


Certifiable Robustness for Nearest Neighbor Classifiers

Austen Z. Fan  

Department of Computer Sciences, University of Wisconsin-Madison, WI, USA

Paraschos Koutris   

Department of Computer Sciences, University of Wisconsin-Madison, WI, USA

Abstract

ML models are typically trained using large datasets of high quality. However, training datasets often contain inconsistent or incomplete data. To tackle this issue, one solution is to develop algorithms that can check whether a prediction of a model is *certifiably robust*. Given a learning algorithm that produces a classifier and given an example at test time, a classification outcome is certifiably robust if it is predicted by every model trained across all possible worlds (repairs) of the uncertain (inconsistent) dataset. This notion of robustness falls naturally under the framework of certain answers. In this paper, we study the complexity of certifying robustness for a simple but widely deployed classification algorithm, k -Nearest Neighbors (k -NN). Our main focus is on inconsistent datasets when the integrity constraints are functional dependencies (FDs). For this setting, we establish a dichotomy in the complexity of certifying robustness w.r.t. the set of FDs: the problem either admits a polynomial time algorithm, or it is **coNP-hard**. Additionally, we exhibit a similar dichotomy for the counting version of the problem, where the goal is to count the number of possible worlds that predict a certain label. As a byproduct of our study, we also establish the complexity of a problem related to finding an optimal subset repair that may be of independent interest.

2012 ACM Subject Classification Theory of computation → Database theory; Theory of computation → Incomplete, inconsistent, and uncertain databases

Keywords and phrases Inconsistent databases, k -NN classification, certifiable robustness

Digital Object Identifier 10.4230/LIPIcs.ICDT.2022.6

Related Version *Full Version*: <https://arxiv.org/abs/2201.04770> [7]

Funding This research was supported in part by National Science Foundation grants CRII-1850348 and III-1910014, as well as a gift by Google.

1 Introduction

Machine Learning (ML) has been widely adopted as a central tool in business analytics, medical decisions, autonomous driving, and many other domains. In supervised learning settings, ML models are typically trained using large datasets of high quality. However, real-world training datasets often contain incorrect or incomplete data. For example, attribute values may be missing from the dataset, attribute values may be wrong, or the dataset can violate integrity constraints. Several approaches to tackle this problem have been proposed in the literature, including data cleaning [17, 23] and robust learning methods [4, 26, 5].

In this work, we study an alternative but complementary approach using the framework of *certain answers*, which has been a focus of the database community in the last two decades [1, 2]. Under this framework, an inconsistent or incomplete training dataset is modeled as a set of possible worlds called *repairs*. We can think of a repair as a way to clean the dataset such that it is consistent (w.r.t. a set of integrity constraints) and complete. In the context of query answering, certain answers are the output tuples that exist in the query result of every possible world. In other words, we will obtain a certain answer in the output no matter how we choose to repair the dataset.



© Austen Z. Fan and Paraschos Koutris;
licensed under Creative Commons License CC-BY 4.0
25th International Conference on Database Theory (ICDT 2022).

Editors: Dan Olteanu and Nils Vortmeier; Article No. 6; pp. 6:1–6:20
Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

	A	B	C	label
t_1	1	0	a	0
t_2	1	2	b	0
t_3	2	0	a	2
t_4	2	5	c	1
t_5	3	1	a	0
t_6	4	2	d	2

The notion of certain answers in the context of ML is known as certifiable (or certified) robustness [3, 26]. Given a learning algorithm that produces a classifier and a tuple at test time, we say that a predicted label is *certifiably robust* if it is predicted by every model trained across all possible worlds. In other words, certifiably robust predictions come with a formal guarantee that the label is robust to how the training dataset is fixed. Such a guarantee can be beneficial to decide whether we should trust the predictions of the ML model or whether we should spend resources to clean the dataset before feeding it to the learning algorithm.

As a first step towards certifying robustness for other more complex ML algorithms, we focus in this work on k -Nearest Neighbor classification (k -NN). In this problem, we start with a d -dimensional dataset equipped with a distance function. Given a test point x , the classifier first finds the k points closest to x w.r.t. the given distance function, and assigns to x the label that is a plurality among these. Certified robustness for k -NNs was recently explored by Karlas et. al [13] under the uncertain model where tuples (training points) with the same label are grouped into blocks, and a possible world is constructed by picking independently exactly one tuple from each block. This setting is equivalent to considering the subset repairs of an inconsistent dataset where the only integrity constraint is a primary key, with the additional restriction that training points with the same key must have the same label.

In this paper, we generalize the study of certifying robustness for k -NNs to other inconsistent and uncertain models. Specifically, we consider subset repairs of a dataset where the integrity constraints can be any set of functional dependencies (FDs) and not only a primary key as in [13].

► **Example 1.** Consider the inconsistent dataset in the table, where the integrity constraint is the FD $A \rightarrow B$. Let the distance function between two tuples s, t be $f(s, t) = |s[A] - t[A]| + |s[B] - t[B]|$. Suppose that we want to label the test point $x = (0, 0, d)$ using a 3-NN classifier. This induces the following ordering of the tuples w.r.t. their distance from x (for convenience, we include the labels as well):

$$t_1 : \mathbf{0} < t_3 : \mathbf{2} < t_2 : \mathbf{0} < t_5 : \mathbf{0} < t_6 : \mathbf{2} < t_4 : \mathbf{1}$$

A repair for this inconsistent dataset has to choose one tuple from $\{t_1, t_2\}$ and one tuple from $\{t_3, t_4\}$. There are 4 possible repairs, which form the following 3-neighborhoods around x :

$$\begin{aligned} \{t_1 : \mathbf{0}, t_5 : \mathbf{0}, t_6 : \mathbf{2}\}, & \quad \{t_2 : \mathbf{0}, t_5 : \mathbf{0}, t_6 : \mathbf{2}\} \\ \{t_1 : \mathbf{2}, t_3 : \mathbf{0}, t_5 : \mathbf{0}\}, & \quad \{t_3 : \mathbf{2}, t_2 : \mathbf{0}, t_5 : \mathbf{0}\} \end{aligned}$$

In all repairs, label 0 occurs two times, and hence it is always the majority label. Hence, we can certify that 0 is a robust label for tuple x .

We show that for general sets of FDs the complexity landscape for certified robustness admits a dichotomy: it is computationally intractable for some sets of FDs, and is in polynomial time for the other sets of FDs. We also investigate certifying robustness for

other widely used uncertain models, including Codd-tables [11], or-sets and ?-sets [25]. We establish that in these settings the problem can always be solved in polynomial time.

Our work shows that the logical structure of the errors (or missing values) from a training set can be exploited to construct fast algorithms for certifiable robustness. Tools developed in the database theory community can facilitate the design of these algorithms. We also demonstrate that, even for the relatively simple k -NN classifier, the complexity landscape exhibits a complex behavior that is related to other problems in consistent query answering.

Our Contribution. We now present in more detail the contributions of this work:

- We establish a complexity dichotomy for certifying robustness for k -NNs under subset repairs (Section 4) into P and coNP-complete. The dichotomy depends on the structure of FDs. More precisely, the syntactic condition for the dichotomy is based on the notion of *lhs chains*. In fact, it is the same as the one used for the complexity classification of the problem of counting the number of subset repairs under a set of FDs [21]. In the case where the only FD constraint is a primary key, we show that we can design an even faster algorithm that runs in linear time in the size of the data, improving upon the running time of the algorithm in [13] (Section 5).
- In addition to certified robustness, we study the related problem of *counting* the number of repairs that predict a given label for a test point (Section 8). We establish a dichotomy into the complexity classes FP and #P-complete with the same syntactic condition as the decision problem. The polynomial time algorithm here depends exponentially on the number of classification labels, in contrast to our algorithm for certifiable robustness which has a linear dependence on the number of labels.
- We show that certifying robustness for k -NNs is tightly connected to the problem of finding the *subset repair with the smallest total weight*, when each tuple is assigned a positive weight (Section 7). As a consequence, we obtain a dichotomy result for that problem as well. Note that this problem is a symmetric variant of the problem in [20], which asks instead for the repair with the maximum total weight.
- Finally, we investigate the complexity of certifiable robustness for k -NNs for three widely used uncertain models: Codd-tables, ?-sets and or-sets (Section 9). We show that for all the above models, certifying robustness for k -NN classification admits a polynomial time algorithm.

2 Related Work

Certain Query Answering. There has been a long line of research in *Certain Query Answering* (CQA) in the database community. Data consistency might be violated, for example, during data integration or in a data warehouse. It is then natural to ask: given a query and such inconsistent data, can we answer the query with a certain answer? Arenas, Bertossi, and Chomicki [1] first define the notion of a repair, which refers to a consistent subinstance that is minimally different from the inconsistent data. A certain answer to the query is defined as an answer that will result from every repair. Beginning from the work of Fuxman and Miller [8], more general dichotomy results in CQA have been proven [14, 15, 16]. A dichotomy theorem for a class of queries and integrity constraints says CQA is either in polynomial time or coNP-complete, usually depending on an efficiently checkable criterion for tractability. Certain answers have also been studied in the context of incomplete models of data [19, 22, 10].

Subset Repairs. Livshits et. al [21] studied the problem of counting the number of subset repairs w.r.t. a given set of FDs, establishing a complexity dichotomy. The syntactic condition for tractability (existence of a lhs chain) is the same as the one we establish for certifiable robustness in k -NN classification. Livshits et. al [20] also studied the problem of finding a maximum-weight subset repair w.r.t. a given set of FDs, and showed that the complexity observes a dichotomy. In this paper we study the symmetric problem of finding a minimum-weight subset repair, and show that the problem also exhibits a complexity dichotomy, albeit the condition for tractability is again the existence of a lhs chain.

Certifiable Robustness in ML. Robust learning methods are used to learn over inconsistent or incomplete datasets. For example, [5] discusses a sound verification technique which proves whether a prediction is robust to data poisoning in decision-tree models. There is also a line of work on smoothing techniques for ML robustness [3, 12, 24, 18], where added random noise, usually with a Gaussian distribution, will sometimes boost the robustness of the model against adversarial attacks such as label-flipping. Our approach is different in that we prove a dichotomy for k -NN certifiable robustness, i.e. *either* we can assert that the dataset will always lead to the same prediction efficiently *or* it is **coNP-complete** to do so, with an efficiently testable criterion. We show how to extend our model to capture scenarios including uncertain labels, weighted tuples, and data poisoning.

3 Preliminaries

In this paper, we consider relation schemas of the form $R(A_1, \dots, A_d)$ with arity d . The attributes A_1, \dots, A_d take values from a (possibly infinite) domain \mathbb{D} . Given a tuple t in an instance I over R , we will use $t[A_i]$ to denote its value at attribute A_i . It will be convenient to interpret an instance I as a training set of points in the d -dimensional space \mathbb{D}^d . We will use the terms point/tuple interchangeably in the paper.

An *uncertain instance* \mathcal{I} over a schema $R(A_1, \dots, A_d)$ is a set of instances over the schema. We will often refer to each instance in \mathcal{I} as a *possible world*. We will see later different ways in which we can define uncertain instances implicitly.

For each tuple t that occurs in some possible world in \mathcal{I} , we associate a *label* $L(t)$ which takes values from a finite set \mathcal{Y} . We will say that the uncertain instance \mathcal{I} equipped with the labeling function L is a *labeled uncertain instance* over the schema $R(A_1, \dots, A_d)$. We similarly define a labeled instance I .

Certifiable Robustness. In this work, we will focus on the classification task. Let \mathcal{L} be a learning algorithm that takes as training set a labeled instance I over the schema $R(A_1, \dots, A_d)$, and returns a *classifier*, which is a total function $\mathcal{L}_I : \mathbb{D}^d \rightarrow \mathcal{Y}$.

► **Definition 2** (Certifiable Robustness). *Let \mathcal{I} be a labeled uncertain instance over $R(A_1, \dots, A_d)$ and \mathcal{L} be a classification learning algorithm with labels in \mathcal{Y} . We say that a (test) point $x \in \mathbb{D}^d$ is certifiably robust in \mathcal{I} if there exists a label $\ell \in \mathcal{Y}$ such that for every possible world $I \in \mathcal{I}$, $\mathcal{L}_I(x) = \ell$. The label ℓ is called a certain label for x .*

In other words, suppose we call ℓ a *possible label* for x if there exists some possible world $I \in \mathcal{I}$ for which $\mathcal{L}_I(x) = \ell$, then certifiable robustness simply means that there is only one possible label for x .

Nearest Neighbor Classifiers. In k -NN classification, we are given a labeled instance I over $R(A_1, \dots, A_d)$, along with a distance function f over \mathbb{D}^d . Given a point $x \in \mathbb{D}^d$, the classifier first finds the k -neighborhood $\mathcal{N}_k(x, I)$, which consists of the k points closest to x w.r.t. the distance function f . Then, the classifier assigns to x the label that is a plurality among $\mathcal{N}_k(x, I)$. When $k = 1$, the classifier returns the label of the point that is closest to x w.r.t. the distance function f . When $|\mathcal{Y}| = 2$, we are performing binary classification. We will also consider the generalization of k -NN where each tuple has a positive weight, and the classifier assigns the label with the largest total weight.

For this work, we require the following tie-breaking mechanisms: (i) if there are two labels in $\mathcal{N}_k(x, I)$ with the maximum number, then we say x is not certifiably robust for k -NN classification, and (ii) if there are more tuples with the same distance to the test point that will make $\mathcal{N}_k(x, I)$ not well-defined, we will break ties according to a predefined ordering of the tuples in the instance. By a slight abuse of notation, throughout the proof when we say $\mathcal{L}_I(x) = \ell$, we mean the number of tuples labeled ℓ is *strictly* more than that of any other labels for any choices made to pick $\mathcal{N}_k(x, I)$.

Functional Dependencies. A functional dependency (FD) is an expression of the form $X \rightarrow Y$, where X and Y are sets of attributes from R . An instance I over R satisfies $X \rightarrow Y$ if for every two tuples in I , if they agree on X they must also agree on Y . We say that I satisfies a set of functional dependencies Σ if it satisfies every functional dependency in Σ . For an attribute A and set of FDs Σ , we define $\Sigma - A$ to be the FD set where we have removed from any FD in Σ the attribute A . An *FD schema* \mathbf{R} is a pair $(R(A_1, \dots, A_d), \Sigma)$, where Σ is a set of FDs defined over R .

Repairs. Given Σ , assume that we have an inconsistent instance D that violates the functional dependencies in Σ . We say that D' is a *repair* of D if it is a maximal subset of D that satisfies Σ . In other words, we can create a repair by removing the smallest possible number of tuples from D such that all the integrity constraints are satisfied. We will use $I_\Sigma(D)$ to denote the set of all possible repairs of D w.r.t. the FD schema Σ . If the instance D is consistent, namely it does not violate any functional dependency, then $I_\Sigma(D)$ is defined to be D itself.

3.1 Problem Definitions

Although our algorithms work for any distance function such that $f(x, x')$ can be computed in time $O(1)$ (assuming the dimension d is fixed), for the hardness results we need a more precise formalization. We consider two variants of the problem. In the first variant, we will consider a specific distance function, the p -norm when the domain is $\mathbb{D} = \mathbb{R}$. Recall that for any $p \geq 1$, the p -norm is

$$\|x - x'\|_p = \left(\sum_{i=1}^d |x[A_i] - x'[A_i]|^p \right)^{1/p}$$

For the following definitions, we fix the dimension d and the label set \mathcal{Y} . Formally, we can now define the following decision problem, parameterized by an FD schema \mathbf{R} and an integer $k > 0$.

► **Definition 3** ($\text{CR-NN}_p(\mathbf{R}, k)$). *Given an inconsistent labeled instance D over an FD schema \mathbf{R} and a test point x , is x certifiably robust in $I_\Sigma(D)$ for k -NN classification w.r.t. the p -norm?*

6:6 Certifiable Robustness for Nearest Neighbor Classifiers

Note that k is fixed in the above problem. We can also define the decision problem where k is instead an input parameter, denoted as $\text{CR-NN}_p\langle\mathbf{R}\rangle$.

In the second variant of the problem, instead of fixing a distance function, we will directly provide as input to the problem a strict ordering of the points in the dataset D w.r.t. their distance from the test point x . From an algorithmic point of view this does not make much difference, since we can compute the ordering in time $O(|D| \log |D|)$ for any distance function that can be computed in time $O(1)$.

► **Definition 4** ($\text{CR-NN}_{<}\langle\mathbf{R}, k\rangle$). *Given an inconsistent labeled instance D over an FD schema \mathbf{R} and a strict ordering of the points in D w.r.t. their distance from a test point x , is x certifiably robust in $I_\Sigma(D)$ for k -NN classification?*

Similarly we also define the problem $\text{CR-NN}_{<}\langle\mathbf{R}\rangle$ with the parameter k as part of the input. Note here that there is a straightforward many-one polynomial time reduction from $\text{CR-NN}_p\langle\mathbf{R}, k\rangle$ to $\text{CR-NN}_{<}\langle\mathbf{R}, k\rangle$.

Finally, we define the counting variant of the problem. Given an inconsistent instance D and a label $\ell \in \mathcal{Y}$, we want to count how many repairs of D will predict label ℓ .

► **Definition 5** ($\#\text{CR-NN}_{<}\langle\mathbf{R}\rangle$). *Given an inconsistent labeled instance D over an FD schema \mathbf{R} , a strict ordering of the points in D w.r.t. their distance from a test point x , and a label $\ell \in \mathcal{Y}$, output the number of repairs in $I_\Sigma(D)$ for which the k -NN classifier assigns label ℓ to x .*

Similarly, one can define the counting question $\#\text{CR-NN}_p\langle\mathbf{R}\rangle$.

4 Main Results

In this section, we present and discuss our key results. The main dichotomy theorem relies on the notion of lhs chains for an FD schema, as defined in [21].

► **Definition 6** (lhs Chain). *A set of FDs Σ has a left-hand-side chain (lhs chain for short) if for every two FDs $X_1 \rightarrow Y_1$ and $X_2 \rightarrow Y_2$ in Σ , either $X_1 \subseteq X_2$ or $X_2 \subseteq X_1$.*

One can determine efficiently whether an FD schema is equivalent to one with an lhs chain or not [21].

► **Example 7.** Consider the relational schema $R(A, B, C, D)$. The FD set $\{A \rightarrow C, B \rightarrow C\}$ does not have an lhs-chain, since neither of the two left-hand-sides of the FDs are contained in each other. The FD set $\{AB \rightarrow C, B \rightarrow D\}$ on the other hand has an lhs chain.

We are now ready to state our main theorem.

► **Theorem 8** (Main Theorem). *Let \mathbf{R} be an FD schema. Then, the following hold:*

- *If \mathbf{R} is equivalent to an FD schema with an lhs chain, then $\text{CR-NN}_{<}\langle\mathbf{R}\rangle$ (and thus $\text{CR-NN}_p\langle\mathbf{R}\rangle$) is in P .*
- *Otherwise, for any integer $k \geq 1$, $\text{CR-NN}_p\langle\mathbf{R}, k\rangle$ (and thus $\text{CR-NN}_{<}\langle\mathbf{R}, k\rangle$) is coNP-complete.*

Moreover, it can be decided in polynomial time which of the two cases holds.

We show the polynomial time algorithm in Section 5 and the hardness proof in Section 6. We should discuss three things at this point. First, the polynomial time algorithm works for any distance function, as long as we can compute the distance between any two points in time $O(1)$. Indeed, the distance function is only needed to compute the order of tuples in the

dataset according to their distance from the test point. Second, we show the intractability result for the p -norm distance function, which is widely used in practice. It is likely that the problem remains hard for other popular distance functions as well. Third, as we will see in the next section, the tractable cases are polynomial in k , the size of the neighborhood. This is important, since k is often set to be a function of the training set size, for example \sqrt{n} . For the intractable cases, the problem is already hard even for $k = 1$.

Uncertain Labels. The above dichotomy theorem holds even when we allow inconsistent labels. We can model inconsistent labels by modifying the labelling function $L(t)$ to take values from $\mathcal{P}(\mathcal{Y})$, the power set of the finite label set \mathcal{Y} . The set of possible worlds is then defined to be the set of possible worlds of the inconsistent instance D paired with a labelling function L' such that $L'(t) \in L(t)$ for all $t \in D$. The definition of certifiable robustness carries over to this set of possible worlds.

We can simulate uncertain labels by adding an extra attribute (label) to the schema and an FD $A_1, \dots, A_d \rightarrow \text{label}$. It is easy to see that the new schema is equivalent to one with an lhs chain if and only if the original one is. Thus, we conclude that *uncertain labels do not change the complexity with respect to certifiable robustness*.

Counting. For the counting variant of certifying robustness for k -NNs, we show an analogous dichotomy result.

► **Theorem 9.** *Let \mathbf{R} be an FD schema. Then, the following hold:*

- *If \mathbf{R} is equivalent to an FD schema with an lhs chain, then $\#CR\text{-}NN_{<}(\mathbf{R})$ is in FP.*
- *Otherwise, even $\#CR\text{-}NN_{<}(\mathbf{R}, 1)$ is $\#P$ -complete.*

Moreover, it can be decided in P which of the two cases holds.

5 Tractable Cases

In this section, we prove that if the FD schema \mathbf{R} has an lhs chain, then there is a polynomial time algorithm for $CR\text{-}NN_{<}(\mathbf{R})$ in the size of the inconsistent dataset, the parameter k and the number of possible labels. Then, we show that when \mathbf{R} consists of a single primary key we can construct an even faster algorithm that runs in *linear time* w.r.t the number of tuples, number of labels, and k .

For this section, let D be an inconsistent labeled instance and x be the test point. Assume w.l.o.g. that $\mathcal{Y} = \{1, 2, \dots, m\}$ and let n be the number of tuples in D . We assume that the points in D are already in strict order w.r.t. their distance from x : $t_1 < t_2 < \dots < t_n$.

5.1 An Algorithm for Certifiable Robustness

Note that in the following analysis we fix an FD schema \mathbf{R} with an lhs chain. The algorithm first constructs a repair I by choosing greedily points using the given ordering as long as they do not conflict with each other. This step can be implemented in time $O(n)$ by, say, building a hash map per FD which maps for each tuple the value of the LHS attribute(s) to the value of the RHS attribute(s). Suppose w.l.o.g. that $\mathcal{L}_I(x) = 1$.

As a second step, for every label $\ell \in \{2, \dots, m\}$, we will attempt to construct a repair I' of D such that the number of ℓ -labeled points is at least as many as the number of 1-labeled points in the k -neighborhood of x . Such a repair will be a witness that some other label is possible for x , hence x is not certifiably robust.

6:8 Certifiable Robustness for Nearest Neighbor Classifiers

It will be helpful now to define the following terms for a subinstance $I \subseteq D$, $\tau \in \{1, \dots, n\}$, and a label $\ell \in \mathcal{Y}$:

$$\begin{aligned}\mathcal{N}_\tau^\leq(I) &= \{t_j \in I \mid j \leq \tau\} \\ C_\tau^\leq(\ell, I) &= |\{t \in \mathcal{N}_\tau^\leq(I) \mid L(t) = \ell\}| \end{aligned}$$

We are now ready to present the core component of our algorithm. This component will be executed for every label $\ell > 1$ and $\tau \in \{1, \dots, n\}$. Thus, it will run $O(|\mathcal{Y}| \cdot n)$ times. Define the following quantity for a subinstance $J \subseteq D$, an FD set Δ , and i where $0 \leq i \leq k$:

$$M_i[J, \Delta] = \max\{C_\tau^\leq(\ell, K) - C_\tau^\leq(1, K) \mid K \in I_\Delta(J) \text{ s.t. } |\mathcal{N}_\tau^\leq(K)| = i\}.$$

Here for simplicity we adopt a slight abuse of notation where, although $M_i[J, \Delta]$ depends on τ , τ is not explicitly shown in the notation $M_i[J, \Delta]$. If there is no repair K for J such that $|\mathcal{N}_\tau^\leq(K)| = i$, we define $M_i[J, \Delta] = -\infty$. The key observation is that if $M_k[D, \Sigma] \geq 0$ then ℓ is a possible label for x and hence robustness is falsified. The algorithm computes this quantity using a combination of dynamic programming and recursion on the structure of the FD set.

The Recursive Algorithm. Given $J \subseteq D$ and a set of FDs Δ , we want to compute $M_i[J, \Delta]$ for every $i = 0, \dots, k$. First, we remove from Δ any trivial FDs. Then we distinguish three cases:

Base Case: *the set of FDs is empty.* In this case, $I_\Delta(J) = \{J\}$. For every $i \neq |\mathcal{N}_\tau^\leq(J)|$, $M_i[J, \Delta] = -\infty$. For $i = |\mathcal{N}_\tau^\leq(J)|$ (if $|\mathcal{N}_\tau^\leq(J)| \leq k$), we have $M_i[J, \Delta] = C_\tau^\leq(\ell, J) - C_\tau^\leq(1, J)$, so we can compute this in a straightforward way.

Consensus FD: *there exists an FD $\emptyset \rightarrow A$.* In this case, we recursively call the algorithm to compute $M_i[\sigma_{A=a}(J), \Delta - A]$ for every $a \in \pi_A(J)$. Then, for every $i = 0, \dots, k$:

$$M_i[J, \Delta] = \max_{a \in \pi_A(J)} M_i[\sigma_{A=a}(J), \Delta - A]$$

Common Attribute: *there exists a common attribute A in the lhs of all FDs.* In this case, we recursively call the algorithm to compute $M_i[\sigma_{A=a}(J), \Delta - A]$ for every $a \in \pi_A(J)$. Let $S = \pi_A(J) = \{a_1, \dots, a_{|S|}\}$. Then, for every $i = 0, \dots, k$:

$$M_i[J, \Delta] = \max_{\sum_{a \in S} i_a = i} \sum_{a \in S} M_{i_a}[\sigma_{A=a}(J), \Delta - A]$$

We next discuss how to do the above computation using dynamic programming. We can view $M_{i_a}[\sigma_{A=a}(J), \Delta - A]$ as a matrix with rows indexed by value a of attributes A and columns indexed by i_a with $0 \leq i_a \leq k$. The task is to pick one entry from each row so that the sum of entries is maximized and the column indices of entries sum to i . The dynamic programming computes an $|S| \times (k+1)$ matrix \mathcal{M} where the (i, j) -entry represents the maximum of $\sum_{u=1}^i M_{i_u}[\sigma_{A=a_u}(J), \Delta - A]$ such that $\sum_{u=1}^i i_u = j$ and finally returns the entries $\mathcal{M}[|S|, i]$.

The algorithm runs in polynomial time with respect to the size of D , the parameter k and the number of labels $|\mathcal{Y}|$. For detailed analysis on the correctness of the algorithm and its running time, see [7].

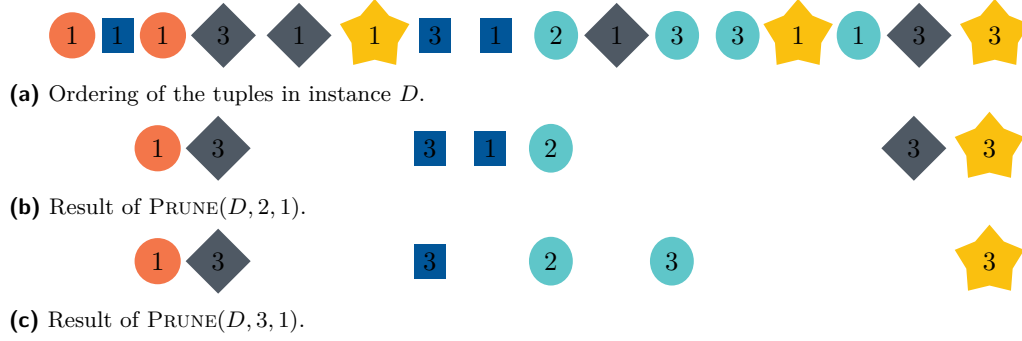
Weighted Majority. The algorithm can also handle the case where we compute the predicted label by weighted majority, where each tuple t is assigned a weight w_t . The only thing we need to modify is the definition of $C_\tau^\leq(\ell, I)$, which now becomes $\sum_{t \in \mathcal{N}_\tau^\leq(I): L(t) = \ell} w_t$.

■ **Algorithm 1** Dynamic Programming.

```

1 for  $j = 0, \dots, k$  do
2    $\mathcal{M}[1, j] \leftarrow M_j[\sigma_{A=a_1}(J), \Delta - A]$ ;
3 for  $i = 2, \dots, |S|$  do
4   for  $j = 0, \dots, k$  do
5      $\mathcal{M}[i, j] \leftarrow \max_u \{ \mathcal{M}[i-1, u] + M_{j-u}[\sigma_{A=a_i}(J), \Delta - A] \}$ ;

```



■ **Figure 1** Running example for the single primary key algorithm. Tuples with the same color/shape belong in the same block.

5.2 A Faster Algorithm for Single Primary Key

The algorithm given in the above section, though in polynomial time, is not very efficient. In this section, we give a faster algorithm for certifiable robustness when the FD schema is equivalent to one with a single primary key. Recall that in this case we need to pick exactly one tuple from the set of tuples that share the same key.

As in the previous section, we will first run k -NN on an arbitrarily chosen repair to obtain a possible label for x (this part needs only linear time). Without any loss of generality, assume that the predicted label for x is 1. For every target label $\ell \in \{2, \dots, m\}$, we will attempt to construct a repair such that ℓ occurs at least as many times as 1 in the k -neighborhood of x . If such a repair exists, then robustness is falsified.

For a tuple t , we denote $\text{key}(t)$ to be its key. The *block* of a tuple t is the set of tuples with the same key. Further, define $C(\ell, I) = |\{t \in \mathcal{N}_k(x, I) \mid L(t) = \ell\}|$.

Suppose now we want to find a repair I such that $C(\ell_2, I) \geq C(\ell_1, I)$. Define $\text{PRUNE}(D, \ell_2, \ell_1)$ to be the dataset obtained from D if we remove any tuple $t \in D$ such that there exists another tuple $t' \in D$ in the same block with:

1. $t < t'$ and $L(t) = \ell_1$; or
2. $t > t'$ and $L(t') = \ell_2$.

► **Lemma 10.** *Let $\ell_1, \ell_2 \in \mathcal{Y}$ and $D^* = \text{PRUNE}(D, \ell_2, \ell_1)$. Then, there exists a repair I of D s.t. $C(\ell_2, I) \geq C(\ell_1, I)$ if and only if there exists a repair I' of D^* with $C(\ell_2, I') \geq C(\ell_1, I')$.*

Lemma 10 tells us that it suffices to consider D^* instead of D . For the proof, see [7].

D^* has the following nice properties:

- every block has at most one tuple with a label from $\{\ell_1, \ell_2\}$.
- any tuple with label in $\{\ell_1, \ell_2\}$ is always the last tuple in its block (i.e. the one furthest away from x).

6:10 Certifiable Robustness for Nearest Neighbor Classifiers

The pruning procedure can be implemented in linear time $O(n)$. Algorithm 2 FASTSCAN now attempts to find the desired repair. It runs in linear time with respect to the size of D and the number of labels $|\mathcal{Y}|$ and, moreover, its time complexity does not depend on k . For detailed analysis on the correctness of the algorithm and its running time, see [7].

Algorithm 2 FASTSCAN.

Input: instance D , test point x , labels ℓ_1, ℓ_2
Output: whether there exists repair I s.t. $C(\ell_2, I) \geq C(\ell_1, I)$

```

1  $D \leftarrow \text{PRUNE}(D, \ell_2, \ell_1)$  ;
2  $n_1, n_2 \leftarrow 0$  ;
3  $B, B^\square \leftarrow \{\}$  ;
4 for  $i \leftarrow 1$  to  $|D|$  do
5    $B \leftarrow B \cup \{\text{key}(t_i)\}$  ;
6   if  $L(t_i) = \ell_2$  then
7      $n_2 \leftarrow n_2 + 1$ ;
8   if  $t_i$  is the only tuple of its block and  $L(t_i) = \ell_1$  then
9      $n_1 \leftarrow n_1 + 1$ ;
10  if  $t_i$  is the last tuple of its block then
11     $B^\square \leftarrow B^\square \cup \{\text{key}(t_i)\}$ ;
12  if  $|B^\square| \leq k \leq |B|$  and  $n_2 \geq n_1$  then
13    return true;
14 return false;
```

► **Theorem 11.** *There exists an $O(|\mathcal{Y}|n)$ algorithm for $\text{CR-NN}_{<}(\mathbf{R})$ when the FD schema \mathbf{R} is equivalent to one with a single primary key.*

When $|\mathcal{Y}| = 2$, the algorithm essentially reduces to the MinMax algorithm in [13]. For $|\mathcal{Y}| \geq 3$ it outperforms the SortScan algorithm [13], since the latter algorithm has an exponential dependence on $|\mathcal{Y}|$ and k . Our algorithm also can deal with the case where two tuples in the same block have different labels, which is not something the MinMax and SortScan algorithms can handle.

► **Example 12.** We now illustrate our algorithm by a simple example where $k = 3$ and $\mathcal{Y} = \{1, 2, 3\}$. Figure 1a represents an inconsistent instance D , where nodes with the same shape/color are in the same block. Their distances to a given test point are increasing from left to right. A repair that chooses the first two tuples assigns label 1 to x , hence 1 is a possible label. Figures 1b and 1c illustrate the pruned instances $\text{PRUNE}(D, 2, 1)$ and $\text{PRUNE}(D, 3, 1)$, respectively. Take Figure 1c for example: when FASTSCAN reaches the iteration where $i = 3$, we have $n_2 = 2, n_1 = 1, |B^\square| = 3$ and $|B| = 3$, so $|B^\square| = k = |B|$ and $n_2 \geq n_1$. Indeed, by choosing the first, second, third, and last two tuples in Figure 1c, we see that label 1 is not robust (against label 3).

6 Hardness Result

In this section, we establish the main intractability result.

► **Theorem 13.** *Let \mathbf{R} be an FD schema that is not equivalent to any FD schema with an lhs chain. Then, the problem $\text{CR-NN}_p(\mathbf{R}, 1)$ is coNP-complete for any $p > 1$.*

Proof. The coNP membership of the problem $\text{CR-NN}_p(\mathbf{R}, 1)$ follows from the observation that if one is not certainly robust, then it can be checked efficiently two given repairs (certificate) indeed lead to two different prediction labels. To prove this hardness result, we will describe a reduction from the SAT-3-restricted problem, inspired by the construction of [27] for the edge dominating set problem. In this variant of SAT, each clause has at most three literals, while every variable occurs two times and its negation once.

Let ϕ be a SAT-3-restricted formula. Suppose that ϕ has m clauses C_1, C_2, \dots, C_m and n variables x_1, x_2, \dots, x_n . Our construction consists of three steps.

Step 1. In the first step, we construct a directed labeled graph $G = (V, E)$ with labels in $\{0, 1\}$:

- The set of vertices $V = \{C_i \cup x_j^k \cup y_j^k \text{ where } 1 \leq i \leq m, 0 \leq k \leq 2 \text{ and } 1 \leq j \leq n\}$.
- For each clause C_i , where $i = 1, \dots, m$, we add the following labeled edge:

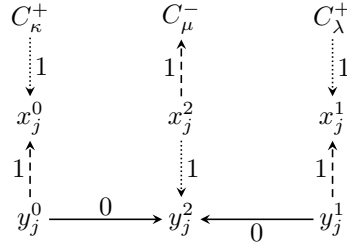
$$(C_i^+, C_i^-) \rightarrow 0$$

That is, we add the directed edge which points from C_i^+ to C_i^- to the set of edges E and label it as 0.

- Suppose that variable x_j , where $j = 1, \dots, n$, appears positive in clauses C_κ, C_λ , and negative in clause C_μ . Then, we introduce the following labeled edges:

$$\begin{aligned} (C_\kappa^+, x_j^0), (y_j^0, x_j^0) &\rightarrow 1 \\ (C_\lambda^+, x_j^1), (y_j^1, x_j^1) &\rightarrow 1 \\ (x_j^2, C_\mu^-), (x_j^2, y_j^2) &\rightarrow 1 \\ (y_j^0, y_j^2), (y_j^1, y_j^2) &\rightarrow 0 \end{aligned}$$

Figure 2 shows the above construction. Note that G is a directed bipartite graph, since no vertex has both incoming and outgoing edges. Hence, one can equivalently view each maximal matching of G as a subset repair of an instance with FD schema $(R(A, B), \{A \rightarrow B, B \rightarrow A\})$ and vice versa (attributes A and B correspond to the two sides of the bipartite graph).



■ **Figure 2** Variable gadget for the hardness reduction from the SAT-3-restricted problem.

▷ **Sublemma 1.** ϕ is satisfiable if and only if there exists a maximal matching for G that includes only edges with label 1.

Proof of Sublemma 1. \Rightarrow Assume that the variable assignment ψ makes ϕ satisfiable. Fix any order of variables $x_1 \dots, x_n$. We form a set of edges M as follows. For any variable x_j visited in the above order, we distinguish two cases:

- $\psi(x_j) = \text{true}$: we pick (x_j^2, y_j^2) . If C_κ^+ is not yet matched, pick (C_κ^+, x_j^0) , otherwise pick (y_j^0, x_j^0) . Similarly for C_λ^+ .
- $\psi(x_j) = \text{false}$: pick (y_j^0, x_j^0) and (y_j^1, x_j^1) . If C_μ^- is not yet matched, pick (x_j^2, C_μ^-) , otherwise pick (x_j^2, y_j^2) .

6:12 Certifiable Robustness for Nearest Neighbor Classifiers

By construction, M contains only edges with label 1.

▷ **Claim 1** (M is a matching). Since x_j, y_j occur only in a variable gadget, they will have at most one adjacent edge from M . By construction, each clause C_κ^+, C_μ^- will also have at most one adjacent edge from M .

▷ **Claim 2** (M is a maximal matching). First, observe that by our construction, all x_j^0, x_j^1, x_j^2 are matched for any $j = 1, \dots, n$. Second, notice that if $\psi(x_j) = \text{true}$, the edge (x_j^2, y_j^2) will be chosen; if $\psi(x_j) = \text{false}$, the edges (y_j^0, x_j^0) and (y_j^1, x_j^1) will be chosen. Thus, the edges $(y_j^0, y_j^2), (y_j^1, y_j^2)$ can not be added to M . Finally, consider the edge (C_i^+, C_i^-) corresponding to clause C_i . If there exists a positive literal which satisfies C_i , then consider the earliest x_j in the linear order of variables. By construction, (C_i^+, x_j^ν) is in the matching, where $\nu \in \{0, 1\}$. Otherwise, C_i is satisfied by a negative literal: consider the earliest such x_k in the linear order. Then (x_k^2, C_i^-) is in M . In either case, (C_i^+, C_i^-) cannot be added to increase the size of the matching.

⇐ Assume a maximal matching M that avoids 0-labeled edges. For every variable x_j , if M contains (x_j^2, y_j^2) , we assign $\psi(x_j) = \text{true}$, otherwise $\psi(x_j) = \text{false}$. We claim that ψ is a satisfying assignment for ϕ . Indeed, take any clause C_i . Since $(C_i^+, C_i^-) \notin M$, there exists some edge in M that conflicts with it. If this edge is of the form (C_i^+, x_j^ν) , then it must be that $(x_j^2, y_j^2) \in M$. But then $\psi(x_j) = \text{true}$, and since x_j occurs positively in C_i the clause is satisfied. If this edge is of the form (x_k^2, C_i^-) , then $(x_j^2, y_j^2) \notin M$. Thus, $\psi(x_j) = \text{false}$, and since x_j occurs negatively in C_i it is again satisfied. ◁

Step 2. A maximal matching of G can be viewed equivalently as a repair of a labeled instance D_0 with FD schema $\mathbf{S} = (R(A, B), \{A \rightarrow B, B \rightarrow A\})$. In the second step, we will transform the instance D_0 of \mathbf{S} to a labeled instance D_1 of the target FD schema \mathbf{R} . We will do this using the concept of *fact-wise reductions*. A fact-wise reduction from \mathbf{S} to \mathbf{R} is a function Π that maps a tuple from an instance of \mathbf{S} to a tuple of an instance of \mathbf{R} such that (i) it is injective, (ii) it preserves consistency and inconsistency (i.e. a tuple in D_0 violates \mathbf{S} if and only if the corresponding tuple in D_1 violates \mathbf{R}), and (iii) it can be computed in polynomial time. In fact, we will use exactly the same fact-wise reduction as the one used in [21] to reduce an instance in \mathbf{S} to one in \mathbf{R} , where \mathbf{R} is not equivalent to an FD schema with an lhs chain. It will be necessary to present this reduction in detail, since its structure will be needed for the third step of our reduction.

W.l.o.g., we can assume the FD schema is minimal. Since it does not have an lhs chain, there are two FDs $X \rightarrow A$ and $X' \rightarrow A'$ such that $X \subsetneq X'$ and $X' \subsetneq X$. Let \oplus be a fresh constant. We map $t = R(u, v)$ with label ℓ to a tuple $\Pi(t)$ with label also ℓ such that:

$$\Pi(t)[A_i] = \begin{cases} \oplus & \text{if } A_i \in (X \cap X')^{+, \Sigma} \\ u & \text{if } A_i \in X \setminus (X \cap X')^{+, \Sigma} \\ v & \text{if } A_i \in X' \setminus (X \cap X')^{+, \Sigma} \\ (u, v) & \text{otherwise.} \end{cases}$$

Here, $X^{+, \Sigma}$ denotes the closure of an attribute set X w.r.t. the FD set Σ . By [21] we know that Π is a fact-wise reduction. Let D_1 be the resulting instance of \mathbf{R} .

▷ **Sublemma 2.** ϕ is satisfiable if and only if there exists a repair for D_1 in \mathbf{R} that includes only tuples with label 1.

Proof of Sublemma 2. This follows from Sublemma 1 and the fact that Π is a fact-wise reduction. ◁

Step 3. In the last step of the reduction, we will present an encoding $\llbracket \cdot \rrbracket$ that embeds the values of the attributes in D_1 to values in \mathbb{N} , hence allowing us to compute distances with the p -norm. The resulting tuples will also be labeled from $\mathcal{Y} = \{0, 1\}$.

Let $\alpha = d \cdot (2m + 8n)$, where d is the number of attributes. First, let $\llbracket \oplus \rrbracket = 0$. For a vertex $u \in V$, let

$$\llbracket u \rrbracket = \begin{cases} i & \text{if } u = C_i^+ \\ m + i & \text{if } u = C_i^- \\ 2m + 3j + \nu & \text{if } u = y_j^\nu \\ \alpha + 3j + \nu & \text{if } u = x_j^\nu \end{cases}$$

It is easy to see that the above embedding is injective, meaning that if $\llbracket u \rrbracket = \llbracket v \rrbracket$ then $u = v$. As for the edges, consider any ordering e_1, e_2, \dots and simply let $\llbracket e_i \rrbracket = i$. Note that the number of edges in G is $|E| = 8n + 2m$. This embedding is also injective. Let $D_2 = \llbracket D_1 \rrbracket$ denote the instance we obtain by encoding every value of D_1 as above. Since the encoding is injective, this is also trivially a fact-wise reduction, hence Sublemma 2 holds for D_2 as well.

▷ **Sublemma 3.** ϕ is satisfiable if and only if $x = R(0, 0, \dots, 0)$ has no certain label in D_2 .

Proof of Sublemma 3. We first need the following two claims.

▷ **Claim 1.** *Any tuple with label 0 is closer to x than any tuple with label 1.* Indeed, a tuple has label 1 if and only if it contains in an attribute a value of the form $\llbracket x_j^\nu \rrbracket$. Hence, any tuple with label 1 has distance $> \alpha$ from x . On the other hand, each attribute in a 0-labeled tuple has value at most $2m + 8n$. Hence, the distance from any tuple with label 0 is bounded by $d^{1/p} \cdot (2m + 8n) \leq d \cdot (2m + 8n)$.

▷ **Claim 2.** *0 is a possible label for x .* Indeed, the tuple corresponding to the edge (C_1^+, C_1^-) is the closest one to x and has label 0. Hence, any repair that includes this tuple will assign the label 0 to x .

⇒ Assume that the variable assignment ψ makes ϕ satisfiable. Then, we know that there exists a repair for D_2 that avoids any tuple with label 0. This repair will then assign label 1 to x , which implies that x is not a certain label since by **Claim 2** 0 is a possible label for x .

⇐ Assume a repair that assigns a label 1 to x – hence making x have no certain label. Since by **Claim 1** all 0-labeled tuples are closer than the 1-labeled tuples, this means that all tuples in the repair must have label 1. But then, ϕ is satisfiable. ◀

This completes the proof. ◀

Finally, we extend the intractability result from $\text{CR-NN}_p(\mathbf{R}, 1)$ to $\text{CR-NN}_p(\mathbf{R}, k)$ for any integer $k \geq 1$. For the proof, see [7].

▶ **Theorem 3.** *Let \mathbf{R} be an FD schema that is not equivalent to any FD schema with an lhs chain. Then, for any integer $k \geq 1$, $\text{CR-NN}_p(\mathbf{R}, k)$ is coNP-hard for any $p > 1$.*

7 Optimal Repairs Revisited

In this section, we investigate the complexity landscape of a related problem to certifying robustness for k -NN classification, which may be of independent interest. In [20], the authors studied the OPT-REPAIR problem: each tuple t is associated with a positive weight $w_t \geq 0$,

and we want to find the optimal subset repair that removes the set of tuples with the smallest total weight. Note that this is equivalent to finding the repair with the largest total weight.

Here, we consider the symmetric problem, MIN-REPAIR: we want to find *the subset repair that has tuples with the smallest total weight*. In this case, we interpret the tuple weights as a measure of how “wrong” we think a tuple is. We can parametrize this problem with a given FD schema \mathbf{R} , as in MIN-REPAIR(\mathbf{R}). MIN-REPAIR captures as a special case the following problem, denoted as FORBIDDEN-REPAIR(\mathbf{R}): given an inconsistent instance D over \mathbf{R} and a subinstance $S \subseteq D$, does there exist a subset repair $I \subseteq D$ such that $I \cap S = \emptyset$?

► **Lemma 4.** *There exists a many-one polynomial time reduction from FORBIDDEN-REPAIR(\mathbf{R}) to MIN-REPAIR(\mathbf{R}).*

Proof. One can set the weight of any tuple in S to be 1, otherwise 0. Then, there exists a repair that avoids the forbidden set S if and only if there exists a repair with total weight equal to 0. ◀

From the hardness proof of Theorem 13, we immediately obtain the following intractability result.

► **Theorem 5.** *Let \mathbf{R} be an FD schema that is not equivalent to any FD schema with an lhs chain. Then, FORBIDDEN-REPAIR(\mathbf{R}) is NP-hard. As a result, MIN-REPAIR(\mathbf{R}) is also NP-hard.*

It turns out that the forbidden set repair problem is directly connected with certifying robustness for 1-NN classification.

► **Lemma 6.** *There exists a many-one polynomial time reduction from FORBIDDEN-REPAIR(\mathbf{R}) to the complement of CR-NN $_{<}(\mathbf{R}, 1)$.*

Proof. Let $D, S \subseteq D$ be the inputs to the FORBIDDEN-REPAIR problem. We will construct a labeled instance for the classification problem using only two labels, $\mathcal{Y} = \{0, 1\}$. The input instance is exactly D . For labeling, if $t \in S$ then $L(t) = 0$, otherwise $L(t) = 1$. Finally, we construct an ordering of the tuples in D such that $t < t'$ whenever $t \in S, t' \in D \setminus S$.

We first claim that any repair of D that avoids S is a repair that assigns a label of 1 to x . Indeed, the 1-neighborhood of x for such a repair will consist of a tuple with label 1 by construction. On the other hand, any repair that includes a tuple from S assigns label 0 to x . Since there exists at least one repair that includes a tuple from S , there exists no repair that avoids the forbidden set S if and only if x is certifiably robust (with label 0). ◀

The above reduction gives a polynomial time algorithm for FORBIDDEN-REPAIR(\mathbf{R}) whenever \mathbf{R} is equivalent to an FD schema with an lhs chain. We now present Algorithm 3 that works for the more general MIN-REPAIR problem.

The algorithm works similarly to the OptSRepair algorithm in [20] with two differences. First, we only need to consider the cases where the FD schema has a common lhs or a consensus FD (i.e., FD with an empty lhs). Second, in the case of the consensus FD, where we partition the instance, we take the repair that has the *minimum* total weight instead of the largest one.

► **Theorem 7.** *Let \mathbf{R} be an FD schema that is equivalent to an FD schema with an lhs chain. Then, MIN-REPAIR(\mathbf{R}) is in P.*

Algorithm 3 MIN-REP(Σ, D).

```

1 if  $\Sigma$  is trivial then
2   | return  $D$ 
3 remove trivial FDs from  $\Sigma$  ;
4 if  $\Sigma$  has a common lhs attribute  $A$  then
5   | return  $\cup_{a \in \pi_A(D)} \text{MIN-REP}(\Sigma - A, \sigma_{A=a}(D))$ 
6 if  $\Sigma$  has a consensus FD  $\emptyset \rightarrow A$  then
7   |  $m \leftarrow \arg \min_{a \in \pi_A(D)} w(\text{MIN-REP}(\Sigma - A, \sigma_{A=a}(D)))$  ;
8   | return  $\text{MIN-REP}(\Sigma - A, \sigma_{A=m}(D))$ 

```

The dichotomy we obtain for MIN-REPAIR coincides with the one we obtain for CR-NN. However, it is different from the dichotomy observed for the OPT-REPAIR problem in [20]; in fact, every FD schema that admits a polynomial time algorithm for MIN-REPAIR also admits a polynomial time algorithm for OPT-REPAIR, but not vice versa. Specifically, the FD schema $(R(A, B), \{A \rightarrow B, B \rightarrow A\})$ is hard for MIN-REPAIR, but tractable for OPT-REPAIR. The reason is that finding a maximum-weight (maximal) matching in a bipartite graph is polynomially solvable, but finding a minimum-weight maximal matching is NP-hard.

8 Certifiable Robustness by Counting

In this section, we consider the counting version of certifiable robustness for k -NN classifiers. The counting problem of certifiable robustness asks the following: among all possible repairs, how many will predict label ℓ for a fixed $\ell \in \mathcal{Y}$?

We show that the counting problem still remains in polynomial time if the FD set \mathbf{R} is equivalent to an lhs chain by generalizing the algorithm in Section 5.1. Formally, the class of counting problems that can be computed in polynomial time is called FP.

► **Theorem 8.** *If the FD schema \mathbf{R} is equivalent to some FD schema with an lhs chain, then the counting problem $\#CR\text{-NN}_{<}(\mathbf{R})$ is in FP.*

We show now how to generalize the algorithm in Section 5.1 to perform counting. Let x be the test point and $\mathcal{Y} = \{1, 2, \dots, m\}$. It suffices to show how to count for the label $\ell = 1$. Recall the definition of $\mathcal{N}_\tau^\leq(x, I)$ and $C_\tau^\leq(\ell, I)$. Similarly as in Section 5.1, we will compute a high-dimensional matrix M “layer by layer” and read off the answer from it.

We now make our key definition. Fix a threshold $\tau > 0$, and define the following quantity for a (possibly inconsistent) subinstance $J \subseteq D$ and integers i, c_2, c_3, \dots, c_m , where $0 \leq i \leq k$ and $-k \leq c_j \leq k$ for all $j \in \{2, 3, \dots, m\}$:

$$M_i[J, c_2, \dots, c_m, \Delta] = \{ \#K \mid K \in I_\Delta(J) \text{ s.t. } |\mathcal{N}_\tau^\leq(x, K)| = i \text{ and } C_\tau^\leq(j, K) - C_\tau^\leq(1, K) = c_j \forall j \in \{2, 3, \dots, m\} \}.$$

For simplicity of notation, let \mathbf{c} denote the vector (c_2, c_3, \dots, c_m) with the understanding that \mathbf{c}_j could represent any new vector $(c_{2_j}, c_{3_j}, \dots, c_{m_j})$. Sometimes we might suppress the FD set Δ to write $M_i[J, \mathbf{c}]$ when the context is clear. The interpretation for the entry $M_i[J, \mathbf{c}]$ is that it records the number of repairs in J with i many tuples which are among τ -th closest to x such that the differences between the number of 1-tuples and the number of j -tuples are exactly c_j , $j \in \{2, 3, \dots, m\}$. If there is no such $K \in I_\Sigma(J)$, we define $M_i[J, \mathbf{c}] = 0$. The algorithm computes the entries of M by a combination of dynamic programming and

6:16 Certifiable Robustness for Nearest Neighbor Classifiers

recursion on the FD schema \mathbf{R} . Note that after computing M , the answer to $\#CR\text{-}NN_{<}(\mathbf{R})$ is exactly the sum of all entries $M_k[D, \mathbf{c}]$ where $c_j \geq 1$ for all $j \in \{2, 3, \dots, m\}$. The algorithm has three disjoint cases when computing $M_i[J, \mathbf{c}]$:

Base Case: *the set of FDs is empty.* In this case, $M_i[J, \mathbf{c}] = 0$ unless $|\mathcal{N}_{\bar{r}}^{\leq}(x, I)| = i \leq k$ and $C_{\bar{r}}^{\leq}(j, J) - C_{\bar{r}}^{\leq}(\ell, J) = c_j$ for all $j \in \{2, 3, \dots, m\}$, in which case the entry is 1. This step clearly can be computed efficiently.

Consensus FD: *there exists an FD $\emptyset \rightarrow A$.* In this case, we recursively call the algorithm to compute $M_i[\sigma_{A=a}(J), \mathbf{c}, \Delta - A]$ for every value $a \in \pi_A(J)$. Then, we calculate $M_i[J, \mathbf{c}] = \sum_{a \in \pi_A(J)} M_i[\sigma_{A=a}(J), \mathbf{c}, \Delta - A]$.

Common Attribute: *there exists a common attribute A in the lhs of all FDs.* In this case, we recursively call the algorithm to compute $M_j[\sigma_{A=a}(J), \mathbf{c}_j, \Delta - A]$ for every value $a \in \pi_A(J)$ and all j, \mathbf{c}_j such that $0 \leq j \leq i$ and $0 \leq c_{l_j} \leq c_l$ where $2 \leq l \leq m$. Let $S = \pi_A(J) = \{a_1, \dots, a_{|S|}\}$. We then compute

$$M_i[J, \mathbf{c}] = \sum_{j=1}^{|S|} \sum M_{i_j}[\sigma_{A=a_j}(J), \mathbf{c}_j, \Delta - A]$$

where the first summation is over all possible solutions of i_j 's and c_{l_j} 's such that $\sum_{j=1}^{|S|} i_j = i$ and $\sum_{j=1}^{|S|} c_{l_j} = c_l$ for all $l \in \{2, 3, \dots, m\}$. We now show how to compute $M_i[J, \mathbf{c}]$ by a direct generalization of Algorithm 1. The dynamic programming computes a $(m+1)$ -dimensional matrix \mathcal{M} where its (p, q, \mathbf{c}_i) -entry represents the sum $\sum_{j=1}^p M_{i_j}[\sigma_{A=a_j}(J), \mathbf{c}_j]$ over all possible solutions i_j 's and \mathbf{c}_j 's such that $\sum_{j=1}^p i_j = q$ and $\sum_{j=1}^p c_{l_j} = c_{l_i}$ for $2 \leq l \leq m$. Note that \mathcal{M} has $|S| \cdot (k+1) \cdot (2k+1)^{m-1} = O(n \cdot k^m)$ entries. Let $K = \{k, k-1, \dots, -k\}$ and $\mathbf{c}_i - \mathbf{c}_j := (c_{2_i} - c_{2_j}, \dots, c_{m_i} - c_{m_j})$. We are now ready to present our dynamic programming algorithm:

■ **Algorithm 4** Dynamic Programming (Counting).

```

1 for  $j = 0, \dots, k$  and  $\mathbf{c} \in K^m$  do
2    $\mathcal{M}_j[1, \mathbf{c}] \leftarrow M_j[\sigma_{A=a_1}(J), \mathbf{c}, \Delta - A]$ ;
3 for  $i = 2, \dots, |S|$  do
4   for  $j = 0, \dots, k$  and  $\mathbf{c} \in K^m$  do
5      $\mathcal{M}_j[i, \mathbf{c}] \leftarrow \sum_{p, \mathbf{c}_q} \{\mathcal{M}_p[i-1, \mathbf{c}_q] + M_{j-p}[\sigma_{A=a_i}(J), \mathbf{c} - \mathbf{c}_q, \Delta - A]\}$ ;

```

We show in [7] that the counting algorithm runs in polynomial time with respect to the size of D and the parameter k . However, the running time depends exponentially on the number of labels $|\mathcal{Y}|$. An open question remains whether this exponential dependency is essential.

The hardness part for counting is an immediate corollary with the previous result in [21].

► **Theorem 9.** *If the FD schema \mathbf{R} is not equivalent to some FD schema with an lhs chain, then $\#CR\text{-}NN_{<}(\mathbf{R})$ is $\#P$ -complete.*

Proof. By Theorem 3.2 in [21], it is $\#P$ -hard to count the number of repairs for the FD schema \mathbf{R} . Now, given any instance D , we can pick any ordering of the points and assign the same label ℓ to every tuple. Then, the number of repairs that predict label ℓ is the same as the total number of repairs. ◀

The approximate counting of certifying robustness of k -NN classifiers is also hard.

► **Theorem 10.** *If the FD schema \mathbf{R} is not equivalent to some FD schema with an lhs chain, then $\#CR\text{-}NN_{<}(\mathbf{R}) \equiv_{AP} \#SAT$.*

Here, \equiv_{AP} refers to the equivalency up to approximation-preserving reductions [6].

Proof. By the same fact-wise reduction in Lemma 2 and labelling in Theorem 9, we have $\#MAXIMALBIS \leq \#SREP \leq \#CR\text{-}NN_{<}(\mathbf{R})$ where $\#MAXIMALBIS$ is the problem of counting the number of maximal independent sets in a bipartite graph and $\#SREP$ is the problem of counting the number of subset repairs of an inconsistent instance. By Theorem 1 in [9], $\#MAXIMALBIS \equiv_{AP} \#SAT$ and thus we obtain $\#CR\text{-}NN_{<}(\mathbf{R}) \equiv_{AP} \#SAT$. ◀

9 Other Uncertainty Models

In this section, we study the complexity of certifying robustness for k -NNs under three simple uncertainty models: ?-sets, or-sets, and Codd tables. We show that for these models we can certify robustness in polynomial time. Throughout the section, we fix the relational schema to be $R(A_1, \dots, A_d)$.

?-Sets with Size Constraints. For a given instance D over the schema, we mark an uncertain subset $D_?$ of the tuples in D . Then, for a positive integer $m \geq 1$, we define the set of possible worlds as:

$$\mathcal{I}_? = \{I \mid D \setminus D_? \subseteq I \subseteq D, |D \setminus I| \leq m\}.$$

In other words, we can construct a possible world by removing any – but at most m – tuples from $D_?$. When $m = |D_?|$, this definition captures the notion of ?-tables as defined in [25]. When $D_? = D$, it captures *data poisoning* scenarios, where the goal is to protect against attacks that can poison up to m tuples of the training set.

For this setting, we can show that certifiable robustness for k -NNs can be computed in almost linear time in the size of the input.

► **Lemma 11.** *We can certify robustness in $\mathcal{I}_?$ for k -NNs in time $O((|\mathcal{Y}| + \log n) \cdot n)$ where n is the size of the dataset.*

Proof. For the sake of simplicity, let $\mathcal{Y} = \{1, 2, \dots\}$.

Let x be the test point. As a first step, we order the tuples in increasing order of $f(x, t)$ – this can be done in time $O(n \log n)$. We assume w.l.o.g. that all distances are distinct. Let the resulting order be t_1, t_2, \dots, t_n .

Since it always holds that $D \in \mathcal{I}_?$, we can first compute the predicted label of x in D in time $O(k)$. W.l.o.g. assume that $\mathcal{L}_D(x) = 1$. For every label $\ell > 1$, the algorithm will now try to construct a possible world where ℓ occurs at least as many times as 1 in the k -neighborhood of x . To make the algorithm exposition easier, define for every tuple $t \in D_?$ a priority value $\rho(t)$ as follows.

$$\rho(t) = \begin{cases} 2, & \text{if } L(t) = \ell \\ 0, & \text{if } L(t) = 1 \\ 1, & \text{otherwise.} \end{cases}$$

■ **Algorithm 5** Certifiable Robustness for $\mathcal{I}_?$ -Sets.

Input: test point x
Output: is k -NN certifiably robust in $\mathcal{I}_?$

- 1 $J \leftarrow \{t_1, \dots, t_k\}$;
- 2 $\Delta \leftarrow 0$;
- 3 **for** $i = k + 1$ **to** n **do**
- 4 $\Delta \leftarrow \Delta + 1$;
- 5 **if** $|\{t \in J \mid L(t) = \ell\}| \geq |\{t \in J \mid L(t) = 1\}|$ **then**
- 6 **return true**;
- 7 **if** $J \cap D_? = \emptyset$ **or** $\Delta > m$ **then**
- 8 **return false**;
- 9 $J \leftarrow J \cup \{t_i\}$;
- 10 remove from J the tuple $\arg \min_{t \in J \cap D_?} \{\rho(t)\}$;

We now present our Algorithm 5. Intuitively, it iterates over the tuples in order of proximity to x . For each tuple t_i , it attempts to construct the most promising k -neighborhood that includes tuples from $\{t_1, \dots, t_i\}$. Each loop in the algorithm can be executed in time $O(1)$. Indeed, we can implement this by keeping three sets with tuples depending on their labels, where we can do insertion and deletion in constant time.

Note that the running time of Algorithm 5 is independent of k, m and $|D_?|$, but it does depend linearly on the number of labels. ◀

Or-Sets. In this uncertain model, each attribute value of a tuple is an *or-set* consisting of finite values (e.g., $\langle 2, 5, 10 \rangle$). Each possible world in \mathcal{I}_{or} is formed by choosing exactly one value from each or-set, independent of the choices across all other or-sets. We can express \mathcal{I}_{or} as subset repairs for the following schema: add to R an extra attribute *id*. Then, assign a unique value for this attribute to each tuple t , and create a tuple for each “realization” of this tuple with the same *id*. This will increase the input size of the problem, but by at most a polynomial factor (since d is fixed). Moreover, the FD schema is clearly equivalent to an lhs chain; in fact, this is the single primary key case. As a consequence, we have the following proposition.

▶ **Proposition 12.** *We can certify robustness in \mathcal{I}_{or} for k -NNs in P .*

Codd tables. In a *Codd table*, a missing value is represented as Null paired with a domain *Dom* from which that value can be chosen. A repair is any possible completion of the table. The first observation is that, by adding a new identifier attribute *ID*, we can think of a Codd table as an inconsistent instance with a single primary key, where each block has a consistent label (i.e., every tuple in the block has the same label). However, if *Dom* is given as an infinite interval, then Algorithm 2 does not apply directly since it may not be possible to write all tuple completions in an increasing order (there are uncountably many). Indeed, in [13] Karlas et al. consider only the situation where *Dom* is given as a finite discrete set.

Formally, the distance between a tuple t in the Codd table and a test point x is given by the function $f(x, t) = g_t(y_1, y_2, \dots, y_n)$ where the y_i 's are the missing entries in t . In order to be able to certify robustness, we need to assume that the minimum and maximum of this function can be computed efficiently. This is a mild assumption, since for example for the p -norm the minimum and maximum is achieved when each summand is minimized or maximized respectively.

Now, we observe that since every block has the same label, we can modify Algorithm 2 so that every resulting repair consists of only the first or the last tuple in a block without changing its correctness properties. With this observation in hand, we can now simply define the extremal set $S = \{\min f(x, t) \cup \max f(x, t) \mid t \in D\}$ consisting of the minimum and maximum of each block and then run Algorithm 2 on S . Since S is at most twice the size of S , this implies a linear time algorithm (w.r.t. the number of tuples, labels, and k) for certifiable robustness on Codd tables. We have shown the following proposition.

► **Proposition 13.** *We can certify robustness in a Codd table for k -NNs in linear time even if the domain Dom is given as an infinite interval.*

10 Conclusion

In this paper, we study the complexity of certifiable robustness for k -NN classification under FD constraints. We show a dichotomy in the complexity of the problem depending on the structure of the FDs, and we prove that the same dichotomy condition also holds for the counting version of the problem. We envision this work to be a first step towards the long-term goal of investigating the complexity of certifying robustness for other widely used classification algorithms, such as decision trees, Naive Bayes classifiers and linear classifiers.

References

- 1 Marcelo Arenas, Leopoldo E. Bertossi, and Jan Chomicki. Consistent query answers in inconsistent databases. In *PODS*, pages 68–79. ACM Press, 1999. doi:10.1145/303976.303983.
- 2 Jan Chomicki. Consistent query answering: Five easy pieces. In *ICDT*, volume 4353 of *Lecture Notes in Computer Science*, pages 1–17. Springer, 2007. doi:10.1007/11965893_1.
- 3 Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 2019. URL: <http://proceedings.mlr.press/v97/cohen19c.html>.
- 4 Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM J. Comput.*, 48(2):742–864, 2019. doi:10.1137/17M1126680.
- 5 Samuel Drews, Aws Albarghouthi, and Loris D’Antoni. Proving data-poisoning robustness in decision trees. In *PLDI*, pages 1083–1097. ACM, 2020. doi:10.1145/3385412.3385975.
- 6 Martin E. Dyer, Leslie Ann Goldberg, Catherine S. Greenhill, and Mark Jerrum. The relative complexity of approximate counting problems. *Algorithmica*, 38(3):471–500, 2004. doi:10.1007/s00453-003-1073-y.
- 7 Austen Z. Fan and Paraschos Koutris. Certifiable robustness for nearest neighbor classifiers. *CoRR*, abs/2201.04770, 2022. URL: <http://arxiv.org/abs/2201.04770>.
- 8 Ariel Fuxman and Renée J. Miller. First-order query rewriting for inconsistent databases. *J. Comput. Syst. Sci.*, 73(4):610–635, 2007. doi:10.1016/j.jcss.2006.10.013.
- 9 Leslie Ann Goldberg, Rob Gysel, and John Lapinskas. Approximately counting locally-optimal structures. *J. Comput. Syst. Sci.*, 82(6):1144–1160, 2016. doi:10.1016/j.jcss.2016.04.001.
- 10 Sergio Greco, Cristian Molinaro, and Francesca Spezzano. *Incomplete Data and Data Dependencies in Relational Databases*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2012. doi:10.2200/S00435ED1V01Y201207DTM029.
- 11 Tomasz Imielinski and Witold Lipski Jr. Incomplete information in relational databases. *J. ACM*, 31(4):761–791, 1984. doi:10.1145/1634.1886.
- 12 Jinyuan Jia, Xiaoyu Cao, Binghui Wang, and Neil Zhenqiang Gong. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. In *ICLR*. OpenReview.net, 2020. URL: <https://openreview.net/forum?id=BkeWw6VFwr>.

- 13 Bojan Karlas, Peng Li, Renzhi Wu, Nezihe Merve Gürel, Xu Chu, Wentao Wu, and Ce Zhang. Nearest neighbor classifiers over incomplete information: From certain answers to certain predictions. *Proc. VLDB Endow.*, 14(3):255–267, 2020. doi:10.5555/3430915.3442426.
- 14 Phokion G. Kolaitis and Enela Pema. A dichotomy in the complexity of consistent query answering for queries with two atoms. *Inf. Process. Lett.*, 112(3):77–85, 2012. doi:10.1016/j.ipl.2011.10.018.
- 15 Paraschos Koutris and Dan Suciu. A dichotomy on the complexity of consistent query answering for atoms with simple keys. In *ICDT*, pages 165–176. OpenProceedings.org, 2014. doi:10.5441/002/icdt.2014.19.
- 16 Paraschos Koutris and Jef Wijsen. The data complexity of consistent query answering for self-join-free conjunctive queries under primary key constraints. In *PODS*, pages 17–29. ACM, 2015. doi:10.1145/2745754.2745769.
- 17 Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J. Franklin, and Ken Goldberg. Active-clean: Interactive data cleaning for statistical modeling. *Proc. VLDB Endow.*, 9(12):948–959, 2016. doi:10.14778/2994509.2994514.
- 18 Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 5458–5467. PMLR, 2020. URL: <http://proceedings.mlr.press/v119/kumar20b.html>.
- 19 Leonid Libkin. Incomplete information and certain answers in general data models. In *PODS*, pages 59–70. ACM, 2011. doi:10.1145/1989284.1989294.
- 20 Ester Livshits, Benny Kimelfeld, and Sudeepa Roy. Computing optimal repairs for functional dependencies. *ACM Trans. Database Syst.*, 45(1):4:1–4:46, 2020. doi:10.1145/3196959.3196980.
- 21 Ester Livshits, Benny Kimelfeld, and Jef Wijsen. Counting subset repairs with functional dependencies. *J. Comput. Syst. Sci.*, 117:154–164, 2021. doi:10.1016/j.jcss.2020.10.001.
- 22 Simon Razniewski and Werner Nutt. Completeness of queries over incomplete databases. *Proc. VLDB Endow.*, 4(11):749–760, 2011. URL: <http://www.vldb.org/pvldb/vol14/p749-razniewski.pdf>.
- 23 Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré. Holoclean: Holistic data repairs with probabilistic inference. *Proc. VLDB Endow.*, 10(11):1190–1201, 2017. doi:10.14778/3137628.3137631.
- 24 Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and J. Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 8230–8241. PMLR, 2020. URL: <http://proceedings.mlr.press/v119/rosenfeld20b.html>.
- 25 Anish Das Sarma, Omar Benjelloun, Alon Y. Halevy, and Jennifer Widom. Working models for uncertain data. In *ICDE*, page 7. IEEE Computer Society, 2006. doi:10.1109/ICDE.2006.174.
- 26 Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *NIPS*, pages 3517–3529, 2017. URL: <https://proceedings.neurips.cc/paper/2017/hash/9d7311ba459f9e45ed746755a32dcd11-Abstract.html>.
- 27 M. Yannakakis and F. Gavril. Edge dominating sets in graphs. *SIAM Journal on Applied Mathematics*, 38(3):364–372, 1980. doi:10.1137/0138030.