Measuring Approximate Functional Dependencies: a Comparative Study

Marcel Parciak*†, Sebastiaan Weytjens†, Niel Hens†, Frank Neven†, Liesbet M. Peeters*† and Stijn Vansummeren†
*UHasselt, BIOMED, Agoralaan, 3590 Diepenbeek, Belgium
†UHasselt, Data Science Institute, Agoralaan, 3590 Diepenbeek, Belgium

***Abstract*—Approximate functional dependencies (AFDs) are functional dependencies (FDs) that "almost" hold in a relation. While various measures have been proposed to quantify the level to which an FD holds approximately, they are difficult to compare and it is unclear which measure is preferable when one needs to discover FDs in real-world data, i.e., data that only approximately satisfies the FD. In response, this paper formally and qualitatively compares AFD measures. We obtain a formal comparison through a novel presentation of measures in terms of Shannon and logical entropy. Through our comparison, we identify two novel, previously unexplored measures. Qualitatively, we perform a sensitivity analysis w.r.t. structural properties of input relations and study the effectiveness of AFD measures for ranking AFDs on real world data. Based on this analysis, we give clear recommendations for the AFD measures to use in practice.**

***Index Terms*—functional dependencies, data cleaning, data profiling**

I. Introduction

There are many benefits to knowing the set of functional dependencies (FDs for short) that are expected to hold in the instances of a database schema: FDs aid in ensuring data consistency and help in data cleaning [11], [39]; facilitate data integration [43]; and can be exploited for query optimization [26], [24], among other tasks. In many data science scenarios, however, the set of design FDs is unknown or incomplete [31]. As such, a variety of techniques have been proposed to reverse engineer this set of design FDs from a given relation instance: see e.g., [30], [41], [2], [5], [31].

Reverse engineering is particularly challenging in the practical setting where the relation instance itself does not satisfy the target set of design FDs. This may happen, e.g., due to errors during data entry, resulting in a corrupt instance. To recover the set of design FDs in such a setting, it clearly does not suffice to simply enumerate the FDs that are satisfied in the relation instance. Instead, one must also consider *approximate functional dependencies* (AFDs), that is, FDs that "almost hold" in the relation. A key decision to then make is when an FD "almost" holds. This decision is reflected in the adoption of an *AFD measure*, which formally quantifies the extent to which an FD holds approximately in a given relation by attributing a score in the interval $[0, 1]$. Higher values indicate a higher degree of FD satisfaction. As such, an AFD measure provides a way of *ranking* the search space of all possible FDs where higher-scoring FDs are ranked before lower-scoring ones. A good AFD measure, then, is one that ranks FDs in the relation's target set of design FDs higher than those that are not in the target set, and does this consistently for relation instances that occur in practice. Given a good AFD measure one may discover the set of design FDs by ranking the search space, and returning all FDs smaller than a given threshold.

Many AFD measures have been proposed in the literature over the past decades [23], [18], [27], [28], [34], [36], [43], [21], [3]. Unfortunately, these measures vary widely in nature and there has been little study so far in comparing them. As such there is no clear guideline for deciding which AFD measure(s) to adopt. In response, we aim to address the following two central questions in this paper:

(a) *How do the different AFD measures compare?*
(b) *In practice, which AFD measures consistently rank FDs in the target set of design FDs higher than those not in the target set?*

We adopt the following methodology to answer (a): In Section IV we present a survey of known AFD measures in a uniform formal framework. Specifically, we observe that there are three classes of measures: (i) measures that quantify the fraction of violations (we refer to them as VIOLATION); (ii) measures based on Shannon entropy (SHANNON) [13]; and (iii) measures based on logical entropy (LOGICAL) [14].

From this classification, we propose two new measures that apply a design principle known from one class to measures in a different class. We then proceed to evaluate the measures' sensitivity to structural properties of the input relation, in Section V. Specifically, we study whether a measure can distinguish between an FD $\boldsymbol{X} \rightarrow \boldsymbol{Y}$ in relation instances that were generated to approximately satisfy the FD versus relation instances where $\boldsymbol{X}$ and $\boldsymbol{Y}$ were randomly generated. We evaluate this on synthetic data, controlling the following parameters: (1) the error level; (2) *LHS-uniqueness*: the normalized number of unique values occurring in $\boldsymbol{X}$; and (3) *RHS-skew*: the skewness of the distribution of values occurring

We thank Dan Suciu for helpful discussions. S. Vansummeren was supported by the Bijzonder Onderzoeksfonds (BOF) of Hasselt University under Grant No. BOF20ZAP02. This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. This work was supported by Research Foundation—Flanders (FWO) for ELIXIR Belgium (I002819N). The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation – Flanders (FWO) and the Flemish Government.

in $\pi_{\boldsymbol{Y}}(R)$. This evaluation sheds light on the measure's distinguishing power as well as on their biases w.r.t. error level, LHS-uniqueness and RHS-skew. In general, measures are more robust if they are inversely proportional to the error level and are insensitive to LHS-uniqueness and RHS-skew, as we discuss in Section V.

We adopt the following methodology to answer (b): We compare the effectiveness of AFD measures on real-world data. Because existing benchmarks for exact FD discovery are designed to gauge algorithmic efficiency and do not contain the "ground truth" set of design FDs to compare against, we propose a new AFD discovery benchmark. This benchmark, denoted RWD, is obtained by manually creating design FDs for existing benchmarks. Our analysis in Section VI on RWD shows that well-ranking measures exist within each class of measures (VIOLATION, SHANNON and LOGICAL). Importantly, these measures are only effective when they are correctly normalized—which is not always done in the literature. Furthermore, the best-ranking measure for VIOLATION (measure g_3' which we formally introduce in Section IV) is sensitive to RHS-skew and therefore performs worse than the best-ranking measures for SHANNON (our newly proposed measure RFI'^{+}) and LOGICAL (a little-known measure called μ^+), which have comparable performance as well as equal structural sensitivity properties. However, RFI'^{+} is slow to compute whereas μ^+ is very efficient. Therefore, we recommend μ^+ for practical AFD discovery. More in-depth discussion and recommendations are given in Section VII, where Table II summarizes our comparison.

In summary, our contributions are as follows. (1) A survey of AFD measures, using a new and uniform presentation and including two new proposed measures. (2) Sensitivity analysis of measure performance w.r.t. structural properties of the input relations. (3) Creation of a real-world benchmark for AFD discovery. (4) Analysis of measure ranking power on this benchmark. (5) Clear recommendations for measure adoption in AFD discovery.

The general organisation of the paper has already been outlined. We discuss related work and motivate the paper's scope in Section II. We introduce the necessary background in Section III.

II. Related Work and Paper Scope

Relaxing FDs. We note that in the literature there are two distinct ways to relax the notion of a FD [8]: relax the constraint that an FD $\boldsymbol{X} \rightarrow \boldsymbol{Y}$ needs to be fully satisfied; or replace the way in which tuples are compared on their $\boldsymbol{X}$-values by a similarity function rather than strict equality. We focus on the former setting in this paper.

AFD measures versus discovery algorithms. The focus in this paper is on understanding AFD measures and how they allow us to rank the search space of possible FDs for AFD discovery. A related but orthogonal issue that we do not consider in this paper is the study of AFD *discovery algorithms*. AFD discovery algorithms usually fix a way to quantify the approximateness of an FD—typically through the choice of an AFD measure that fixes a ranking of the search space—but then combine a multitude of techniques to do the actual discovery. This includes pruning the ranked search space for efficiency reasons (e.g., [25], [27], [28], [34]) or complementing the measures' ranking with heuristics for application-specific purposes (e.g., [21], [47]). Of course, improved knowledge of AFD measures may aid in improving discovery algorithms in the future. For this reason, when we introduce measures in Section IV we also indicate in which discovery algorithms they are used. Because discovery algorithms do not directly expose the way in which they rank the search space of FDs, we do not experimentally compare against discovery algorithms.

Correlation. When an FD holds in a relation, there is clearly a statistical correlation among the FD's attributes. Conversely, correlated attributes may (but need not) indicate the presence of an FD. The techniques that are typically used to test statistical correlation, such as the χ^2 test or mean-square contingency [21], however, only measure the strength of correlation (e.g., $\boldsymbol{X}$ and $\boldsymbol{Y}$ are correlated) but do not indicate the direction in which functional dependence ($\boldsymbol{X} \rightarrow \boldsymbol{Y}$ or $\boldsymbol{Y} \rightarrow \boldsymbol{X}$) is likely to hold. As such, these techniques do not form appropriate AFD measures [36] and are not further considered here.

Exact FD discovery In the context of exact FD discovery, some works consider the problem of ranking exact FDs according to relevance, where the challenge lies in quantifying relevance [46]. We are not concerned with exact FD discovery, but with measures for quantifying the extent to which FDs hold approximately. Discovery of AFDs should also not be confused with the approximate discovery of exact FDs as e.g., done in [5]. There, only a subset of all FDs that satisfy input R are computed in return for performance improvements.

Existing comparisons of AFD measures. Giannella and Robertson [18] compare a limited number of measures on theoretical examples as well as on 4 real world datasets. In their experiments, they report on average differences between pairs of measures and do not compare with a ground truth set of FDs. They therefore do not empirically compare the effectiveness of the measures considered here. In their survey concerning FD relaxations, Caruccio et al. [8] also survey some of the AFD measures that are considered here, but do not provide a qualitative comparison.

UNI-DETECT [44] is a framework to automatically detect four common types of errors in relations: numeric-outliers, misspellings, uniqueness and FD violations. For FD violations, UNI-DETECT uses the FD-compliance-ratio which is similar to g_2, and it is shown that UNI-DETECT outperforms the naive algorithm ranking AFDs based on the measures ρ, g_1 and g_2. However, no comparative study of the measures for AFD discovery is provided.

Discovery of Conditional FDs. Conditional FDs (CFDs for short) generalize FDs: they are FDs that only hold on a subset

of the data [6]. CFDs are widely used in data cleaning [15], [16]. Discovery of (approximate) CFDs amounts to (i) selecting suitable subsets of the data and (ii) discovering (A)FDs in these subsets [38]. In particular, Geerts and Rammelaere [38] propose a generic (A)CFD discovery algorithms in which any (A)FD discovery algorithm can be plugged. Insights into the discovery of (A)FDs, as is our focus here, are therefore also useful for the discovery of (A)CFDs.

Additional AFD measures. Pfahringer and Kramer propose an AFD measure based on how well the considered FD can be used to compress the input table [35]. Their approach applies only to tables in which all attributes have a binary domain, however, which is why we do not consider it here.

Simovici et al. [42] study so-called impurity measures on sets and partitions and observe that each impurity measure induces an FD error measure. For an impurity measure i and threshold α, they define the *purity dependency* $\boldsymbol{X} \rightarrow_{i,\alpha} \boldsymbol{Y}$ to hold in R if the corresponding FD error measure value is at most α. They study relational decomposition for purity dependencies, as well as algorithms to discover the purity dependencies that hold in a relation R, given a fixed threshold α and a fixed set of impurity measures. No comparative study of impurity measures for AFD discovery is provided.

III. Preliminaries

We assume given a fixed set of attributes, where each attribute X has a domain $dom(X)$ of possible data values. We use uppercase letters X, Y, Z to denote attributes and boldface type like $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ to denote sets of attributes. Lowercase $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ denote tuples over these sets. Formally, as usual, a tuple over $\boldsymbol{X}$ is a mapping $\boldsymbol{x}$ that assigns each attribute $X \in \boldsymbol{X}$ to a value $\boldsymbol{x}(X) \in dom(X)$. We write $\boldsymbol{x} \colon \boldsymbol{X}$ to indicate that $\boldsymbol{x}$ is a tuple over $\boldsymbol{X}$, and write $\boldsymbol{x}|_{\boldsymbol{Y}}$ for the restriction of $\boldsymbol{x}$ to $\boldsymbol{Y} \subseteq \boldsymbol{X}$. We use juxtapostion like $\boldsymbol{XY}$ to denote the union $\boldsymbol{X} \cup \boldsymbol{Y}$ of two sets of attributes, and also apply this notation to tuples: if $\boldsymbol{x} \colon \boldsymbol{X}$ and $\boldsymbol{y} \colon \boldsymbol{Y}$ with $\boldsymbol{X}$ and $\boldsymbol{Y}$ disjoint, then $\boldsymbol{xy}$ is the tuple that equals $\boldsymbol{x}$ on all attributes in $\boldsymbol{X}$ and $\boldsymbol{y}$ on all attributes in $\boldsymbol{Y}$, i.e. $\boldsymbol{xy}|_{\boldsymbol{X}} = \boldsymbol{x}$ and $\boldsymbol{xy}|_{\boldsymbol{Y}} = \boldsymbol{y}$.

We will work with bag-based relations. Formally, a relation over $\boldsymbol{X}$ (also called $\boldsymbol{X}$-relation) is a mapping R that assigns a natural number $R(\boldsymbol{x}) \in \mathbb{N}$ to each tuple $\boldsymbol{x} \colon \boldsymbol{X}$. We also call $R(\boldsymbol{x})$ the frequency of $\boldsymbol{x}$ in R. We require relations to be finite in the sense that $R(\boldsymbol{x})$ can be non-zero for at most a finite number of $\boldsymbol{x}$. In what follows, we write $\boldsymbol{x} \in R$ to denote that $R(\boldsymbol{x}) > 0$ and stress that R is an $\boldsymbol{X}$-relation by means of the notation $R(\boldsymbol{X})$. If $\boldsymbol{Y} \subseteq \boldsymbol{X}$ then we denote by $dom_R(\boldsymbol{Y})$ the set of $\boldsymbol{Y}$-tuples that occur in R, $dom_R(\boldsymbol{Y}) = \{\boldsymbol{x}|_Y \mid \boldsymbol{x} \in R\}$. We denote by $|R|$ the total number of tuples in R, i.e. $|R| = \sum_{\boldsymbol{x} \colon \boldsymbol{X}} R(\boldsymbol{x})$. We denote bag-based relational projection and selection as usual by $\pi_{\boldsymbol{Y}}(R)$ and $\sigma_{\boldsymbol{X}=\boldsymbol{x}}(R)$, respectively.

Functional Dependencies. A *functional dependency* (FD for short) is an expression φ of the form $\boldsymbol{X} \rightarrow \boldsymbol{Y}$. An FD is *linear* if $|\boldsymbol{X}| = |\boldsymbol{Y}| = 1$ and *non-linear* otherwise. A relation $R(\boldsymbol{W})$ with $\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{W}$ *satisfies* φ if for all tuples $\boldsymbol{w}, \boldsymbol{w'} \in R$ we have that $\boldsymbol{w}|_{\boldsymbol{Y}} = \boldsymbol{w'}|_{\boldsymbol{Y}}$ whenever $\boldsymbol{w}|_{\boldsymbol{X}} = \boldsymbol{w'}|_{\boldsymbol{X}}$. We write $R \models \varphi$ to indicate that R satisfies φ, and $R \not\models \varphi$ to indicate that it violates φ. In what follows, we always implicitly assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are disjoint when considering FDs.

Dependency Discovery A *schema* is a finite set of FDs. In the *exact FD discovery problem* we are given a relation R that satisfies all FDs in some fixed design schema $\Delta(R)$, but have no knowledge of $\Delta(R)$ itself. We are then asked to recover $\Delta(R)$ by deriving the largest set $\Lambda \supseteq \Delta(R)$ of FDs that are satisfied by R. In the *approximate FD discovery problem*, we are given a relation R that does not satisfy $\Delta(R)$ and again we are asked to recover $\Delta(R)$. Here, we assume that R is obtained by means of a noisy channel process as follows. From a clean relation R' that satisfies $\Delta(R)$, R is obtained by modifying certain values in tuples in R. We consider an *error* each cell for which R differs from the clean version R'. Note that by running exact FD discovery algorithms on R, we will still be able to recover satisfied FDs in $\Delta(R)$. Our interest in this paper is in *approximate FD discovery*, i.e., deriving the FDs in $\Delta(R)$ that, because of errors introduced, are violated in R and therefore cannot be discovered by exact FD discovery.

In Section IV we will survey various measures that have been proposed to quantify the level to which an FD holds approximately. As we will see, many of these measures are based on exploiting notions of entropy. However, some employ *Shannon* entropy while others employ *logical* entropy. We introduce these notions next.

Probabilities. Both notions of entropy are defined w.r.t. a given joint probability distribution. In our setting, this probability distribution is defined by the relation under consideration. Let $R(\boldsymbol{W})$ be a relation. The joint probability distribution $p_R(\boldsymbol{W})$ over $\boldsymbol{W}$ induced by R is defined by $p_R(\boldsymbol{W} = \boldsymbol{w}) = \frac{R(\boldsymbol{w})}{|R|}$. As such, $p_R(\boldsymbol{W} = \boldsymbol{w})$ is the probability of observing $\boldsymbol{w}$ when randomly drawing a tuple from R. We note that this probability distribution is only well-defined when R is non-empty. Because the empty relation vacuously satisfies all FDs, we will implicitly assume without loss of generality in the rest of this paper that relations are non-empty.

To simplify notation in what follows, we simply write $p_R(\boldsymbol{w})$, $p_R(\boldsymbol{y})$ and $p_R(\boldsymbol{y} \mid \boldsymbol{x})$ instead of $p_R(\boldsymbol{W} = \boldsymbol{w})$, $p_R(\boldsymbol{Y} = \boldsymbol{y})$ and $p_R(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x})$, respectively. The notions of marginal and conditional distributions derived from $p_R(\boldsymbol{w})$ are defined as follows. For the remainder of the section, let $\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{W}$ be disjoint subsets of $\boldsymbol{W}$. Then, $p_R(\boldsymbol{y})$ denotes the marginal probability distribution on $\boldsymbol{Y}$-tuples in R, while $p_R(\boldsymbol{y} \mid \boldsymbol{x})$ is the conditional distribution on $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$:

$$p_R(\boldsymbol{y}) = \sum_{\boldsymbol{w} : \boldsymbol{W} \text{ s.t. } \boldsymbol{w}|_{\boldsymbol{Y}} = \boldsymbol{y}} p_R(\boldsymbol{w}), \qquad p_R(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{p_R(\boldsymbol{xy})}{p_R(\boldsymbol{x})}.$$

It is readily verified that $p_R(\boldsymbol{Y})$ equals the distribution induced by $\pi_Y(R)$, while $p_R(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x})$ equals the distribution induced by $\pi_Y \sigma_{\boldsymbol{X}=\boldsymbol{x}}(R)$.

Shannon Entropy. We write $H_R(\boldsymbol{X})$ for the *Shannon entropy* of $\boldsymbol{X}$ in R, defined as usual [13] by[1]

$$H_R(\boldsymbol{X}) = -\sum_{\boldsymbol{x}\colon \boldsymbol{X}} p_R(\boldsymbol{x}) \log p_R(\boldsymbol{x}).$$

$H_R(\boldsymbol{X})$ reflects the average level of uncertainty inherent in the possible tuples over $\boldsymbol{X}$ in $\pi_X(R)$. The *conditional entropy* $H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ is the uncertainty in $\boldsymbol{Y}$ given $\boldsymbol{X}$, defined as

$$H_R(\boldsymbol{Y} \mid \boldsymbol{X}) = -\sum_{\boldsymbol{x}\colon \boldsymbol{X}, \boldsymbol{y}\colon \boldsymbol{Y}} p_R(\boldsymbol{xy}) \log \frac{p_R(\boldsymbol{xy})}{p_R(\boldsymbol{x})}.$$

Equivalently, denoting by $H_R(\boldsymbol{Y} \mid \boldsymbol{x})$ the Shannon entropy of $\boldsymbol{Y}$ in the conditional distribution $p_R(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x})$, we see that $H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ is the expected value of $H_R(\boldsymbol{Y} \mid \boldsymbol{x})$, taken over all $\boldsymbol{x}$, i.e., $H_R(\boldsymbol{Y} \mid \boldsymbol{X}) = \mathbb{E}_{\boldsymbol{x}}\left[H_R(\boldsymbol{Y} \mid \boldsymbol{x})\right]$.

Logical Entropy. The *logical entropy* of $\boldsymbol{X}$ in R is the probability that two tuples $\boldsymbol{w}$ and $\boldsymbol{w'}$, drawn randomly with replacement from R according to p_R, differ in some attribute in $\boldsymbol{X}$ [14]. That is,

$$h_R(\boldsymbol{X}) := 1 - \sum_{\boldsymbol{x}\colon \boldsymbol{X}} p_R(\boldsymbol{x})^2.$$

Here, $p_R(\boldsymbol{x})^2$ is the probability that two random tuples are exactly equal to $\boldsymbol{x}$ on X.

We denote by $h_R(\boldsymbol{Y} \mid \boldsymbol{x})$ the logical entropy of $\boldsymbol{Y}$ in the conditional distribution $p_R(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x})$, i.e.,

$$h_R(\boldsymbol{Y} \mid \boldsymbol{x}) = 1 - \sum_{\boldsymbol{y}\colon \boldsymbol{Y}} p_R(\boldsymbol{y} \mid \boldsymbol{x})^2.$$

The *logical conditional entropy* of $\boldsymbol{Y}$ given $\boldsymbol{X}$ in R, denoted $h_R(\boldsymbol{Y} \mid \boldsymbol{X})$ is the probability that two tuples drawn at random with replacement from R according to p_R are equal in all attributes of $\boldsymbol{X}$ but differ in some attribute of $\boldsymbol{Y}$,

$$h_R(\boldsymbol{Y} \mid \boldsymbol{X}) := \sum_{\boldsymbol{x},\boldsymbol{y}} p_R(\boldsymbol{xy})[p_R(\boldsymbol{x}) - p_R(\boldsymbol{xy})].$$

Here, the factor $p_R(\boldsymbol{xy})$ expresses the probability of observing $\boldsymbol{xy}$ in the first tuple and the factor $p_R(\boldsymbol{x}) - p_R(\boldsymbol{x}, \boldsymbol{y})$ is the probability that the second tuple has the same value for $\boldsymbol{x}$ but differs in $\boldsymbol{y}$.

Note that, in contrast to the case of Shannon entropy where $H_R(\boldsymbol{Y} \mid \boldsymbol{X}) = \mathbb{E}_{\boldsymbol{x}}[H_R(\boldsymbol{Y} \mid \boldsymbol{x})]$, in logical entropy $h_R(\boldsymbol{Y} \mid \boldsymbol{X}) \neq \mathbb{E}_{\boldsymbol{x}}[h_R(\boldsymbol{Y} \mid \boldsymbol{x})]$. We discuss the relationship between logical and Shannon entropy further in Appendix A

IV. AFD Measures

In this section, we survey the literature on AFD measures. While our focus in this paper is on measures and how they compare, and not on AFD discovery algorithms, for purpose of illustration we also mention for each measure the AFD discovery algorithms in which this measures is used.

[1]Here and in the sequel we use the common convention that $0 \log 0 = 0$ and $\frac{0}{0} = 0$.

AFD measures. Formally, an *AFD measure*, short for *approximate* FD measure, is a function that maps pairs (φ, R), with φ an FD and R a relation, to a number in the interval $[0, 1]$ that indicates the level to which φ holds in R. Higher values are intended to indicate that R makes fewer violations to φ, and we require that $f(\varphi, R) = 1$ if R perfectly satisfies φ. It is important to note that instead of defining AFD measures, some papers in the literature define *error measures* where a high value indicates a high number of errors against the FD. In what follows, we routinely re-define such error measures e into an AFD measure f_e by setting $f_e(\varphi, R) := 1 - e(\varphi, R)$.

Every AFD measure f naturally gives rise to an associated AFD discovery algorithm as follows. From an abstract viewpoint, an AFD discovery algorithm simply consists of a fixed AFD measure f and a threshold $\epsilon \in [0, 1]$. Given a relation $R(\boldsymbol{W})$ the algorithm returns all FDs over $\boldsymbol{W}$ whose f-value lies in the range $[\epsilon, 1[$. (In particular, this excludes the FDs satisfied by R.)

Interpretation and baselines. As with all threshold-based algorithms, a key difficulty for AFD discovery algorithms lies in determining the correct threshold ϵ to use. At its core, this question boils down to how we should interpret the significance of the values returned by f. It is tempting to see the values of f as a percentage with $f(\varphi, R) = 1$ indicating that R perfectly satisfies φ and $f(\varphi, R) = 0$ indicating that R completely fails to satisfy φ. This interpretation, however, is only valid if the measure has a notion of R "completely failing to satisfy" φ. In particular, this is only possible when there are relations for which $f(\varphi, R) = 0$. In what follows, we call a relation R with $f(\varphi, R) = 0$ a *baseline of f for φ*. If f has a baseline for every FD φ then we say that f *has baselines*, otherwise we call f *without baselines*. Having baselines is a necessary condition for interpreting measure scores as percentages.

Conventions. Throughout this section, assume that R is a $\boldsymbol{W}$-relation, let $\boldsymbol{X}, \boldsymbol{Y}$ be disjoint subsets of $\boldsymbol{W}$ and let $\varphi = \boldsymbol{X} \to \boldsymbol{Y}$. We convene that for all measures f that we describe, we trivially set $f(\varphi, R) := 1$ if $R \models \varphi$. So, the definitions that follow only apply when $R \not\models \varphi$. In that case, observe that R must be non-empty, that $|dom_R(\boldsymbol{X})| \neq |R|$ and that $|dom_R(\boldsymbol{Y})| > 1$ since otherwise R trivially satisfies $\boldsymbol{X} \to \boldsymbol{Y}$. As a consequence, $H_R(\boldsymbol{Y}) > 0$ and $h_R(\boldsymbol{Y}) > 0$. This ensures that the denominator of fractions in the formulas that follow are never zero.

A. Co-occurrence ratio

Ilyas et al. [21] consider the derivation of AFDs (called *soft* FDs in their paper) as well as general correlations between attributes. To derive AFDs, they consider the ratio between the number of distinct $\boldsymbol{X}$-tuples and the number of distinct $\boldsymbol{XY}$-tuples occurring in R. We denote this measure by ρ, formally defined as:

$$\rho(\boldsymbol{X} \to \boldsymbol{Y}, R) := \frac{|\,dom_{\boldsymbol{X}}(R)|}{|\,dom_{\boldsymbol{XY}}(R)|}.$$

This ratio is 1 when R satisfies $\boldsymbol{X} \to \boldsymbol{Y}$ and decreases when more $\boldsymbol{y}$-tuples occur with the same $\boldsymbol{x}$-tuple. Note that ρ is a set-based measure, as it ignores the multiplicities of the tuples in R. It is also without baselines, as $|dom_{\boldsymbol{X}}(R)| > 0$ for any non-empty relation R and as, by convention, $\rho(\varphi, R) = 1$ when R is empty.

B. g-measures

Kivinen and Mannila [23] introduced three error measures on set-based relations. Generalized to bag-based relations, and converted to AFD measures, these are the following.

The measure g_1. The measure g_1 is based on logical entropy. Specifically, Kivinen and Manila defined g_1 to reflect the (normalized) number of violating pairs in R. Here, a pair $(\boldsymbol{w}, \boldsymbol{w}')$ of R-tuples is a *violating pair* if they are equal on $\boldsymbol{X}$ but differ on $\boldsymbol{Y}$. Formally, if we denote the bag of violating pairs in $R \times R$ by $G_1(\boldsymbol{X} \to \boldsymbol{Y}, R)$ then, converted to an AFD measure instead of an error measure

$$g_1(\boldsymbol{X} \to \boldsymbol{Y}, R) := \frac{|R|^2 - |G_1(\boldsymbol{X} \to \boldsymbol{Y}, R)|}{|R|^2} = 1 - \frac{|G_1(\boldsymbol{X} \to \boldsymbol{Y}, R)|}{|R|^2} = 1 - h_R(\boldsymbol{Y} \mid \boldsymbol{X}).$$

In other words, g_1 is maximized when the logical conditional entropy is minimized.

The measure g_1 is without baselines: because pairs of the form $(\boldsymbol{w}, \boldsymbol{w})$ are never violating, it is straightforward to see that the total number of violating pairs is bounded from above by $|R|^2 - \sum_{\boldsymbol{w}} R(\boldsymbol{w})^2$. We denote by g_1' the normalized version of g_1,

$$g_1'(\boldsymbol{X} \to \boldsymbol{Y}, R) := 1 - \frac{|G_1(\boldsymbol{X} \to \boldsymbol{Y}, R)|}{|R|^2 - \sum_{\boldsymbol{w}} R(\boldsymbol{w})^2}.$$

The baselines of g_1' are hence those relations for which the set $G_1(\boldsymbol{X} \to \boldsymbol{Y}, R)$ consists of all possible violating pairs.

Both g_1 and g_1' have been used as the basis of AFD discovery algorithms. In particular, g_1 is the basis of FDX [47] while g_1' is the basis of PYRO [25]. Adaptations of g_1' are also used in the context of denial constraints [33] and roll-up dependencies [7].

The measure g_2. Kivinen and Manila defined g_2 to reflect the probability that a random tuple participates in a violating pair. We define $G_2(\boldsymbol{X} \to \boldsymbol{Y}, R)$ to be the set of all tuples in R that participate in a violating pair,

$$G_2(\boldsymbol{X} \to \boldsymbol{Y}, R) := \{\boldsymbol{w} \in R \mid \exists \boldsymbol{w}' \in R, (\boldsymbol{w}, \boldsymbol{w}') \in G_1(\boldsymbol{X} \to \boldsymbol{Y}, R)\}.$$

Then, g_2, converted to an AFD measure instead of an error measure as originally proposed, computes the probability that a tuple, drawn randomly from R according to p_R, is not part of a violating pair,

$$g_2(\boldsymbol{X} \to \boldsymbol{Y}, R) := 1 - \sum_{\boldsymbol{w} \in G_2(\boldsymbol{X} \to \boldsymbol{Y}, R)} p_R(\boldsymbol{w}).$$

The FD-compliance-ratio that is used as one of the building blocks in UNI-DETECT [44] is based on g_2.

The measure g_3. The measure g_3 computes the relative size of a maximal subrelation of R for which $\boldsymbol{X} \to \boldsymbol{Y}$ holds. Specifically, define $R'(\boldsymbol{W})$ to be a subrelation of $R(\boldsymbol{W})$, denoted $R' \subseteq R$, if $R'(\boldsymbol{w}) \leq R(\boldsymbol{w})$ for all $\boldsymbol{w}\colon \boldsymbol{W}$. Let $G_3(\boldsymbol{X} \to \boldsymbol{Y}, R)$ denote the set of all subrelations of R that satisfy $\boldsymbol{X} \to \boldsymbol{Y}$,

$$G_3(\boldsymbol{X} \to \boldsymbol{Y}, R) := \{R' \mid R' \subseteq R, R' \models \boldsymbol{X} \to \boldsymbol{Y}\}.$$

Then g_3 is defined as the maximum relative size of a subrelation satisfying $\boldsymbol{X} \to \boldsymbol{Y}$:

$$g_3(\boldsymbol{X} \to \boldsymbol{Y}, R) := \max_{R' \in G_3(\boldsymbol{X} \to \boldsymbol{Y}, R)} \frac{|R'|}{|R|}.$$

Note that $1 - g_3(\boldsymbol{X} \to \boldsymbol{Y}, R)$ can naturally be interpreted as the minimum fraction of tuples that need to be removed for $\boldsymbol{X} \to \boldsymbol{Y}$ to hold in R.

The measure g_3 is without baselines. Indeed, for any non-empty R we can always obtain a subrelation $R' \in G_3(\varphi, R)$ of size $|dom_{\boldsymbol{X}}(R)|$ by arbitrarily fixing one $\boldsymbol{y}$-value for each $\boldsymbol{x}$-value. As such, g_3 is bounded from below by $\frac{|dom_{\boldsymbol{X}}(R)|}{|R|} > 0$. Gianella and Robertson [18] proposed a normalized variant g_3' of g_3, defined as follows:

$$g_3'(\boldsymbol{X} \to \boldsymbol{Y}, R) := \max_{R' \in G_3(\boldsymbol{X} \to \boldsymbol{Y}, R)} \frac{|R'| - |dom_R(\boldsymbol{X})|}{|R| - |dom_R(\boldsymbol{X})|}.$$

This variant has baselines, namely all relations R for which no subrelation $R' \in G_3(\varphi, R)$ is larger than $|dom_R(\boldsymbol{X})|$.

The unnormalized measure g_3 is used in multiple AFD discovery algorithms [23], [2], [20], [22]. Furthermore, the 'per-tuple' probability of an FD as defined in [43] is precisely g_3. Berzal et al. [3] use it as the basis for relational decomposition based on AFDs instead of FDs. Exact and approximate solutions for the computation of g_3 in the context of non-crisp FDs are proposed in [17]. We note that g_3 has been generalized to other dependencies as well: e.g., conditional FDs [12], [37], inclusion dependencies [29], and conditional matching dependencies [45]. By contrast, apart from [18] we know of no other work that considers the normalized version g_3'.

C. Fraction of information

Fraction of Information. Cavallo and Pittarelli [9] introduced *fraction of information* (FI) as a way to generalize FDs from deterministic to probabilistic databases. Usage of FI as an AFD measure was later studied by Giannelli and Robertson [18] . FI is based on Shannon entropy and is formally defined as

$$\text{FI}(\boldsymbol{X} \to \boldsymbol{Y}, R) := \frac{H_R(\boldsymbol{Y}) - H_R(\boldsymbol{Y} \mid \boldsymbol{X})}{H_R(\boldsymbol{Y})}.$$

The numerator $H_R(\boldsymbol{Y}) - H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ is known as *mutual information* [13], which we further denote by $I_R(\boldsymbol{X}; \boldsymbol{Y})$.

We can understand FI as follows. $H_R(\boldsymbol{Y})$ measures the uncertainty of observing $\boldsymbol{Y}$ while $H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ measures the uncertainty of observing $\boldsymbol{Y}$ after observing $\boldsymbol{X}$. FI hence represents the proportional reduction of uncertainty about $\boldsymbol{Y}$ that is achieved by knowing $\boldsymbol{X}$. When R satisfies $\boldsymbol{X} \to \boldsymbol{Y}$

there is no uncertainty about $\boldsymbol{Y}$ after observing $\boldsymbol{X}$ and hence $H_R(\boldsymbol{Y} \mid \boldsymbol{X}) = 0$ and so FI is 1. Conversely, when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent random variables in p_R then there is no reduction in uncertainty, and hence $H_R(\boldsymbol{Y} \mid \boldsymbol{X}) = H_R(\boldsymbol{Y})$ and so FI is 0. Thus, the baselines of FI for $\boldsymbol{X} \to \boldsymbol{Y}$ are those relations R for which $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent in p_R.

Bias. Mandros et al. [27], [28] and Pennerath et al. [34] proposed two refinements to FI specifically for AFD discovery, called *reliable FI* (RFI) and *smoothed FI* (SFI), respectively. They are motivated in proposing these refinements by the following observation. Consider a relation $S(\boldsymbol{W})$ and assume that we are given relation $R(\boldsymbol{W})$ of size n that is obtained by sampling n tuples from S according to distribution p_S. Further assume that we do not have access to S and wish to determine $\text{FI}(\boldsymbol{X} \to \boldsymbol{Y}, S)$ based on R. Then a result by Roulston [40] states that the expected value of $I_R(\boldsymbol{X}; \boldsymbol{Y})$, taken over all R obtained in this manner, equals

$$I_S(\boldsymbol{X}; \boldsymbol{Y}) + \frac{|\mathit{dom}_S(\boldsymbol{XY})| - |\mathit{dom}_S(\boldsymbol{X})| - |\mathit{dom}_S(\boldsymbol{Y})| + 1}{2n}.$$

In other words, we may expect $I_R(\boldsymbol{X}; \boldsymbol{Y})$ to overestimate $I_S(\boldsymbol{X}; \boldsymbol{Y})$ and the magnitude of overestimation depends on the size of the active domains of $\boldsymbol{XY}$, $\boldsymbol{X}$, and $\boldsymbol{Y}$ in S, as well as on n. Because in addition $H_R(\boldsymbol{Y})$ underestimates $H_S(\boldsymbol{Y})$ [40], we may conclude that $\text{FI}(\boldsymbol{X} \to \boldsymbol{Y}, R)$ is expected to overestimate $\text{FI}(\boldsymbol{X} \to \boldsymbol{Y}, S)$ and the magnitude of overestimation depends on the active domain sizes and the size of S. This overestimation is problematic since even if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent in p_S, and $\text{FI}(\boldsymbol{X} \to \boldsymbol{Y}, S)$ is hence 0, $\text{FI}(\boldsymbol{X} \to \boldsymbol{Y}, R)$ will be quite large.

Reliable FI. Reliable FI corrects for this bias by subtracting the mutual information value that is expected under random $(\boldsymbol{X}; \boldsymbol{Y})$-permutations.

Definition 1. *Relation R' is an $(\boldsymbol{X}; \boldsymbol{Y})$-*permutation *of R, denoted $R \sim_{\boldsymbol{X};\boldsymbol{Y}} R'$ if (i) $|R| = |R'|$; (ii) $\pi_{\boldsymbol{X}}(R) = \pi_{\boldsymbol{X}}(R')$; (iii) $\pi_{\boldsymbol{Y}}(R) = \pi_{\boldsymbol{Y}}(R')$; and (iv) $\pi_{\boldsymbol{Z}}(R) = \pi_{\boldsymbol{Z}}(R')$ where $\boldsymbol{Z} = \boldsymbol{W} \setminus \boldsymbol{XY}$.*

In particular, R' and R have the same marginal distributions on $\boldsymbol{X}$ and on $\boldsymbol{Y}$, $p_{R'}(\boldsymbol{X}) = p_R(\boldsymbol{X})$ and $p_{R'}(\boldsymbol{Y}) = p_R(\boldsymbol{Y})$. In what follows, for a measure f, we denote by $\mathbb{E}_R[f(\boldsymbol{X} \to \boldsymbol{Y}, R))]$ the expected value of $f(\boldsymbol{X} \to \boldsymbol{Y}, R)$ where the expectation is taken over all $(\boldsymbol{X}; \boldsymbol{Y})$-permutations of R.

Reliable fraction of information is then defined as

$$\text{RFI}(\boldsymbol{X} \to \boldsymbol{Y}, R) := \text{FI}(\boldsymbol{X} \to \boldsymbol{Y}, R) - \mathbb{E}_R[\text{FI}(\boldsymbol{X} \to \boldsymbol{Y}, R)].$$

Because the number of permutations of R is finite, we may compute $\mathbb{E}_R[\text{FI}(\boldsymbol{X} \to \boldsymbol{Y}, R)]$, and hence also $\text{RFI}(\boldsymbol{X} \to \boldsymbol{Y}, R)$, by simply computing $\text{FI}(\boldsymbol{X} \to \boldsymbol{Y}, R')$ for every permutation R' of R and taking the average. More efficient algorithms are proposed in [28], [27]. Even with these improved algorithms computing RFI remains inefficient, however, as we show in Section VI.

Strictly speaking, RFI is not an AFD measure since it can become negative when $\text{FI}(\varphi, R) < \mathbb{E}_R[\text{FI}(\varphi, R)]$. Because such negative RFI values indicate that there is weak evidence to conclude that φ is an AFD, we turn RFI into an actual AFD measure RFI^+ by setting

$$\text{RFI}^+(\boldsymbol{X} \to \boldsymbol{Y}, R) := \max(\text{RFI}(\boldsymbol{X} \to \boldsymbol{Y}, R), 0).$$

The baselines of RFI^+ for $\boldsymbol{X} \to \boldsymbol{Y}$ are hence all relations whose FI value is smaller or equal than the expected value under random permutations.

Smoothed FI. Smoothed FI uses *laplace smoothing* to reduce bias. Laplace smoothing is a well-known statistical technique to reduce estimator variance. It is parameterized by a value $\alpha > 0$. Specifically, for a relation $S(\boldsymbol{XY})$ let $S^{(\alpha)}$ denote the α-smoothed version of S, defined by $S^{(\alpha)}(\boldsymbol{xy}) := S(\boldsymbol{xy}) + \alpha$ for every $\boldsymbol{x} \in \mathit{dom}_S(\boldsymbol{X})$ and $\boldsymbol{y} \in \mathit{dom}_S(\boldsymbol{Y})$. Note in particular, that it is possible that $S(\boldsymbol{xy}) = 0$, in which case $S^{(\alpha)}(\boldsymbol{xy}) = \alpha$. Then the smoothed FI of R is simply the normal FI of the α-smoothed version of $\pi_{\boldsymbol{XY}}(R)$:

$$\text{SFI}_\alpha(\boldsymbol{X} \to \boldsymbol{Y}, R) := \text{FI}(\boldsymbol{X} \to \boldsymbol{Y}, \pi^{(\alpha)}_{\boldsymbol{XY}}(R)).$$

We note that, because $\pi^{(\alpha)}_{\boldsymbol{XY}}(R)$ contains a tuple $\boldsymbol{xy}$ for every possible combination of $\boldsymbol{x} \in \mathit{dom}_{\boldsymbol{X}}(R)$ and $\boldsymbol{y} \in \mathit{dom}_{\boldsymbol{Y}}(R)$, it can be many times larger than R. SFI is therefore also relatively inefficient to compute, as we show in Section VI.

AFD discovery algorithms based on RFI and SFI are presented in [27], [28] and [34], respectively.

D. Probabilistic dependency, τ and μ

Piatetsky-Shapiro and Matheus [36] proposed *probabilistic dependency* as another probabilistic generalization of a functional dependency. They also introduced a normalized version of probabilistic dependency, which is equivalent to the Goodman and Kruskal τ measure of association [19]. Finally, they also propose a rescaled version of τ. All three notions are defined as follows. It is worth noting, that, apart from [36], we are not aware of any work that considers these measures for AFD discovery in the database context, let alone designs AFD discovery algorithms for them.

Probabilistic dependency. The *probabilistic dependency of $\boldsymbol{Y}$ on $\boldsymbol{X}$ in R*, denoted by $\mathit{pdep}(\boldsymbol{X} \to \boldsymbol{Y}, R)$, represents the conditional probability that two tuples drawn randomly with replacement from R are equal on $\boldsymbol{Y}$, given that they are equal on $\boldsymbol{X}$. Formally,

$$\mathit{pdep}(\boldsymbol{X} \to \boldsymbol{Y}, R) := \sum_{\boldsymbol{x}} p_R(\boldsymbol{x})\, \mathit{pdep}(\boldsymbol{Y} \mid \boldsymbol{x}, R),$$

where $\mathit{pdep}(\boldsymbol{Y} \mid \boldsymbol{x}, R)$ is the probability that two random $\boldsymbol{Y}$-tuples drawn with replacement from the conditional distribution $p_R(\boldsymbol{Y} \mid \boldsymbol{x})$ are equal:

$$\mathit{pdep}(\boldsymbol{Y} \mid \boldsymbol{x}, R) := \sum_{\boldsymbol{y}} p_R(\boldsymbol{y} \mid \boldsymbol{x})^2 = 1 - h_R(\boldsymbol{Y} \mid \boldsymbol{x}).$$

Probabilistic dependency is hence a measure based on logical entropy. It can be understood as follows. Suppose that we are given two tuples that equal $\boldsymbol{x}$ on $\boldsymbol{X}$. Then $\mathit{pdep}(\boldsymbol{Y} \mid \boldsymbol{x}, R)$ is the probability that these tuples are also equal on $\boldsymbol{Y}$, and

$pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)$ is the expected value of $pdep(\boldsymbol{Y} \mid \boldsymbol{x}, R)$ over all $\boldsymbol{x}$.

We note that probabilistic dependency can also be seen as a generalization of the measure g_2. Whereas g_2 computes the probability that a random tuple cannot be extended to a violating pair, probabilistic dependency computes the average conditional probability that a given $\boldsymbol{X}$-tuple $\boldsymbol{x}$ cannot be extended to a violating pair, where the average is taken over all values of $\boldsymbol{X}$.

The measure τ. It is straightforward to see that $pdep(\boldsymbol{X} \to \boldsymbol{Y}, R) > 0$, always. As such, *pdep* is a measure without baselines. In fact, Piatetsky-Shapiro and Matheus [36] show that we always have

$$pdep(\boldsymbol{X} \to \boldsymbol{Y}, R) \geq pdep(\boldsymbol{Y}, R)$$

where $pdep(\boldsymbol{Y}, R)$, called *probabilistic self-dependency*, is defined as the probability that two random tuples in R have equal $\boldsymbol{Y}$ attributes,

$$pdep(\boldsymbol{Y}, R) := \sum_{\boldsymbol{y}} p_R(\boldsymbol{y})^2 = 1 - h_R(\boldsymbol{Y}).$$

To account for the relationship between $pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)$ and $pdep(\boldsymbol{Y}, R)$, Piatetsky-Shapiro and Matheus propose to normalize $pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)$ w.r.t. $pdep(\boldsymbol{Y}, R)$. The resulting measure is equivalent to the Goodman and Kruskal τ (tau) measure of association [19], which is defined as

$$\tau(\boldsymbol{X} \to \boldsymbol{Y}, R) := \frac{pdep(\boldsymbol{X} \to \boldsymbol{Y}, R) - pdep(\boldsymbol{Y}, R)}{1 - pdep(\boldsymbol{Y}, R)}.$$

Piatetsky-Shapiro and Matheus explain τ in the following way [36]. Suppose we are given a tuple drawn randomly from R according to p_R, and we need to guess its $\boldsymbol{Y}$ value. One strategy is to make guesses randomly according to the marginal distribution of $\boldsymbol{Y}$, i.e. guess value $\boldsymbol{Y} = \boldsymbol{y}$ with probability $p_R(\boldsymbol{y})$. Then the probability for a correct guess is $pdep(\boldsymbol{Y}, R)$. If we also know that item has $\boldsymbol{X} = \boldsymbol{x}$, we can improve our guess using conditional probabilities of $\boldsymbol{Y}$, given that $\boldsymbol{X} = \boldsymbol{x}$. Then our probability for success, averaged over all values of $\boldsymbol{X}$, is $pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)$, and $\tau(\boldsymbol{X} \to \boldsymbol{Y}, R)$ is the relative increase in our probability of successfully guessing $\boldsymbol{Y}$, given $\boldsymbol{X}$. The baselines of τ for $\boldsymbol{X} \to \boldsymbol{Y}$ are hence those relations where this relative increase is zero.

The measure μ. Piatetsky-Shapiro and Matheus [36] note that *pdep* and τ have the following undesirable property.

Theorem 1 (Piatetsky-Rotem-Shapiro [36])**.** *Given a random relation R of size $N \geq 2$ containing attributes $\boldsymbol{X}$ and $\boldsymbol{Y}$, where $\boldsymbol{X}$ has $K = |dom_R(\boldsymbol{X})|$ distinct values in its active domain, the expected values of pdep and τ under random permutations of R are*

$$\mathbb{E}_R[pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)] = pdep(\boldsymbol{Y}, R) + \frac{K-1}{N-1}(1 - pdep(\boldsymbol{Y}, R)),$$

$$\mathbb{E}_R[\tau(\boldsymbol{X} \to \boldsymbol{Y}, R)] = \frac{|dom_R(\boldsymbol{X})| - 1}{|R| - 1}.$$

Thus, for a fixed distribution of $\boldsymbol{Y}$ values, $\mathbb{E}_R[pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)]$ depends only on the number of distinct $\boldsymbol{X}$ values and not on their relative frequency. Moreover, the formula for $\mathbb{E}_R[\tau(\boldsymbol{X} \to \boldsymbol{Y}, R)]$ tells us that if we have two candidate AFDs with the same right hand side, $\boldsymbol{X} \to \boldsymbol{Y}$ and $\boldsymbol{Z} \to \boldsymbol{Y}$, then if $|dom_R(\boldsymbol{Z})| > |dom_R(\boldsymbol{X})|$, we may expect τ to score $\boldsymbol{Z} \to \boldsymbol{Y}$ better than $\boldsymbol{X} \to \boldsymbol{Y}$, regardless of any intrinsic better relationship between $\boldsymbol{Z}$ and $\boldsymbol{Y}$ over $\boldsymbol{X}$ and $\boldsymbol{Y}$ in R. In response, Piatetsky-Shapiro and Matheus compensate for this effect by introducing the measure μ which normalizes $pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)$ with respect to $\mathbb{E}_R[pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)]$ instead of $pdep(\boldsymbol{Y}, R)$:[2]

$$\begin{aligned}\mu(\boldsymbol{X} \to \boldsymbol{Y}, R) &:= \frac{pdep(\boldsymbol{X} \to \boldsymbol{Y}, R) - \mathbb{E}_R[pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)]}{1 - \mathbb{E}_R[pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)]} \\ &= 1 - \frac{1 - pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)}{1 - pdep(\boldsymbol{Y}, R)} \frac{|R| - 1}{|R| - |dom_R(\boldsymbol{X})|}\end{aligned}$$

Strictly speaking, μ is not a measure since it returns negative values when $pdep(\boldsymbol{X} \to \boldsymbol{Y}, R) > \mathbb{E}_R[pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)]$. Because such negative μ values indicate that there is weak evidence to conclude that φ is an AFD, we turn μ into an actual AFD measure μ^+ by setting

$$\mu^+(\boldsymbol{X} \to \boldsymbol{Y}, R) := \max(\mu(\boldsymbol{X} \to \boldsymbol{Y}, R), 0).$$

The baselines of μ^+ for $\boldsymbol{X} \to \boldsymbol{Y}$ are hence all relations where the $pdep(\boldsymbol{X} \to \boldsymbol{Y})$ value is smaller or equal to the expected value under random permuations.

E. Classes of AFD measures.

Among the measures introduced previously, we discern the following three classes (see also the second row in Table II):

(1) The class of measures that have a notion of "violation" and quantify the number of violations, consisting of ρ, g_2, g_3, and g_3'. We denote this class by VIOLATION (V).
(2) The class of measures based on Shannon entropy, consisting of $\mathrm{FI}, \mathrm{RFI}^+$, and SFI. We denote this class by SHANNON (S).
(3) The class of measures based on logical entropy, consisting of $g_1, g_1', pdep, \tau$, and μ^+ and denoted by LOGICAL (L).

The new measure g_1^S. The measure g_1 is based on logical entropy. We may hence view $1 - H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ as the Shannon equivalent of g_1, where logical entropy is replaced by Shannon entropy. Giannella and Robertson [18] observed that $1 - H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ has the range $[-\infty, 1]$ instead of $[0, 1]$ and therefore disregard it as an AFD measure. Nevertheless, we propose the following Shannon variant for our comparison, obtained by limiting $1 - H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ to be positive:

$$g_1^S(\boldsymbol{X} \to \boldsymbol{Y}, R) := \max(1 - H_R(\boldsymbol{Y} \mid \boldsymbol{X}), 0).$$

[2]This fraction is ill-defined if the denominator $1 - \mathbb{E}_R[pdep(\varphi, R)] = 0$. This only happens, however, when $R \models \varphi$, which we have assumed not to be the case throughout this section, since we have already convened to set $\mu(\varphi, R) = 1$ whenever $R \models \varphi$. See the Appendix for a proof.

The new measure RFI$'^+$. We also observe a conceptual similarity between μ and RFI: μ corrects for the bias of τ under random permutations while RFI corrects for the bias of FI under random permutations. Despite this conceptual similarity, note that the corrections are done differently: μ corrects by taking the *normalized* difference between *pdep* and $\mathbb{E}_R[pdep]$ while RFI corrects by taking the *absolute* difference between FI and $\mathbb{E}_R[\text{FI}]$. As such RFI is not a normalized measure. Since it is natural to ask what the normalized variant of RFI is and how it behaves, we define

$$\text{RFI}'^+(\varphi, R) := \max\left(\frac{\text{FI}(\varphi, R) - \mathbb{E}_R[\text{FI}(\varphi, R)]}{1 - \mathbb{E}_R[\text{FI}(\varphi, R)]}, 0\right).$$

V. Sensitivity Analysis

In this section, we investigate the sensitivity of measures w.r.t. structural properties of the input relation. Specifically, we want to get insight into the measures' ability to distinguish between an FD $\boldsymbol{X} \to \boldsymbol{Y}$ in relation instances that were generated to satisfy the FD, but subsequently had errors introduced so that the FD no longer holds exactly, versus relation instances where $\boldsymbol{X}$ and $\boldsymbol{Y}$ were randomly generated. A good FD measure should be able to consistently distinguish between these two cases, giving high scores to the former and low scores to the latter with a clear separation in scores between the two cases.

A. Methodology

There are various properties of the input relation R that may affect the measures' power to distinguish between these two aforementioned cases. We study the effect on the distinguishing power for $\boldsymbol{X} \to \boldsymbol{Y}$ w.r.t. the following.

(1) The *error rate*, i.e., the amount of errors introduced.

A reasonable requirement for a good AFD measure is that it should be inversely proportional to the error rate: an increase in the number of introduced errors should result in a decrease of the measure value. As discussed below, this is not the case for all considered measures.

(2) Statistics of left-hand-side $\boldsymbol{X}$ of the FD and the right-hand-side $\boldsymbol{Y}$ of the FD. Specifically, define *left-hand-side (LHS) uniqueness* of the input relation R as the ratio $|dom_R(\boldsymbol{X})|/|R|$, quantifying the uniqueness of $\boldsymbol{X}$. Define the *right-hand-side (RHS) skew of* R, as the skewness of the distribution $p_R(\boldsymbol{Y})$. It has previously been observed in the literature [36], [34], [27], [28] that measures can be biased w.r.t. LHS-uniqueness and RHS-skew in the sense that measures may give very high scores to relations with high LHS-uniqueness or RHS-skew values. However, these statistics alone do *not* provide a good signal for quantifying the approximateness of $\boldsymbol{X} \to \boldsymbol{Y}$ because they look only at $\boldsymbol{X}$ or only at $\boldsymbol{Y}$ but not at their correlation. We note that this is in particular true for non-linear FDs: as the number of attributes in $\boldsymbol{X}$ grows, LHS uniqueness naturally tends to 1. Thus, a measure that is biased w.r.t. LHS-uniqueness risks returning too many non-linear FDs, especially on tables with many attributes.

We have created three synthetic benchmarks, denoted Err, Uniq, and Skew to study the measures' sensitivity to errors, LHS uniqueness, and RHS-skew, respectively. Each synthetic benchmark $\mathcal{B}$ consists of relations $R(\boldsymbol{XY})$ with $\boldsymbol{X} = \{X\}$ and $\boldsymbol{Y} = \{Y\}$ and is partitioned into two subsets: (1) $\mathcal{B}^-$ containing relations R where $\boldsymbol{X} \to \boldsymbol{Y} \notin \Delta(R)$; and (2) $\mathcal{B}^+$ containing relations where $\boldsymbol{X} \to \boldsymbol{Y} \in \Delta(R)$. (Recall that $\Delta(R)$ denotes the design schema of R.) Each subset employs a distinct random process to generate relations. For relations in $\mathcal{B}^-$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ values are generated independently at random, while relations in $\mathcal{B}^+$ are generated by first constructing a relation R such that $R \models \boldsymbol{X} \to \boldsymbol{Y}$, and then passing R through a controlled noisy error channel.

Generation process. The generation process of a relation R depends on a number of parameters that are drawn uniformly at random from the following ranges: $|R| \in [100; 10000]$; $|dom_R(\boldsymbol{X})| \in [\frac{1}{5}|R|, \frac{3}{4}|R|]$, $|dom_R(\boldsymbol{Y})| \in [5, \frac{1}{2}|dom_R(\boldsymbol{X})|]$; and error rate $\eta \in [0.5\%, 2\%]$. Values for $\boldsymbol{X}$ and $\boldsymbol{Y}$ are drawn according to the Beta distribution[3], $B(\alpha, \beta)$, which is a family of continuous probability distributions defined on the interval $[0, 1]$ in terms of two positive parameters α and β that control the shape of the distribution. We consider the ranges $\alpha \in (0, 1]$ and $\beta \in [1, 10]$. For $\alpha = \beta = 1$ the distribution is uniform and for any other values it is reverse J-shaped with a right tail. The *skewness* is defined as $\frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$ and is known to measure the asymmetry of the probability distribution about its mean. In particular, the skew is zero for the uniform distribution and increasing values indicate longer tails with lower mass, that is, a higher mass near the left end of the interval $[0, 1]$. We sample values for α and β such that the skewness is at most one (except for Skew below where we consider skew values up to 10).

So, for every relation R the parameters $|R|, |dom_R(\boldsymbol{X})|, |dom_R(\boldsymbol{Y})|, \alpha_{\boldsymbol{X}}, \beta_{\boldsymbol{X}}, \alpha_{\boldsymbol{Y}}, \beta_{\boldsymbol{Y}}, \eta$ are chosen uniformly at random under the conditions described above. To generate a table R in $\mathcal{B}^-$, the following procedure is repeated $|R|$ times: sample $\boldsymbol{x} \in dom_R(\boldsymbol{X})$ (resp., $\boldsymbol{y} \in dom_R(\boldsymbol{Y})$) according to $B(\alpha_{\boldsymbol{X}}, \beta_{\boldsymbol{X}})$ (resp., $B(\alpha_{\boldsymbol{Y}}, \beta_{\boldsymbol{Y}})$) and add $(\boldsymbol{x}, \boldsymbol{y})$ to R. To generate a table R in $\mathcal{B}^+$, we first construct a dictionary D by, for each value $\boldsymbol{x} \in dom_R(\boldsymbol{X})$, assigning a value $D(\boldsymbol{x}) \in dom_R(\boldsymbol{Y})$ drawn at random according to $B(\alpha_{\boldsymbol{Y}}, \beta_{\boldsymbol{Y}})$. Then, we populate R by adding $|R|$ tuples $(\boldsymbol{x}, D(\boldsymbol{x}))$ where $\boldsymbol{x} \in dom_R(\boldsymbol{X})$ is drawn at random according to $B(\alpha_{\boldsymbol{X}}, \beta_{\boldsymbol{X}})$. By construction, R satisfies the FD $\boldsymbol{X} \to \boldsymbol{Y}$. We then pass R through a controlled error channel such that, denoting by R' the obtained relation, R' does not satisfy $\boldsymbol{X} \to \boldsymbol{Y}$ anymore. Concretely, we modify $k = \lfloor \eta |R| \rfloor$ tuples $\boldsymbol{w} = (\boldsymbol{x}, D(\boldsymbol{x}))$, where η indicates the error rate, by randomly picking any $\tilde{\boldsymbol{w}} \in R$ with $\tilde{\boldsymbol{w}}|_{\boldsymbol{Y}} \neq \boldsymbol{w}|_{\boldsymbol{Y}}$ and make $\tilde{\boldsymbol{w}}|_{\boldsymbol{Y}}$ the new value for $\boldsymbol{w}|_{\boldsymbol{Y}}$. We point out that this does not introduce any new $\boldsymbol{Y}$-values and keeps $dom_R(\boldsymbol{Y})$ stable. We also experimented with other error channels that introduce new $\boldsymbol{Y}$ values, but the results were similar and are therefore omitted. Note that $\boldsymbol{X}$ is not modified, and therefore

[3] https://en.wikipedia.org/wiki/Beta_distribution

$p_{R'}(\boldsymbol{X}) = p_R(\boldsymbol{X})$. We note that the generation process is related to the one from Zhang et al. [47] but with the addition of value distributions for both $\boldsymbol{X}$ and $\boldsymbol{Y}$ based on the Beta distribution.

The three synthetic benchmarks are created by controlling one of the parameters in the parameter set as follows. Every benchmark $\mathcal{B}$ consists of 2500 $\mathcal{B}^-$ tables and 2500 $\mathcal{B}^+$ tables.

Benchmark ERR. We iteratively increase the error rate η from 0% to 10% in 50 steps and generate 50 relations in ERR$^+$ per step, varying all other parameters as described above. ERR is then extended with 2500 tables generated in ERR$^-$.

Benchmark UNIQ. We construct UNIQ by iteratively increasing LHS-uniqueness from $\frac{1}{5}|R|$ to $10|R|$ in 50 steps and by generating for every step 50 tables in UNIQ$^+$ and UNIQ$^-$.

Benchmark SKEW. We construct SKEW by iteratively increasing RHS-skew from 0 to 10 in 50 steps and by generating for every step 50 tables in SKEW$^+$ and SKEW$^-$.

B. Results

We describe the results on the basis of Figure 1, which, for each benchmark $\mathcal{B}$ and measure f plots $\delta(f, \mathcal{B})$, the difference between average measure values on $\mathcal{B}^+$ and average measure values on $\mathcal{B}^-$,

$$\delta(f, \mathcal{B}) := \text{avg}_{R \in \mathcal{B}^+} f(\boldsymbol{X} \to \boldsymbol{Y}, R) \; - \; \text{avg}_{R \in \mathcal{B}^-} f(\boldsymbol{X} \to \boldsymbol{Y}, R).$$

We also call $\delta(f, \mathcal{B})$ the *separation* of f on $\mathcal{B}$. When it is small, f cannot distinguish between cases where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are sampled independently at random and where data is generated according to our generation process for $\mathcal{B}^+$. Values for g_1 and g_1' are grouped in Figure 1 as their separations are identical. Our conclusions are summarized in Table II.

Error rate. The top row of Figure 1 plots the separation on ERR as a function of error rate η. For g_1 and g_1', the separation is zero, while for SFI it is nearly zero. This means that these measures have limited distinguishing power and are not well-suited as a yardstick for assessing the amount of errors w.r.t. an FD. For all other measures, there is a clear separation, albeit less pronounced for FI and RFI$^+$. As expected, when the error level increases the separation decreases, save for g_1, g_1', and SFI where it remains constant. The measures hence become less certain of having found an AFD as the error rate increases. While FI and RFI$^+$ also decrease as η increases, this decrease is less steep than for the other measures.

LHS-uniqueness. The middle row of Figure 1 shows the separation on UNIQ as a function of LHS-uniqueness. For g_1, g_1' and SFI we see the same behavior as on ERR: their separation is (nearly) zero; they hence lack distinguishing power. Because it would be misleading to label g_1, g_1' and SFI as being insensitive to LHS-uniqueness, we indicate in Table II that LHS-uniqueness is inapplicable with the symbol _. The distinguishing power of g_3', RFI$'^+$, and μ^+ is not affected by LHS-uniqueness as the separation remains large for all values of LHS-uniqueness. We do observe that the separation decreases slightly for very large LHS-uniqueness levels, indicating that these measures become less confident to have found an FD $\boldsymbol{X} \to \boldsymbol{Y}$ in a relation R when $\pi_{\boldsymbol{X}}(R)$ contains fewer duplicates.

For all other measures the separation drops as LHS-uniqueness increases, tending to zero at maximum LHS-uniqueness levels. These measures are hence biased w.r.t. LHS-uniqueness. As discussed before, it will therefore prove problematic to discover non-linear AFDs by means of these measures.

RHS-skew. The bottom row of Figure 1 shows the separation on SKEW as a function of RHS-skew. The measures g_1, g_1', and SFI exhibit the same behavior as before, with (nearly) zero separation. We indicate the corresponding cells in Table II with the symbol _. The distinguishing power of all VIOLATION measures, as well as g_1^S, and *pdep* drops when RHS-skew increases. These measures are thus biased w.r.t. RHS-skew. By contrast, FI, RFI$^+$, RFI$'^+$, τ, and μ^+ correct for this behavior and are insensitive to RHS-skew.

Conclusion. The measures g_1, g_1' and SFI are the least suitable AFD measures since, by contrast to the other measures, they do not clearly separate relations in $\mathcal{B}^+$ from relations in $\mathcal{B}^-$ for any of the three considered sensitivity parameters. The measures g_3', RFI$'^+$, and μ^+ have a built-in mechanism that corrects for LHS-uniqueness which is a most desirable property when discovering non-linear FDs. The SHANNON measures (save g_1^S and SFI), τ, and μ^+ correct for RHS-skew. The most desirable measures are therefore RFI$'^+$ and μ^+ as they both are insensitive to LHS-uniqueness and RHS-skew, and are inversely proportional to the error level.

VI. EVALUATION ON REAL-WORLD DATA

In this section, we compare the effectiveness of the described AFD measures on RWD: a new benchmark consisting of real-world tables which hence exhibits data distributions as well as data errors that occur in practice. Because of space limitations, we focus on our most important findings. Complementary information and additional experiments may be found in the Appendix.

A. Overview

FDs in relations with NULLs. The relations that we consider in this section come from practical domains and often also contain NULL values. Because it is unclear whether two distinct occurrences of a NULL should be considered the same value, or distinct values, there is no clear semantics of FDs in the presence of NULL values. We therefore ignore NULL values when checking FD satisfaction and calculating measure scores. That is, if $R(\boldsymbol{W})$ is a relation with NULLs and $\varphi = \boldsymbol{X} \to \boldsymbol{Y}$ an FD, then we consider φ to be satisfied if it is satisfied in the subrelation R' of R consisting of all tuples $\boldsymbol{w} \in R$ for which $\boldsymbol{w}(A) \neq \text{NULL}$ for all $A \in \boldsymbol{XY}$. Similarly, the score of measure f on (φ, R) is computed by computing $f(\varphi, R')$ instead.

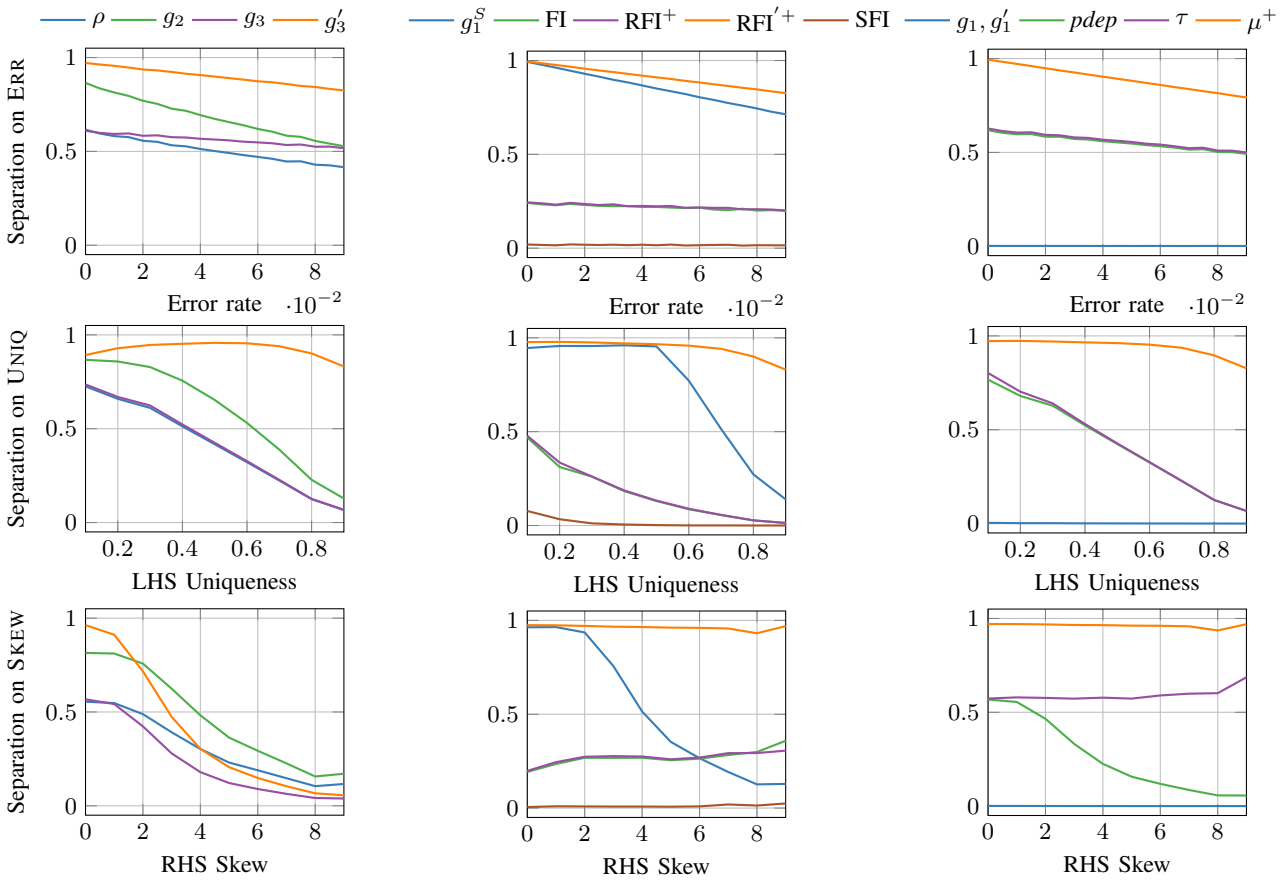

Fig. 1: Increasing error rates, LHS-uniqueness levels or RHS-skew levels impacts most measures' ability to seperate between $\mathcal{B}^+$ and $\mathcal{B}^-$. The plots show the separation on ERR (top), UNIQ (middle) and SKEW (bottom) benchmarks.

Real world data benchmark (RWD). We created the RWD benchmark as follows. We started by considering all relations mentioned in [4], which collects the real-world relations most commonly used in the dependency discovery literature. This base set was extended with the relation Adult used, e.g., in [10], [25], [46]. Since design schemas for these relations are unavailable, we manually created them as follows. First, in order to ensure semantically sound design schemas, we restricted our attention to the subset of relations that have a generally interpretable domain. Further, to keep the manual annotation endeavor manageable, we restricted ourselves to relations that have no more than 50 columns and to linear FDs. This results in 10 relations, listed in Table I. For each relation R, we enumerate all candidate linear FDs (i.e., pairs $(X, Y) : \exists \boldsymbol{w} \in R, \boldsymbol{w}(X) \neq \mathsf{NULL} \wedge \boldsymbol{w}(Y) \neq \mathsf{NULL}$). We manually validate whether a candidate FD is semantically meaningful, and is hence part of the design schema or not, but only if its g_3'-score is ≥ 0.5. We feel that this is a reasonable way to keep the endeavor manageable, as a g_3'-score < 0.5 means that we should remove more than 50% of the tuples to obtain a subrelation that satisfies the candidate FD; making it improbable that it should be in the design schema. With this approach, we identified 939 candidate FDs. Two individuals manually inspected each candidate. Non-matching decisions (i.e. one saw a candidate as valid whereas the other did not) were discussed until a consensus was reached. We observe that each validated semantically meaningful FD candidate has a g_3'-score ≥ 0.99.

In this manner, we derive for each benchmark relation R its design schema $\Delta(R)$. This set of FDs is partitioned into two sets: $\mathit{PFD}(R) := \{\varphi \in \Delta(R) \mid R \models \varphi\}$ the set of *perfect* (design) FDs and $\mathit{AFD}(R) := \{\varphi \in \Delta(R) \mid R \not\models \varphi\}$ the set of *approximate* (design) FDs. In particular, $\mathit{AFD}\, R$ forms the ground truth of FDs to discover during AFD discovery on R.

Table I shows statistics of the obtained benchmark. In total, we obtain 143 design FDs across all relations in RWD, of which 126 are perfect design FDs and 17 are approximate design FDs. To appreciate the difficulty of the AFD discovery task, it is worth pointing out that the search space during AFD discovery consists of 1634 candidate FDs across all relations in RWD. Out of these, only a small number (17) are AFDs, which emphasizes the intrinsic difficulty of AFD discovery and illustrates the need for good measures to distinguish AFDs from the rest of the search space.

Methodology. We are interested in comparing the suitability of afds measures for the purpose of AFD discovery.

Therefore, we compare AFD measures as follows. Remember from Section IV that every AFD measure f and every threshold $\epsilon \in [0, 1]$ naturally induces a discovery algorithm A_f^ϵ which, on input relation $R(\boldsymbol{W})$, returns all FDs φ over

TABLE I: Overview of relations in RWD benchmark. The #insp column indicates the number of manually inspected candidates when determining the design schema.

Relation R		#rows	#attrs	#insp	#*PFD*(R)	#*AFD*(R)
R_1	adult	32561	15	111	2	0
R_2	claims	97231	13	42	2	2
R_3	dblp10k	10000	34	368	75	2
R_4	hospital	114919	15	74	22	7
R_5	tax	1000000	15	95	3	0
R_6	gath. agent	72737	18	55	5	2
R_7	gath. area	137710	11	43	3	2
R_8	gathering	90991	35	64	0	1
R_9	ident. taxon	562958	3	2	0	1
R_{10}	ident.	91799	38	85	14	0

$\boldsymbol{W}$ with $R \not\models \varphi$ and $f(\varphi, R) \in [\epsilon, 1[$. In this respect, every measure hence defines a class $DISC_f$ of discovery algorithms, namely $DISC_f = \{A_f^\epsilon \mid 0 \leq \epsilon < 1\}$. Given a subset $\mathcal{B}$ of benchmark relations, we compare the effectiveness of measures on $\mathcal{B}$ by computing the area under the precision-recall[4] curve (AUC-PR) of $DISC_f$ for each measure f, where the PR-curve is the set $\{(rcl(A, \mathcal{B}), prec(A, \mathcal{B})) \mid A \in DISC_f\}$. Here, $rcl(A, \mathcal{B})$ and $prec(A, \mathcal{B})$ denote recall and precision of A on $\mathcal{B}$, respectively. It is known that PR curves are well-suited to visualize the tradeoff between precision and recall at various values of ϵ when the prediction classes are very imbalanced, which is the case here. So, the measure with the highest AUC-PR score is the measure providing the best such tradeoff.

Furthermore, to obtain a more fine-grained view of measure performance on the level of each Relation R individually we also report the *rank at max recall*: $r@mr(f, R) := |A_f^\epsilon|$, with $\epsilon = min(f(AFD(R))$. Intuitively, $r@mr(f, R)$ indicates how many candidate AFDs need to be examined when processing them in decreasing order of f-score to find all of $AFD(R)$.

Since SFI is parameterized by a parameter α it is not one measure but a collection of measures. We performed experiments with the same values of α as in the original SFI paper [34], namely $\alpha \in \{0.5, 1, 2\}$. Because the performance of $\alpha = 0.5$ consistently dominates the performance of $\alpha \in \{1, 2\}$, we only report the performance of SFI for $\alpha = 0.5$ in what follows.

We implemented all measures in a Python library. This library, together with the benchmark datasets is publicly available [32]. Given a candidate FD $\boldsymbol{X} \rightarrow \boldsymbol{Y}$, computing the score for most measures is straightforward, requiring only the evaluation of the given formula. For RFI, RFI$'^+$ and SFI, which are the most complex to compute, we use the currently best known algorithms, for RFI and RFI$'^+$ the one of [27], for SFI the one of [34].

Efficient computation and RWD$^-$. We observe massive differences in the measures' computing runtimes. In particular, SFI, RFI$^+$ and RFI$'^+$ require an unreasonable amount of computation time. In 24 hours using a single CPU core, SFI was able to calculate values for 1430 while RFI$^+$ and RFI$'^+$ processed only 250 of all 1634 AFD candidates. All other measures finished processing all candidates in roughly two minutes. For practical applications, the runtime of SFI, RFI$^+$ and RFI$'^+$ hence present major obstacles. An overview of all runtimes is included in Appendix F in Table IV.

In fact, the computational complexity of RFI$^+$ and RFI$'^+$ did not allow us to compute values for RFI$^+$ and RFI$'^+$ on all candidate FDs in RWD in a reasonable amount of time. In approximately 168 hours we obtained values of RFI$^+$ and RFI$'^+$ for 1229 candidate AFDs, including all design AFDs, out of a total of 1634. We denote this set of 1229 candidates FDs by RWD$^-$ in what follows. To ensure fair comparison among all measures, we report all comparison metrics (AUC, *r@mr*, ...) relative to RWD$^-$. Although the results are similar, for the sake of completeness, we include in Appendix F, the performance metrics relative to the whole RWD excluding the measures RFI$^+$ and RFI$'^+$.

B. Results

AUC. Table 2a lists the AUC scores for RWD$^-$ at the benchmark and relation level, where the AUC value is expressed as a percentage. The last column shows the fraction of relations on which a measure reached maximal AUC score, allowing us to judge how consistent a measure is. When looking at the benchmark level, we observe that there are effective measures in each measure class. RFI$'^+$ (SHANNON, AUC = 0.971) is the most effective measure, closely followed by μ^+ (LOGICAL, AUC = 0.946) and somewhat further followed by g_3' (VIOLATION, AUC = 0.901). All other measures have significantly lower AUC values. When the correct number of AFDs is not known beforehand and a specific threshold needs to be set uniformly for all relations, RFI$'^+$, μ^+ and g_3' hence provide the best tradeoff between precision and recall. We find it striking to note that the unnormalized variants of these measures (i.e., FI, *pdep*, and g_3, respectively) perform significantly worse, which highlights the importance of normalisation when designing measures. For RFI$'^+$ and μ^+ in particular, we note that the normalisation w.r.t the expected value of FI resp. *pdep* under random permutations performs significantly better than computing the absolute difference w.r.t this absolute value (RFI$^+$), respectively normalising w.r.t. $pdep(\boldsymbol{Y})$ (for τ).

The AUC scores at the relation level give a more detailed picture. In particular, when looking at the last column in Table 2a we see that RFI$^+$ yields the highest AUC score on each relation, while μ^+ does so for 90% of the relations, and g_3' for "only" 80% of the relations. In particular, μ^+ performs worse than RFI$'^+$ only on relation R_7, where its AUC score is no worse than that of the other measures. g_3' also performs worse than RFI$'^+$ on R_7 and additionally performs worse than both RFI$'^+$ and μ^+ on R_6.

Surprisingly, FI, which has a low AUC score $= 0.415$ at the benchmark level has a highest AUC score on 90% of the relations, like μ^+. It does particularly poor on relation R_3 (dblp, AUC=0.054), which explains its AUC score on the benchmark level. Similarly to g_3', τ has a highest AUC on 80% of the relations, but it also performs very poor on R_3,

[4]https://en.wikipedia.org/wiki/Precision_and_recall

	RWD$^-$	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	best
ρ	41.7	100	03.3	04.1	100	100	100	66.7	100	100	100	70
g_2	50.4	100	51.1	04.1	100	100	100	66.7	100	100	100	70
g_3	67.4	100	100	14.8	100	100	79.2	66.7	100	100	100	70
g_3'	90.1	100	100	100	100	100	79.2	66.7	100	100	100	80
g_1^S	10.9	100	100	02.7	100	100	63.3	66.7	10.0	100	100	60
FI	41.5	100	100	05.4	100	100	100	91.7	100	100	100	90
RFI$^+$	49.4	100	100	18.2	50.8	100	13.3	66.7	25.0	100	100	50
RFI$'^+$	97.1	100	100	100	100	100	100	91.7	100	100	100	100
SFI	32.0	100	100	00.4	31.8	100	05.3	58.3	06.2	100	100	50
g_1	42.5	100	58.7	01.5	100	100	61.3	66.7	100	100	100	60
g_1'	42.5	100	54.8	01.5	100	100	61.3	66.7	100	100	100	60
pdep	64.7	100	100	07.8	100	100	79.2	66.7	100	100	100	70
τ	63.0	100	100	08.4	100	100	100	66.7	100	100	100	80
μ^+	94.6	100	100	100	100	100	100	66.7	100	100	100	90

(a)

	R_2	R_3	R_4	R_6	R_7	R_8	R_9
$AFD(R)$	2	2	7	2	2	1	1
ρ	45	25	7	2	3	1	1
g_2	70	25	7	2	3	1	1
g_3	2	8	7	3	3	1	1
g_3'	2	2	7	3	3	1	1
g_1^S	2	39	7	6	3	5	1
FI	2	21	7	2	3	1	1
RFI$^+$	2	8	32	11	3	2	1
RFI$'^+$	2	2	7	2	3	1	1
SFI	2	283	66	21	4	8	1
g_1	9	70	7	7	3	1	1
g_1'	16	70	7	7	3	1	1
pdep	2	18	7	3	3	1	1
τ	2	18	7	2	3	1	1
μ^+	2	2	7	2	3	1	1

(b)

	LHS-uniq.		RHS-skew	
	R_3	R_6	R_3	R_6
ρ	0.88		0.83	
g_2	0.88		0.83	
g_3	0.88	0.04	1.24	4.42
g_3'		0.04		4.42
g_1^S	0.59	0.12	0.94	4.79
FI	0.87		0.64	
RFI$^+$	0.45	0.03	1.11	4.74
SFI	0.4	0.12	1.04	3.8
g_1	0.85	0.78	1.09	3.8
g_1'	0.85	0.78	1.09	3.8
pdep	0.88	0.04	0.81	4.42
τ	0.88		0.81	
$AFD(R)$	0.07		1.43	
rest	0.24		1.59	

(c)

Fig. 2: Measure performance on RWD$^-$. (a) Heatmap of PR-AUC scores (red indicates low value). (b) Rank at max recall (heatmap per column where red indicates high rank). (c) LHS-uniqueness and RHS-skew of mislabeled FD candidates.

explaning its lower AUC score at the benchmark level. We note that our observation from Section V on synthetic data, namely that g_1, g_1' and SFI do not have good distinguishing power, is confirmed on RWD$^-$: these measures perform the poorest, attaining maximal AUC score in only 60% resp. 50% of the relations. Measures g_1^S and RFI$^+$ perform equally poor.

In summary, RFI$'^+$, μ^+ and g_3' provide the best tradeoff between precision and recall with RFI$'^+$ performing better than μ^+ (marginally, on one relation), and μ^+ performing better than g_3' (again marginally, on one relation).

Rank at max recall is shown in Table 2b. The first row indicates the total number of design AFDs to discover (the smallest attainable *r@mr* value). We note in particular that best measures, g_3', RFI$'^+$ and μ^+ have optimal *r@mr*, save on relation R_7 where they differ only by 1 from the optimum and still have minimal *r@mr* among all measures. In addition, g_3' has non-optimal *r@mr* on R_6, where it is off by 1. At maximum recall, these measures hence have very high precision. As such, with these measures it suffices to only inspect a small number of highly ranked to recover the true design FDs that were obscured by errors. By contrast, all the other measures have relations where the *r@mr* is an order of magnitude larger than the optimum, or more, yielding low precision at maximum recall.

LHS-uniqueness and RHS-skew. From Figures 2a and 2b we observe that there are four kinds of relations in RWD: "trivial" relations for which every measure attains optimal AUC and *r@mr* (relations R_1, R_5,R_9,R_{10}), "easy" relations for which nearly all measures do so (R_2 and R_8), "challenging" relations where only a minority of measures reach optimal scores (R_3 and R_6), and "out-of-reach" relations where no measure attains the optimum (R_7).

Next, we investigate what properties of the input data makes a relation challenging by analyzing 'mislabeled' candidate FDs in R_3 and R_6. We refer to a candidate as *mislabeled* analogous to our definition of *r@mr*: from the candidates counted for $r@mr(f, R)$ we exclude all $AFD(R)$ to obtain our mislabeled candidate FDs. In other words, the mislabeled candidate FDs are the highest ranked mistakes made by a measure. Figure 2c shows the average LHS-uniqueness and RHS-skew values of all mislabeled candidate FDs per measure. For comparison, the bottom two rows show the average LHS-uniqueness and RHS-skew over the set of all design AFDs and the set of all candidate FDs not in the design set.

We start with analyzing R_3. From Figures 2a and 2b we recall that measures g_3', RFI$'^+$, and μ^+ attain optimal AUC and *r@mr* here, while the AUC scores of all other measures are extremely low and their *r@mr* is very high. In Figure 2c, we observe that these other measures have LHS-uniqueness values for mislabeled candidate FDs that are much higher than the average for design AFDs (0.07) or non-FDs (0.2). We postulate that this is what makes R_3 challenging for these measures. The group of measures ρ, g_2, g_3, FI, g_1, g_1', *pdep* and τ all have mislabeled LHS-uniqueness > 0.8 and we recall from Section V that the distinguishing power of these measures is small at high LHS-uniqueness values. In addition, from Figure 1 (middle row) we observe that RFI$^+$ (0.45) and SFI (0.4) have small separation (and hence limited distinguishing power) already at modest values of LHS-uniqueness. For g_1^S, the situation is less clear. We note, however that its LHS-uniqueness value (0.59) is much larger than the average for design AFDs in R_3(0.074).

On R_6, ρ, g_2, FI, RFI$'^+$, τ and μ^+ attain optimal AUC and *r@mr*. In Table 2c we observe high RHS-skew values (>3.7) for all other measures, compared to the values of design FDs and non-fds (both 0.4). We postulate this is what makes R_6 challenging for g_3, g_3', g_1^S, and *pdep*: recall from Section V that these measures are sensitive to RHS-skew. In contrast, we know from Section V that SFI, g_1 and g_1' consistently have (almost) zero separation, independent of RHS-skew. Similarly, RFI$^+$ is insensitive to RHS-skew, but its separation is limited, as shown in Figure 1 (middle row).

We conclude that high LHS-uniqueness and RHS-skew values are observable in practice and sensitivity to these structural properties may explain the lower performance of some measures. Insensitivity to LHS-uniqueness and RHS-skew are therefore desirable properties to aim for when designing measures.

TABLE II: Properties of considered AFD measures. The symbol ✓ stands for *applies*, an empty cell denotes *does not apply*. The symbol _ stands for *not applicable* (cf. Section V). Further, f' refers to the normalization of the measure f as discussed in Section IV and f^+ to the adaptation of f that maps all negative values to zero.

	ρ	g_2	g_3	g_3'	g_1^S	FI	RFI$^+$	RFI$'^+$	SFI	g_1	g_1'	*pdep*	τ	μ^+
Considered in	[21]	[23], [44]	[2], [3], [20] [22], [43], [23]	[18]	new	[9], [18] [27], [28], [34]	[27], [28]	new	[34]	[23], [47]	[25]	[36]	[19], [36]	[36]
Class (V/S/L)	V	V	V	V	S	S	S	S	S	L	L	L	L	L
Has baselines		✓		✓	✓	✓	✓	✓	✓		✓		✓	✓
Efficiently computable	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓
Inversely proportional to error level	✓	✓	✓	✓		✓	✓	✓				✓	✓	✓
Insensitive to LHS-unique				✓				✓	–	–	–			✓
Insensitive to RHS-skew						✓	✓	✓	–	–	–		✓	✓
AUC on RWD	0.417	0.504	0.674	0.901	0.109	0.415	0.494	0.971	0.320	0.425	0.425	0.647	0.630	0.946

VII. Conclusions

An overview of our results is given in Table II. We find that well-ranking measures exist within each class: g_3' in VIOLATION, RFI$'^+$ in SHANNON, and μ^+ in LOGICAL. We further observe that measures are only effective when correctly normalized—which is not always done in the literature. Indeed, g_3 is widely known and cited [23], [2], [20], [18], [22], [17], [3] but to the best of our knowledge only [18] considers the correctly normalized version g_3'. The sensitivity of g_3' to RHS-skew remains a structural weakness, hampering its effectiveness as illustrated by its lower performance in practice on R_6(Fig 2).

FI is the defining measure of SHANNON and suffers from sensitivity to LHS-uniqueness as illustrated by its behavior on R_3. Its corrections SFI [34] and RFI [27], [28] were aimed at removing bias from FI, but our experiments on SYN reveal that their distinguishing power is greatly diminished; especially they overcompensate their correction of FI w.r.t. LHS-uniqueness. This is reflected by their behavior on RWD, where they are among the worst performing measures. Our novel correction RFI$'^+$ of RFI is the best performing measure on RWD and is insensitive to both LHS-uniqueness and RHS-skew. Its main drawback is the slow computation by current algorithms rendering it essentially useless in practice.

Our recommendation for most suitable AFD measure is therefore the little-known measure μ. It has comparable performance to RFI$'^+$ as well as equal structural sensitivity properties, but can be efficiently computed.

When the number of attributes in the LHS increases, LHS-uniqueness is expected to increase naturally to 1. Since g_3', RFI$'^+$, and $\mu^+$are the only measures among the considered ones that are insensitive to LHS-uniqueness, they are the most promising measures for discovering non-linear AFDs.

Another finding worth noting is that we illustrated on RWD, perhaps contrary to popular belief, that by only inspecting a small number of top-ranked candidate FDs (according to g_3', RFI$'^+$, μ), one already succeeds in finding a large number of true design FDs that were obscured by errors. This means in particular, that a domain expert does not need to wade through hundreds of high-ranked candidate FDs but can restrict attention to a handful.

Inspired by Arocena et al. [1], we study in Appendix G the measures' sensitivity to different kinds and different levels of errors by passing the relations in RWD through a controlled error channel. We again find that RFI$'^+$ and μ^+ perform best, closely followed by g_3'. We also find that the effectiveness of all measures quickly deteriorates for increasing error levels making them essentially useless for error levels above 5%.

References

[1] P. C. Arocena, B. Glavic, G. Mecca, R. J. Miller, P. Papotti, and D. Santoro, "Messing up with BART: error generation for evaluating data-cleaning algorithms," *Proc. VLDB Endow.*, vol. 9, no. 2, pp. 36–47, 2015. [Online]. Available: http://www.vldb.org/pvldb/vol9/p36-arocena.pdf

[2] L. Berti-Équille, H. Harmouch, F. Naumann, N. Novelli, and S. Thirumuruganathan, "Discovery of genuine functional dependencies from relational data with missing values," *Proc. VLDB Endow.*, vol. 11, no. 8, pp. 880–892, 2018. [Online]. Available: http://www.vldb.org/pvldb/vol11/p880-berti-equille.pdf

[3] F. Berzal, J. Cubero, F. Cuenca, and J. Medina, "Relational decomposition through partial functional dependencies," *Data and Knowledge Engineering*, vol. 43, no. 2, pp. 207–234, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169023X02000563

[4] J. Birnick, T. Bläsius, T. Friedrich, F. Naumann, T. Papenbrock, and M. Schirneck, "Hitting set enumeration with partial information for unique column combination discovery," *Proc. VLDB Endow.*, vol. 13, no. 11, pp. 2270–2283, 2020.

[5] T. Bleifuß, S. Bülow, J. Frohnhofen, J. Risch, G. Wiese, S. Kruse, T. Papenbrock, and F. Naumann, "Approximate discovery of functional dependencies for large datasets," in *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, S. Mukhopadhyay, C. Zhai, E. Bertino, F. Crestani, J. Mostafa, J. Tang, L. Si, X. Zhou, Y. Chang, Y. Li, and P. Sondhi, Eds. ACM, 2016, pp. 1803–1812. [Online]. Available: https://doi.org/10.1145/2983323.2983781

[6] P. Bohannon, W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis, "Conditional functional dependencies for data cleaning," in *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, R. Chirkova, A. Dogac, M. T. Özsu, and T. K. Sellis, Eds. IEEE Computer Society, 2007, pp. 746–755. [Online]. Available: https://doi.org/10.1109/ICDE.2007.367920

[7] T. Calders, R. T. Ng, and J. Wijsen, "Searching for dependencies at multiple abstraction levels," *ACM Trans. Database Syst.*, vol. 27, no. 3, pp. 229–260, 2002. [Online]. Available: https://doi.org/10.1145/581751.581752

[8] L. Caruccio, V. Deufemia, and G. Polese, "Relaxed functional dependencies - A survey of approaches," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 147–165, 2016. [Online]. Available: https://doi.org/10.1109/TKDE.2015.2472010

[9] R. Cavallo and M. Pittarelli, "The theory of probabilistic databases," in *VLDB'87, Proceedings of 13th International Conference on Very Large Data Bases, September 1-4, 1987, Brighton, England*, P. M. Stocker, W. Kent, and P. Hammersley, Eds. Morgan Kaufmann, 1987, pp. 71–81.

[10] F. Chiang and R. J. Miller, "Discovering data quality rules," *Proc. VLDB Endow.*, vol. 1, no. 1, pp. 1166–1177, 2008. [Online]. Available: http://www.vldb.org/pvldb/vol1/1453980.pdf

[11] X. Chu, I. F. Ilyas, and P. Papotti, "Holistic data cleaning: Putting violations into context," in *29th IEEE International Conference on Data Engineering, ICDE 2013, Brisbane, Australia, April 8-12, 2013*, C. S. Jensen, C. M. Jermaine, and X. Zhou, Eds. IEEE Computer Society, 2013, pp. 458–469. [Online]. Available: https://doi.org/10.1109/ICDE.2013.6544847

[12] G. Cormode, L. Golab, F. Korn, A. McGregor, D. Srivastava, and X. Zhang, "Estimating the confidence of conditional functional dependencies," in *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2009, Providence, Rhode Island, USA, June 29 - July 2, 2009*, U. Çetintemel, S. B. Zdonik, D. Kossmann, and N. Tatbul, Eds. ACM, 2009, pp. 469–482. [Online]. Available: https://doi.org/10.1145/1559845.1559895

[13] S. Edwards, "Thomas m. cover and joy a. thomas, elements of information theory (2nd ed.), john wiley & sons, inc. (2006)," *Inf. Process. Manag.*, vol. 44, no. 1, pp. 400–401, 2008. [Online]. Available: https://doi.org/10.1016/j.ipm.2007.02.009

[14] D. Ellerman, *New Foundations for Information Theory - Logical Entropy and Shannon Entropy*, ser. SpringerBriefs in Philosophy. Springer, 2021.

[15] W. Fan, F. Geerts, and X. Jia, "A revival of integrity constraints for data cleaning," *Proc. VLDB Endow.*, vol. 1, no. 2, pp. 1522–1523, 2008. [Online]. Available: http://www.vldb.org/pvldb/vol1/1454220.pdf

[16] W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis, "Conditional functional dependencies for capturing data inconsistencies," *ACM Trans. Database Syst.*, vol. 33, no. 2, pp. 6:1–6:48, 2008. [Online]. Available: https://doi.org/10.1145/1366102.1366103

[17] P. Faure-Giovagnoli, J. Petit, and V. Scuturici, "Assessing the existence of a function in a dataset with the g3 indicator," in *38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia, May 9-12, 2022*. IEEE, 2022, pp. 607–620. [Online]. Available: https://doi.org/10.1109/ICDE53745.2022.00050

[18] C. Giannella and E. L. Robertson, "On approximation measures for functional dependencies," *Inf. Syst.*, vol. 29, no. 6, pp. 483–507, 2004. [Online]. Available: https://doi.org/10.1016/j.is.2003.10.006

[19] L. A. Goodman and W. H. Kruskal, "Measures of association for cross classifications," *Journal of the American Statistical Association*, vol. 49, no. 268, pp. 732–764, 1954. [Online]. Available: http://www.jstor.org/stable/2281536

[20] Y. Huhtala, J. Kärkkäinen, P. Porkka, and H. Toivonen, "TANE: an efficient algorithm for discovering functional and approximate dependencies," *Comput. J.*, vol. 42, no. 2, pp. 100–111, 1999. [Online]. Available: https://doi.org/10.1093/comjnl/42.2.100

[21] I. F. Ilyas, V. Markl, P. J. Haas, P. Brown, and A. Aboulnaga, "CORDS: automatic discovery of correlations and soft functional dependencies," in *Proceedings of the ACM SIGMOD International Conference on Management of Data, Paris, France, June 13-18, 2004*, G. Weikum, A. C. König, and S. Deßloch, Eds. ACM, 2004, pp. 647–658. [Online]. Available: https://doi.org/10.1145/1007568.1007641

[22] R. S. King and J. J. Legendre, "Discovery of functional and approximate functional dependencies in relational databases," *Adv. Decis. Sci.*, vol. 7, no. 1, pp. 49–59, 2003. [Online]. Available: https://doi.org/10.1155/S117391260300004X

[23] J. Kivinen and H. Mannila, "Approximate inference of functional dependencies from relations," *Theor. Comput. Sci.*, vol. 149, no. 1, pp. 129–149, 1995. [Online]. Available: https://doi.org/10.1016/0304-3975(95)00028-U

[24] J. Kossmann, T. Papenbrock, and F. Naumann, "Data dependencies for query optimization: a survey," *VLDB J.*, vol. 31, no. 1, pp. 1–22, 2022. [Online]. Available: https://doi.org/10.1007/s00778-021-00676-3

[25] S. Kruse and F. Naumann, "Efficient discovery of approximate dependencies," *Proc. VLDB Endow.*, vol. 11, no. 7, pp. 759–772, 2018. [Online]. Available: http://www.vldb.org/pvldb/vol11/p759-kruse.pdf

[26] H. Liu, D. Xiao, P. Didwania, and M. Y. Eltabakh, "Exploiting soft and hard correlations in big data query optimization," *Proc. VLDB Endow.*, vol. 9, no. 12, pp. 1005–1016, 2016. [Online]. Available: http://www.vldb.org/pvldb/vol9/p1005-liu.pdf

[27] P. Mandros, M. Boley, and J. Vreeken, "Discovering reliable approximate functional dependencies," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*. ACM, 2017, pp. 355–363. [Online]. Available: https://doi.org/10.1145/3097983.3098062

[28] ——, "Discovering dependencies with reliable mutual information," *Knowl. Inf. Syst.*, vol. 62, no. 11, pp. 4223–4253, 2020. [Online]. Available: https://doi.org/10.1007/s10115-020-01494-9

[29] F. D. Marchi, S. Lopes, and J. Petit, "Unary and n-ary inclusion dependency discovery in relational databases," *J. Intell. Inf. Syst.*, vol. 32, no. 1, pp. 53–73, 2009. [Online]. Available: https://doi.org/10.1007/s10844-007-0048-x

[30] T. Papenbrock, J. Ehrlich, J. Marten, T. Neubert, J. Rudolph, M. Schönberg, J. Zwiener, and F. Naumann, "Functional dependency discovery: An experimental evaluation of seven algorithms," *Proc. VLDB Endow.*, vol. 8, no. 10, pp. 1082–1093, 2015. [Online]. Available: http://www.vldb.org/pvldb/vol8/p1082-papenbrock.pdf

[31] T. Papenbrock and F. Naumann, "A hybrid approach to functional dependency discovery," in *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, F. Özcan, G. Koutrika, and S. Madden, Eds. ACM, 2016, pp. 821–833. [Online]. Available: https://doi.org/10.1145/2882903.2915203

[32] M. Parciak, S. Weytjens, N. Hens, F. Neven, L. Peeters, and S. Vansummeren, "Artifacts related to "approximately measuring functional dependencies: a comparitive study"," Available at https://github.com/MarcelPa/AFD_comparative_study, 2022.

[33] E. H. M. Pena, E. C. de Almeida, and F. Naumann, "Discovery of approximate (and exact) denial constraints," *Proc. VLDB Endow.*, vol. 13, no. 3, pp. 266–278, 2019. [Online]. Available: http://www.vldb.org/pvldb/vol13/p266-pena.pdf

[34] F. Pennerath, P. Mandros, and J. Vreeken, "Discovering approximate functional dependencies using smoothed mutual information," in *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, R. Gupta, Y. Liu, J. Tang, and B. A. Prakash, Eds. ACM, 2020, pp. 1254–1264. [Online]. Available: https://doi.org/10.1145/3394486.3403178

[35] B. Pfahringer and S. Kramer, "Compression-based evaluation of partial determinations," in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, August 20-21, 1995*, U. M. Fayyad and R. Uthurusamy, Eds. AAAI Press, 1995, pp. 234–239. [Online]. Available: http://www.aaai.org/Library/KDD/1995/kdd95-027.php

[36] G. Piatetsky-Shapiro and C. J. Matheus, "Measuring data dependencies in large databases," in *Proceedings of the 2nd International Conference on Knowledge Discovery in Databases*. AAAI Press, 1993, pp. 162––173.

[37] J. Rammelaere and F. Geerts, "Explaining repaired data with cfds," *Proc. VLDB Endow.*, vol. 11, no. 11, pp. 1387–1399, 2018. [Online]. Available: http://www.vldb.org/pvldb/vol11/p1387-rammelaere.pdf

[38] ——, "Revisiting conditional functional dependency discovery: Splitting the "c" from the "fd"," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part II*, ser. Lecture Notes in Computer Science, vol. 11052. Springer, 2018, pp. 552–568. [Online]. Available: https://doi.org/10.1007/978-3-030-10928-8_33

[39] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré, "Holoclean: Holistic data repairs with probabilistic inference," *Proc. VLDB Endow.*, vol. 10, no. 11, pp. 1190–1201, 2017. [Online]. Available: http://www.vldb.org/pvldb/vol10/p1190-rekatsinas.pdf

[40] M. S. Roulston, "Estimating the errors on measured entropy and mutual information," *Physica D: Nonlinear Phenomena*, vol. 125, no. 3, pp. 285–294, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167278998002693

[41] P. Schirmer, T. Papenbrock, S. Kruse, F. Naumann, D. Hempfing, T. Mayer, and D. Neuschäfer-Rube, "Dynfd: Functional dependency discovery in dynamic datasets," in *Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019*, M. Herschel, H. Galhardas, B. Reinwald, I. Fundulaki, C. Binnig, and Z. Kaoudi, Eds. OpenProceedings.org, 2019, pp. 253–264. [Online]. Available: https://doi.org/10.5441/002/edbt.2019.23

[42] D. A. Simovici, D. Cristofor, and L. Cristofor, "Impurity measures in databases," *Acta Informatica*, vol. 38, no. 5, pp. 307–324, 2002. [Online]. Available: https://doi.org/10.1007/s002360100078

[43] D. Z. Wang, X. L. Dong, A. D. Sarma, M. J. Franklin, and A. Y. Halevy, "Functional dependency generation and applications

in pay-as-you-go data integration systems," in *12th International Workshop on the Web and Databases, WebDB 2009, Providence, Rhode Island, USA, June 28, 2009*, 2009. [Online]. Available: http://webdb09.cse.buffalo.edu/papers/Paper18/webdb09.pdf

[44] P. Wang and Y. He, "Uni-detect: A unified approach to automated error detection in tables," in *SIGMOD*. ACM, 2019, pp. 811–828. [Online]. Available: https://doi.org/10.1145/3299869.3319855

[45] Y. Wang, S. Song, L. Chen, J. X. Yu, and H. Cheng, "Discovering conditional matching rules," *ACM Trans. Knowl. Discov. Data*, vol. 11, no. 4, pp. 46:1–46:38, 2017. [Online]. Available: https://doi.org/10.1145/3070647

[46] Z. Wei and S. Link, "Discovery and ranking of functional dependencies," in *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*. IEEE, 2019, pp. 1526–1537. [Online]. Available: https://doi.org/10.1109/ICDE.2019.00137

[47] Y. Zhang, Z. Guo, and T. Rekatsinas, "A statistical perspective on discovering functional dependencies in noisy data," in *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*, D. Maier, R. Pottinger, A. Doan, W. Tan, A. Alawini, and H. Q. Ngo, Eds. ACM, 2020, pp. 861–876. [Online]. Available: https://doi.org/10.1145/3318464.3389749

APPENDIX A
DISCUSSION ON LOGICAL VS SHANNON ENTROPY

The notion of logical entropy arises in mathematical philosophy [14], where it is observed to provide a theory of information based on logic. Importantly, formulas and equalities concerning logical entropy can be converted into corresponding formulas and equalities concerning Shannon entropy by the so-called dit-bit transform (see [14]). Logical and Shannon entropy are hence highly similar, but measure different things: logical entropy measures the probability of two random tuples to be distinguished, while Shannon entropy measures average uncertainty.

APPENDIX B
PROOF OF WELL-DEFINEDNESS OF μ

In this section we prove the following result mentioned in Footnote 7.

Lemma 1. *If* $\mathbb{E}_R[pdep(\varphi, R)] = 1$ *then* $R \models \varphi$.

Proof. Assume that $\mathbb{E}_R[pdep(\varphi, R)] = 1$. Let $R_1, \ldots, R_n$ be an enumeration of all permuations of R. Then

$$\mathbb{E}_R[pdep(\varphi, R)] = \frac{\sum_{i=1}^{N} pdep(\varphi, R_i)}{N}$$

Hence, $\mathbb{E}_R[pdep(\varphi, R)] = 1$ iff $\sum_{i=1}^{N} pdep(\varphi, R_i) = N$. Because the range of *pdep* is the interval $[0, 1]$ this sum can equal N if, and only if, $pdep(\varphi, R_i) = 1$ for every R_i, including R itself. Suppose, for the purpose of contradiction, that $R \not\models \varphi$. Then, the value of *pdep* is given by the formula in Section IV-D, i.e.

$$\begin{aligned} pdep(\boldsymbol{X} \to \boldsymbol{Y}, R) &= \sum_{\boldsymbol{x}} p_R(\boldsymbol{x})[1 - pdep(\boldsymbol{Y} \mid \boldsymbol{x}, R) \\ &= 1 - \mathbb{E}_{\boldsymbol{x}}[h_R(\boldsymbol{Y} \mid \boldsymbol{x})] \end{aligned}$$

where the second equality is due to Lemma 3. Since $pdep(\boldsymbol{X} \to \boldsymbol{Y}, R) = 1$, this means in particular that $\mathbb{E}_{\boldsymbol{x}}[h_R(\boldsymbol{Y} \mid \boldsymbol{x})] = 0$, which by reasoning similar as above can only happen if $h_R(\boldsymbol{Y} \mid \boldsymbol{x}) = 0$ for every $\boldsymbol{x} \in \pi_{\boldsymbol{X}}(R)$. This means, that for every $\boldsymbol{x} \in \pi_{\boldsymbol{X}}(R)$, the probability to draw two distinct $\boldsymbol{Y}$-tuples in $\pi_{\boldsymbol{Y}}(\sigma_{\boldsymbol{X}=\boldsymbol{x}}(R))$ is zero. But that can only happen if there is only one $\boldsymbol{Y}$-value $\pi_{\boldsymbol{Y}}(\sigma_{\boldsymbol{X}=\boldsymbol{x}}(R))$, in which case $R \models \varphi$ and we obtain our desired contradiction. □

APPENDIX C
FORMAL COMPARISON OF THE AFD MEASURES

This section presents a formal comparison of the measures introduced in Section IV, which motivates the definition of the new measures g_1^S and RFI'^{+}.

Among the measures introduced in Section IV, we discern the following three classes (see also the second row in Table II):

(1) The class of measures that have a notion of "violation" and quantify the number of violations, consisting of ρ, g_2, g_3, and g_3'. We denote this class by VIOLATION.
(2) The class of measures based on Shannon entropy, consisting of $\text{FI}, \text{RFI}^{+}$, and SFI. We denote this class by SHANNON.
(3) The class of measures based on logical entropy, consisting of $g_1, g_1', pdep, \tau$, and μ^{+} and denoted by LOGICAL.

We discuss the similarities in the design of LOGICAL measures and those in the VIOLATION and SHANNON class by means of Table III, which clusters measures into groups that we find similar and where we rewrite measures into equivalent form when this is necessary to stress the similarities.

Theorem 2. *The alternate formulas given in Table III are equivalent to their definition given in Sections IV-A–IV-D.*

The proof may be found in Appendix D.

① We have already observed that g_1 is a measure based on logical entropy, $g_1(\boldsymbol{X} \to \boldsymbol{Y}, R) = 1 - h_R(\boldsymbol{Y} \mid \boldsymbol{X})$. We find it interesting to observe that Giannella and Robertson [18] considered an axiomatisation of FD error measures, and showed that Shannon entropy $H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ is, up to a multiplicative constant, the unique unnormalized error measure that satisfies their axioms. As such, we may view $1 - H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ as the Shannon equivalent of g_1. Unfortunately, however, $1 - H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ is not an AFD measure: the value of $H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ is unbounded and $1 - H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ hence has range $[-\infty, 1]$ instead of $[0, 1]$. Giannella and Robertson [18] therefore turn $1 - H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ into an AFD measure by moving to FI, which normalizes $H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ w.r.t. $H_R(\boldsymbol{Y})$. This is no longer the conceptual Shannon counterpart of g_1. However, as further discussed below, it is nevertheless natural to ask what the conceptual Shannon counterpart of g_1 is and how it behaves. We thus propose the following Shannon variant g_1^S of g_1, obtained by limiting $1 - H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ to be positive:

$$g_1^S(\boldsymbol{X} \to \boldsymbol{Y}, R) := \max(1 - H_R(\boldsymbol{Y} \mid \boldsymbol{X}), 0).$$

	LOGICAL measure	VIOLATION/SHANNON
①	$g_1 = 1 - h_R(\boldsymbol{Y} \mid \boldsymbol{X})$	$1 - H_R(\boldsymbol{Y} \mid \boldsymbol{X})$
②	$\mathit{pdep} = \sum_{\boldsymbol{x}} p_R(\boldsymbol{x})\,(1 - h_R(\boldsymbol{Y} \mid \boldsymbol{x}))$ $= 1 - \sum_{\boldsymbol{x}} p_R(\boldsymbol{x}) h_R(\boldsymbol{Y} \mid \boldsymbol{x})$	$g_3 = \sum_{\boldsymbol{x}} p_R(\boldsymbol{x}) \max_{\boldsymbol{y}} p_R(\boldsymbol{y} \mid \boldsymbol{x})$ $g_2 = 1 - \sum_{\boldsymbol{w} \in G_2(\boldsymbol{X} \to \boldsymbol{Y}, R)} p_R(\boldsymbol{w})$
③	$\tau = 1 - \frac{\mathbb{E}_{\boldsymbol{x}}[h_R(\boldsymbol{Y} \mid \boldsymbol{x})]}{h_R(\boldsymbol{Y})}$	$\mathrm{FI} = 1 - \frac{H_R(\boldsymbol{Y} \mid \boldsymbol{X})}{H_R(\boldsymbol{Y})}$ $= 1 - \frac{\mathbb{E}_{\boldsymbol{x}}[H_R(\boldsymbol{Y} \mid \boldsymbol{x})]}{H_R(\boldsymbol{Y})}$
④	$\mu = \frac{\mathit{pdep}(\varphi, R) - \mathbb{E}_R[\mathit{pdep}(\varphi, R)]}{1 - \mathbb{E}_R[\mathit{pdep}(\varphi, R)]}$	$\mathrm{RFI} = \mathrm{FI}(\varphi, R) - \mathbb{E}_R[\mathrm{FI}(\varphi, R)]$

TABLE III: Overview of similarities between LOGICAL measures and measures in VIOLATION/ SHANNON.

② We have already observed in Section IV-D that we may view *pdep* as a generalisation of g_2. We may also view it as an alternate to g_3. Indeed, *pdep* equals the expected value of $1 - h_R(\boldsymbol{Y} \mid \boldsymbol{x})$—expressing the probability of $\boldsymbol{x}$ not participating in a violating pair—while g_3 equals the expected value of $\max_{\boldsymbol{y}} p_R(\boldsymbol{y} \mid \boldsymbol{x})$—expressing the largest subgroup of non-violating tuples in $\pi_{\boldsymbol{Y}} \sigma_{\boldsymbol{X}=\boldsymbol{x}}(R)$, where in both cases expectation is taken over all $\boldsymbol{x}$.

③ As can be seen by the rewritten formulas in line 3 of Table III, FI is simply the Shannon entropy-based version of τ.

④ The similarity between τ and FI extends to a conceptual similarity between μ and RFI: μ corrects for the bias of τ under random permutations while RFI corrects for the bias of FI under random permutations. Despite this conceptual similarity, note that the corrections are done differently: μ corrects by taking the *normalized* difference between *pdep* and $\mathbb{E}_R[\mathit{pdep}]$ while RFI corrects by taking the *absolute* difference between FI and $\mathbb{E}_R[\mathrm{FI}]$. As such RFI is not a normalized measure. Since it is natural to ask what the normalized variant of RFI is and how it behaves, we define

$$\mathrm{RFI}'^{+}(\varphi, R) := \max\left(\frac{\mathrm{FI}(\varphi, R) - \mathbb{E}_R[\mathrm{FI}(\varphi, R)]}{1 - \mathbb{E}_R[\mathrm{FI}(\varphi, R)]}, 0\right).$$

We attribute both new measures g_1^S and RFI'^{+} to SHANNON, and compare their behavior to that of the other measures in the following sections.

APPENDIX D
PROOFS OF EQUIVALENCE

In this section we prove that the alternate measure formulations shown in Table III are correct, hence proving Theorem 2. The theorem is proved as a sequence of lemmas. Throughout this section, assume that $R \not\models \boldsymbol{X} \to \boldsymbol{Y}$.

Lemma 2. $g_3(\boldsymbol{X} \to \boldsymbol{Y}, R) = \sum_{\boldsymbol{x}} p_R(\boldsymbol{x}) \max_{\boldsymbol{y}} p_R(\boldsymbol{y} \mid \boldsymbol{x})$.

Proof. We reason as follows.

$$\begin{aligned} g_3(\boldsymbol{X} \to \boldsymbol{Y}, R) &= \max_{R' \in G_3(\boldsymbol{X} \to \boldsymbol{Y}, R)} \frac{|R'|}{|R|} \\ &= \max_{R' \in G_3(\boldsymbol{X} \to \boldsymbol{Y}, R)} \sum_{\boldsymbol{w} \in R'} p_R(\boldsymbol{w}) \\ &= \sum_{\boldsymbol{x}} \max_{\boldsymbol{y}} p_R(\boldsymbol{x}\boldsymbol{y}) \\ &= \sum_{\boldsymbol{x}} p_R(\boldsymbol{x}) \max_{\boldsymbol{y}} p_R(\boldsymbol{y} \mid \boldsymbol{x}). \end{aligned}$$

Here, the first equality is the definition of g_3. The second equality follows by definition of p_R. The third equality follows from the following observation: a relation $R' \subseteq R$ can only be maximal if $R'(\boldsymbol{w}) = R(\boldsymbol{w})$ whenever $R'(\boldsymbol{w}) > 0$ for all $\boldsymbol{w} \in R$. That is, either we keep all occurrences of $\boldsymbol{w}$ or we remove all of them. So, maximizing $\sum_{\boldsymbol{w} \in R'} p_R(\boldsymbol{w})$ corresponds to, for every $\boldsymbol{x}$, keeping that $\boldsymbol{y}$ that maximizes $p_R(\boldsymbol{x}\boldsymbol{y})$. Thereby, effectively removing all other tuples $\boldsymbol{x}\boldsymbol{y}'$ with $\boldsymbol{y} \neq \boldsymbol{y}'$. The last equality then follows from the definition of conditional probability. □

Lemma 3. $\mathit{pdep}(\boldsymbol{Y} \mid \boldsymbol{x}, R) = 1 - h_R(\boldsymbol{Y} \mid \boldsymbol{x})$ *and therefore*

$$\begin{aligned} \mathit{pdep}(\boldsymbol{X} \to \boldsymbol{Y}, R) &= \sum_{\boldsymbol{x}} p_R(\boldsymbol{x})(1 - h_R(\boldsymbol{Y} \mid x)) \\ &= 1 - \sum_{\boldsymbol{x}} p_R(\boldsymbol{x}) h_R(\boldsymbol{Y} \mid \boldsymbol{x}) \\ &= 1 - \mathbb{E}_{\boldsymbol{x}}[h_R(\boldsymbol{Y} \mid \boldsymbol{x})] \end{aligned}$$

Proof. We first observe

$$\begin{aligned} \mathit{pdep}(\boldsymbol{Y} \mid \boldsymbol{x}, R) &= \sum_{\boldsymbol{y}} p_R(\boldsymbol{y} \mid \boldsymbol{x})^2 \\ &= 1 - (1 - \sum_{\boldsymbol{y}} p_R(\boldsymbol{y} \mid \boldsymbol{x})^2) \\ &= 1 - h_R(Y \mid \boldsymbol{x}). \end{aligned}$$

Hence,

$$\begin{aligned} \mathit{pdep}(\boldsymbol{X} \to \boldsymbol{Y}, R) &= \sum_{\boldsymbol{x}} p_R(\boldsymbol{x})\, \mathit{pdep}(\boldsymbol{Y} \mid \boldsymbol{x}, R) \\ &= \sum_{\boldsymbol{x}} p_R(\boldsymbol{x})(1 - h_R(\boldsymbol{Y} \mid x)) \\ &= \sum_{\boldsymbol{x}} p_R(\boldsymbol{x}) - \sum_{\boldsymbol{x}} p_R(\boldsymbol{x}) h_R(\boldsymbol{Y} \mid \boldsymbol{x}) \\ &= 1 - \sum_{\boldsymbol{x}} p_R(\boldsymbol{x}) h_R(\boldsymbol{Y} \mid \boldsymbol{x}). \end{aligned}$$ □

Lemma 4. $\tau(X \to Y, R) = 1 - \frac{\mathbb{E}_{\boldsymbol{x}}[h_R(\boldsymbol{Y}|\boldsymbol{x})]}{h_R(\boldsymbol{Y})}$

Proof. We reason as follows.

$$\begin{aligned} \tau(X \to Y, R) &= \frac{\mathit{pdep}(\boldsymbol{X} \to \boldsymbol{Y}, R) - \mathit{pdep}(\boldsymbol{Y}, R)}{1 - \mathit{pdep}(\boldsymbol{Y}, R)} \\ &= \frac{(1 - \mathbb{E}_{\boldsymbol{x}}[h_R(\boldsymbol{Y} \mid \boldsymbol{x})]) - (1 - h_R(\boldsymbol{Y}))}{1 - (1 - h_R(\boldsymbol{Y}))} \\ &= \frac{h_R(\boldsymbol{Y}) - \mathbb{E}_{\boldsymbol{x}}[h_R(\boldsymbol{Y} \mid \boldsymbol{x})]}{h_R(\boldsymbol{Y})} \\ &= 1 - \frac{\mathbb{E}_{\boldsymbol{x}}[h_R(\boldsymbol{Y} \mid \boldsymbol{x})]}{h_R(\boldsymbol{Y})} \end{aligned}$$ □

Here, the second equality is by Lemma 3 and the fact that $\mathit{pdep}(\boldsymbol{Y}, R) = 1 - h_R(\boldsymbol{Y})$ by definition.

Relating this to logical entropy we observe

Lemma 5. $\mu(\boldsymbol{X} \to \boldsymbol{Y}, R) = 1 - \frac{\mathbb{E}_{\boldsymbol{x}}[h_R(\boldsymbol{Y}|\boldsymbol{x})]}{h_R(\boldsymbol{Y})} \frac{|R|-1}{|R|-|\,\mathit{dom}(\boldsymbol{X},R)|}$

Proof. We reason as follows.

$$\begin{aligned} &\mu(\boldsymbol{X} \to \boldsymbol{Y}, R) \\ &:= \frac{\mathit{pdep}(\boldsymbol{X} \to \boldsymbol{Y}, R) - \mathbb{E}_R[\mathit{pdep}(\boldsymbol{X} \to \boldsymbol{Y}, R)]}{1 - \mathbb{E}_R[\mathit{pdep}(\boldsymbol{X} \to \boldsymbol{Y}, R)]} \\ &= 1 - \frac{1 - \mathit{pdep}(\boldsymbol{X} \to \boldsymbol{Y}, R)}{1 - \mathit{pdep}(\boldsymbol{Y}, R)} \frac{|R| - 1}{|R| - |\,\mathit{dom}(\boldsymbol{X}, R)|} \\ &= 1 - \frac{1 - (1 - \mathbb{E}_{\boldsymbol{x}}[h_R(\boldsymbol{Y} \mid \boldsymbol{x})])}{1 - (1 - h_R(\boldsymbol{Y}))} \frac{|R| - 1}{|R| - |\,\mathit{dom}(\boldsymbol{X}, R)|} \\ &= 1 - \frac{\mathbb{E}_{\boldsymbol{x}}[h_R(\boldsymbol{Y} \mid \boldsymbol{x})]}{h_R(\boldsymbol{Y})} \frac{|R| - 1}{|R| - |\,\mathit{dom}(\boldsymbol{X}, R)|}. \end{aligned}$$ □

Lemma 6.

$$\mathit{FI}(\boldsymbol{X} \to \boldsymbol{Y}, R) = 1 - \frac{H_R(\boldsymbol{Y} \mid \boldsymbol{X})}{H_R(\boldsymbol{Y})}.$$

Proof. To show the claimed equality, we reason as follows. Recall that we implicitly assume throughout the paper that R is non-empty. By definition

$$\text{FI}(\boldsymbol{X} \to \boldsymbol{Y}, R) := \begin{cases} 1 & \text{if } |\,\mathit{dom}_R(\boldsymbol{Y})| = 1, \\ \frac{H_R(\boldsymbol{Y}) - H_R(\boldsymbol{Y}|\boldsymbol{X})}{H_R(\boldsymbol{Y})} & \text{otherwise.} \end{cases}$$

We now make a case analysis.

- If $|\,\mathit{dom}_R(\boldsymbol{Y})| = 1$ then $H_R(\boldsymbol{Y}) = 0$. Moreover, if $H_R(\boldsymbol{Y}) = 0$, also $H_R(\boldsymbol{Y} \mid \boldsymbol{X}) = 0$. As such,

$$1 - \frac{H_R(\boldsymbol{Y} \mid \boldsymbol{X})}{H_R(\boldsymbol{X})} = 1 - \frac{0}{0} = 1 - 0 = 1 = \text{FI}(\boldsymbol{X} \to \boldsymbol{Y}, R),$$

as desired.

- If $|\,\mathit{dom}_R(\boldsymbol{Y})| > 1$ then

$$\begin{aligned} \text{FI}(\boldsymbol{X} \to \boldsymbol{Y}, R) &= \frac{H_R(Y) - H_R(\boldsymbol{Y} \mid \boldsymbol{X})}{H_R(\boldsymbol{Y})} \\ &= 1 - \frac{H_R(\boldsymbol{Y} \mid \boldsymbol{X})}{H(\boldsymbol{Y})} \end{aligned}$$ □

Appendix E
Additional information on the sensitivity analysis

We provide more detail on the material presented in Section V-B. Specifically, for each synthetic benchmark $\mathcal{B}$ Figure 3 shows separate plots of the average measure values on $\mathcal{B}^+$ (in solid lines) and $\mathcal{B}^-$ (in dashed lines). Values for g_1 and g_1' are grouped together as their average measure values are the same on both $\mathcal{B}^+$ and $\mathcal{B}^-$. From Figure 3 we make the following additional observations w.r.t. LHS-uniqueness and RHS-skew.

LHS-uniqueness. For ρ, g_2, g_1^S, g_3, FI, *pdep*, and τ the average measure values on UNIQ$^-$ increase, eventually approaching the measure values on UNIQ$^+$. For RFI$^+$ and SFI, by contrast, the average measure values over UNIQ$^+$ decrease towards zero for increasing LHS-uniqueness, eventually reaching the value on UNIQ$^-$. Note that this decrease is already observable for small LHS-uniqueness values. This means that all mentioned measures are biased w.r.t. LHS-uniqueness (either by inflating the scores of non-FDs or deflating the scores of FDs).

RHS-skew. The distinguishing power of all VIOLATION measures, as well as g_1^S, and *pdep* drops when RHS-skew increases. Indeed, over SKEW$^+$ the average measure values remains relatively constant as RHS-skew increases, while the average measure values increases and approaches the values over SKEW$^+$. These measures are thus biased w.r.t. RHS-skew: their score for $\boldsymbol{X} \to \boldsymbol{Y}$ increases solely on the basis of $\boldsymbol{Y}$ and independent of $\boldsymbol{X}$ even if relations are generated by a process that sampled $\boldsymbol{X}$ and $\boldsymbol{Y}$ independently at random. We observe that the SHANNON measures (save g_1^S), τ, and μ^+ correct for this behavior. Furthermore, FI and τ slightly decrease at higher levels of RHS-skew.

Appendix F
Additional information on the evaluation on Real-World Data

We provide more detail on the material presented in Section VI.

Precision-Recall Curves. The precision-recall curves shown in Figure 4 supplement our statements based on the area under the curve values. They constitute the graphical counterpart for Table 2a. It is clear that for each class one of the measures stands out; g_3' for VIOLATION, RFI$'^+$ for SHANNON and μ^+ for LOGICAL.

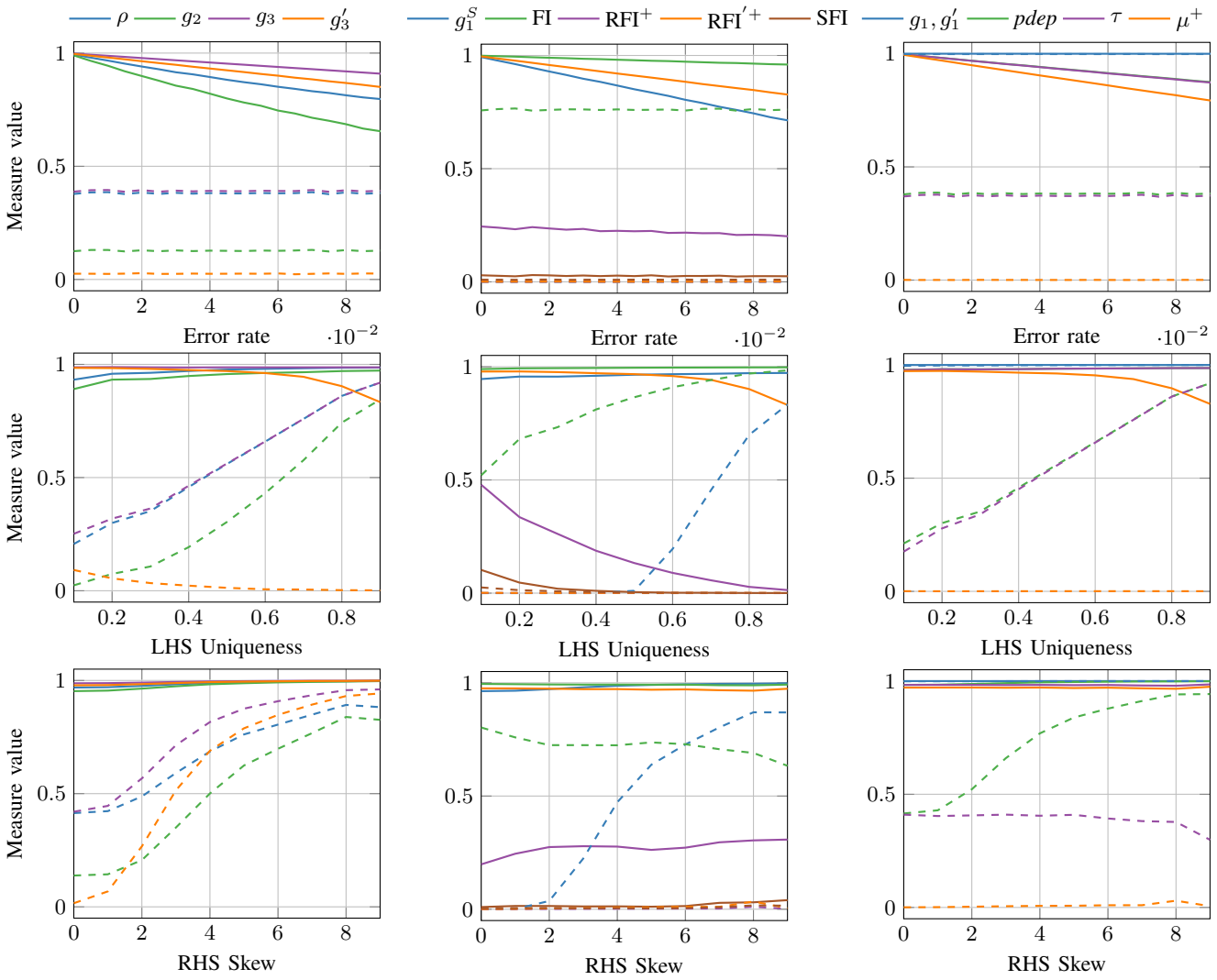

Fig. 3: Average measures values of the three experiments on SYN data. The top row shows ERR$^+$ (solid lines) and ERR$^-$ (dashed lines) for different error levels. Middle row shows UNIQ$^+$ (solid lines) and UNIQ$^-$ (dashed lines) for different LHS-uniqueness levels. Bottom row shows SKEW$^+$ (solid lines) and SKEW$^-$ (dashed lines) for different RHS-skew levels. Some of the lines are hidden behind each other. The ERR$^-$ lines of g_1^S and RFI$^+$ are behind RFI$'^+$. The ERR$^+$ line of *pdep* is behind τ. The ERR$^-$ line of g_1, g_1' is behind its own ERR$^+$ line. The UNIQ$^-$ line of RFI$^+$ is behind RFI$'^+$. The UNIQ$^+$ line of *pdep* is behind τ. The SKEW$^-$ lines of RFI$^+$ and RFI$'^+$ are behind SFI. The SKEW$^-$ line of g_1, g_1' is behind the SKEW$^+$ line of itself.

Measure Runtimes. Table IV shows the runtimes of each measure in more detail. We observe that ρ is the fastest measure with 110 seconds to calculate a value for all 1634 candidate FDs. In general, the measures of the VIOLATION class are faster compared to the others. Measures of the LOGICAL class take on average roughly 21 seconds longer to compute values for all candidate FDs. In the SHANNON class, the differences are much larger. While g_1^S and FI achive runtimes comparable to the measures from the LOGICAL class, RFI$^+$, RFI$'^+$ and SFI were not able to calculate values for all candidates. In that, SFI is able to calculate a value for 1430 (roughly 90%) candidates while RFI$^+$ and RFI$'^+$ finish only 250 (roughly 15%) candidates.

RWD. Next we consider the whole of RWD and disregard the measures RFI$^+$ and RFI$'^+$ for which the scores of not all candidate FDs could be computed. A comparison between the AUC values of RWD$^-$ and RWD in Table V show only minor differences. SFI shows the largest differences (-0.036), followed by g_1 and g_1' (-0.026 and -0.027 respectively) and FI with a difference of -0.019. All other measures show differences below -0.01. We also note that all differences are negative, which means that each measure achieved better results when the NULL values of RFI$^+$ and RFI$'^+$ are left out. We do not observe any notable changes for any of the tables of RWD.

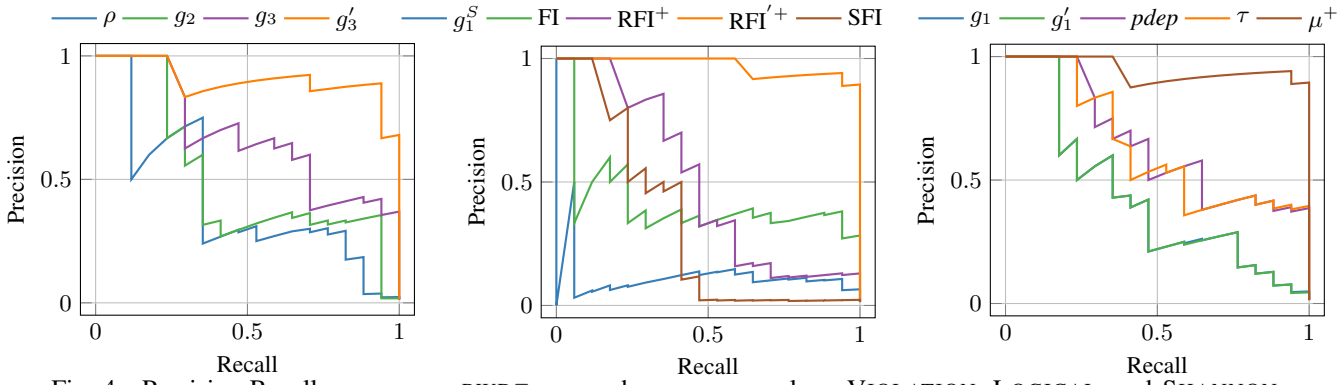

Fig. 4: Precision-Recall curves over RWD$^-$ grouped per measure class: VIOLATION, LOGICAL, and SHANNON.

TABLE IV: RFI$^+$, RFI$'^+$ and SFI are significantly slower than the other measures. The table shows the runtimes, capped at 24 hours, and the number of measures AFD candidates within the runtime.

	runtime	candidates
ρ	110s	1634
g_2	130s	1634
g_3	118s	1634
g_3'	128s	1634
g_1^S	137s	1634
FI	154s	1634
RFI$^+$	24h	250
RFI$'^+$	24h	250
SFI	24h	1430
g_1	134s	1634
g_1'	135s	1634
pdep	135s	1634
τ	151s	1634
μ^+	157s	1634

RWD without RWD$^-$. To get more insight into the candidate FDs that could not be calculated by RFI$^+$ and RFI$'^+$, we investigated the measure results for this exact subset: RWD $\setminus$ RWD$^-$. This subset contains 405 candidates that are exclusively non-AFDs. For this reason, we cannot assess precision, recall or other metrics. However, we can derive some intuition about how difficult those candidates are by investigating the measure values of the other measures. The upper part of Table VI shows an overview of summary statistics of RWD $\setminus$ RWD$^-$.

We observe that both average and median of most measure values are low. Still, it is noteworthy that judging by very high max values there seems to be a small subset of candidate FDs that could lead to changes the AUCs of RFI$^+$ and RFI$'^+$. Ultimately, the only way to confirm that suspicion would be to use RFI$^+$ and RFI$'^+$ to calculate values for all candidates of RWD $\setminus$ RWD$^-$.

In addition, we investigated the properties of RWD $\setminus$ RWD$^-$, especially with regard to the properties we used in Section V. The lower part of Table VI shows an overview of summary statistics of the number of tuples, LHS-uniqueness and RHS-skew. Both LHS-uniqueness and RHS-skew do not show unexpected values. LHS-uniqueness is rather low, RHS-skew is around 1.0, both of which match properties of RWD in general. The number of tuples is rather large, indicating that having a lot of values increases the runtimes of RFI$^+$ and RFI$'^+$.

APPENDIX G
EVALUATION ON RWD WITH EXTRA ERRORS

To study the measures' sensitivity to different kinds and different levels of errors on real world data, we created the benchmark RWD^e. We obtain RWD^e by passing the relations $R \in$ RWD through a controlled error channel such that, denoting by R' the obtained relation, some FDs in $PFD(R)$ do not hold anymore in R' and hence become part of $AFD(R')$. Existing AFDs are always maintained, i.e., $AFD(R) \subseteq AFD(R')$.

We actually have multiple error channels, which are parameterized by an error level $\eta \in [0, 1]$ and an error type. When passing R through the channel we consider all $\boldsymbol{X} \to \boldsymbol{Y} \in PFD(R)$ and modify $k = \lfloor \eta |R| \rfloor$ $\boldsymbol{Y}$-values. To avoid interference, we select at most one FD $\boldsymbol{X} \to \boldsymbol{Y}$ for every unique $\boldsymbol{Y}$ per relation, ensuring that $\boldsymbol{Y}$ does not appear in $AFD(R)$, and that no FD $\boldsymbol{Y} \to \boldsymbol{Z}$ has previously been selected. The procedure to modify the $\boldsymbol{Y}$ values is determined by the chosen type of data error for which we consider three categories inspired by Arocena et al. [1]: copy error, typo and bogus value. For a chosen tuple $\boldsymbol{w} \in R$, only $\boldsymbol{w}|_{\boldsymbol{Y}}$ is changed, where the change depends on the data error type:

(i) copy: Randomly pick any $\tilde{\boldsymbol{w}} \in R$ with $\tilde{\boldsymbol{w}}|_{\boldsymbol{Y}} \neq \boldsymbol{w}|_{\boldsymbol{Y}}$ and make $\tilde{\boldsymbol{w}}|_{\boldsymbol{Y}}$ the new value for $\boldsymbol{w}|_{\boldsymbol{Y}}$.
(ii) typo: To every $\boldsymbol{y} \in dom_R(\boldsymbol{Y})$, we associate three new values representing three common typos. From these, one is chosen each time at random as the new value for $\boldsymbol{w}|_{\boldsymbol{Y}}$.
(iii) bogus: $\boldsymbol{w}|_{\boldsymbol{y}}$ is assigned a unique newly generated value.

TABLE V: The differences between RWD$^-$ and RWD without RFI$^+$ and RFI$'^+$ are minimal. The table shows the same overview as Table 2a for RWD. RFI$^+$ and RFI$'^+$ show the same values, as the candidates do not change compared to RWD$^-$.

	RWD	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	best (pct)
ρ	0.411	**1.000**	0.029	0.038	**1.000**	**1.000**	**1.000**	0.667	**1.000**	**1.000**	**1.000**	70
g_2	0.497	**1.000**	0.507	0.038	**1.000**	**1.000**	**1.000**	0.667	**1.000**	**1.000**	**1.000**	70
g_3	0.669	**1.000**	**1.000**	0.148	**1.000**	**1.000**	0.792	0.667	**1.000**	**1.000**	**1.000**	70
g_3'	0.901	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.792	0.667	**1.000**	**1.000**	**1.000**	80
g_1^S	0.108	**1.000**	**1.000**	0.027	**1.000**	**1.000**	0.633	0.667	0.100	**1.000**	**1.000**	60
FI	0.396	**1.000**	**1.000**	0.049	**1.000**	**1.000**	**1.000**	**0.917**	**1.000**	**1.000**	**1.000**	**90**
RFI$^+$	0.494	**1.000**	**1.000**	0.182	0.508	**1.000**	0.133	0.667	0.250	**1.000**	**1.000**	50
RFI$'^+$	**0.971**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.917**	**1.000**	**1.000**	**1.000**	**100**
SFI	0.284	**1.000**	**1.000**	0.003	0.296	**1.000**	0.050	0.583	0.056	**1.000**	**1.000**	50
g_1	0.399	**1.000**	0.587	0.014	**1.000**	**1.000**	0.598	0.667	**1.000**	**1.000**	**1.000**	60
g_1'	0.398	**1.000**	0.548	0.014	**1.000**	**1.000**	0.598	0.667	**1.000**	**1.000**	**1.000**	60
pdep	0.642	**1.000**	**1.000**	0.076	**1.000**	**1.000**	0.792	0.667	**1.000**	**1.000**	**1.000**	70
τ	0.623	**1.000**	**1.000**	0.082	**1.000**	**1.000**	**1.000**	0.667	**1.000**	**1.000**	**1.000**	80
μ^+	**0.946**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.667	**1.000**	**1.000**	**1.000**	**90**

TABLE VI: The candidate FDs that RFI$^+$ and RFI$'^+$ could not calculate have the potential to decrease their performance. The upper part shows summary statistics of all measures for RWD $\setminus$ RWD$^-$. The lower part shows summary statistics of properties of the candidates, indicating that a high tuple count increases the runtime of RFI$^+$ and RFI$'^+$.

	mean	std	min	median	max
ρ	0.163	0.285	0.000	0.010	0.997
g_2	0.117	0.248	0.000	0.000	0.995
g_3	0.215	0.299	0.000	0.041	0.998
g_3'	0.142	0.236	0.000	0.013	0.982
g_1^S	0.069	0.219	0.000	0.000	0.995
FI	0.343	0.339	0.000	0.226	1.000
SFI	0.056	0.142	0.000	0.004	0.871
g_1	0.791	0.260	0.071	0.954	1.000
g_1'	0.789	0.262	0.071	0.953	1.000
pdep	0.187	0.290	0.000	0.021	0.998
τ	0.164	0.281	0.000	0.012	0.998
μ^+	0.089	0.198	0.000	0.000	0.981
tuples	377989	432051	10000	114869	1000000
LHS-uniqueness	0.095	0.235	0.000	0.000	0.981
RHS-skew	0.967	1.260	0.000	0.687	8.639

We point out that copy does not introduce any new values and keeps $dom_R(\boldsymbol{Y})$ stable, while typo (resp., bogus) introduces a number of new values independent of (resp., dependent on) the error level. $\boldsymbol{X}$ is not modified, and therefore $p_{R'}(\boldsymbol{X}) = p_R(\boldsymbol{X})$. To ensure that increasing error levels do not accidentally reduce errors, we ensure that, for each $\boldsymbol{x}$: $\boldsymbol{X}$ we pick at most $\lfloor N_{\boldsymbol{x}}/2 \rfloor$ tuples $\boldsymbol{w}$ with $\boldsymbol{w}|_{\boldsymbol{X}} = \boldsymbol{x}$ to modify, where $N_{\boldsymbol{X}}$ is the number of times that $\boldsymbol{x}$ occurs in $\pi_{\boldsymbol{X}}(R)$. *PFD*s for which this cannot be guaranteed are omitted. The number of new AFDs that can be constructed therefore depends on the error level.

We consider four error levels: 1%, 2%, 5% and 10%. For each type of data error t and each error level η, we obtain a new benchmark RWD$^e[t, \eta]$. Consequently, we generate 12 RWD^e tables per RWD table R for which $|PFD(R)| > 0$ (so, tables R_8 and R_9 are excluded). Overall the number of AFDs increases from 17 in RWD to 39 in RWD^e[copy, 1%]. That number is the same for the other error types but can drop a little for higher noise levels as explained above. Similar to RWD$^-$, we retain only AFD candidates where each measure calculates a value in a reasonable amount of time. A complete overview of the number of additional AFDs per relation and per error level is given in [32]. Per combination of parameters, the ground truth then consists of the thus constructed AFDs together with the AFDs from R.

AUC. Table VII lists AUC scores over RWD^e per error type and for different error levels. We observe that μ' has the highest AUC score in 6 out of 12 cases, followed by RFI$'^+$ that scores highest in 4 our of 12 cases. Both μ^+ and RFI$'^+$ are the top-two scoring measures consistently. The only exception is RWD^e[typo, 10%] where RFI$^+$ is the highest scoring measure. We also observe that μ^+ scores highest on all RWD^e datasets with $\eta = 2\%, 5\%$ while RFI$'^+$ scores highest on all RWD^e with $\eta = 1\%$. There is no clear observation for $\eta = 10\%$. Regarding the error types, we observe that in general the AUCs of copy are higher on the same error levels. For RWD^e[bogus, 10%] g_3,g_1^S and FI are exceptions. We observe a similar, yet less pronounced, trend for typo compared to bogus, which does not hold for $\eta = 10\%$. We remark that for some measures the AUC score on RWD^e is larger at the 1% error level than for RWD$^-$. This is not completely unexpected as the ground truth for both is different. A suprising result is presented for g_1^S, where the AUC of each RWD^e dataset is higher than for RWD$^-$.

TABLE VII: μ^+ and RFI$'^+$ score best RWD^e with one exception. The table shows the AUCs for each noise type and noise level of RWD^e, the dataset derived from RWD by introducing errors. in the table header, the sizes n of each of the RWD^e datasets is shown. For comparison, the first column repeats the AUCs for RWD$^-$.

n	RWD$^-$ 1229	copy, 1 1204	copy, 2 1206	copy, 5 1218	copy, 10 1143	bogus, 1 1130	bogus, 2 1127	bogus, 5 1111	bogus, 10 992	typo, 1 1188	typo, 2 1194	typo, 5 1189	typo, 10 915
ρ	0.417	0.345	0.229	0.161	0.100	0.241	0.172	0.101	0.069	0.288	0.204	0.132	0.073
g_2	0.504	0.341	0.258	0.205	0.169	0.235	0.238	0.180	0.215	0.265	0.235	0.169	0.102
g_3	0.674	0.595	0.435	0.323	0.251	0.513	0.342	0.250	0.265	0.540	0.362	0.256	0.163
g_3'	0.901	0.586	0.474	0.361	0.302	0.550	0.404	0.261	0.271	0.561	0.461	0.283	0.184
g_1^S	0.109	0.468	0.345	0.263	0.223	0.364	0.288	0.220	0.251	0.403	0.302	0.224	0.148
FI	0.415	0.467	0.367	0.288	0.259	0.367	0.338	0.243	0.303	0.399	0.336	0.238	0.169
RFI$^+$	0.494	0.403	0.401	0.378	0.410	0.327	0.289	0.244	0.224	0.392	0.377	0.333	**0.387**
RFI$'^+$	**0.971**	**0.775**	**0.626**	**0.513**	**0.468**	**0.622**	**0.497**	**0.363**	**0.372**	**0.689**	**0.523**	**0.383**	**0.304**
SFI	0.320	0.238	0.240	0.230	0.256	0.082	0.080	0.074	0.102	0.169	0.170	0.169	0.238
g_1	0.425	0.369	0.312	0.252	0.213	0.277	0.271	0.215	0.177	0.316	0.296	0.215	0.119
g_1'	0.425	0.368	0.311	0.251	0.212	0.276	0.271	0.215	0.176	0.315	0.296	0.215	0.119
pdep	0.647	0.517	0.391	0.283	0.239	0.412	0.325	0.235	0.253	0.443	0.344	0.237	0.150
τ	0.630	0.630	0.473	0.333	0.275	0.459	0.365	0.260	0.299	0.491	0.379	0.261	0.182
μ^+	**0.946**	**0.773**	**0.637**	**0.550**	**0.476**	**0.618**	**0.508**	**0.374**	**0.352**	**0.668**	**0.552**	**0.400**	0.294

TABLE VIII: On average, μ^+ and RFI$'^+$ score most reliably over RWD^e. The table shows the percentages per error type where a measure has the lowest rank to reach a recall of 1.0.

	copy	bogus	typo
ρ	3.1	0.0	0.0
g_2	9.4	0.0	9.7
g_3	21.9	**42.3**	48.4
g_3'	25.0	**42.3**	41.9
g_1^S	25.0	19.2	32.3
FI	40.6	38.5	35.5
RFI$^+$	25.0	26.9	25.8
RFI$'^+$	43.8	**50.0**	**58.1**
SFI	12.5	15.4	16.1
g_1	12.5	7.7	9.7
g_1'	12.5	7.7	9.7
pdep	21.9	26.9	29.0
τ	**53.1**	30.8	41.9
μ^+	**75.0**	**42.3**	**54.8**

We do see that, as expected, the AUC score for each of the measures deteriorates at increasing error levels to an absolute low at error level 10%. The exception is RFI$'$ whose performance increases at higher error levels for error types copy but not for the other error types. When error types and error levels are unknown but expected to be small, μ' therefore remains the best choice of AFD-measure. Furthermore, it is evident from Table V that AFD-measures are not very effective when error levels are greater than 5%.

Within VIOLATION g_3' remains the best measure. Within SHANNON we see RFI$'^+$ scoring best while SFI performs worst for most datasets. The measures g_1^S, FI and RFI$^+$ show very similar scores. Within LOGICAL we observe that τ is usually an improvement over *pdep*, and μ^+ is an improvement over τ. Further, g_1 and g_1' perform consistently worse than any other measure in LOGICAL.

Rank at max recall. We show in Table VIII a qualitative comparison between measures by listing, for each measure f and error type t, its *winning number*, which is defined as follows. Consider a particular $(\text{relation}, t, \eta)$ combination in RWD^e. A measure f *wins* this triple if its *r@mr* is minimal among all measures on this triple. The winning number of f for error type t is then the number of times f wins, taken over all triples of type t. Here, we see again that both RFI$'^+$ and μ^+ score very well. Across all error types, μ^+ is the only measure that is in included in the top two values. It scores best on copy, where its rank at max recall is minimal for 75% of all relations. τ is the runner-up with 53.1% of the relations. In bogus, RFI$'^+$ achives a minimal rank at max recall for 50.0% of the relations. g_3, g_3' and μ^+ follow with 42.3%. For typo, RFI$'^+$ ranks candidates best on 58.1% relations. μ^+ is behind that with 54.8%.