If this is the case, a solution is found, the quality of this solution is calculated and the algorithm moves to the next phase. If this is not the case, this loop is repeated until the generalised instance satisfies k-anonymity.

Phase 2: Local search. The local search algorithm iteratively tries to replace the current solution by a better solution in the neighbourhood. The algorithm terminates when no better solution is found. The neighborhood is defined by checking for every quasi-identifier in the current solution whether slightly changing it, is a feasible solution (satisfies k-anonymity) and improves the solution quality. A slight change in this case is adding a value (if the quasi-identifier is a single value) or removing a value from the list (if the quasi-identifier is already a generalized list).

Algorithm 1 GRASP

```
for i = 1, ..., MaxIter do
    Solution ← ConstructGreedyRandomizedSolution(Input);
    Solution ← LocalSearch(Solution);
    BestSolution ← UpdateSolution(Solution, BestSolution);
end for
return BestSolution;
```

GRASP. We iterate these two phases for a specified number of iterations. After each iteration, we check if the new solution is better than the current best solution and if this is the case, we update the current best solution. After the specified number of iterations, the algorithm terminates and the current best solution is returned.

5.3.2 Choice of parameters. The input parameters in our algorithm are k, the level of desired privacy, α, the degree of randomness we give to the algorithm and the number of iterations the algorithm can perform. We show the effect of changing the input parameters on the *german* dataset by evaluating the metrics NCP, pureness and execution time.

Degree of privacy k. We see that if we increase k, the level of privacy guarantees for each individual, the other metrics deteriorate. The Normalized Certainty Penalty, which measures how much information value we lose by making the data k-anonymous, increases when we increase the value of k. This makes sense as the data quality degrades more when we add more privacy guarantees and therefore require more instances to be identical. Furthermore, the average pureness, and thus the counterfactual validity, also decreases. The trade-off between privacy (measured by k) and information loss (measured by NCP) has been confirmed by the literature [3, 53], but we are the first to show this trade-off between k and counterfactual validity (measured by pureness). Furthermore, the average execution time also increases if we increase k, as more privacy guarantees have to be implemented. For the remainder of the experiments, we use a k of 10 as this is a common number to baseline k-anonymity performance [12, 13]. We include the results for other values of k for both our algorithm (CF-K) and Mondrian in the Appendix.

Parameter α. The parameter α is a measure of the randomness of the algorithm, as it determines the number of closest neighbors from which we randomly select one. We see that increasing α will increase the NCP but will lower the pureness. This is to be expected as we look at further neighbors when α is larger, so this will increase the information loss, but also creates more room to improve the pureness in the local search. Increasing α decreases the execution time, which is reasonable as we will satisfy k-anonymity faster by taking further neighbors The optimal value of α will depend on the dataset and

Just Accepted

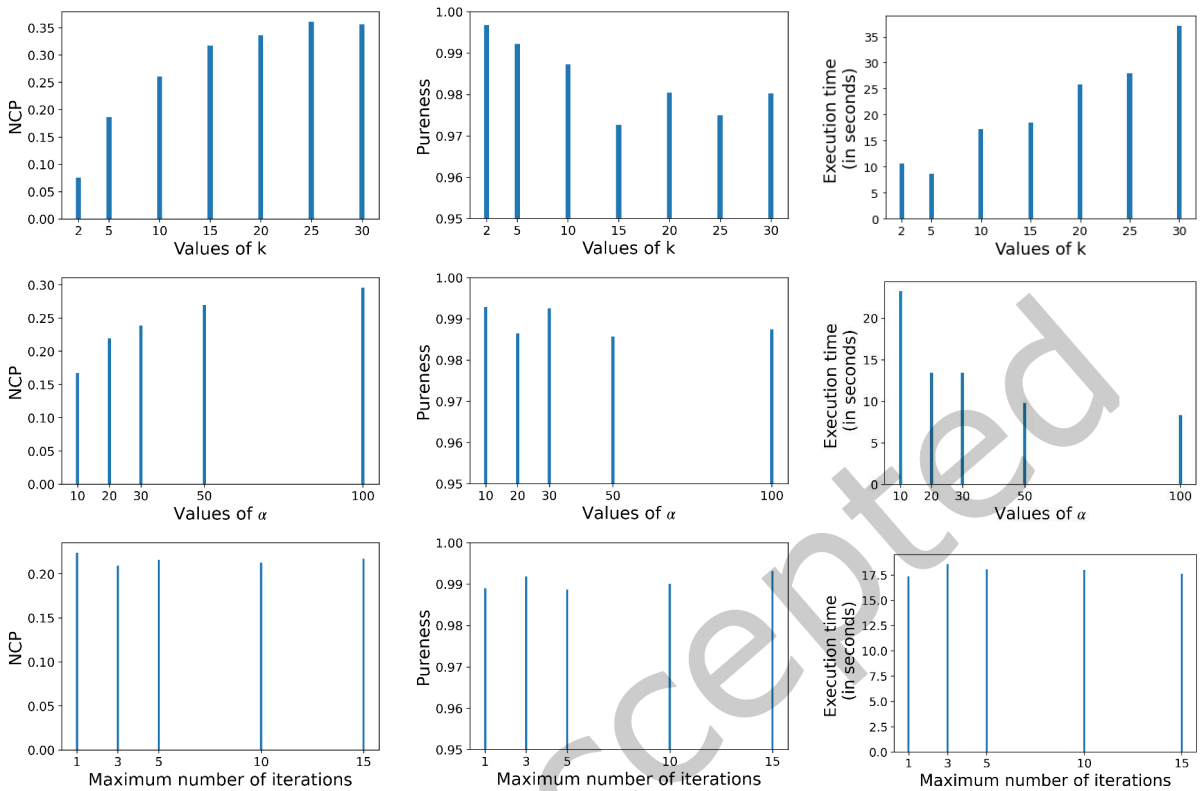


Fig. 4. Evaluation of parameters

Just Accepted

how highly one values the different metrics. To avoid a multiple comparisons problem, we fix α at 20 for the rest of the experiments.

Number of iterations. Increasing the number of iterations improves both the NCP and the pureness, but also increases the execution time. A trade-off has to be made between solution quality and execution time in determining the optimal number of iterations. We fix the number of iterations at 3 for the rest of the experiments.

6 RESULTS

6.1 Results per dataset

When we compare the results of Table 6 with the privacy risks of each dataset reported in Table 5, we see that explanations of the datasets with the highest privacy risks (*German* and *Informs*) have the highest information loss (in terms of NCP) when they are made anonymous. We measured the privacy risk by calculating the number of people in the dataset that are in equivalence classes smaller than 10 (before anonymizing), and for *German* and *Informs*, this will be the case for every person. For other datasets, such as *Adult*, only around 15% of individuals are in equivalence classes smaller than 10, so only a small portion of counterfactual instances will have to be anonymized. The average information loss (measured by NCP) for the anonymous explanations of this dataset is therefore much lower. The C_{DM} metric is harder to compare across datasets, as the size of the anonymous set has a large influence here. Therefore, we add an extra row where we divide C_{DM} by the number of anonymized explanations. This gives us the average size of the equivalence class for all the anonymized explanations. We see that for *Adult*, some equivalence classes can be really large, but the average NCP is low, which is more important

Table 6. Results of CF-K over all the datasets (k = 10).

Dataset	Adult	CMC	German	Heart	Hospital	Informs
NCP (mean)	0.55%	3.84%	21.41%	2.81%	3.42%	9.97%
Pureness (mean)	99.81%	93.15%	98.52%	100%	91.39%	85.33%
Execution time (mean)	24.78s	16.20s	13.31s	3.93s	17.76s	32.20s
C_{DM}	87,181	5,366	1,010	790	17,115	9,023
$\frac{C_{DM}}{\#explanations}$	110.78	13.2	16.83	14.11	22.94	13.65
CM	0.82	0.28	0.03	0.32	0.77	0.12

for our problem. This consequently implies that the data did not have to be significantly degraded, but the generalized quasi-identifiers still encompass a substantial number of individuals. We also see that in the *Heart* dataset, the counterfactual validity measured by the pureness is always 100%. We expect this to be the case if the quasi-identifiers, which are *Age* and *Sex* in this case, have a small influence on the outcome of the machine learning model. We verify this by examining the feature importance ranking of the used model, and indeed see that the quasi-identifiers are ranked very low. This could explain why generalizing them has no effect on the counterfactual validity. For all datasets, the pureness is above 85%, which makes the generalized counterfactual explanations pretty valid. We see that although CM and pureness are related, they can give very different results per dataset. CM assesses the majority label of the whole equivalence class, while pureness will evaluate how many value combinations in the k-anonymous counterfactual instance will lead to the desired target outcome. As already said, for our use case, pureness is more relevant as this will actually assess how often the counterfactual explanations points us in the 'right' direction. Also note that for pureness, higher values are better, while for CM, lower values are preferred (less penalties).

We can also assess how the results vary for different values of k in the Appendix. We see that if k increases, in general the information loss becomes higher and the pureness becomes lower. This is in line with the results of Section 5.3.2, and again shows the trade-off between privacy and explainability.

With regard to the execution time, we see that it will be fast enough for most applications, and is in line with the order of magnitude of generating counterfactual explanations [10]. If further speed-ups are necessary, this can be realised by decreasing the number of iterations, further optimization of the algorithm or using a stronger computer. All measurements were taken on a Dell Latitude 7400 laptop with 16GB of RAM and Intel®Core^{TM} i7-8665U CPU.

6.2 Comparison with Mondrian

Table 7. Results of the Mondrian algorithm (k = 10)

Dataset	Adult	CMC	German	Heart	Hospital	Informs
NCP (mean)	15.97%	7.05%	59.55%	53.01%	26.03%	36.31%
Pureness (mean)	90.30%	69.15%	90.50%	100%	63.77%	72.40%
Execution time (mean)	7.11s	0.87s	0.38s	0.23s	1.19s	1.11s
C_{DM} **(mean)**	120,227	6,318	963	1,044	16,534	9,177
$\frac{C_{DM}}{\#explanations}$	152.77	15.56	16.05	18.64	22.16	13.88
CM (mean)	0.83	0.24	0.17	0.41	0.80	0.40

We compare CF-K with an alternative strategy: making the whole dataset k-anonymous, and taking the counterfactual explanations out of this anonymized dataset. This differs from our strategy where we directly make the counterfactual instances and explanations k-anonymous. We use an open source implementation of Mondrian[15] to compare CF-K with. Mondrian is a top-down greedy data anonymization algorithm that has been shown to be one of the best performers [3, 29].

[15]https://github.com/danielegiampaoli/Mondrian_K-anonymization

Just Accepted

For all instances in the test set (max 1,000) with an unfavorable outcome, we compare the k-anonymous counterfactual explanation generated by CF-K with the k-anonymous counterfactual explanation based on an instance selected from the anonymized (by Mondrian) test set. When we compare the results in Table 6, with the results in Table 7, we see that for all datasets CF-K succeeds in achieving a better (and thus lower) average NCP than the Mondrian implementation on the whole dataset. This is in line with our hypothesis that only k-anonymizing the counterfactual instances can result in lower information loss, as unused training instances do not need to be used in the calculations for the best clustering. Furthermore, the average counterfactual validity (measured by pureness) in all datasets is higher when using k-anonymous explanations than when using an explanation from a k-anonymous dataset (except for the *Heart* dataset, where the average counterfactual validity is 100% for both implementations). Counterfactual validity can only be calculated on an explanation, and not on a dataset, so methods to make the dataset k-anonymous can not optimize for this metric. Therefore, our methodology to make the explanations k-anonymous, was needed to be able to take this metric into account. With regard to the C_{DM} metric: in four out of the six datasets, CF-K results in the smallest classes, while in two out of the six datasets, Mondrian will achieve slightly smaller equivalence classes. However, even in those cases, the average information loss measured by NCP will be lower when using CF-K, and as explained, it makes more sense to focus on this metric. The results for the CM metric show that for most datasets, the CF-K algorithm results in equivalence classes that are a bit more uniform with respect to the class label. However, as mentioned, pureness is more suited to measure the actual validity of the counterfactual explanations. We see in the Appendix, that the results for other values of k (5 and 20) are in line. For the Mondrian algorithm, the evaluation metrics also deteriorate when the level of privacy protection (k) is increased, and CF-K still outperforms Mondrian in terms of NCP and pureness for all values of k.

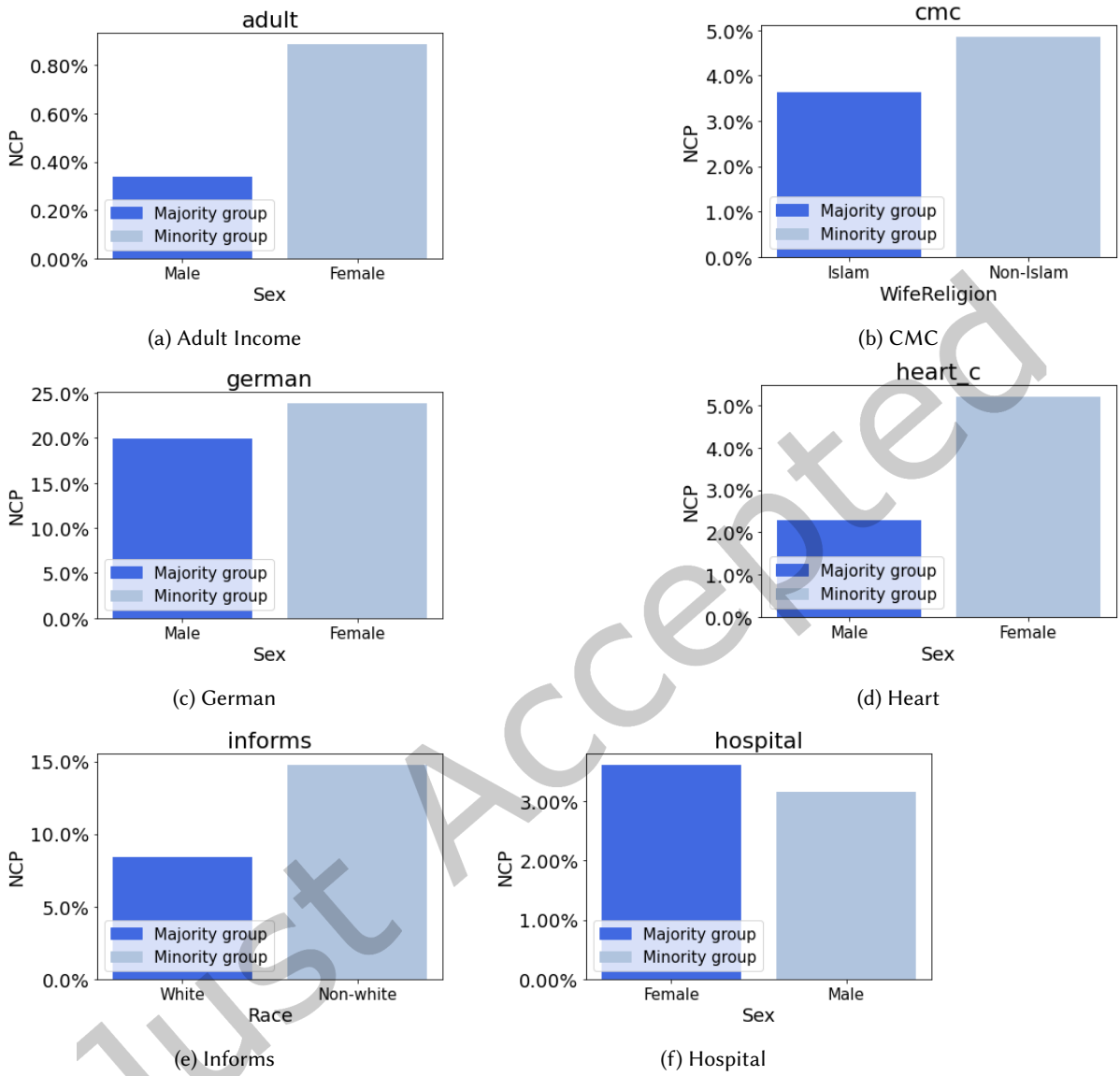
6.3 Does this have fairness implications?

A minority group is defined as a group whose characteristics such as race, religion, gender, ... etc. are fewer in numbers than the main group of that classification. Nowadays, it is often used to refer to people that experience a relative disadvantage based on their group membership [23]. We define the minority and majority group for each dataset based on the sensitive attribute, mentioned in Table 5. The minority group is the category of that sensitive attribute that is the least present in the training set. We see in Figure 4 that when we make the explanations more private (increase k), the explanation quality decreases and they become less useful. Unfortunately, this effect is larger for minority groups which can lead to potential issues regarding fairness. As can be seen in Figure 5, in every examined dataset (except for *Hospital*), the average NCP is higher for the minority group. For the average counterfactual validity, we found no difference between both groups. So we see in Figure 5 that the quality of explanations of the minority group has to be reduced more to achieve the same level of privacy. This can be explained by the fact that they often have more unique quasi-identifiers, as there are less people that share their public characteristics (definition of a minority group), so their quasi-identifiers have to be generalised more to be anonymous. For the *Hospital* dataset, the average information loss is slightly higher for the majority group. We hypothesize that this is due to the higher percentage of individuals with the desired target outcome (high income) for the minority group (men) than for the majority group (women), and hence it will be more difficult to find *pure* explanations for the latter. When explanations are used in high-stakes settings, it is undesirable that minority groups are offered lower quality explanations, but also that there is a higher risk of leaking their private information when no precautions are taken [41]. Other research showed another possible trade-off between fairness and privacy, as the privacy risks of different demographic groups are disparately affected by fairness-aware machine learning [8]. These results show that different ethical objectives can work against each other and that one has to make sure that minority groups are not adversely affected in unexpected ways.

6.4 Comparison with perturbation-based counterfactual algorithms

We mentioned before that using *native* counterfactuals increases desirable properties such as plausibility, compared to counterfactual algorithms based on perturbations. CF-K is essentially slightly perturbing the native counterfactuals, so will the returned counterfactual explanations still be more plausible than the explanations from perturbed-based algorithms? Plausibility estimates the closeness of the counterfactual to the data manifold, by measuring the closeness to the nearest instance(s) in the training data [5, 9]. We report the average distance to the nearest and the 5-nearest neighbors for all settings of NICE (*none*, *proximity*, *sparsity*, *plausibility*). As explained before, only the *None* setting refers to a *native* counterfactual that will be grounded in the dataset, and the other settings will be perturbation-based counterfactual algorithms that aim

Just Accepted



(a) Adult Income (b) CMC (c) German (d) Heart (e) Informs (f) Hospital

Fig. 5. Comparison of the average NCP between the majority and minority group

Just Accepted

to optimize for proximity, sparsity, and plausibility [5]. We see in the benchmark study of Brughmans et al. [5] that the native counterfactual algorithms such as WIT and NICE (*None*) will result in the best plausibility scores, followed by NICE (*plausibility*), which is to be expected as it is designed to optimize for this metric. NICE (*plausibility*) outperformed all other perturbation-based algorithms, so this algorithm is chosen to compare with.

	NICE (none)	NICE (sparse)	NICE (prox)	NICE (plaus)	CF-K (k=5)	CF-K (k=10)	CF-K (k=20)
1NN	0	2.77	2.94	2.48	0.84	**1.22**	1.32
5NN	2.64	3.73	3.81	3.54	2.72	2.80	2.83

Table 8. Plausibility results for various settings of the NICE algorithm and CF-K, lower values are better (closer to the data manifold). NICE (*None*) returns *native* counterfactuals, the other settings of NICE (*sparse*, *prox* and *plaus*) return perturbed counterfactuals, and CF-K returns the anonymized version of the *native* counterfactuals (which will thus be slightly perturbed).

We also calculate the distance to the nearest and the 5-nearest neighbors for the anonymous counterfactual instances generated by CF-K (for different privacy settings).[16] The results for the *German* dataset can be seen in Table 8. NICE (*None*) still reports the best results, but CF-K significantly outperforms the other perturbation-based counterfactual algorithms, even NICE (*plausibility*). Furthermore, these other settings of NICE still start from an instance in the training set, so while they are less likely to return real instances, it is still a possibility. This is why for an optimal level of plausibility and a guarantee of privacy, it is better to use CF-K. Furthermore, we are also interested in the relationship between plausibility and privacy. When we increase the level of privacy protection, what is the effect on the plausibility of the k-anonymous explanations? We see in Table 8 that the plausibility metrics will deteriorate when we increase the level of privacy protection, which shows another side of the privacy-explainability trade-off.

7 DISCUSSION AND FUTURE RESEARCH

Transparency in machine learning has become a major topic, yet there is little research on the resulting potential risks to user privacy [41]. Although research has shown that offering model explanations may come at the cost of user privacy [48, 52], none of the currently offered model explanation technologies offer any privacy guarantees. Once such explanation systems are deployed on high-stakes data, such as financial transactions or patient health records, a formal investigation of privacy risks is necessary. In this research, we introduce the *explanation linkage attack*, constituting the privacy risk that some counterfactual explanation techniques pose to the privacy of data subjects, because adversaries can infer their private attributes. We are the first to apply k-anonymity on counterfactual explanations instead of on the complete dataset, and show that applying k-anonymity only on the counterfactual explanations can achieve lower *information loss* and higher *counterfactual validity*. Furthermore, we see that if we increase the privacy constraints, the quality of the explanations becomes worse, which demonstrates the trade-off between *privacy* and *transparency*.

Other researchers [41, 48] have stated that assessing the privacy/explainability trade-off for minority groups is a promising avenue for future exploration, which is what we explored in Section 6.3. We noticed that the average information loss tends to be higher for minority groups, and this difference increases with the level of privacy, hereby introducing a new element of unfairness.

A debate on explanation quality could also be a promising avenue for future research. For k-anonymous counterfactual explanations that have a pureness of 100%, generalized quasi-identifiers might actually be an advantage instead of a drawback. Think about the following scenario: Would you prefer the explanation *'If you would have been a teacher and would have earned $10K more, then you would have received the loan'* or the explanation *'If you would have been a teacher* or a nurse *and would have earned $10K more, then you would have received the loan'*, if both explanations are valid? While generalizing instances in a dataset means less information value, this trade-off is less clear in counterfactual explanations; generalizing them might give you more options to achieve the required target outcome and thus be *more* valuable. However, this is only the case when the counterfactual explanations are entirely *valid* and have a pureness of 100%. A discussion on explanation quality was not the goal of this study, so we leave this as an avenue for future research.

[16]We measure the distance to a generalized counterfactual instance in a conservative way: We sample 100 times a value combination out of the generalized counterfactual instance (as we did to calculate pureness), and calculate its distance to its nearest and 5-nearest neighbors. For one generalized counterfactual instance, we then take the average distance over the 100 samples.

Just Accepted

We also foresee another way to implement privacy constraints in future research, where the explanation technique itself is adapted to have privacy guarantees, instead of enforcing it in post-processing. Other authors propose a methodology where they search for a group of counterfactual explanations for a group of instances [7]. They do not include any privacy guarantees yet, but this kind of set-up could be used to create anonymized explanations as well. This could also have other desired side effects such as more robust explanations.

A last direction for future research we envision is applying other privacy schemes to counterfactual explanations. Beyond k-anonymity, other widely accepted protection schemes include l-diversity [34], t-closeness [30] and differential privacy [11]. K-anonymity can be prone to privacy risks, for example when the attacker has background knowledge or when there is little diversity in the private attributes. l- diversity tries to solve these issues by requiring that the private attribute(s) should have a minimum of l properly depicted values. T-closeness goes even further and requires that the distance between the distribution of the private attribute in any equivalence class and the distribution in the whole table is less than a threshold t. Differential privacy offers a broader approach that captures the increased risk to one's privacy incurred by participating in a database, and counters this by introducing controlled noise into the data. Up until now, we assumed that all the attributes in the dataset except the quasi-identifiers, are private attributes, so l-diversity and t-closeness might not be that straightforward to implement. It is also important to note that we are explicitly searching for *no* diversity in the target variable, as we want k-anonymous counterfactual explanations that are as *pure* as possible. It will be interesting to see how applying these other privacy schemes (l-diversity, t-closeness and differential privacy) affect the explanations, and whether they will have the same implications regarding the explanation quality and fairness.

ACKNOWLEDGMENTS

This research was funded by Research Foundation—Flanders grant number 11N7723N.

REFERENCES

[1] Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. 2020. Model extraction from counterfactual explanations. *arXiv preprint arXiv:2009.01884* (2020).

[2] André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. 2021. Evaluating Robustness of Counterfactual Explanations. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 01–09.

[3] Vanessa Ayala-Rivera, Patrick McDonagh, Thomas Cerqueus, Liam Murphy, et al. 2014. A systematic comparison and evaluation of k-anonymization algorithms for practitioners. *Transactions on data privacy* 7, 3 (2014), 337–370.

[4] Roberto J Bayardo and Rakesh Agrawal. 2005. Data privacy through optimal k-anonymization. In *21st International conference on data engineering (ICDE'05)*. IEEE, 217–228.

[5] Dieter Brughmans, Pieter Leyman, and David Martens. 2023. Nice: an algorithm for nearest instance counterfactual explanations. *Data Mining and Knowledge Discovery* (2023), 1–39.

[6] Tobias Budig, Selina Herrmann, and Alexander Dietz. 2021. Trade-offs Between Privacy-Preserving and Explainable Machine Learning in Healthcare. *cii Student Papers-2021* (2021), 59.

[7] Emilio Carrizosa, Jasone Ramírez-Ayerbe, and Dolores Romero. 2021. *Generating collective counterfactual explanations in score-based classification via mathematical optimization.* Technical Report. Technical report, IMUS, Sevilla, Spain, https://www. researchgate. net ....

[8] Hongyan Chang and Reza Shokri. 2021. On the privacy risks of algorithmic fairness. IEEE, 292–303.

[9] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. 2020. Multi-objective counterfactual explanations. In *Parallel Problem Solving from Nature–PPSN XVI: 16th International Conference, PPSN 2020, Leiden, The Netherlands, September 5-9, 2020, Proceedings, Part I*. Springer, 448–469.

[10] Raphael Mazzine Barbosa de Oliveira and David Martens. 2021. A framework and benchmarking study for counterfactual generating methods on tabular data. *Applied Sciences* 11, 16 (2021), 7274.

[11] Cynthia Dwork. 2006. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*. Springer, 1–12.

[12] Khaled El Emam and Fida Kamal Dankar. 2008. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association* 15, 5 (2008), 627–637.

[13] Khaled El Emam, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, et al. 2009. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association* 16, 5 (2009), 670–682.

[14] Mark Elliot and Angela Dale. 1999. Scenarios of attack: the data intruder's perspective on statistical disclosure risk. *Netherlands Official Statistics* 14, Spring (1999), 6–10.

Just Accepted

[15] European Commission. 2021. Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence. European Commission, Online, accessed February 24, 2022.

[16] Thomas A Feo and Mauricio GC Resende. 1995. Greedy randomized adaptive search procedures. *Journal of global optimization* 6, 2 (1995), 109–133.

[17] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 1322–1333.

[18] Gabriel Ghinita, Panagiotis Karras, Panos Kalnis, and Nikos Mamoulis. 2007. Fast data anonymization with low information loss. In *Proceedings of the 33rd international conference on Very large data bases*. 758–769.

[19] Gabriel Ghinita, Panagiotis Karras, Panos Kalnis, and Nikos Mamoulis. 2009. A framework for efficient data anonymization under privacy and accuracy constraints. *ACM Transactions on Database Systems (TODS)* 34, 2 (2009), 1–47.

[20] Aristides Gionis and Tamir Tassa. 2008. k-Anonymization with minimal loss of information. *IEEE Transactions on Knowledge and Data Engineering* 21, 2 (2008), 206–219.

[21] Sofie Goethals, David Martens, and Theodoros Evgeniou. 2022. The non-linear nature of the cost of comprehensibility. *Journal of Big Data* 9, 1 (2022), 30.

[22] Riccardo Guidotti. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* (2022), 1–55.

[23] Joseph F. Healey, Andi Stepnick, and Eileen O'Brien. 2019. *Race, ethnicity, gender, & class : the sociology of group conflict and change* (eigth ed.). SAGE.

[24] Vijay S Iyengar. 2002. Transforming data to satisfy privacy constraints. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 279–288.

[25] Mark T Keane and Barry Smyth. 2020. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai). In *International Conference on Case-Based Reasoning*. Springer, 163–178.

[26] Slava Kisilevich, Yuval Elovici, Bracha Shapira, and Lior Rokach. 2008. kACTUS 2: Privacy preserving in classification tasks using k-Anonymity. In *Annual Workshop on Information Privacy and National Security*. Springer, 63–81.

[27] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2019. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. *arXiv preprint arXiv:1907.09294* (2019).

[28] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. 2005. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. 49–60.

[29] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. 2006. Mondrian multidimensional k-anonymity. In *22nd International conference on data engineering (ICDE'06)*. IEEE, 25–25.

[30] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2006. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*. IEEE, 106–115.

[31] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2021. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–36.

[32] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil Jain, and Jiliang Tang. 2022. Trustworthy AI: A computational perspective. *ACM Transactions on Intelligent Systems and Technology* 14, 1 (2022), 1–59.

[33] Grigorios Loukides and Aris Gkoulalas-Divanis. 2012. Utility-preserving transaction data anonymization with low information loss. *Expert systems with applications* 39, 10 (2012), 9764–9777.

[34] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 3–es.

[35] Catherine Marsh, Chris Skinner, Sara Arber, Bruce Penhale, Stan Openshaw, John Hobcraft, Denise Lievesley, and Nigel Walford. 1991. The case for samples of anonymized records from the 1991 census. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 154, 2 (1991), 305–340.

[36] David Martens. 2022. *Data Science Ethics: Concepts, Techniques, and Cautionary Tales*. Oxford University Press.

[37] David Martens and Foster Provost. 2014. Explaining data-driven document classifications. *MIS quarterly* 38, 1 (2014), 73–100.

[38] Adam Meyerson and Ryan Williams. 2004. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 223–228.

[39] Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.

[40] Francesca Naretto, Anna Monreale, and Fosca Giannotti. 2022. Evaluating the Privacy Exposure of Interpretable Global Explainers. In *2022 IEEE 4th International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 13–19.

[41] Neel Patel, Reza Shokri, and Yair Zick. 2022. Model explanations with differential privacy. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1895–1904.

[42] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 809–818.

Just Accepted

[43] Martin Pawelczyk, Himabindu Lakkaraju, and Seth Neel. 2022. On the Privacy Risks of Algorithmic Recourse. *arXiv preprint arXiv:2211.05427* (2022).
[44] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. FACE: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 344–350.
[45] Md Ileas Pramanik, Raymond YK Lau, and Wenping Zhang. 2016. K-anonymity through the enhanced clustering method. In *2016 IEEE 13th International Conference on e-Business Engineering (ICEBE)*. IEEE, 85–91.
[46] Pengrui Quan, Supriyo Chakraborty, Jeya Vikranth Jeyakumar, and Mani Srivastava. 2022. On the amplification of security and privacy risks by post-hoc explanations in machine learning models. *arXiv preprint arXiv:2206.14004* (2022).
[47] Maria Rigaki and Sebastian Garcia. 2020. A survey of privacy attacks in machine learning. *arXiv preprint arXiv:2007.07646* (2020).
[48] Reza Shokri, Martin Strobel, and Yair Zick. 2019. Privacy Risks of Explaining Machine Learning Models. *CoRR* abs/1907.00164 (2019). http://arxiv.org/abs/1907.00164
[49] Reza Shokri, Martin Strobel, and Yair Zick. 2020. Exploiting transparency measures for membership inference: a cautionary tale. In *The AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI). AAAI*, Vol. 13.
[50] MS Simi, K Sankara Nayaki, and M Sudheep Elayidom. 2017. An extensive study on data anonymization algorithms based on k-anonymity. In *IOP Conference Series: Materials Science and Engineering*, Vol. 225. IOP Publishing, 012279.
[51] Djordje Slijepčević, Maximilian Henzl, Lukas Daniel Klausner, Tobias Dam, Peter Kieseberg, and Matthias Zeppelzauer. 2021. k-Anonymity in practice: How generalisation and suppression affect machine learning classifiers. *Computers & Security* 111 (2021), 102488.
[52] Kacper Sokol and Peter A Flach. 2019. Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety. In *SafeAI@ AAAI*.
[53] M Sumana and KS Hareesha. 2010. Anonymity: an assessment and perspective in privacy preserving data mining. *International Journal of Computer Applications* 6, 10 (2010).
[54] Latanya Sweeney. 2000. Simple demographics often identify people uniquely. *Health (San Francisco)* 671, 2000 (2000), 1–34.
[55] Latanya Sweeney. 2002. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 571–588.
[56] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.
[57] Henk CA Van Tilborg and Sushil Jajodia. 2014. *Encyclopedia of cryptography and security*. Springer Science & Business Media.
[58] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.
[59] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.
[60] Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Ada Wai-Chee Fu. 2006. Utility-based anonymization for privacy preservation with less information loss. *ACM SIGKDD Explorations Newsletter* 8, 2 (2006), 21–30.
[61] Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Ada Wai-Chee Fu. 2006. Utility-based anonymization using local recoding. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 785–790.

Just Accepted

A RESULTS WITH DIFFERENT VALUES FOR k

Table 9. Results of CF-K (k = 5).

Dataset	Adult	CMC	German	Heart	Hospital	Informs
NCP (mean)	0.18%	2.07%	14.53%	0.93%	1.97%	7.34%
Pureness (mean)	99.69%	96.24%	99.85%	100%	95.31%	89.16%
Execution time (mean)	16.82s	12.67s	6.25s	1.92s	14.93s	25.04s
C_{DM}	83,990	3,584	576	450	12,809	4,755
$\frac{C_{DM}}{\#explanations}$	106.72	8.83	9.6	8.04	17.17	7.19
CM	0.81	0.24	0.07	0.25	0.71	0.11

Table 10. Results of Mondrian (k = 5)

Dataset	Adult	CMC	German	Heart	Hospital	Informs
NCP (mean)	9.28%	4.95%	42.02%	24.44%	14.72%	29.40%
Pureness (mean)	92.30%	79.85%	93.82%	100%	71.05%	74.39%
Execution time (mean)	16.65s	1.56s	0.56s	0.36s	2.48s	2.64s
C_{DM} **(mean)**	116,132	4,348	402	486	12,524	4,607
$\frac{C_{DM}}{\#explanations}$	147.56	10.71	6.07	8.68	16.79	6.97
CM (mean)	0.82	0.24	0.22	0.32	0.73	0.31

Table 11. Results of CF-K (k = 20).

Dataset	Adult	CMC	German	Heart	Hospital	Informs
NCP (mean)	0.84%	7.46%	27.38%	5.02%	5.35%	12.86%
Pureness (mean)	99.61%	88.85%	99.08%	100%	88.85%	82.46%
Execution time (mean)	32.69s	19.29s	25.44s	4.15s	27.61s	86.101s
C_{DM}	93,597	9,509	1,541	1,329	24,442	15,976
$\frac{C_{DM}}{\#explanations}$	118.93	23.42	25.68	23.73	32.76	24.17
CM	0.84	0.25	0.02	0.39	0.85	0.14

Table 12. Results of Mondrian (k = 20)

Dataset	Adult	CMC	German	Heart	Hospital	Informs
NCP (mean)	23.32%	12.35%	71.39%	57.44%	38.63%	42.62%
Pureness (mean)	85.01%	59.40%	89.52%	100%	58.08%	72.88%
Execution time (mean)	16.39s	2.21s	0.72s	0.42s	2.77s	3.06s
C_{DM} **(mean)**	126,238	14,557	1,832	2,003	25,566	20,242
$\frac{C_{DM}}{\#explanations}$	160.40	35,85	30.53	35.77	34.27	30.62
CM (mean)	0.82	0.20	0.13	0.43	0.87	0.33

Just Accepted