Measuring approximate functional dependencies: a comparative study

Peer-reviewed author version

PARCIAK, Marcel; WEYTJENS, Sebastiaan; HENS, Niel; NEVEN, Frank; PEETERS, Liesbet & VANSUMMEREN, Stijn (2025) Measuring approximate functional dependencies: a comparative study. In: Vldb Journal, 34 (4) (Art N° 56).

DOI: 10.1007/s00778-025-00931-x
Handle: http://hdl.handle.net/1942/46453

Noname manuscript No.
(will be inserted by the editor)

Measuring Approximate Functional Dependencies: a Comparative Study

Marcel Parciak · Sebastiaan Weytjens · Niel Hens · Frank Neven · Liesbet M. Peeters · Stijn Vansummeren

Accepted: 16/06/2025

Abstract Approximate functional dependencies (abbreviated: 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. Qualitatively, we perform a sensitivity analysis w.r.t. structural properties of input relations. Quantitatively, we study the effectiveness of AFD measures for ranking linear AFDs on real world data. Based on this analysis, we give clear recommendations for the AFD measures to use in practice.

Keywords functional dependencies · data cleaning · data profiling

M. Parciak
UHasselt, BIOMED & Data Science Institute
Agoralaan, Building D, 3590 Diepenbeek, Belgium
E-mail: marcel.parciak@uhasselt.be

S. Weytjens
UHasselt, Data Science Institute
Agoralaan, Building D, 3590 Diepenbeek, Belgium
E-mail: sebastiaan.weytjens@uhasselt.be

N. Hens
UHasselt, Data Science Institute
Agoralaan, Building D, 3590 Diepenbeek, Belgium
E-mail: niel.hens@uhasselt.be

F. Neven
UHasselt, Data Science Institute
Agoralaan, Building D, 3590 Diepenbeek, Belgium
E-mail: frank.neven@uhasselt.be

L. M. Peeters
UHasselt, BIOMED & Data Science Institute
Agoralaan, Building D, 3590 Diepenbeek, Belgium
E-mail: liesbet.peeters@uhasselt.be

S. Vansummeren
UHasselt, Data Science Institute
Agoralaan, Building D, 3590 Diepenbeek, Belgium
E-mail: stijn.vansummeren@uhasselt.be

1 Introduction

Functional dependencies (FDs) describe a strong relation between two sets of relational attributes, indicating that the values of one set determine the values of the other in a given relation. Knowing FDs for a given instance of a database schema aids in ensuring data consistency, and helps in data cleaning and data profiling [1, 12, 40]; facilitates data integration [43]; and can be exploited for query optimization [25, 27], among other tasks. In many data science scenarios, however, the set of design FDs is unknown or incomplete [32]. As such, a variety of techniques have been proposed to reverse engineer this set of design FDs from a given relation instance [3, 6, 31, 32, 42].

In practical settings, however, database instances may get corrupted with respect to the target set of design FDs. This is due to, for example, errors during data entry. Reverse engineering FDs from such instances is particularly challenging, as it clearly does not suffice to simply enumerate the FDs that are satisfied in an instance. Instead, one must also consider *approximate functional dependencies* (AFDs) that is, FDs that "almost hold" in the relation instance. In this respect, 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. Given a AFD measure one may discover the set of original design FDs by ranking the search space, and returning all FDs larger than a given threshold. A good AFD measure, then, is one that ranks the 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.

Many AFD measures have been proposed in the literature over the past decades [4, 19, 22, 24, 28, 29, 36, 37, 43]. Unfortunately, these measures vary widely in nature and there has been little study so far in comparing them. As such, the answers to the following two questions remain unclear.

(a) *How do the different AFD measures compare?*
(b) *Which AFD measure(s) should one use when aiming to discover AFDs in real-world data?*

In this paper, we provide a complete answer to (a) and a partial answer to (b) for the special setting when we aim to discover *linear* AFDs in real-world data.[1] Our answer to (b) is based on experimental comparison on a newly established benchmark with real-world data. Importantly, we show theoretically and in our synthetic data analysis that the challenges affecting measure performance on linear AFDs are also manifested for non-linear AFDs. Thus, our study identifies the most suitable measures for linear AFD discovery while suggesting that they are also the most promising candidates for non-linear AFD discovery.

It is important to state that we only study AFD measures and do not consider the design of AFD *discovery algorithms*. AFD discovery algorithms usually fix an AFD measure but then combine a multitude of techniques to do the actual discovery. This includes pruning the ranked search space for efficiency reasons (e.g., [26, 28, 29, 36]) or complementing the measures' ranking with heuristics for application-specific purposes (e.g., [22, 47]). Of course, the improved knowledge of AFD measures that we provide here may aid in improving discovery algorithms in the future.

We next summarize how we address (a) and (b), as well as our main results.

1.1 Conceptual comparison

To answer question (a) we adopt the following methodology. First and foremost, we survey 12 known AFD measures and present them in a uniform formal framework. We proceed to qualitatively compare these measures along the following three axes. Table 9 in Section 9 summarizes all studied measures and our findings.

(Axis 1) First, we consider the measures' interpretability. In particular, we note that some measures are equipped with *baselines*, which are relation instances on which the measure yields a score of zero (indicating that the relation completely fails to satisfy an FD), while others do not. Having baselines is a precondition for correctly interpreting the measure score, as we discuss in Section 3. A stronger property is being *normalized*, which allows to interpret the score as a percentage between completely failing to satisfy an AFD (score of 0) and completely satisfying it (score of 1). We find in general that normalized measures perform better than non-normalized ones.

(Axis 2) Second, we exhibit common design principles among measures. We observe that existing measures can be divided into three broad classes: (i) measures that quantify the fraction of violations (we refer to measures in this class as Violation); (ii) measures based on Shannon entropy (Shannon) [14]; and (iii) measures based on logical entropy (Logical) [15]. By means of a formal comparison, we highlight striking similarities between measures in each of these classes, allowing us to relate and link measures *across classes*. For example, this allows us to say that *pdep*, a particular measure in Logical, is the logical entropy variant of g_2, a particular violation-fraction measure in Violation. By means of this comparison, we propose two new measures that do not appear in the literature, but which can be viewed as Shannon-variants of existing measures, hence applying a design principle known from one class to measures in a different class. As such, we obtain 14 measures in total for our study.

(Axis 3) We evaluate the measures' ability to distinguish between an FD $\boldsymbol{X} \rightarrow \boldsymbol{Y}$ in relation instances that were generated to satisfy the FD, but subsequently had errors introduced so that the FD no longer holds exactly (we call such instance *dependently-generated*), versus relation instances where $\boldsymbol{X}$ and $\boldsymbol{Y}$ were randomly generated. Because FDs are supposed to indicate correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$, a good AFD measure should at a minimum 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. We note, however, that there are various properties of the input relation as well as the FD itself that affect the measures' power to distinguish between these two cases. In particular, we evaluate the measures' sensitivity w.r.t. the following structural properties of the input. (i) The er-

[1] An FD is linear if it has a single attribute on the left.

ror rate, i.e., the amount of errors introduced. (ii) *LHS -uniqueness*: the normalized number of unique values occurring in $\pi_{\boldsymbol{X}}(R)$; (iii) *RHS-skew*: the skewness of the distribution of values occurring in $\pi_{\boldsymbol{Y}}(R)$. Finally, we also evaluate the measure's sensitivity w.r.t $\#LHS$, which is the number of attributes in the left-hand-side $\boldsymbol{X}$ of the FD. In general, we find that measures are more robust if they are inversely proportional to the error level and are insensitive to LHS-uniqueness and RHS-skew. Further, for non-linear FDs insensitivity to LHS-uniqueness becomes more important as #LHS increases. Through our study on synthetic data, we hence identify the measures that are robust in this respect, and those that are not.

1.2 Practical comparison

We adopt the following methodology to answer (b). Our study of question (a) focuses on the measures' ability to distinguish between dependently-generated instances and randomly-generated instances for a fixed FD $\boldsymbol{X} \to \boldsymbol{Y}$. In practice, however, we need to discover AFDs from a given relation R, not determine whether R is dependently or randomly generated. For successful practical AFD discovery, hence, we require that a measure ranks AFDs that are semantically meaningful before meaningless ones, while also being practical to compute.

To gauge the measure's performance in this respect, we first construct a new AFD discovery benchmark. This is necessary because existing benchmarks are for exact FD discovery, which are designed to gauge algorithmic efficiency. As such, they do not contain a semantically meaningful "ground truth" set of design FDs that need to be discovered. Our benchmark, denoted RWD, identifies a set of linear and semantically meaningful FDs to discover. It is obtained by inspecting real-world datasets from existing benchmarks and manually creating the set of design FDs for them.

On RWD, we first give insight into the measures' computational efficiency. In particular, while most measures are quick to rank the entire search space of possible FDs, the SHANNON measures reliable and smooth fraction of information fail to rank the entire search space within a reasonable amount of time.

We then compare the quality of the measures' ranking ability on RWD^-, which is a subset of RWD on which all measures were able to compute a score within a reasonable time threshold to identify the best-ranking AFD measure(s). Our analysis in Section 7.2 of measure performance on RWD^- shows that well-ranking measures exist within each of our newly identified measures classes (VIOLATION, SHANNON and LOGICAL). Furthermore, the best-ranking measure for VIOLATION (measure g_3') is sensitive to RHS-skew and therefore performs worse compared to the best-ranking measures for SHANNON (measure RFI'^+) and LOGICAL (measure μ^+), which have comparable performances as well as equal structural sensitivity properties. However, RFI'^+ has the disadvantage of being unreasonably slow to compute whereas μ^+ is very efficient.

We subsequently complement our findings by disregarding slow-to-compute measures and analysing the remaining measures on the full RWD benchmark in Section 7.3. This ensures that our findings from RWD^- were not influenced by the exclusion of measures due to computational constraints.

The number of actual AFDs in RWD (i.e. semantically meaningful FDs that are not satisfied in the data) remains relatively low. To gauge how consistent the AFD measures are, we therefore synthetically increase the number of AFDs in RWD by passing relations through a controlled error channel causing previously satisfied FDs to no longer hold. This final benchmark, denoted RWD^{e}, allows us to inspect the measures' ability to cope with an increasing amount of errors while having realistic data distributions.

Recommendation. Based on all of theses investigations, we recommend μ^+ for practical linear AFD discovery. We give more in-depth discussion and recommendations in Section 9, where Table 9 summarizes our comparison.

In summary, our contributions are as follows. (1) A survey of AFD measures, using a new and uniform presentation. (2) Formal classification of measures, and comparison and linking of the measures across classes. (3) Sensitivity analysis of measure performance w.r.t. structural properties of the input relations. (4) Creation of three real-world benchmarks for linear AFD discovery. (5) Analysis of measure ranking power on these benchmarks. (6) Clear recommendations for measure adoption in linear AFD discovery. (7) Discussion of the importance of insensitivity to LHS-uniqueness for non-linear AFD discovery.

Parts of this paper have been published in the proceedings of the 40th IEEE International Conference on Data Engineering [34]. This paper extends the conference version by adding (1) the interpretability analysis (Axis 1 of the conceptual comparison); (2) the formal comparison and linking between classes (Axis 2); (3) an extension of the generated relation instances with respect to #LHS (Axis 3); (4) a deeper practical comparison, also discussing computational efficiency, validation of our findings on the novel benchmark RWD

Table 1: Glossary of notation used in this paper.

X, Y	an attribute
$dom(X)$	the domain of an attribute
$\boldsymbol{X}, \boldsymbol{Y}$	a finite set of attributes
$\boldsymbol{x}$	a tuple over set $\boldsymbol{X}$ of attributes
$\boldsymbol{x}\colon \boldsymbol{X}$	$\boldsymbol{x}$ is a tuple over $\boldsymbol{X}$
$\boldsymbol{x}\vert_{\boldsymbol{Y}}$	restriction of $\boldsymbol{x}$ to $\boldsymbol{Y}$ with $\boldsymbol{Y} \subseteq \boldsymbol{X}$
$\boldsymbol{XY}$	union of two sets of attributes, i.e. $\boldsymbol{X} \cup \boldsymbol{Y}$
$\boldsymbol{xy}$	idem for tuples: i.e. $\boldsymbol{xy}\vert_{\boldsymbol{X}} = \boldsymbol{x}$ and $\boldsymbol{xy}\vert_{\boldsymbol{Y}} = \boldsymbol{y}$
R	a relation
$R(\boldsymbol{X})$	R is relation over $\boldsymbol{X}$
$R(\boldsymbol{x}) \in \mathbb{N}$	frequency of $\boldsymbol{x}$ in R
$\boldsymbol{x} \in R$	$\boldsymbol{x}$ is a tuple in R with $R(\boldsymbol{x}) > 0$
$dom_R(\boldsymbol{Y})$	$\{\boldsymbol{x}\vert_Y \mid \boldsymbol{x} \in R\}$
$\vert R\vert$	total number of tuples contained in R, i.e. $\sum_{\boldsymbol{x}\colon \boldsymbol{X}} R(\boldsymbol{x})$
$\pi_{\boldsymbol{X}}(R)$	bag-based relational projection of R onto $\boldsymbol{X}$
$\sigma_{\boldsymbol{X}=\boldsymbol{x}}(R)$	bag-based relational selection of R on $\boldsymbol{X} = \boldsymbol{x}$
$\varphi := \boldsymbol{X} \to \boldsymbol{Y}$	a functional dependency
$R \models \varphi$	R satisfies a functional dependency φ
$R \not\models \varphi$	R violates a functional dependency φ
$\Delta(R)$	the (fixed) schema of R

over RWD$^-$, and analysis for increasing error levels on the novel benchmark RWD^e. We believe that together this significantly contributes to our overall comparison of AFD measures and their effectiveness under different conditions.

This paper is organized as follows. We introduce the necessary background in Section 2. We survey and formally introduce AFD measures in Section 3 where we also compare them w.r.t. interpretability. We classify and formally compare AFD measures in Section 4. We relate measure behavior on linear FDs to behavior on non-linear FDs in Section 5. We investigate sensitivity w.r.t. structural properties in Section 6. We compare measures on real-world data in Section 7. We discuss related work in Section 8. Finally, we provide recommendations, discuss future work, and conclude in Section 9.

2 Preliminaries

We summarize the notation used in this paper in Table 1. 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 $\boldsymbol{x}\vert_{\boldsymbol{Y}}$ for the restriction of $\boldsymbol{x}$ to $\boldsymbol{Y} \subseteq \boldsymbol{X}$. We use juxtaposition 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. $\boldsymbol{xy}\vert_{\boldsymbol{X}} = \boldsymbol{x}$ and $\boldsymbol{xy}\vert_{\boldsymbol{Y}} = \boldsymbol{y}$.

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 require relations to be finite in the sense that $R(\boldsymbol{x})$ can be non-zero for at most a finite number of $\boldsymbol{x}$. 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 $R(\boldsymbol{X})$. $|R|$ denotes the total number of tuples in R, $\pi_{\boldsymbol{Y}}(R)$ denotes bag-based projection on $\boldsymbol{Y}$, and $\sigma_{\boldsymbol{X}=\boldsymbol{x}}(R)$ denotes bag-based selection. If $\boldsymbol{Y} \subseteq \boldsymbol{X}$ then we denote by $dom_R(\boldsymbol{Y})$ the set $\{\boldsymbol{x}\vert_Y \mid \boldsymbol{x} \in R\}$.

Functional Dependencies. A *functional dependency* (an FD for short) is an expression φ of the form $\boldsymbol{X} \to \boldsymbol{Y}$. 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 $\boldsymbol{w}\vert_{\boldsymbol{Y}} = \boldsymbol{w'}\vert_{\boldsymbol{Y}}$ whenever $\boldsymbol{w}\vert_{\boldsymbol{X}} = \boldsymbol{w'}\vert_{\boldsymbol{X}}$. In what follows, we always implicitly assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are disjoint when considering FDs. An FD is *linear* if $|\boldsymbol{X}| = 1$ and *non-linear* otherwise. An FD is *unitary* if $|\boldsymbol{Y}| = 1$. It is well-known that non-unitary FDs with $|\boldsymbol{Y}| > 1$ can be expressed as a set of unitary FDs. Whenever convenient, we may therefore assume w.l.o.g. that $|\boldsymbol{Y}| = 1$.

Dependency Discovery. A *schema* is a finite set of unitary 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 $\Delta(R)$, the noisy relation 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 3 we 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 *Shannon* or *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 will 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: $p_R(\boldsymbol{y})$ 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 $\boldsymbol{X} = \boldsymbol{x}$. Thus,

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

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 [14] by[2]

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

$H_R(\boldsymbol{X})$ 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

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

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.,

$$H_R(\boldsymbol{Y} \mid \boldsymbol{X}) = \mathbb{E}_{\boldsymbol{x}} \left[H_R(\boldsymbol{Y} \mid \boldsymbol{x}) \right].$$

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

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 $\boldsymbol{X}$ [15]. That is,

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

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.,

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

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 $p_R(\boldsymbol{x}) - p_R(\boldsymbol{xy})$ 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 $h_R(\boldsymbol{Y} \mid \boldsymbol{X}) \neq \mathbb{E}_{\boldsymbol{x}}[h_R(\boldsymbol{Y} \mid \boldsymbol{x})]$.

Discussion on logical and Shannon entropy. The notion of logical entropy arises in mathematical philosophy [15], 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 [15]). 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.

3 AFD Measures

In this section, we survey the literature on AFD measures.

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 (inefficient) 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 unitary FDs over $\boldsymbol{W}$ whose f-value lies in the range $[\epsilon, 1[$. In particular, this excludes the unitary FDs satisfied by R. As already mentioned in the Introduction, practical discovery algorithms will apply efficient pruning strategies to make the discovery efficient, and in particular aim to discover only "minimal" unitary AFDs, i.e. AFDs $\boldsymbol{X} \to Y$ for which no other AFD $\boldsymbol{X}' \to Y$ with $\boldsymbol{X}' \subset \boldsymbol{X}$ is discovered. Since our focus in this paper is on understanding measure ranking power, we disregard this aspect and continue to work with this conceptual but inefficient notion of AFD discovery algorithm.

Interpretation and baselines. A key difficulty lies for AFD discovery algorithms, as with all threshold-based algorithms, 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.

Set-based measures. Some measures do not depend on the multiplicity of tuples in the input relation, in the sense that for all $R(\boldsymbol{W})$ and $S(\boldsymbol{W})$, if $dom_R(\boldsymbol{W}) = dom_S(\boldsymbol{W})$ then $f(\varphi, R) = f(\varphi, S)$. We refer to such measures as *set-based* measures in what follows.

Conventions. Throughout this section, we define measures in full generality for arbitrary FDs, not only unitary FDs. Throughout, let R be 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.

3.1 Co-occurrence ratio

Ilyas et al. [22] 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 is 1 if 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.

3.2 g-measures

Kivinen and Mannila [24] 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

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

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 [26]. Adaptations of g_1' are also used in the context of denial constraints [35] and roll-up dependencies [8].

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 baselines of g_2 are those relations R in which every tuple is part of a violating pair.

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 [19] 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 as baselines 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 [3, 21, 23, 24]. Furthermore, the 'per-tuple' probability of an FD, as defined in [43], is precisely g_3. Berzal et al. [4] 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 [18]. We note that g_3 has been generalized to other dependencies as well: e.g., conditional FDs [13, 38], inclusion dependencies [30], and conditional matching dependencies [45]. By contrast, the normalized version g_3' only appears in [19].

3.3 Fraction of Information

Cavallo and Pittarelli [10] 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 [19]. FI is based on Shannon entropy and is formally defined as

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

The numerator $H_R(\boldsymbol{Y}) - H_R(\boldsymbol{Y} \mid \boldsymbol{X})$ is known as *mutual information* [14], which we denote by $I_R(\boldsymbol{X}; \boldsymbol{Y})$ in what follows.

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, 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. [28, 29] and Pennerath et al. [36] 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 $\mathrm{FI}(\boldsymbol{X} \to \boldsymbol{Y}, S)$ based on R. Then a result by Roulston [41] states that the expected value of $I_R(\boldsymbol{X}; \boldsymbol{Y})$, taken over all R obtained in this manner, equals

$$I_S(\boldsymbol{X}; \boldsymbol{Y}) + \frac{B_{\boldsymbol{XY}} - B_{\boldsymbol{X}} - B_{\boldsymbol{Y}} + 1}{2n}$$

where $B_{\boldsymbol{X}} := |\mathit{dom}_S(\boldsymbol{X})|$. 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. Additionally, because $H_R(\boldsymbol{Y})$ underestimates $H_S(\boldsymbol{Y})$ [41], we may conclude that $\mathrm{FI}(\boldsymbol{X} \to \boldsymbol{Y}, R)$ is expected to overestimate $\mathrm{FI}(\boldsymbol{X} \to \boldsymbol{Y}, S)$ and the magnitude of overestimation depends on the active domain sizes and the size of S. This overestimation is problematic since $\mathrm{FI}(\boldsymbol{X} \to \boldsymbol{Y}, R)$ will be quite large, even if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent in p_S, resulting in $\mathrm{FI}(\boldsymbol{X} \to \boldsymbol{Y}, S)$ being 0.

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 both on $\boldsymbol{X}$ and $\boldsymbol{Y}$, $p_{R'}(\boldsymbol{X}) = p_R(\boldsymbol{X})$ 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

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

Because the number of permutations of R is finite, we may compute $\mathbb{E}_R[\mathrm{FI}(\boldsymbol{X} \to \boldsymbol{Y}, R)]$, and therefore also $\mathrm{RFI}(\boldsymbol{X} \to \boldsymbol{Y}, R)$, by simply computing $\mathrm{FI}(\boldsymbol{X} \to \boldsymbol{Y}, R')$ for every permutation R' of R and taking the average. More efficient algorithms are proposed in [28, 29]. Even with these improved algorithms, computing RFI remains inefficient, as we show in Section 7.

Strictly speaking, RFI is not an AFD measure since it can become negative when $\mathrm{FI}(\varphi, R) < \mathbb{E}_R[\mathrm{FI}(\varphi, R)]$. 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

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

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 $\pi_{\boldsymbol{XY}}(R)$:

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

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 7.

AFD discovery algorithms based on RFI and SFI are presented in [28, 29] and [36], respectively.

3.4 Probabilistic dependency, τ and μ

Piatetsky-Shapiro and Matheus [37] 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 [20]. Finally, they also propose a rescaled version of τ. All three notions are defined as follows. It is worth noting that, apart from [37], 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,

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

where $pdep(\boldsymbol{Y} \mid \boldsymbol{x}, R)$ 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:

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

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 $pdep(\boldsymbol{Y} \mid \boldsymbol{x}, R)$ 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 we average over all values of $\boldsymbol{X}$.

The measure τ. For *pdep* 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 [37] 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 $pdep(\boldsymbol{Y}, R)$ and $pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)$, Piatetsky-Shapiro and Matheus propose to normalize $pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)$ with respect to $pdep(\boldsymbol{Y}, R)$. The resulting measure is equivalent to the τ (tau) measure of association [20], which is defined by Goodman and Kruskal 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 [37]. 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 [37] note that *pdep* and τ have the following undesirable property.

Theorem 1 (Piatetsky-Rotem-Shapiro [37]) *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*

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

and

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

Therefore, assuming 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 $pdep(\boldsymbol{Y}, R)$:

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

Note that this fraction is ill-defined if the denominator $1 - \mathbb{E}_R[pdep(\varphi, R)] = 0$. In the following lemma we show that 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 $R \models \varphi$.

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 3.4, i.e.

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

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. □

Strictly speaking, μ is not a measure since it returns negative values when $pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)$ is larger than $\mathbb{E}_R[pdep(\boldsymbol{X} \to \boldsymbol{Y}, R)]$. 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.

4 Classes of AFD measures

Looking at the previous definitions, we observe three different notions that are used to formally define a measure. Hence, we discern the following three classes (see also the second row in Table 9):

(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 FI, 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).

We discuss the similarities in the design of Logical measures and those in the Violation and Shannon class by means of Table 2, 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 2 are equivalent to their definition given in Sections 3.1–3.4.*

The interested reader may find the proof in Appendix A. Next, we discuss the similarities found in Table 2.

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 [19] considered an axiomatisation of FD error measures, and showed that the 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, where logical entropy is replaced by Shannon entropy. Giannella and Robertson [19] 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. In [19], they 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:

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

② We have already observed in Section 3.4 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 $1 - h_R(\boldsymbol{Y} \mid \boldsymbol{x})$ by expressing the probability of $\boldsymbol{x}$ not participating in a violating pair. Likewise, g_3 equals the expected value of $\max_{\boldsymbol{y}} p_R(\boldsymbol{y} \mid \boldsymbol{x})$ by expressing the largest subgroup of non-violating tuples in $\pi_{\boldsymbol{Y}}\sigma_{\boldsymbol{X}=\boldsymbol{x}}(R)$. In both cases, the expectation is taken over all $\boldsymbol{x}$.

③ As shown by the rewritten formulas in line 3 of Table 2, 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

Table 2: Overview of similarities between LOGICAL measures and measures in VIOLATION/ SHANNON.

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

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 $\mathbb{E}_R[\mathrm{FI}]$. 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

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

Conclusion. By means of the harmonized comparison above, we have identified two new measures that are SHANNON versions of existing measures. For completeness, we include both of these measures in our study, and compare their behavior to that of the other measures in the following.

5 On linear vs non-linear FDs

In the following sections we will proceed with our sensitivity analysis and experimental comparison. There, we will focus primarily on comparing measure behavior on *linear* FDs. Our primary focus on linear FDs is motivated by the following observation: each introduced measure f computes its score $f(\boldsymbol{X} \to \boldsymbol{Y}, R)$ without ever reasoning about the number of attributes in $\boldsymbol{X}$, or $\boldsymbol{Y}$. Indeed measures look at the data distributions of $\pi_{\boldsymbol{X}}(R)$, $\pi_{\boldsymbol{Y}}(R)$, $\pi_{\boldsymbol{XY}}(R)$, the number of violations, and so on ..., but never at $|\boldsymbol{X}|$ nor $|\boldsymbol{Y}|$.

In fact, each measure treats $\boldsymbol{X}$ and $\boldsymbol{Y}$ as if they were a single attribute. This can be formalized as follows. Consider a relation $R(\boldsymbol{W})$ and an FD $\boldsymbol{X} \to \boldsymbol{Y}$. Let A and B be two new attributes not occurring in $\boldsymbol{W}$, let $\boldsymbol{W}' = \boldsymbol{W} \setminus \boldsymbol{XY} \cup \{A, B\}$. Intuitively, the set of attributes $\boldsymbol{X}$ in $\boldsymbol{W}$ is replaced by the single attribute A, and $\boldsymbol{Y}$ by B. Define the linear FD $\varphi' = A \to B$ and let $R'(\boldsymbol{W}')$ be the relation obtained as follows. For every tuple $t \in R$ create a tuple $t' \in R'$ by first setting $t' = t|_{\boldsymbol{W} \setminus \boldsymbol{XY}}$ and subsequently setting $t'(A)$ (resp. $t'(B)$) equal to $t|_{\boldsymbol{X}}$ (resp. $t|_{\boldsymbol{Y}}$). This can e.g. be done by converting all $\boldsymbol{X}$-values in t into strings and concatenating them to become the single value $t'(A)$. Under this transformation of R into R' it is straightforward to verify that $f(\boldsymbol{X} \to \boldsymbol{Y}, R) = f(A \to B, R')$ for all considered measures f. Conversely, it is also possible to start from a concrete linear FD φ' and relation R' and construct non-linear FD φ and relation R such that again $f(\boldsymbol{X} \to \boldsymbol{Y}, R) = f(A \to B, R')$: split A into multi-attribute entries $\boldsymbol{X}$ and similarly B into $\boldsymbol{Y}$. In this respect, f hence does not distinguish between $\boldsymbol{X}$ and $\boldsymbol{Y}$ having multiple attributes, or being a single attribute. In what follows, we refer to this property of the measures as the *linear-indistinguishability* property.

Because of the linear-indistinguishability property, if we want to understand how different measures compare in how they rank a set $\{\varphi_1, \ldots, \varphi_\ell\}$ of candidate FDs in a relation R, it suffices to understand how they compare in ranking the corresponding set of *linear* FDs $\{\varphi'_1, \ldots, \varphi'_\ell\}$ on R', with R' constructed as discussed above. Conversely, if we see that a measure exhibits certain behavior on linear FDs, we can be sure that there are data instances where the measure exhibits the same behavior on non-linear FDs.

For this reason, we will focus primarily on linear FDs in our following discussions.

6 Sensitivity Analysis

In this section, we investigate the sensitivity of measures w.r.t. structural properties of the input relation as well as properties of the FD itself. Specifically, we want to get insight into the measures' ability to distinguish between a unitary FD $\varphi = \boldsymbol{X} \to Y$ 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 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.

6.1 Methodology

There are various properties of the input relation R as well as properties of the FD $\boldsymbol{X} \to Y$ that may affect the measures' power to distinguish between these two aforementioned cases. In what follows we say that a relation R is *dependently generated* if it was generated to satisfy $\boldsymbol{X} \to Y$ but later had errors introduced. It is *randomly generated* if it was constructed by picking values for $\boldsymbol{X}$ and Y at random. A detailed description of the generation process is given below. We study the effect on the measure's power do distinguish between dependently and randomly generated instances w.r.t. the following.

(1) The *error rate*, i.e., the amount of errors introduced in R. 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) The *uniqueness of the left-hand-side (LHS)* $\boldsymbol{X}$ of the FD in R, defined as the ratio $|dom_R(\boldsymbol{X})|/|R|$. The more this ratio approaches 1, the more $\boldsymbol{X}$ acts as a key. However, because this statistic looks only at $\boldsymbol{X}$ without taking Y into account, it not necessarily provides a good signal for concluding that R is dependently generated. Indeed, if the possible domain $dom(\boldsymbol{X})$ of $\boldsymbol{X}$-values is large and $|R|$ is relatively small, then even randomly generated relations may have large LHS-uniqueness. As a practical example, consider the setting where R is supposed to describe publications, with attribute *title* containing a publications' title. The domain of possible titles is large, making it unlikely that two publications in R share a title. The attribute will hence have a high LHS-uniqueness, but on this alone it is difficult to determine whether R is dependently generated. Nevertheless, as we will see, certain measures are heavily biased towards LHS-uniqueness, while others are not.

(3) The *right-hand-side (RHS) skew of* R, defined as the skewness of the distribution $p_R(Y)$. The larger the RHS-skew, the fewer distinct values occur in $\pi_Y(R)$, and hence the smaller the chance of violating the FD. However, similar to (2), because this statistic looks only at Y without taking $\boldsymbol{X}$ into account, it again not necessarily provides a good signal for concluding that R is dependently generated. For example, consider a relation describing World-War I casualties, where attribute *sex* records the deceased's sex. Because soldiers then were primarily male, this attribute is highly skewed towards men (although there were also woman casualties). The attribute hence has high RHS skew. Based on RHS-skew alone, it is difficult to determine whether R is dependently generated. Nevertheless, as we will see, certain measures are also heavily biased towards RHS-skew, while others are not.

(4) The number of attributes in $\boldsymbol{X}$. Ideally, distinguishing between dependently and randomly generated instances should not depend on $|\boldsymbol{X}|$. In what follows we refer to the number of attributes in the LHS as $\#LHS$.

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

- $|R| \in [100; 10000]$;
- $|\boldsymbol{X}| \in [1; 5]$;
- $|dom_R(\boldsymbol{X})| \in [\frac{1}{5}|R|, \frac{3}{4}|R|]$;
- $|dom_R(Y)| \in [5, \frac{1}{2}|dom_R(\boldsymbol{X})|]$; and
- error rate $\eta \in [0.5\%, 2\%]$.

Values for $\boldsymbol{X}$ and Y are drawn according to the Beta distribution, $B(\alpha, \beta)$, 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 we generate the parameters $|R|$, $|dom_R(\boldsymbol{X})|$, $|dom_R(Y)|$, $\alpha_{\boldsymbol{X}}$, $\beta_{\boldsymbol{X}}$, α_Y, β_Y, η are chosen uniformly at random under the conditions described above. To generate a table R in $\mathcal{B}^-$, we repeat the following procedure $|R|$ times: sample $\boldsymbol{x} \in dom_R(\boldsymbol{X})$ (respectively $y \in dom_R(Y)$) according to $B(\alpha_{\boldsymbol{X}}, \beta_{\boldsymbol{X}})$ (resp., $B(\alpha_Y, \beta_Y)$) and add $(\boldsymbol{x}, y)$ 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 $D(\boldsymbol{x}) \in dom_R(Y)$ drawn at random according to $B(\alpha_Y, \beta_Y)$. 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 $\boldsymbol{X} \to Y$. We then pass R through a controlled error channel such that, denoting by R' the obtained relation, R' does not satisfy $\boldsymbol{X} \to Y$ 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 $\tilde{\boldsymbol{w}}|_Y \neq \boldsymbol{w}|_Y$ and make $\tilde{\boldsymbol{w}}|_Y$ the new value for $\boldsymbol{w}|_Y$. We point out that this does not introduce any new Y-values and keeps $dom_R(Y)$ stable. We also experimented with other error channels that introduce new Y values, but the results were similar and are therefore omitted. Note that $\boldsymbol{X}$ is not modified, and therefore $p_{R'}(\boldsymbol{X}) = p_R(\boldsymbol{X})$. 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 Y based on the Beta distribution.

The first three synthetic benchmarks are created by setting $|\boldsymbol{X}| = 1$ and controlling one of the other parameters in the parameter set as follows. The last benchmark controls only $|\boldsymbol{X}|$.

Benchmark Err. Fixing $|\boldsymbol{X}| = 1$ 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. Fixing $|\boldsymbol{X}| = 1$ we iteratively increase LHS-uniqueness from $\frac{1}{5}|R|$ to $10|R|$ in 50 steps and generate 50 relations in both Uniq$^+$ and Uniq$^-$.

Benchmark Skew. Fixing $|\boldsymbol{X}| = 1$ we iteratively increase RHS-skew from 0 to 10 in 50 steps to construct Skew and generate 50 relations in both Skew$^+$ and Skew$^-$ per step.

Benchmark NonLin. We iteratively increase $|\boldsymbol{X}|$ from 2 to 5 while keeping $dom_R(X) <= \frac{1}{100}|R|$ for every LHS attribute $X \in \boldsymbol{X}$. We generate 250 relations per step in both NonLin$^+$ and NonLin$^-$, resulting in 2000 relations in total.

Note that the first three benchmarks consists of 2500 $\mathcal{B}^-$ tables and 2500 $\mathcal{B}^+$ tables while the last has 2000 table in each.

6.2 Results

We describe the results on the basis of Figure 1. Figure 1 plots on rows 1, 2 and 4 for each benchmark $\mathcal{B}$ and measure f the difference $\delta(f, \mathcal{B})$ between average measure values on $\mathcal{B}^+$ and average measure values on $\mathcal{B}^-$,

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

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 Y are sampled independently at random and where data is generated according to our generation process for $\mathcal{B}^+$. On rows 3 and 5, Figure 1 shows a more detailed view of the results on Uniq and Skew, namely separate plots of the average measure values on $\mathcal{B}^+$ (in solid lines) and $\mathcal{B}^-$ (in dashed lines). In both figures, values for g_1 and g_1' are grouped together as their measure values are indistinguishable from each other. Our conclusions are summarized in Table 9.

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 second 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 9 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

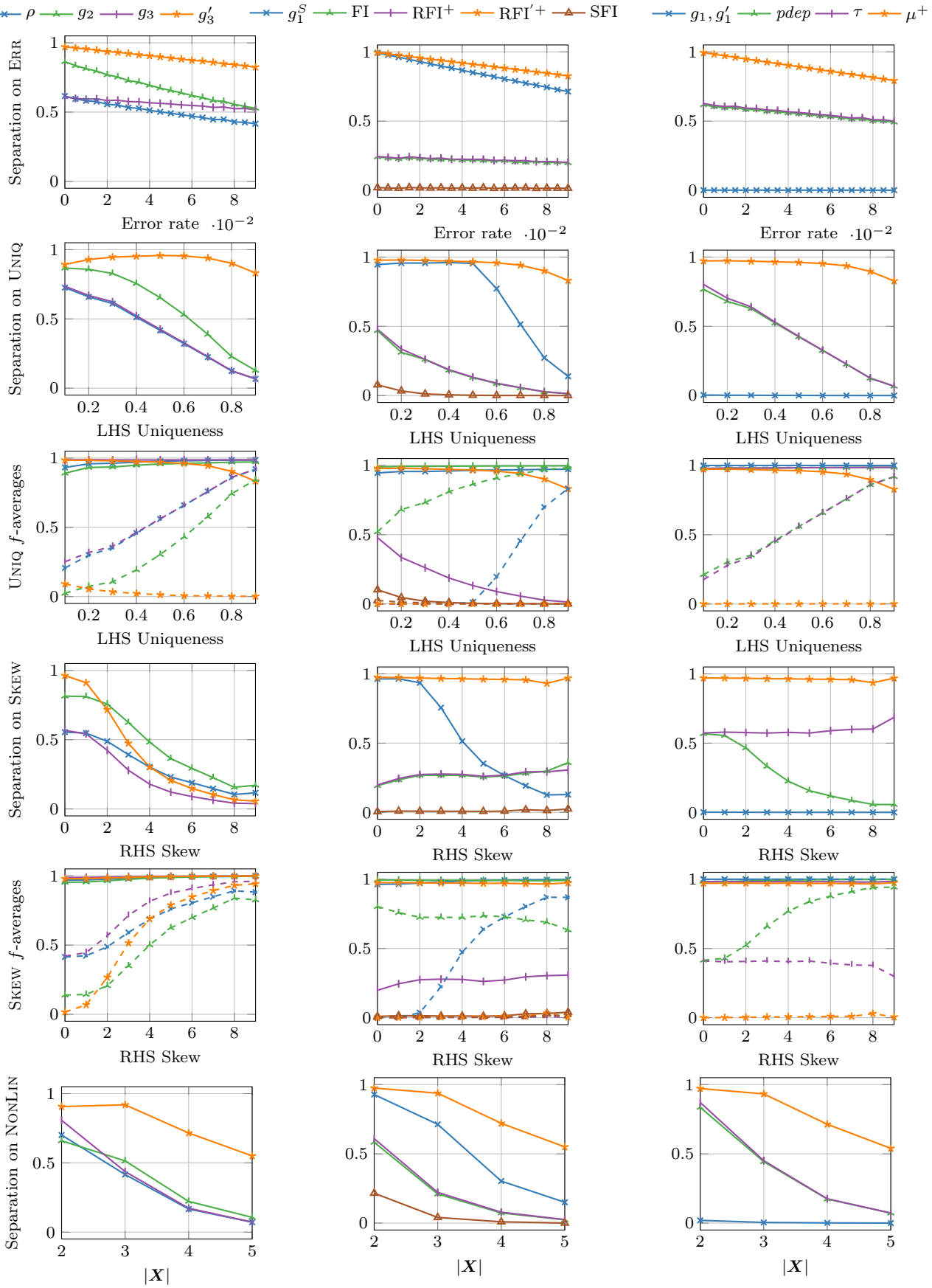

Fig. 1: Increasing error rates, LHS-uniqueness levels or RHS-skew levels impacts most measures' ability to seperate between $\mathcal{B}^+$ and $\mathcal{B}^-$. The plots show the separation on Err (row 1), Uniq (row 2) and Skew (row 4). Row 3 shows the average measure values of Uniq$^+$ (solid) and Uniq$^-$ (dashed), idem for Skew$^+$ and Skew$^-$ in Row 5. Row 6 shows the separation regarding #LHS.

decreases slightly for very large LHS-uniqueness levels, indicating that these measures become less confident to have found an FD $\boldsymbol{X} \to Y$ in a relation R when $\pi_{\boldsymbol{X}}(R)$ contains fewer duplicates. For other measures the separation drops as LHS-uniqueness increases, tending to zero at maximum LHS-uniqueness levels.

The third row of Figure 1 shows that 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.

We conclude that ρ, g_2, g_1^S, g_3, FI, *pdep*, and τ are biased w.r.t. LHS-uniqueness. As discussed later, it will therefore prove problematic to discover non-linear AFDs by means of these measures.

RHS-skew. The fourth 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 9 with the symbol _. The distinguishing power of all VIOLATION measures, as well as g_1^S, and *pdep*, drops when RHS-skew increases.

The fifth row of Figure 1 confirms these observations, all VIOLATION measures, as well as g_1^S, and *pdep* exhibit a drop in distinguishing power as RHS-skew increases. 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$^+$.

Thus, these measures are biased w.r.t. RHS-skew: their score for $\boldsymbol{X} \to Y$ increases solely on the basis of Y and independent of $\boldsymbol{X}$ even if relations are generated by a process that sampled $\boldsymbol{X}$ and Y independently at random. By contrast, FI, RFI$^+$, RFI$'^+$, τ, and μ^+ correct for this behavior and are insensitive to RHS-skew.

#LHS. We first observe that increasing $|\boldsymbol{X}|$ naturally implies that the possible domain of $\boldsymbol{X}$-tuples increases, exponentially in $|\boldsymbol{X}|$: if $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ and each X_i has a domain with N possible values, $\boldsymbol{X}$ has a domain with N^n possible values. Consequently, increasing $|\boldsymbol{X}|$ results in higher LHS-uniqueness in our experiments: on average it grew from 0.006 ($|\boldsymbol{X}| = 1$) to 0.91 ($|\boldsymbol{X}| = 5$) while the LHS-uniqueness of each individual LHS column remained constant. Consequently, as seen in the sixth row of Figure 1, separation follows similar trends as seen in our LHS-uniqueness experiments. We attribute the stronger decline in separation of all measures to higher LHS-uniqueness observed with higher $|\boldsymbol{X}|$. Insensitivity towards high LHS-uniqueness is thus an important property for AFD-measures, in particular when handling non-linear AFDs.

6.3 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.

7 Evaluation on Real-World Data

In this section, we compare the effectiveness of the described AFD measures for discovering linear AFDs in real-world tables which exhibit data distributions as well as data errors that occur in practice.

7.1 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 $\varphi = X \to Y$ a linear and unitary 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 $\boldsymbol{w}(A) \neq$ NULL for all $A \in XY$. Similarly, the score of measure f on (φ, R) is computed by computing $f(\varphi, R')$ instead.

*Real world data benchmark (*RWD*).* In the following, we elaborate how we created the RWD benchmark. We started by considering all relations mentioned in [5], 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 [11, 26, 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 3. For each relation R, we enumerate all candidate linear and unitary 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 constraint ourselves to linear FDs to keep the manual validation of all FD candidates feasible. We manually validate whether a candidate FD is semantically meaningful, and is hence part of the design schema or not, if its g_3-score is ≥ 0.5. While we risk missing semantically meaningful FDs this way, note that a g_3-score < 0.5 means that we need to remove more than 50% of the tuples to obtain a subrelation that satisfies the candidate FD, making it an improbable candidate for the design schema. We observe that each validated semantically meaningful candidate FD has a g_3-score ≥ 0.99. In other words, increasing the g_3-threshold to 0.99 would result in the same results as we present in this paper, strengthening our impression that it is unlikely that we have missed semantically meaningful FDs. We identified 1170 candidate FDs to inspect. 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.

In this manner, we derive for each benchmark relation R its design schema $\Delta(R)$. This set of FDs is partitioned into two sets:

$$PFD(R) := \{\varphi \in \Delta(R) \mid R \models \varphi\}, \text{ and}$$
$$AFD(R) := \{\varphi \in \Delta(R) \mid R \not\models \varphi\}.$$

We will refer to these sets as the *perfect* (design) FDs and *approximate* (design) FDs, respectively. In particular, $AFD(R)$ forms the ground truth of FDs to discover during AFD discovery on R.

Table 3 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.

Example 1 (Running Example) The following are three attribute pairs, the first two found in R_3 and the last one found in R_6. R_3 contains bibliographical information about computer science journals and proceedings while R_6 describes data about persons that have collected biosamples.

φ_{E1} := p2booktitlefull → p2booktitle,
φ_{E2} := p2title → p2type,
φ_{E3} := PersonFullName → _datasetguid

In φ_{E1}, *p2booktitlefull* refers to the full title of the book a publication is part of (e.g. the conference proceedings book title, or the journal title) while *p2booktitle* is its abbreviation. For example, "International Conference on Database Theory" is abbreviated as "ICDT". Hence, we expect every full book title to have only one abbreviation and define $\varphi_{E1} \in AFD(R_3)$.

Further, in φ_{E2}, *p2title* is the title of a publication, and *p2type* is the publication type, such as book, journal article, or proceedings. Two publications of different types may share the same title, e.g. when an article is first published in a conference and later as an extended version in a journal. Hence, we define $\varphi_{E2} \notin AFD(R_3)$.

Finally, in φ_{E3}, *PersonFullName* is a person's full name, and *_datasetguid* is a globally unique identifier for the dataset in which the person's collected biosample(s) appear. Since in principle these samples can appear in multiple datasets, $\varphi_{E3} \notin AFD(R_6)$.

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 3 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 $\boldsymbol{W}$ 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[3] 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.

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

Table 3: 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

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 FDs 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 [36], 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 [33]. Given a candidate FD, computing the measure value is straightforward for most measures, 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 [28], for SFI the one of [36].

Measure Runtimes. We observe significant differences in the time required to compute a score for each FD candidate. Table 4 shows the runtimes of each measure. 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 achieve runtimes comparable to the measures from the LOGICAL class, RFI^+, RFI'^+ and SFI were not able to calculate values for all candidates within 24 hours. Specifically, SFI is able to calculate a value for 1430 (roughly 90%) candidates within 24 hours, while RFI^+ and RFI'^+ finish only 250 (roughly 15%) candidates. Inspecting the rightmost column of Table 4 we observe that all measures except SFI, RFI^+ and RFI'^+ run within a tenth of second on average per candidate. SFI averages at a minute per candidate, RFI^+ and RFI'^+ take more than five minutes. Such runtimes are prohibitory in many real-world scenarios.

Table 4: 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	s/candidate
ρ	110s	1634	0.067
g_2	130s	1634	0.080
g_3	118s	1634	0.072
g_3'	128s	1634	0.078
g_1^S	137s	1634	0.084
FI	154s	1634	0.094
RFI^+	24h	250	345.600
RFI'^+	24h	250	345.600
SFI	24h	1430	60.420
g_1	134s	1634	0.082
g_1'	135s	1634	0.083
pdep	135s	1634	0.083
τ	151s	1634	0.092
μ^+	157s	1634	0.096

Defining RWD^-. 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 FDs, including all design FDs, 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 our comparison metrics (AUC, $r@mr$, ...) relative to RWD^-, first. Afterwards,

we compare all AFD measures except RFI^+ and RFI'^+ on RWD to gain intuition on how the two excluded measures could perform on all of RWD.

7.2 Results of RWD^-

AUC. Figure 2 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.

At the benchmark level, we observe that there are effective measures in each measure class. Overall, 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. $pdep(Y)$ (for τ).

The AUC scores at the relation level give a more detailed picture. In particular, the last column in Figure 2 shows 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. μ^+ performs worse than RFI'^+ only on relation R_7, where its AUC score still outperforms 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 the highest AUC score on 90% of the relations, like μ^+. It does particularly poor on relation R_3 (dblp, AUC=5.4%), 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, explaning its lower AUC score at the benchmark level. We note that our observation from Section 6 on synthetic data, namely that g_1, g'_1 and SFI have poor distinguishing power, holds 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.

Table 5: AFD-measure values of φ_{E1} and φ_{E2}.

measure f	$f(\varphi_{E1})$	$f(\varphi_{E2})$	$f(\varphi_{E1}) - f(\varphi_{E2})$
ρ	0.9951	0.9994	-0.0044
g_2	0.9830	0.9986	-0.0156
g_3	0.9983	0.9994	-0.0011
g'_3	0.9979	0.9954	0.0024
g_1^S	0.9979	0.9987	-0.0008
FI	0.9992	0.9987	0.0005
RFI^+	0.3446	0.1245	0.2201
RFI'^+	0.9975	0.9893	0.0082
SFI	0.0050	0.0793	-0.0743
g_1	1.0000	1.0000	-0.0000
g'_1	1.0000	1.0000	-0.0000
pdep	0.9971	0.9993	-0.0022
τ	0.9971	0.9986	-0.0015
μ^+	0.9964	0.9893	0.0071

Example 2 (Running Example) Table 5 shows measure values for φ_{E1} and φ_{E2} introduced in Example 1. The last column in shows the difference in measure values. We observe that most AFD measures obtain values > 0.98, except for RFI^+ and SFI which report much lower values. We desire $f(\varphi_{E1}) - f(\varphi_{E2}) > 0$ as this means that an AFD measure ranks the semantically meaningful φ_{E1} before the non-meaningful φ_{E2}. g'_3, FI, RFI^+, RFI'^+ and μ^+ perform well in this example.

Rank at max recall. In Figure 3a we show the *r@mr*. The first row indicates the total number of design AFDs to discover (the smallest attainable *r@mr* value), the last column sums all candidates that need to be inspected to retrieve all of $AFD(R)$. We observe that the best measures, g'_3, RFI'^+ and μ^+ have optimal *r@mr*, save on relation R_7 where they differ 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 retain a precision of 100%, save on relation R_7 (66%), and g'_3 also on relation R_6 (66%). As summarized in the column RWD^-, these measures hence require us to inspect only a small number of highly ranked AFDs 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, yielding low precision at maximum recall.

LHS-uniqueness and RHS-skew. From Figures 2 and 3a 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"

	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

Fig. 2: Heatmap of PR-AUC scores expressed as a precentage (red indicates low value). The last column expresses the fraction of relations on which a measure yields a highest AUC score.

	R_2	R_3	R_4	R_6	R_7	R_8	R_9	RWD$^-$
$AFD(R)$	2	2	7	2	2	1	1	17
ρ	45	25	7	2	3	1	1	84
g_2	70	25	7	2	3	1	1	109
g_3	2	8	7	3	3	1	1	25
g'_3	2	2	7	3	3	1	1	19
g_1^S	2	39	7	6	3	5	1	63
FI	2	21	7	2	3	1	1	37
RFI$^+$	2	8	32	11	3	2	1	59
RFI$'^+$	2	2	7	2	3	1	1	18
SFI	2	283	66	21	4	8	1	385
g_1	9	70	7	7	3	1	1	98
g'_1	16	70	7	7	3	1	1	105
pdep	2	18	7	3	3	1	1	35
τ	2	18	7	2	3	1	1	34
μ^+	2	2	7	2	3	1	1	18

(a)

	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	

(b)

Fig. 3: Measure performance on RWD$^-$. (a) Rank at max recall (heatmap per column where red indicates high rank), indicating how many FD candidates need to be examined to retrieve $AFD(R)$. (b) LHS-uniqueness and RHS-skew of mislabeled FD candidates, RFI$'^+$ and μ^+did not mislabel any candidate on R_3 and R_6 and are hence hidden. The last two lines show LHS-uniqueness and RHS-skew averages of all $AFD(R)$ and non-$AFD(R)$.

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 3b 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 2 and 3a we recall that measures g'_3, RFI$'^+$, and μ^+ attain optimal AUC and *r@mr*, while the AUC scores of all other measures are extremely low and their *r@mr* is very high. In Figure 3b, we observe that these other measures have much higher LHS-uniqueness values for mislabeled candidate FDs than the average for design AFDs (0.07) or non-FDs (0.2). We postulate that this makes R_3 challenging for these 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 6 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 Figure 3b we observe high RHS-skew values (>3.7) for all other measures, compared to the values of design FDs and non-FDs (resp. 1.4 and 1.6). We postulate this is what makes R_6 challenging for g_3, g_3', g_1^S, and *pdep*: recall from Section 6 that these measures are sensitive to RHS-skew. In contrast, we know from Section 6 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 the fourth row of Figure 1.

Example 3 (Running Example) Consider the non-AFDs introduced in Example 1 in comparison to the averages of LHS-uniqueness and RHS-skew shown in Table 3b. We observe a relatively high LHS-uniqueness of 0.869 and slightly above average RHS-skew of 2.076 for φ_{E2}. φ_{E3} shows an about average LHS-uniqueness of 0.041 and a relatively high RHS-skew of 4.419. φ_{E2} and φ_{E3} represent common false positive AFDs. Most measures ranked φ_{E2} and φ_{E3} higher than the elements of $AFD(R_3)$ resp. $AFD(R_6)$ in RWD^- (see Figure 2). This illustrates that insensitivity to LHS-uniqueness and RHS-skew is desireable for distinguishing semantically meaningful FDs from independent attribute pairs.

Conclusion. RFI'^+, μ^+ and g_3' provide the best trade-off 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). All three obtain a low *r@mr* values, misranking only a few column combinations higher than true $AFDs$. Further, we observe that high values for LHS-uniqueness and RHS-skew do occur 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.

7.3 Results of RWD

To allow a fair comparison among all measures, we have excluded from our search space for the benchmark RWD^- all candidate AFDs for which we could not calculate the RFI^+ and RFI'^+ values within 168 hours. By excluding these measures, our discussion of the results for the other in the previous section could potentially be biased. In this subsection we therefore investigate this potential bias by analyzing the measure performance on the whole search space comprised by RWD, but disregarding RFI^+ and RFI'^+.

Table 6: Descriptive statistics of the measure values of g_3', FI and μ^+ as well as the tuple count, LHS-uniqueness and RHS-skew of $\text{RWD} \setminus \text{RWD}^-$.

	min	mean	max
g_3'	0.000	0.142	0.982
FI	0.000	0.343	1.000
μ^+	0.000	0.089	0.981
tuples	10000	377989	1000000
LHS-uniqueness	0.000	0.095	0.981
RHS-skew	0.000	0.967	8.639

Figure 4 shows the AUC scores of all measures except RFI^+ and RFI'^+ on RWD. When we compare these scores to the scores for RWD^- (Figure 2) we observe only minor differences.

SFI exhibits the largest differences (-0.036), followed by g_1 and g_1' (-0.026 and -0.027 respectively) and FI (-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 on the larger benchmark.

RWD *without* RWD^-. To get more insight into the candidate FDs that could not be calculated by RFI^+ and RFI'^+, we investigate this exact subset: $\text{RWD} \setminus \text{RWD}^-$. This subset contains 405 candidates that are exclusively non-AFDs. For this reason, we cannot meaningfully assess precision, recall and associated metrics on $\text{RWD}\setminus\text{RWD}^-$. Instead, in Table 6 we compare descriptive statistics of AFD measure values of g_3', FI and μ^+. FI represents the closest comparison to RFI^+ and RFI'^+ from SHANNON while g_3' and μ^+ represent the closest AFD measures in terms of PR-AUC. Further, we present descriptive statistics of the tuple counts, LHS -uniqueness and RHS-skew.

In the upper part of Table 6 we observe that the average of most measure values are low. Still, judging by very high max values of FI there seems to be a small subset of candidate FDs that could lead to changes the AUCs of RFI^+ and RFI'^+. The only way to confirm that suspicion would be to use RFI^+ and RFI'^+ to calculate values for all candidates of $\text{RWD}\setminus\text{RWD}^-$. The lower part of Table 6 shows that LHS-uniqueness and RHS-skew are within expected ranges. The tuple count, however, indicates that, as expected, the runtimes of RFI^+ and RFI'^+ is positively correlated to the relation size.

	RWD	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	best
ρ	41.1	100	02.9	03.8	100	100	100	66.7	100	100	100	70
g_2	49.7	100	50.7	03.8	100	100	100	66.7	100	100	100	70
g_3	66.9	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	10.8	100	100	02.7	100	100	63.3	66.7	10.0	100	100	60
FI	39.6	100	100	04.9	100	100	100	91.7	100	100	100	90
SFI	28.4	100	100	00.3	29.6	100	05.0	58.3	05.6	100	100	50
g_1	39.9	100	58.7	01.4	100	100	59.8	66.7	100	100	100	60
g_1'	39.8	100	54.8	01.4	100	100	59.8	66.7	100	100	100	60
pdep	64.2	100	100	07.6	100	100	79.2	66.7	100	100	100	70
τ	62.3	100	100	08.2	100	100	100	66.7	100	100	100	80
μ^+	94.6	100	100	100	100	100	100	66.7	100	100	100	90

Fig. 4: The differences between RWD$^-$ and RWD without RFI$^+$ and RFI$'^+$ are minimal. The table shows the same overview as Figure 2 but now for RWD.

Conclusion. The comparison of RWD and RWD$^-$ show very little differences in AUC values, indicating that our findings on RWD$^-$ carry over to RWD. We strengthen this conclusion by observing that the structural properties of RWD $\setminus$ RWD$^-$ are within expected ranges. The relatively high number of tuples indicates that the runtime of RFI$^+$ and RFI$'^+$ increases with the number of tuples.

7.4 Results of RWD^e

In this section, we introduce controlled errors into FDs present in RWD via an error channel to simulate practical scenarios characterized by data inconsistencies and noise. This approach enables us to assess how AFD measures perform under real-world conditions, providing deeper insights into their reliability and practical utility when applied to imperfect datasets. We obtain RWD^e by passing each relation $R \in$ 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., $AFD(R) \subseteq AFD(R')$.

Inspired by Arocena et al. [2], we study the measures' sensitivity to different error channels. We parameterize each error channel by an error level $\eta \in [0, 1]$ and an error type. When passing R through the channel we consider all $X \to Y \in PFD(R)$ and modify $k = \lfloor \eta |R| \rfloor$ Y-values. To avoid interference, we select at most one FD $X \to Y$ for every unique Y per relation, ensuring that Y does not appear in $AFD(R)$, and that no FD $Y \to Z$ has previously been selected. We note that picking a single FD for passing through an error channel may produce multiple AFDs, namely every FD that share the same Y. Over all datasets, we observed 30 Y attributes that were part of multiple AFDs. Hence, we do not expect to introduce any bias towards more frequent FDs with our approach.

The procedure to modify the Y values is determined by the chosen type of data error for which we consider three categories: copy error, typo and bogus value. For a chosen tuple $\boldsymbol{w} \in R$, only $\boldsymbol{w}|_Y$ is changed, where the change depends on the data error type proposed by [2]:

(i) copy: Randomly pick any $\tilde{\boldsymbol{w}} \in R$ with $\tilde{\boldsymbol{w}}|_Y \neq \boldsymbol{w}|_Y$ and make $\tilde{\boldsymbol{w}}|_Y$ the new value for $\boldsymbol{w}|_Y$.

(ii) typo: To every $y \in dom_R(Y)$, we associate three new values representing three common typos. We choose one from these three each time at random as the new value for $\boldsymbol{w}|_Y$.

(iii) bogus: $\boldsymbol{w}|_y$ is assigned a unique newly generated value.

We point out that copy does not introduce any new values and keeps $dom_R(Y)$ stable, while typo (resp., bogus) introduces a number of new values independent of (resp., dependent on) the error level. X is not modified, and therefore $p_{R'}(X) = p_R(X)$. To ensure that increasing error levels do not accidentally reduce errors, we ensure that, for each $x \colon X$ we pick at most $\lfloor N_x/2 \rfloor$ tuples $\boldsymbol{w}$ with $\boldsymbol{w}|_X = x$ to modify, where N_X is the number of times that x occurs in $\pi_X(R)$. *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 73 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. We denote by n the total number of AFD candidates retained in RWD$^e[t, \eta]$. The header row in Table 7 summarizes n as

well as the total number of AFDs for each RWD$^{e}[t, \eta]$. We observe that, with a single exception for copy at 5%, increasing the error level decreases the number of additional AFDs. We further observe that the ranking copy > typo > bogus holds with respect to the number of additional AFDs.

AUC. Table 7 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$'^{+}$ show very small differences, making it difficult to draw general conclusions. Regarding the error types, we observe that in general the AUCs of copy are higher on the same error levels with exceptions for g_3, g_1^S and FI on RWD^{e}[bogus, 10%]. 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^{-}. We do not analyse this anomaly further, as g_1^S remains one of the worst-scoring AFD-measures.

In general, 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%. It is evident from Table 4 that AFD-measures are not very effective when error levels are greater than 5% for copy, for bogus and typo even from 2%. This is surprising as in the first row of Figure 1, we see that separation on ERR looks acceptable at least for g_3', RFI$'^{+}$ and μ^{+}. We hypothesize that increasing the error rate in combination with the other structural properties investigated in Section 6 degrade the effectiveness of all AFD-measures. Further investigation into this hypothesis is left for future research.

Rank at max recall. We show in Table 8 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 (relation, t, η) 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%.

Conclusion. We strengthen our findings that both μ^{+} and RFI$'^{+}$ perform well by applying error-inducing approaches to RWD to create RWD^{e}. We observed that the error types influence measure effectiveness and error rates $\geq 5\%$ are too high for all AFD-measures.

8 Related Work

Relaxing FDs. In the literature there are two distinct ways to relax the notion of an FD [9]: 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 and point the interested reader to [9] for the latter.

Correlation. When an FD holds in a relation, there is also 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 [22], 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 [37] 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 [6]. There, only a subset of all FDs that are satisfied are computed in return for performance improvements.

Existing comparisons of AFD measures. Giannella and Robertson [19] compare a limited number of measures on theoretical examples as well as on four 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 as done here. In their survey concerning FD relaxations, Caruccio et al. [9] also survey some of the AFD measures that are considered here, but do not provide a qualitative comparison.

Table 7: μ^+ and RFI$'^+$ score best on RWD^e with one exception. The table shows the AUCs for each noise type and 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$^-$. Per column, we typeset the best AUC bold and underlined and the second best AUC bold.

	RWD$^-$	copy, 1	copy, 2	copy, 5	copy, 10	bogus, 1	bogus, 2	bogus, 5	bogus, 10	typo, 1	typo, 2	typo, 5	typo, 10
n	1229	1204	1206	1218	1143	1130	1127	1111	992	1188	1194	1189	915
$AFD(R)$	17	73	72	76	59	58	58	58	28	68	68	67	43
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 8: 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**

Discovery of Conditional FDs. Conditional FDs (CFDs for short) generalize FDs: they are FDs that only hold on a subset of the data [7]. CFDs are widely used in data cleaning [16, 17]. The discovery of (approximate) CFDs amounts to (i) selecting suitable subsets of the data and (ii) discovering (A)FDs in these subsets [39]. In particular, Geerts and Rammelaere [39] propose a generic (A)CFD discovery algorithm in which any (A)FD discovery algorithm can be plugged. Insights into AFD measures, as is our focus here, can improve AFD discovery which in turn can be useful for the discovery of (A)CFDs.

Applications of AFDs. Data management tasks may use AFDs to different extents. For example, only semantically meaningful AFDs (i.e., FDs that a human database administrator would include in a database schema design) are useful for data cleaning tasks [26] while for query optimisation one could try to exploit all AFDs present in a relation, even if they are not semantically meaningful. Our real-world benchmark focuses on the former kind of AFDs. Thus, we present a study of AFD measures applicable to discover semantically meaningful AFDs.

9 Conclusions

An overview of our results is given in Table 9. 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 [3, 4, 18, 19, 21, 23, 24] but to the best of our knowledge only [19] considers the correctly normalized version g'_3. The sensitivity of g'_3 to RHS-skew (Figure 1) remains a structural weakness, hampering its effectiveness as illustrated by its lower AUC in practice on R_6 (Figure 2).

FI is the defining measure of SHANNON and suffers from sensitivity to LHS-uniqueness as illustrated by its behavior on R_3 Figure 3). Its corrections SFI [36] and RFI$^+$ [28, 29] were aimed at removing bias from FI, but our experiments on SYN reveal that their distinguish-

Table 9: Properties of considered AFD measures. The symbol ✓ stands for *applies*, the symbol ✗ denotes *does not apply*. The symbol _ stands for *not applicable* (cf. Section 6). Further, f' refers to the normalization of the measure f as discussed in Section 3 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	[22]	[24, 44]	[3, 4, 21] [23, 24, 43]	[19]	new	[10, 19] [28, 29, 36]	[28, 29]	new	[36]	[24, 47]	[26]	[37]	[20, 37]	[37]
Class (V/S/L)	V	V	V	V	S	S	S	S	S	L	L	L	L	L
Has baselines	✗	✓	✗	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓
Is normalized	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✓	✗	✓	✓
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

ing power is insufficient; they especially 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 (Table 4).

Our recommendation for linear AFD discovery is therefore the little-known measure μ^+. It has comparable AUC-performance to RFI'^+ as well as equal structural sensitivity properties, but can be efficiently computed. Because of the linear-indistinguishability property (Section 5) we expect μ^+ to be the most promising measures for discovering non-linear AFDs in real-world data as well. An experimental validation of this hypothesis is left for future research. In particular, we acknowledge that there may be additional properties exhibited for non-linear AFDs in real-world data, not considered here, that necessitate the design of new measures for effective AFD discovery. This is an exciting topic for future research.

On RWD^e, where we introduced errors into FDs to create AFDs, we observed that our findings generally hold for error rates below 5%. However, with error rates of 5% or higher, the performance of all AFD-measures declines notably, rendering them ineffective (Table 7). We leave a more extensive investigation including other, more realistic error types, for future research.

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 the linear 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. Whether this holds for non-linear FDs is left for future research.

Acknowledgements 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.

A Proof of Theorem 2

In this section, we present the proof of Theorem 2 as a sequence of lemmas. We assume that $R \not\models \boldsymbol{X} \to \boldsymbol{Y}$ for lemmas 2–6.

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.

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

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 $p_R(\boldsymbol{xy})$. Thereby, effectively removing all other tuples $\boldsymbol{xy'}$ with $\boldsymbol{y} \neq \boldsymbol{y'}$. The

last equality then follows from the definition of conditional probability. □

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

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

Proof We first observe

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

Hence,

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

□

Lemma 4

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

Proof We reason as follows.

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

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

Lemma 5 *Relating this to logical entropy we observe the following equality.*

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

Proof We reason as follows.

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

□

Lemma 6

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

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

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

We now make a case analysis.

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

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

  as desired.
- If $|\, dom_R(\boldsymbol{Y})| > 1$ then

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

□

References

1. Abedjan, Z., Golab, L., Naumann, F.: Profiling relational data: a survey. VLDB J. **24**(4), 557–581 (2015). DOI 10.1007/s00778-015-0389-y. URL `https://doi.org/10.1007/s00778-015-0389-y`
2. Arocena, P.C., Glavic, B., Mecca, G., Miller, R.J., Papotti, P., Santoro, D.: Messing up with BART: error generation for evaluating data-cleaning algorithms. Proc. VLDB Endow. **9**(2), 36–47 (2015). DOI 10.14778/2850578.2850579. URL `http://www.vldb.org/pvldb/vol9/p36-arocena.pdf`
3. Berti-Équille, L., Harmouch, H., Naumann, F., Novelli, N., Thirumuruganathan, S.: Discovery of genuine functional dependencies from relational data with missing values. Proc. VLDB Endow. **11**(8), 880–892 (2018). DOI 10.14778/3204028.3204032. URL `http://www.vldb.org/pvldb/vol11/p880-berti-equille.pdf`
4. Berzal, F., Cubero, J., Cuenca, F., Medina, J.: Relational decomposition through partial functional dependencies. Data and Knowledge Engineering **43**(2), 207–234 (2002). DOI https://doi.org/10.1016/S0169-023X(02)00056-3. URL `https://www.sciencedirect.com/science/article/pii/S0169023X02000563`
5. Birnick, J., Bläsius, T., Friedrich, T., Naumann, F., Papenbrock, T., Schirneck, M.: Hitting set enumeration with partial information for unique column combination discovery. Proc. VLDB Endow. **13**(11), 2270–2283 (2020)
6. Bleifuß, T., Bülow, S., Frohnhofen, J., Risch, J., Wiese, G., Kruse, S., Papenbrock, T., Naumann, F.: Approximate discovery of functional dependencies for large datasets. In: S. Mukhopadhyay, C. Zhai, E. Bertino, F. Crestani, J. Mostafa, J. Tang, L. Si, X. Zhou, Y. Chang, Y. Li, P. Sondhi (eds.) Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016, pp. 1803–1812. ACM (2016). DOI 10.1145/2983323.2983781. URL `https://doi.org/10.1145/2983323.2983781`

7. Bohannon, P., Fan, W., Geerts, F., Jia, X., Kementsietsidis, A.: Conditional functional dependencies for data cleaning. In: R. Chirkova, A. Dogac, M.T. Özsu, T.K. Sellis (eds.) Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007, pp. 746–755. IEEE Computer Society (2007). DOI 10.1109/ICDE.2007.367920. URL https://doi.org/10.1109/ICDE.2007.367920
8. Calders, T., Ng, R.T., Wijsen, J.: Searching for dependencies at multiple abstraction levels. ACM Trans. Database Syst. **27**(3), 229–260 (2002). DOI 10.1145/581751.581752. URL https://doi.org/10.1145/581751.581752
9. Caruccio, L., Deufemia, V., Polese, G.: Relaxed functional dependencies - A survey of approaches. IEEE Trans. Knowl. Data Eng. **28**(1), 147–165 (2016). DOI 10.1109/TKDE.2015.2472010. URL https://doi.org/10.1109/TKDE.2015.2472010
10. Cavallo, R., Pittarelli, M.: The theory of probabilistic databases. In: P.M. Stocker, W. Kent, P. Hammersley (eds.) VLDB'87, Proceedings of 13th International Conference on Very Large Data Bases, September 1-4, 1987, Brighton, England, pp. 71–81. Morgan Kaufmann (1987)
11. Chiang, F., Miller, R.J.: Discovering data quality rules. Proc. VLDB Endow. **1**(1), 1166–1177 (2008). DOI 10.14778/1453856.1453980. URL http://www.vldb.org/pvldb/vol1/1453980.pdf
12. Chu, X., Ilyas, I.F., Papotti, P.: Holistic data cleaning: Putting violations into context. In: C.S. Jensen, C.M. Jermaine, X. Zhou (eds.) 29th IEEE International Conference on Data Engineering, ICDE 2013, Brisbane, Australia, April 8-12, 2013, pp. 458–469. IEEE Computer Society (2013). DOI 10.1109/ICDE.2013.6544847. URL https://doi.org/10.1109/ICDE.2013.6544847
13. Cormode, G., Golab, L., Korn, F., McGregor, A., Srivastava, D., Zhang, X.: Estimating the confidence of conditional functional dependencies. In: U. Çetintemel, S.B. Zdonik, D. Kossmann, N. Tatbul (eds.) Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2009, Providence, Rhode Island, USA, June 29 - July 2, 2009, pp. 469–482. ACM (2009). DOI 10.1145/1559845.1559895. URL https://doi.org/10.1145/1559845.1559895
14. Edwards, S.: Thomas m. cover and joy a. thomas, elements of information theory (2nd ed.), john wiley & sons, inc. (2006). Inf. Process. Manag. **44**(1), 400–401 (2008). DOI 10.1016/j.ipm.2007.02.009. URL https://doi.org/10.1016/j.ipm.2007.02.009
15. Ellerman, D.: New Foundations for Information Theory - Logical Entropy and Shannon Entropy. SpringerBriefs in Philosophy. Springer (2021). DOI 10.1007/978-3-030-86552-8
16. Fan, W., Geerts, F., Jia, X.: A revival of integrity constraints for data cleaning. Proc. VLDB Endow. **1**(2), 1522–1523 (2008). DOI 10.14778/1454159.1454220. URL http://www.vldb.org/pvldb/vol1/1454220.pdf
17. Fan, W., Geerts, F., Jia, X., Kementsietsidis, A.: Conditional functional dependencies for capturing data inconsistencies. ACM Trans. Database Syst. **33**(2), 6:1–6:48 (2008). DOI 10.1145/1366102.1366103. URL https://doi.org/10.1145/1366102.1366103
18. Faure-Giovagnoli, P., Petit, J., Scuturici, V.: 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, pp. 607–620. IEEE (2022). DOI 10.1109/ICDE53745.2022.00050. URL https://doi.org/10.1109/ICDE53745.2022.00050
19. Giannella, C., Robertson, E.L.: On approximation measures for functional dependencies. Inf. Syst. **29**(6), 483–507 (2004). DOI 10.1016/j.is.2003.10.006. URL https://doi.org/10.1016/j.is.2003.10.006
20. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications. Journal of the American Statistical Association **49**(268), 732–764 (1954). URL http://www.jstor.org/stable/2281536
21. Huhtala, Y., Kärkkäinen, J., Porkka, P., Toivonen, H.: TANE: an efficient algorithm for discovering functional and approximate dependencies. Comput. J. **42**(2), 100–111 (1999). DOI 10.1093/comjnl/42.2.100. URL https://doi.org/10.1093/comjnl/42.2.100
22. Ilyas, I.F., Markl, V., Haas, P.J., Brown, P., Aboulnaga, A.: CORDS: automatic discovery of correlations and soft functional dependencies. In: G. Weikum, A.C. König, S. Deßloch (eds.) Proceedings of the ACM SIGMOD International Conference on Management of Data, Paris, France, June 13-18, 2004, pp. 647–658. ACM (2004). DOI 10.1145/1007568.1007641. URL https://doi.org/10.1145/1007568.1007641
23. King, R.S., Legendre, J.J.: Discovery of functional and approximate functional dependencies in relational databases. Adv. Decis. Sci. **7**(1), 49–59 (2003). DOI 10.1155/S117391260300004X. URL https://doi.org/10.1155/S117391260300004X
24. Kivinen, J., Mannila, H.: Approximate inference of functional dependencies from relations. Theor. Comput. Sci. **149**(1), 129–149 (1995). DOI 10.1016/0304-3975(95)00028-U. URL https://doi.org/10.1016/0304-3975(95)00028-U
25. Kossmann, J., Papenbrock, T., Naumann, F.: Data dependencies for query optimization: a survey. VLDB J. **31**(1), 1–22 (2022). DOI 10.1007/s00778-021-00676-3. URL https://doi.org/10.1007/s00778-021-00676-3
26. Kruse, S., Naumann, F.: Efficient discovery of approximate dependencies. Proc. VLDB Endow. **11**(7), 759–772 (2018). DOI 10.14778/3192965.3192968. URL http://www.vldb.org/pvldb/vol11/p759-kruse.pdf
27. Liu, H., Xiao, D., Didwania, P., Eltabakh, M.Y.: Exploiting soft and hard correlations in big data query optimization. Proc. VLDB Endow. **9**(12), 1005–1016 (2016). DOI 10.14778/2994509.2994519. URL http://www.vldb.org/pvldb/vol9/p1005-liu.pdf
28. Mandros, P., Boley, M., Vreeken, J.: 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, pp. 355–363. ACM (2017). DOI 10.1145/3097983.3098062. URL https://doi.org/10.1145/3097983.3098062
29. Mandros, P., Boley, M., Vreeken, J.: Discovering dependencies with reliable mutual information. Knowl. Inf. Syst. **62**(11), 4223–4253 (2020). DOI 10.1007/s10115-020-01494-9. URL https://doi.org/10.1007/s10115-020-01494-9
30. Marchi, F.D., Lopes, S., Petit, J.: Unary and n-ary inclusion dependency discovery in relational databases. J. Intell. Inf. Syst. **32**(1), 53–73 (2009). DOI 10.1007/s10844-007-0048-x. URL https://doi.org/10.1007/s10844-007-0048-x
31. Papenbrock, T., Ehrlich, J., Marten, J., Neubert, T., Rudolph, J., Schönberg, M., Zwiener, J., Naumann,

F.: Functional dependency discovery: An experimental evaluation of seven algorithms. Proc. VLDB Endow. **8**(10), 1082–1093 (2015). DOI 10.14778/2794367.2794377. URL http://www.vldb.org/pvldb/vol8/p1082-papenbrock.pdf
32. Papenbrock, T., Naumann, F.: A hybrid approach to functional dependency discovery. In: F. Özcan, G. Koutrika, S. Madden (eds.) Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016, pp. 821–833. ACM (2016). DOI 10.1145/2882903.2915203. URL https://doi.org/10.1145/2882903.2915203
33. Parciak, M., Weytjens, S., Hens, N., Neven, F., Peeters, L., Vansummeren, S.: Artifacts related to "approximately measuring functional dependencies: a comparitive study". Available at https://github.com/MarcelPa/AFD_comparative_study (2022)
34. Parciak, M., Weytjens, S., Hens, N., Neven, F., Peeters, L.M., Vansummeren, S.: Approximately measuring functional dependencies: a comparative study. In: Proceedings of the 40th IEEE International Conference on Data Engineering, ICDE 2024, Utrecht, Netherlands, May 13-17, 2024. IEEE Computer Society (2024)
35. Pena, E.H.M., de Almeida, E.C., Naumann, F.: Discovery of approximate (and exact) denial constraints. Proc. VLDB Endow. **13**(3), 266–278 (2019). DOI 10.14778/3368289.3368293. URL http://www.vldb.org/pvldb/vol13/p266-pena.pdf
36. Pennerath, F., Mandros, P., Vreeken, J.: Discovering approximate functional dependencies using smoothed mutual information. In: R. Gupta, Y. Liu, J. Tang, B.A. Prakash (eds.) KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020, pp. 1254–1264. ACM (2020). DOI 10.1145/3394486.3403178. URL https://doi.org/10.1145/3394486.3403178
37. Piatetsky-Shapiro, G., Matheus, C.J.: Measuring data dependencies in large databases. In: Proceedings of the 2nd International Conference on Knowledge Discovery in Databases, pp. 162—173. AAAI Press (1993)
38. Rammelaere, J., Geerts, F.: Explaining repaired data with cfds. Proc. VLDB Endow. **11**(11), 1387–1399 (2018). DOI 10.14778/3236187.3236193. URL http://www.vldb.org/pvldb/vol11/p1387-rammelaere.pdf
39. Rammelaere, J., Geerts, F.: 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, *Lecture Notes in Computer Science*, vol. 11052, pp. 552–568. Springer (2018). DOI 10.1007/978-3-030-10928-8_33. URL https://doi.org/10.1007/978-3-030-10928-8_33
40. Rekatsinas, T., Chu, X., Ilyas, I.F., Ré, C.: Holoclean: Holistic data repairs with probabilistic inference. Proc. VLDB Endow. **10**(11), 1190–1201 (2017). DOI 10.14778/3137628.3137631. URL http://www.vldb.org/pvldb/vol10/p1190-rekatsinas.pdf
41. Roulston, M.S.: Estimating the errors on measured entropy and mutual information. Physica D: Nonlinear Phenomena **125**(3), 285–294 (1999). DOI https://doi.org/10.1016/S0167-2789(98)00269-3. URL https://www.sciencedirect.com/science/article/pii/S0167278998002693
42. Schirmer, P., Papenbrock, T., Kruse, S., Naumann, F., Hempfing, D., Mayer, T., Neuschäfer-Rube, D.: Dynfd: Functional dependency discovery in dynamic datasets. In: M. Herschel, H. Galhardas, B. Reinwald, I. Fundulaki, C. Binnig, Z. Kaoudi (eds.) Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019, pp. 253–264. OpenProceedings.org (2019). DOI 10.5441/002/edbt.2019.23. URL https://doi.org/10.5441/002/edbt.2019.23
43. Wang, D.Z., Dong, X.L., Sarma, A.D., Franklin, M.J., Halevy, A.Y.: 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). URL http://webdb09.cse.buffalo.edu/papers/Paper18/webdb09.pdf
44. Wang, P., He, Y.: Uni-detect: A unified approach to automated error detection in tables. In: SIGMOD, pp. 811–828. ACM (2019). DOI 10.1145/3299869.3319855. URL https://doi.org/10.1145/3299869.3319855
45. Wang, Y., Song, S., Chen, L., Yu, J.X., Cheng, H.: Discovering conditional matching rules. ACM Trans. Knowl. Discov. Data **11**(4), 46:1–46:38 (2017). DOI 10.1145/3070647. URL https://doi.org/10.1145/3070647
46. Wei, Z., Link, S.: Discovery and ranking of functional dependencies. In: 35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019, pp. 1526–1537. IEEE (2019). DOI 10.1109/ICDE.2019.00137. URL https://doi.org/10.1109/ICDE.2019.00137
47. Zhang, Y., Guo, Z., Rekatsinas, T.: A statistical perspective on discovering functional dependencies in noisy data. In: D. Maier, R. Pottinger, A. Doan, W. Tan, A. Alawini, H.Q. Ngo (eds.) Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020, pp. 861–876. ACM (2020). DOI 10.1145/3318464.3389749. URL https://doi.org/10.1145/3318464.3389749