Quantifying the Loss of Acyclic Join Dependencies

Batya Kenig  
Technion, Israel Institute of Technology  
Haifa, Israel  
batyak@technion.ac.il

Nir Weinberger  
Technion, Israel Institute of Technology  
Haifa, Israel  
nirwein@technion.ac.il

ABSTRACT
Acyclic schemes posses known benefits for database design, speeding up queries, and reducing space requirements. An acyclic join dependency (AJD) is lossless with respect to a universal relation if joining the projections associated with the schema results in the original universal relation. An intuitive and standard measure of loss entailed by an AJD is the number of redundant tuples generated by the acyclic join. Recent work has shown that the loss of an AJD can also be characterized by an information-theoretic measure. Motivated by the problem of automatically fitting an acyclic schema to a universal relation, we investigate the connection between these two characterizations of loss. We first show that the loss of an AJD is captured using the notion of KL-Divergence. We then show that the KL-divergence can be used to bound the number of redundant tuples. We prove a deterministic lower bound on the percentage of redundant tuples. For an upper bound, we propose a random database model, and establish a high probability bound on the percentage of redundant tuples, which coincides with the lower bound for large databases.

CCS CONCEPTS
• Information systems → Relational database model; • Theory of computation → Database constraints theory.

KEYWORDS
Acyclic Schemas; Information Theory; Data Dependencies

ACM Reference Format:
Batya Kenig and Nir Weinberger. 2023. Quantifying the Loss of Acyclic Join Dependencies. In Proceedings of the 42nd ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS ’23), June 18–23, 2023, Seattle, WA, USA. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3584372.3588658

1 INTRODUCTION
In the traditional approach to database design, the designer has a clear conceptual model of the data, and of the dependencies between the attributes. This information guides the database normalization process, which leads to a database schema consisting of multiple relation schemas that have the benefit of reduced redundancies, and more efficient querying and updating of the data. This approach requires that the data precisely meet the constraints of the model; various data repair techniques [1, 4] have been developed to address the case that the data does not meet the constraints of the schema exactly. Current data management applications are required to handle data that is noisy, erroneous, and inconsistent. The presumption that such data meet a predefined set of constraints is not likely to hold. In many cases, such applications are willing to tolerate the “loss” entailed by an imperfect database schema, and will be content with a database schema that only “approximately fits” the data. Motivated by the task of schema-discovery for a given dataset, in this work, we investigate different ways of measuring the “loss” of an imperfect database schema, and the relationship between these different measures.

Decomposing a relation schema Ω is the process of breaking the relation schema into two or more relation schemes Ω₁, . . . , Ωₖ whose union is Ω. The decomposition is lossless if, for any relation instance R over Ω, it holds that R = ΠΩ₁(R) ⊲ · · · ⊲ ΠΩₖ(R). The loss of a database schema (Ω₁, . . . , Ωₖ) with respect to a relation instance R over Ω, is the set of tuples (w₁ ⊲ · · · ⊲ wₖ) \ R that are in the join, but not in R (formal definitions in Section 2). We say that such tuples are spurious. Normalizing a relation schema is the process of (losslessly) decomposing it into a database schema where each of its resulting relational schemes have certain properties. The specific properties imposed on the resulting relational schemes define different normal forms such as 3NF [6], BCNF [7], 4NF [10], and 5NF [9, 11].

A data dependency defines a relationship between sets of attributes in a database. A Join Dependency (JD) defines a k-way decomposition (where k ≥ 2) of a relation schema Ω, and is said to hold in a relation instance R over Ω if the join is lossless with respect to R (formal definitions in Section 2). Join Dependencies generalize Multivalued Dependencies (MVDs) that are effectively Join Dependencies where k = 2, which in turn generalize Functional Dependencies (FDs), which are perhaps the most widely studied data dependencies due to their simple and intuitive semantics.

Acyclic Join Dependencies (AJDs) is a type of JD that is specified by an Acyclic Schema [2]. Acyclic schemes have many applications in databases and in machine learning; they enable efficient query evaluation [23], play a key role in database normalization and design [10, 17], and improve the performance of many well-known machine learning algorithms over relational data [14, 20, 21].

Consider how we may measure the loss of an acyclic schema S = (Ω₁, . . . , Ωₖ) with respect to a given relation instance R over Ω. An intuitive approach, based on the definition of a lossless join, is to simply count the number of spurious tuples generated by the join (i.e., and are not included in the relation instance R). In previous work, Kenig et al. [12] presented an algorithm that discovers “approximate Acyclic Schemes” for a given dataset. Building on earlier work by Lee [15, 16], the authors proposed to measure the loss of an acyclic
schema, with respect to a given dataset, using an information-theoretic metric called the $J$-measure, formally defined in Section 2. Lee [15, 16] has shown that this information-theoretic measure characterizes lossless AJDs. That is, the $J$-measure of an acyclic schema with respect to a dataset is 0 if and only if the AJD defined by this schema is lossless with respect to the dataset [15, 16] (i.e., no spurious tuples). Beyond this characterization, not much is known about the relationship between these two measures of loss. In fact, the relationship is not necessarily monotonic, and may vary widely even between two acyclic schemas with similar $J$-measure values [12]. Nevertheless, empirical studies have shown that low values of the $J$-measure generally lead to acyclic schemas that incur a small number of spurious tuples [12].

Our first result is a characterization of the $J$-measure of an acyclic schema $S = \{\Omega_1, \ldots, \Omega_k\}$, with respect to a relation instance $R$, as the KL-divergence between two empirical distributions: the one associated with $R$, and the one associated with $R' \equiv \Pi_{\Omega_1}(R) \otimes \cdots \otimes \Pi_{\Omega_k}(R)$. The empirical distribution is a standard notion used to associate a multivariate probability distribution with a multiset of tuples, and a formal definition is deferred to Section 2. The KL-divergence is a non-symmetric measure of the similarity between two probability distributions $P(\Omega)$ and $Q(\Omega)$. It has numerous information-theoretic applications, and can be loosely thought of as a measure of the information lost when $Q(\Omega)$ is used to approximate $P(\Omega)$. Our result that the $J$-measure is, in fact, the KL-divergence between the empirical distributions associated with the original relation instance and the one implied by the acyclic schema, explains the success of the $J$-measure for identifying “approximate acyclic schemas” in [12], and how the $J$-measure characterizes lossless AJDs [15, 16].

With this result at hand, we address the following problem. Given a relation schema $\Omega$, an acyclic schema $S = \{\Omega_1, \ldots, \Omega_k\}$, and a value $J \geq 0$, compute the minimum and maximum number of spurious tuples generated by $S$ with respect to any relation instance $R$ over $\Omega$, whose KL-divergence from $R' \equiv \Pi_{\Omega_1}(R) \otimes \cdots \otimes \Pi_{\Omega_k}(R)$ is $J$. To this end, we prove a deterministic lower bound on the number of spurious tuples that depends only on $J$. We then consider the problem of determining an upper bound on the number of spurious tuples. As it turns out, this problem is more challenging, as a deterministic upper bound does not hold. We thus propose a random relation model, in which a relation is drawn uniformly at random from all possible empirical distributions of a given size $N$. We then show that a bound analogous to the deterministic lower bound on the relative number of spurious tuples also holds as an upper bound with two differences: First, it holds with high probability over the random choice of relation (and not with probability 1), and second, it holds with an additive term, though one which vanishes for asymptotically large relation instances. The proof of this result is fairly complicated, and as discussed in Section 5, requires applications of multiple techniques from information theory [8] and concentration of measure [5].

Beyond its theoretical interest, understanding the relationship between the information-theoretic KL-divergence, and the tangible property of loss, as measured by the number of spurious tuples, has practical consequences for the task of discovering acyclic schemas that fit a dataset. Currently, the system of [12] can discover acyclic schemas that fit the data well in terms of its $J$-measure. Understanding how the $J$-measure relates to the loss in terms of spurious tuples will enable finding acyclic schemas that generate a bounded number of spurious tuples. This is important for applications that apply factorization as a means of compression, while wishing to maintain the integrity of the data [19].

To summarize, in this paper we: (1) Show that the $J$-measure of Lee [15, 16] is the KL-divergence between the empirical distribution associated with the original relation and the one induced by the acyclic schema. (2) Prove a general lower bound on the loss (i.e., spurious tuples) in terms of the KL-divergence, and present a simple family of relation instances for which this bound is tight, and (3) Propose a random relation model and prove an upper bound on the loss, which holds with high probability, and which converges to the lower bound for large relational instances.

2 BACKGROUND

For the sake of consistency, we adopt some of the notation from [12]. We denote by $[n] = \{1, \ldots, n\}$. Let $\Omega$ be a set of variables, also called attributes. If $X,Y \subseteq \Omega$, then $XY$ denotes $X \cup Y$.

2.1 Data Dependencies

Let $\Omega = \{X_1, \ldots, X_n\}$ denote a set of attributes with domains $\mathcal{D}(X_1), \ldots, \mathcal{D}(X_n)$. We denote by $\Rel(\Omega) \equiv \{R : R \subseteq X_1 \times \cdots \times X_n, \mathcal{D}(X_1) \}$ the set of all possible relation instances over $\Omega$. Fix a relation instance $R \in \Rel(\Omega)$ of size $N = |R|$. For $Y \subseteq \Omega$ we let $R[Y]$ denote the projection of $R$ onto the attributes $Y$. A schema is a set $S = \{\Omega_1, \ldots, \Omega_m\}$ such that $\bigcup_{i=1}^m \Omega_i = \Omega$ and $\Omega_i \not\subseteq \Omega_j$ for $i \neq j$.

We say that the relation instance $R$ satisfies the join dependency $JD(S)$, and write $R \models JD(S)$, if $R \Rightarrow \bigwedge_{i=1}^m R(\Omega_i)$. If $R \not\models JD(S)$, then $S$ incurs a loss with respect to $R$, denoted $\rho(R,S)$, defined:

$$\rho(R,S) \equiv \frac{\sum_{i=1}^m \left| R(\Omega_i) \right| - |R|}{|R|}$$

We call the set of tuples $\bigwedge_{i=1}^m R(\Omega_i) \setminus R$ spurious tuples. Clearly, $R \models JD(S)$ if and only if $\rho(R,S) = 0$.

We say that $R$ satisfies the multivalued dependency (MVD) $\phi = X \Rightarrow Y_1, \ldots, Y_m$, where $m \geq 2$, the $Y_i$s are pairwise disjoint, and $XY_1 \cdots Y_m = \Omega$, if $R = R[X] \Rightarrow \cdots \Rightarrow R[XY_m]$, or if the schema $S = \{XY_1, \ldots, XY_m\}$ is lossless (i.e., $\rho(R,S) = 0$). We review the concept of a join tree from [3]:

Definition 2.1. A join tree or junction tree is a pair $\langle T, \chi \rangle$ where $T$ is an undirected tree, and $\chi$ is a function that maps each node $u \in \nodes(T)$ to a set of variables $\chi(u)$, called a bag, such that the running intersection property holds: for every variable $X$, the set $\{u \in \nodes(T) \mid X \in \chi(u)\}$ is a connected component of $T$. We denote by $\chi(T) \equiv \bigcup \chi(u)$, the set of variables of the join tree.

We often denote the join tree as $T$, dropping $\chi$ when it is clear from the context. The schema defined by $T$ is $S = \{\Omega_1, \ldots, \Omega_m\}$, where $\Omega_1, \ldots, \Omega_m$ are the bags of $T$. We call a schema $S$ acyclic if there exists a join tree whose schema is $S$. When $\Omega_i \not\subseteq \Omega_j$ for all $i \neq j$, then the acyclic schema $S = \{\Omega_1, \ldots, \Omega_m\}$ satisfies $m \leq |\Omega|$ [3]. We say that a relation $R$ satisfies the acyclic join dependency $S$, and denote $R \models AJD(S)$, if $S$ is acyclic and $R \models JD(S)$. An MVD
We call the \( \phi_{u,a} \) (i.e., \( K_P | \ ) \) a probability space, it holds that \( D \) for any pair of probability distributions \( I \) and denoted by conditional mutual information is reduced to the standard mutual information defined as:

\[ H(I;J) = \sum_{x \in \mathcal{D}} p(x) \log \frac{1}{p(x)}. \] (2)

It holds that \( H(X) \leq \log |\mathcal{D}| \), and equality holds if and only if \( P \) is uniform. The definition of entropy naturally generalizes to sets of jointly distributed random variables \( \Omega = \{X_1, \ldots, X_n\} \), by defining the function \( H : 2^{\Omega} \to \mathbb{R} \) as the entropy of the joint random variables in the set. For example,

\[ H(X_1X_2) = \sum_{x_1 \in \mathcal{D}_1, x_2 \in \mathcal{D}_2} p(x_1, x_2) \log \frac{1}{p(x_1, x_2)}. \] (3)

Let \( A, B, C \subseteq \Omega \). The conditional mutual information \( I(A; B | C) \) is defined as:

\[ I(A; B | C) = H(B) + H(A|C) - H(A|B, C). \] (4)

It is known that the conditional independence \( P(A \perp B | C) \) (i.e., \( A \) is independent of \( B \) given \( C \)) holds if and only if \( I(A; B | C) = 0 \). When \( C = \emptyset \), or if \( C \) is a constant (i.e., \( H(C) = 0 \)), then the conditional mutual information is reduced to the standard mutual information, and denoted by \( I(A; B) \).

Let \( X = \{X_1, \ldots, X_n\} \) be a set of discrete random variables, and let \( P(X) \) and \( Q(X) \) denote discrete probability distributions. The KL-divergence between \( P \) and \( Q \) is:

\[ D_{KL}(P \parallel Q) = \mathbb{E}_P \log \frac{P(X)}{Q(X)} = \sum_{x \in \mathcal{D}} P(x) \log \frac{P(x)}{Q(x)}. \] (5)

For any pair of probability distributions \( P, Q \) over the same probability space, it holds that \( D_{KL}(P \parallel Q) \geq 0 \), with equality if and only if \( P = Q \). It is an easy observation that

\[ I(A; B | C) = D_{KL}(P(ABC) \parallel P(A|C)P(B|C)P(C)) \] (6)

for any probability distribution \( P \).

Let \( R \) be a multiset of tuples over the attribute set \( \Omega = \{X_1, \ldots, X_n\} \), and let \( |R| = N \). The empirical distribution associated with \( R \) is the multivariate probability distribution over \( \mathcal{D}_1 \times \cdots \times \mathcal{D}_n \) that assigns a probability of \( P(t) = \frac{K}{N} \) to every tuple \( t \) in \( R \) with multiplicity \( K \). When \( R \) is a relation instance, and hence a set of \( N \) tuples, its empirical distribution is the uniform distribution over its tuples, i.e. \( P(t) = 1/N \) for all \( t \in R \), and so its entropy is \( H(\Omega) = \log N \). For a relation instance \( R \in \Omega(\Omega) \), we let \( P_R \) denote the empirical distribution over \( R \). For \( \alpha \subseteq [n] \), we denote by \( X_\alpha \) the set of variables \( X_i \) \( i \in \alpha \), and denote by \( R(X_\alpha = x_\alpha) \) the subset of tuples \( t \in R \) where \( t[\alpha] = x_\alpha \), for fixed values \( x_\alpha \). When \( P_R \) is the empirical distribution over \( R \) then the marginal probability is \( P_R(X_\alpha = x_\alpha) = \frac{|R(X_\alpha = x_\alpha)|}{N} \).

Lee [15, 16] gave an equivalent formulation of functional, multivalued, and acyclic join dependencies in terms of entropic measures. Let \( (T, \chi) \) be a join tree (Definition 2.1). The \( J \)-measure is defined as:

\[ J(T, \chi) = \sum_{(u, v) \in \text{edges}(T)} H(\chi(u)) - \sum_{(u, v) \in \text{edges}(T)} H(\chi(u) \cap \chi(v)). \] (7)

where \( H \) is the entropy (see (2)) taken over the empirical distribution associated with \( R \). We abbreviate \( J(T, \chi) \) with \( J(T) \), or \( J \) when \( T, \chi \) are clear from the context. Observe that \( J \) depends only on the schema \( \Omega \) defined by the join tree, and not on the tree itself. For a simple example, consider the MVD \( X \rightarrow U|V|W \) and its associated acyclic schema \( X \rightarrow UV, XV, XW \). For the join tree \( XU \rightarrow XW \rightarrow XV \), it holds that \( J = H(XU) + H(XV) + H(XW) - H(XUV) \). Another join tree is \( XV \rightarrow XW \rightarrow VX \), and \( J \) is the same. Therefore, if \( S \) is acyclic, then we write \( J(S) \) to denote \( J(T) \) for any join tree \( S \). When \( S = \{XZ, XY\} \), then the \( J \)-measure reduces to the conditional mutual information (see (4)), to wit \( J(S) = I(Z; Y|X) \).

**Theorem 2.1.** ([16]) Let \( S \) be an acyclic schema over \( \Omega \), and let \( R \) be a relation instance over \( \Omega \). Then \( R \models AJD(S) \) if and only if \( J(S) = 0 \).

In the particular case of a standard MVD, Lee’s result implies that \( R \models X \rightarrow Y | Z \) if and only if \( I(Y; Z|X) = 0 \) in the empirical distribution associated with \( R \).

### 2.3 MVDs and Acyclic Join Dependencies

Beeri et al. [3] have shown that an acyclic join dependency defined by the acyclic schema \( S \), over \( m \) relation schemas, is equivalent to \( m - 1 \) MVDs (called its support). In [12], this characterization was generalized as follows. Let \( (T, \chi) \) be the join tree corresponding to the acyclic schema \( S \). Root the tree \( T \) at node \( u_1 \), orient the tree accordingly, and let \( u_1, \ldots, u_m \) be a depth-first enumeration of nodes \( (T) \). Thus, \( u_i \) is the root, and for every \( i > 1 \), parent\((u_i) = \) some node \( u_j \) with \( j < i \). For every \( i \), we define \( \Omega_i \equiv \chi(u_i), \Omega_{i:j} \equiv \bigcup_{\ell=i}^{j} \Omega_{\ell}, \text{ and } \Delta_{i} \equiv \chi(\text{parent}(u_i)) \cap \chi(u_i) \). By the running intersection property (Definition 2.1) it holds that \( \Delta_{i} = \Omega_{i:i-1} \cap \Omega_{i} \).

**Theorem 2.2.** ([12]) Let \( (T, \chi) \) be a join tree over variables \( \chi(T) = \Omega \), where nodes \( T = \{u_1, \ldots, u_m\} \) with corresponding bags \( \Omega_i = \chi(u_i) \). Then:

\[ \max_{i \in [2, m]} I(\Omega_{i:i-1}; \Omega_{i:m} \mid \Delta_i) \leq J(T, \chi) \leq \sum_{i=2}^{m} I(\Omega_{i:i-1}; \Omega_{i:m} \mid \Delta_i) \] (8)

Since the support of \( (T, \chi) \) are precisely the MVDs \( \{\Delta_i \models \Omega_{i:i-1} \cap \Omega_{i:m} \mid i \in [2, m]\} \), then (8) generalizes the result of Beeri et al. [3].
We denote by $\Delta$ the following two lemmas, interesting on their own, and is deferred to the complete version of this paper [13].

3 CHARACTERIZING ACYCLIC SCHEMAS WITH KL-DIVERGENCE

Let $X \equiv \{X_1, \ldots, X_n\}$ be a set of discrete random variables over domains $\mathcal{D}(X_i)$, and let $P(X_1, \ldots, X_n)$ be a joint probability distribution over $X$. Let $x \in \mathcal{D}(X_1) \times \cdots \times \mathcal{D}(X_n)$. For a subset $Y \subseteq X$, we denote by $x|Y$ the assignment $x$ restricted to the variables $Y$. We denote by $P[Y]$ the marginal probability distribution over $Y$.

Let $(T, \chi)$ be a join tree where $\text{nodes}(T) = \{u_1, \ldots, u_m\}$. Let $S = \{\Omega_1, \ldots, \Omega_m\}$ be the acyclic schema associated with $(T, \chi)$

Proposition 3.1. Let $P(X_1, \ldots, X_n)$ be any joint probability distribution over $n$ variables, and let $(T, \chi)$ be a join tree where $\chi(T) = \{X_1, \ldots, X_n\}$. Then $P \models T$ (Definition 2.2) if and only if $P = P_T$ where:

$$P_T(x_1, \ldots, x_n) \equiv \prod_{i=1}^{m} P(\Omega_i)(x(\Omega_i)) / \prod_{i=1}^{m} P(\Omega_i)(x(\Delta_i))$$

where $P(\Omega_i) (P(\Delta_i))$ denote the marginal probabilities over $\Omega_i$ ( $\Delta_i$).

The proof of Proposition 3.1 appears in [13]. It follows from Definition 2.2, and a simple induction on the number of nodes in $T$.

In this section, we refine the statement of Proposition 3.1 and prove the following variational representation.

Theorem 3.2. For any joint probability distribution $P(X_1, \ldots, X_n)$ and any join tree $(T, \chi)$ with $\chi(T) = \{X_1, \ldots, X_n\}$ it holds that:

$$\mathcal{F}(T) = \min_{P \models T} \text{KL}(P||Q) = \text{KL}(P||P_T)$$

In words, this theorem states that when the join tree $(T, \chi)$ is given, then out of all probability distributions $Q$ over $\Omega = \{X_1, \ldots, X_n\}$ that model $T$ (see (10)), the one closest to $P$ in terms of KL-Divergence, is $P_T$. Importantly, this KL-divergence is precisely $\mathcal{F}(T)$ (i.e., $\mathcal{F}(T) = \text{KL}(P||P_T)$).

While the Theorem holds for all probability distributions $P$, a special case is when $P$ is the empirical distribution associated with relation $R$. The proof of Theorem 3.2 follows from the following two lemmas, interesting on their own, and is deferred to the complete version of this paper [13].

Lemma 3.3. Let $P(X_1, \ldots, X_n)$ be a joint probability distribution over $n$ random variables, and let $T$ be a join tree over $X_1, \ldots, X_n$ with bags $\Omega_1, \ldots, \Omega_m$. Then $P(\Omega_i) = P_T(\Omega_i)$ for every $i \in [1, m]$, and $P(\Delta_i) = P_T(\Delta_i)$ for every $i \in [1, m-1]$.

The proof of Lemma 3.3 follows from an easy induction on $m$, the number of nodes in the join tree $T$, and can be found in [13].

Lemma 3.4. The following holds for any joint probability distribution $P(X_1, \ldots, X_n)$ and any join tree $T$ over variables $X_1, \ldots, X_n$:

$$\arg \min_{P \models T} \text{KL}(P||Q) = P_T$$

Proof. From Lemma 3.3 we have that, for every $i \in [1, m]$, $P_T(\Omega_i) = P(\Omega_i)$ where, $\Omega_i = \chi(u_i)$. Since $\Delta_i \subseteq \Omega_i$, then $P_T(\Delta_i) = P[\Delta_i]$. Now,

$$\min_{Q \models T} \text{KL}(P||Q) = \min_{Q \models T} \mathbb{E}_P \log \frac{P(X_1, \ldots, X_n)}{Q(X_1, \ldots, X_n)}$$

$$= \min_{Q \models T} \mathbb{E}_P \log \frac{P_T(X_1, \ldots, X_n)}{Q(X_1, \ldots, X_n)}$$

$$= \text{KL}(P||P_T) + \min_{Q \models T} \mathbb{E}_P \log \frac{P_T(X_1, \ldots, X_n)}{Q(X_1, \ldots, X_n)}$$

$$= \text{KL}(P||P_T),$$

where the last equality follows from (5). Since $\text{KL}(P_T||Q_T) \geq 0$, with equality if and only if $P_T = Q_T$, then choosing $Q_T$ to be $P_T$ minimizes $\text{KL}(P||Q)$, thus proving the claim. The remainder of the proof, proving (16), follows from Lemma 3.3 which states that $P(\Omega_i) = P_T(\Omega_i)$ for every $i \in [1, m]$, and $P(\Delta_i) = P_T(\Delta_i)$ for every $i \in [m-1]$. The proof is quite technical, and can be found in [13].

4 SPURIOUS TUPLES: A LOWER BOUND BASED ON $\mathcal{F}(T)$

Let $P_R$ be the empirical distribution over a relation instance $R$ with $N$ tuples and $n$ attributes $\Omega = \{X_1, \ldots, X_n\}$. That is, the probability associated with every record in $R$ is $\frac{1}{N}$. Let $T$ be any junction tree over the variables $\Omega$, and let $S = \{\Omega_1, \ldots, \Omega_m\}$ where $\Omega_i = \chi(u_i)$. By Theorem 2.1 and Theorem 3.2, it holds that $R \models \text{AJD}(S)$ if and only if $\mathcal{F}(T) = \text{KL}(P_R||P_T) = 0$. In what follows, given an acyclic schema $S$, we provide a lower bound for $\rho(R, S)$ that is based on its associated junction tree $\mathcal{F}(T)$.

Lemma 4.1. Let $P$ be the empirical distribution over a relation instance $R$ with $N$ tuples, and let $S = \{\Omega_1, \ldots, \Omega_m\}$ denote an acyclic schema with junction tree $(T, \chi)$. Then:

$$\mathcal{F}(T) \leq \log(1 + \rho(R, S)).$$

So if $\rho(R, S) = 0$, then $\mathcal{F}(T) = 0$ as well.

Proof. From Theorem 3.2 we have that:

$$\mathcal{F}(T) = \text{KL}(P||P_T),$$

where

$$P_T = \arg \min_{Q \models T} \text{KL}(P||Q).$$

Let us define

$$Q_{TU} = \arg \min_{Q \models T} \text{KL}(P||Q),$$

where $Q$ is uniform.
and verify that such a distribution $Q_{TU}$ always exists: Let $R' \overset{\text{def}}{=} \Pi_{i=1}^m \Omega_i(R)$. Let $Q_{TU}$ denote the empirical distribution over $R'$. By construction, $Q_{TU}$ is a uniform distribution (i.e., over tuples $R'$), and $Q_{TU} \models T$.

By definition, we have that $|R'| = N(1 + \rho(R, S))$, where $|R| = N$, and $\rho(R, S)$ is the loss of $S$ with respect to $R$ (see (1)). By limiting the minimization region to uniform distributions, the minimum can only increase. Therefore:

$$J(T) = D_{KL}(P||P_T) \leq D_{KL}(P||Q_{TU})$$

(22)

Evaluating the KL-divergence term on the right hand side, and using the fact that $Q_{TU}$ is uniform, we get:

$$D_{KL}(P||Q_{TU}) = \sum_{r \in R} P(r) \log \frac{P(r)}{Q_{TU}(r)}$$

(23)

$$= \sum_{r \in R} \frac{1}{N} \log \left(\frac{1}{N} \cdot \frac{1}{N + N \cdot \rho(R, S)}\right)$$

(24)

$$= \sum_{r \in R} \frac{1}{N} \log \left(1 + \rho(R, S)\right)$$

(25)

$$= \log(1 + \rho(R, S))$$

(26)

Hence, $J(T) \leq \log(1 + \rho(R, S))$, and $\rho(R, S) \geq 2J(T) - 1 \quad \square$

The following simple example shows that the lower bound of (18) is tight. That is, there exists a family of relation instances $R$, and a schema $S$ where $J(S) = \log(1 + \rho(R, S))$.

**Example 4.1.** Let $\Omega = \{A, B\}$, where $\mathcal{D}(A) = \{a_1, \ldots, a_N\}$ and $\mathcal{D}(B) = \{b_1, \ldots, b_N\}$ where $\mathcal{D}(A) \cap \mathcal{D}(B) = \emptyset$. Let $R = \{t_i = (a_1, b_1), \ldots, t_N = (a_N, b_N)\}$ be a relation instance over $\Omega$. By definition, $P_R(t_i) = \frac{1}{N}$. Noting that $H(A) = H(B) = H(AB) = \log N$ (see (2)), we have that $I(A; B) = \log N$ (see (4)). Now, consider the schema $S = \{A, \{\}\}$, and let $R' = \Pi_A(R) \approx \Pi_B(R)$. Clearly, $|R'| = N^2$, and $\rho(R, S) = N - 1$ (see (1)). In particular, we have that $J(S) = I(A; B) = \log N = \log(1 + \rho(R, S))$, and that this holds for every $N \geq 2$.

5 SPURIOUS TUPLES: AN UPPER BOUND BASED ON MUTUAL INFORMATION

In the previous section, we have shown that given a relation $R$ and an acyclic schema $S$ defined by a join tree $T$, it holds that $\log(1 + \rho(R, S)) \geq J(T)$ (Lemma 4.1). In this section, we derive an upper bound on $\rho(R, S)$ in terms of an information-theoretic measure. To this end, recall that Theorem 2.2 shows that if $\{\Omega_1, \ldots, \Omega_m\}$ are the bags of $T$, then $J(T) \leq \sum_{i=1}^m \log \left(1 - \frac{I(\Omega_{i-1}; \Omega_i | \Delta_i)}{I(\Omega_{i-1}; \Omega_i)}\right)$. That is, $J(T)$ is upper bounded by the sum of the conditional mutual information of the $m-1$ MVDs in the upper bound $T$, given by $\Delta_i \Rightarrow \Omega_{i-1}|\Omega_i$, for $i \in [2, m]^{-1}$. In this section, we relate this sum of conditional mutual information of the $m-1$ MVDs in the support $T$ to an approximate upper bound on $\rho(R, S)$, which holds with high probability.

To this end, we begin by relating the relative number of spurious tuples of the relation $R$ of a schema $S$, that is $\rho(R, S)$, with the spurious tuples of each of the MVDs in its support. Concretely, let

$$\phi_i \overset{\text{def}}{=} \Delta_i \rightarrow \Omega_{i-1} | \Omega_{i-1} \Delta_i$$

be the $i$th MVD in the support of $T$. Then, the relative number of spurious tuples for $\phi_i$ is defined as

$$\rho(R, \phi_i) \overset{\text{def}}{=} \frac{\Pi_{\Omega_{i-1}(R)} \rightarrow \Pi_{\Omega_{i-1}(R)} - |R|}{|R|} \quad (28)$$

**Proposition 5.1.** Let a relation $R$ be given, and let $S$ be an acyclic schema over the attributes of $R$ with join tree $T$, whose support are the MVDs $\phi_i \rightarrow \Omega_{1-1} | \Omega_{1-1} \Delta_i$, for $i \in [2, m]$. Then,

$$\log(1 + \rho(R, S)) \leq \sum_{i=2}^m \log(1 + \rho(R, \phi_i))$$

(29)

Proof. We prove by induction on the number of MVDs (or nodes) in the schema. Let $m$ be the number nodes (and $m-1$ be the number of MVDs in the schema). The base case $m-1 \leq 1$ is immediate. Assuming it holds for $m-1 < k$, we prove the claim for $m = 1$. Let $T$ be a join tree representing $k$ MVDs (and hence $k+1$ nodes). Let $u_{k+1}$ be a leaf in this join tree with parent $p = \text{parent}(u_{k+1})$. Let $T'$ be the join tree where nodes $u_{k+1}$ and $p$ are merged to the node $u'$ where $\Omega(u') = \Omega(u_{k+1}) \cup \Omega(p)$. Hence, by the induction hypothesis

$$1 + \rho(R, T') \leq \sum_{i=2}^k \left[1 + \rho(R, \phi_i)\right]$$

(30)

Now, let $R' = \prod_{i=1}^k R[\Omega_i]$. Consider the MVD $\phi_{k+1} = \Omega(u_{k+1}) \cap \Omega(p) \rightarrow \Omega(u_{k+1}) | \Omega_{i-1} \Delta_i$. Then, $R'' \overset{\text{def}}{=} \Pi_{\Omega_k}(R') \rightarrow \Pi_{\Omega_{k+1}}(R')$ and by the induction hypothesis, $|R''| \leq |R| \cdot [1 + \rho(R, \phi_k)]$. By (30),

$$|R'| \leq |R| \cdot \left[1 + \rho(R, \phi_k)\right]$$

(31)

and hence $|R''| \leq |R| \cdot \prod_{i=2}^{k+1}[1 + \rho(R, \phi_i)]$, which proves claim. \quad \square

Proposition 5.1 reduces the problem of upper bounding $\log(1 + \rho(R, S))$ to bounding each of the terms $\log(1 + \rho(R, \phi_i))$ in (29), each of them corresponding to the relative number of spurious tuples of the $m-1$ MVDs in the support of $T$. Considering an arbitrary MVD, which we henceforth denote for simplicity by $\phi \overset{\text{def}}{=} C \rightarrow A|B$ (Lemma 4.1 implies the lower bound $\log(1 + \rho(R, \phi)) \geq I(A; B | C)$, since an MVD is a simple instance of an acyclic schema. However, obtaining an upper bound on $\log(1 + \rho(R, \phi))$ in terms of $I(A; B | C)$ is challenging because the mutual information $I(A; B | C)$ varies wildly for an MVD $\phi$ even when $\rho(R, \phi)$ and the domains sizes $d_A, d_B$ and $d_C$ remain constant (where $d_A \overset{\text{def}}{=} |\Pi_{\Omega_A}(R)|$, and similarly for $d_B$ and $d_C$). Figure 1 illustrates this phenomenon in the simple case in which $d_C = 1$ and $C$ is a degenerated random variable, and $d_A = d_B$. In other words, the value $I(A; B | C)$ depends on the actual contents of the relation instance $R$. However, while $I(A; B | C)$ might not be an accurate upper bound to $\log(1 + \rho(R, \phi))$ for an arbitrary relation, it may hold that it is an approximate upper bound for most relations. Therefore, we next propose a random relation model, in which the tuples of the relation $R$ are chosen at random.

**Definition 5.2 (Random relation model).** Let $\Omega \overset{\text{def}}{=} \{X_1, \ldots, X_n\}$ be a set of attributes with domains $\mathcal{D}(X_1), \ldots, \mathcal{D}(X_n)$, and assume w.l.o.g. that $\mathcal{D}(X_i) \overset{\text{def}}{=} \{d_i\}_{i=1}^{|d_i|} \subset \mathbb{N}_+$. Let $N \in \mathbb{N}_+$ be given
such that $0 < N \leq \prod_{i=1}^{m}{d_i}$. Let $S$ be a set of $N$ tuples chosen uniformly at random from $\times_{i=1}^{m}{[d_i]}$, without replacement. Given $S$, we let $P_S$ denote the empirical distribution over $S$:

$$P_S \left[ \sum_{i=1}^{n}{X_i = t_i} \right] = \begin{cases} \frac{1}{N}, & (t_1, t_2, \ldots, t_n) \in S \\ 0, & \text{otherwise} \end{cases}$$  \hspace{1cm} (32)

for any $(t_1, t_2, \ldots, t_n) \in \times_{i=1}^{n}{[d_i]}$.

In other words, in the random relational model $R$ is chosen uniformly at random from the set of possible relations of size $N$, that is, from the set $R(\Omega) \cap \{R: |R| = N\}$. The next proposition states that the existence of a high probability bound on the relative number of spurious tuples associated with an arbitrary MVD $\phi$, implies the existence of a high-probability upper bound on the relative number of spurious tuples $\log(1 + \rho(R, S))$ associated with an acyclic schema $S$.

**Proposition 5.3.** Let $\epsilon(\phi, N, \delta) \geq 0$ where $\phi \overset{\text{def}}{=} C \rightarrow A|B$ is an MVD, $\delta \in (0, 1)$, and $R$ is a random relation over attributes $ABC$, where $|R| = N$. Let $S = \{\Omega_1, \ldots, \Omega_m\}$ be an acyclic schema over the attributes of $R$ with join tree $T$. If the random relation $R$ satisfies

$$\log(1 + \rho(R, \phi)) \leq I(A; B \mid C) + \epsilon(\phi, N, \delta),$$

with probability larger than $1 - \frac{\delta}{m - 1}$, for all MDVs $\phi_i \overset{\text{def}}{=} \Delta_i \rightarrow \Omega_{1 \rightarrow i-1} \mid \Omega_{i,m}$ in the support of $S$. Then:

$$\log(1 + \rho(R, S)) \leq \sum_{i=2}^{m} I(\Omega_{1 \rightarrow i-1}; \Omega_{i,m} \mid \Delta_i) + \epsilon_i$$  \hspace{1cm} (33)

with probability $1 - \delta$, where $\epsilon_i \overset{\text{def}}{=} \epsilon(\Delta_i \rightarrow \Omega_{1 \rightarrow i-1} \mid \Omega_{i,m}, N, \frac{\delta}{m - 1})$.

**Proof.** For the MVD $\phi_i \overset{\text{def}}{=} \Delta_i \rightarrow \Omega_{1 \rightarrow i-1} \mid \Omega_{i,m}$, it holds that

$$\log(1 + \rho(R, \phi)) \leq I(\Omega_{1 \rightarrow i-1}; \Omega_{i,m} \mid \Delta_i) + \epsilon(\phi_i, N, \frac{\delta}{m - 1})$$  \hspace{1cm} (35)

with probability larger than $1 - \frac{\delta}{m - 1}$, Then, (33) follows from Proposition 5.1, and a union bound over the $m - 1$ MDVs $(\phi_i)_{i=2}^{m}$ in the support of $S$. The bound (34) follows from (8) in Theorem 2.2.

Hence, the problem of deriving an upper bound on $\log(1 + \rho(R, S))$, which holds with high probability, is reduced to the problem of showing that $\log(1 + \rho(R, \phi)) \leq I(A; B \mid C) + \epsilon(\phi, N, \delta)$ holds with high probability for an MVD $\phi \overset{\text{def}}{=} C \rightarrow A|B$, in the setting of the random relational model, assuming that the relation size is fixed to $N$. In other words, it now suffices to prove the probabilistic upper bound for a single MVD in the random relation model (Definition 5.2).

In what follows, we focus on a single MVD, denoted $\phi \overset{\text{def}}{=} C \rightarrow A|B$, and where $d_A$, $d_B$ and $d_C$ are the domain sizes of $A$, $B$, and $C$, respectively. Then, for any $S \subseteq [d_A] \times [d_B] \times [d_C]$

$$P_S \left[ A = a, B = b, C = c \right] = \begin{cases} \frac{1}{N}, & (a, b, c) \in S \\ 0, & \text{otherwise} \end{cases}$$  \hspace{1cm} (36)

for any $(a, b, c) \in [d_A] \times [d_B] \times [d_C]$, and the relation is such that the set $S$ is chosen uniformly at random from $[d_A] \times [d_B] \times [d_C]$ from all possible sets of size $N$. While both the domain sizes $d_A$, $d_B$ and $d_C$ and relation size $N$ are fixed in this model, the mutual information $I(A; B \mid C)$ is a random variable due to the random choice of the set $S$. Specifically, the random relation instance $S \subseteq [d_A] \times [d_B] \times [d_C]$ where $|S| = N$, is a random variable, and each specific realization $S = s$, defines a triplet of random variables $A_s \overset{\text{def}}{=} \Pi_A(s)$, $B_s \overset{\text{def}}{=} \Pi_B(s)$ and $C_s \overset{\text{def}}{=} \Pi_C(s)$. Consequently, every such set $s$ defines various information theoretic measures, such as $H(A_s)$, $I(A_s; B_s \mid C_s)$, and so on. Furthermore, a random choice of $S$ makes these information measures random quantities themselves, for example, $H(A_S)$ and $I(A_S; B_S \mid C_S)$ are random variables. In a similar fashion, if we let $R_S$ denote the random relation defined by $S$, then $\rho(R_S, \phi)$ is again a random variable. Our main result regarding the mutual information of an MVD is as follows:

**Theorem 5.1 (Confidence bound of the random mutual information of an MVD).** Let $\delta \in (0, 1)$. Assume w.l.o.g. that $d_A \geq d_B$, denote $\overline{d} \overset{\text{def}}{=} \max\{d_A, d_C\}$ and assume further that

$$N \geq 256d_A^2 \log \left( \frac{384\overline{d}}{\delta} \right).$$  \hspace{1cm} (37)

Let

$$e^*(\phi, N, \delta) \overset{\text{def}}{=} 60d_A^2 \log^3(\frac{\delta N d_C}{\delta})$$

If $R_S$ is drawn from the random relation model of Definition 5.2, then

$$\log(1 + \rho(R_S, \phi)) \leq I(A_S; B_S \mid C_S) + e^*(\phi, N, \delta)$$  \hspace{1cm} (38)

with probability larger than $1 - \delta$.

The proof of Theorem 5.1 is fairly complicated, and is discussed in detail in Section 5.1. Theorem 5.1 shows that, with high probability, the upper bound $\log(1 + \rho(R_S, \phi)) \leq I(A_S; B_S \mid C_S)$ approximately holds, up to an additive factor of $\tilde{O}(\sqrt{d_A} \max\{d_A, d_C\}/N)$, where the $\tilde{O}(\cdot)$ hides logarithmic terms. This result is suitable for large domain sizes, and when the number of tuples $N$ is proportional to the domain sizes. More accurately, the bound holds whenever $N = \Omega(d_A^2 d_C)$, when $\tilde{O}(\cdot)$ hides logarithmic terms (condition (37)), which is a mild condition when targeting a low fraction of spurious tuples. So, when $\delta$ is fixed to some desired reliability, and the qualifying condition (37) holds, then the deviation term in the claim of Theorem 5.1 is given by $e^*(\phi, N, \delta) = \tilde{O}(\max\{d_A^2, d_A d_C\}/N)$. Hence, when $N$ increases as $N = \omega(\max\{d_A^2, d_A d_C\})$, then $e^*(\phi, N, \delta)$ vanishes. For example, if $d_A = d_B = d_C = d$, then the deviation term is $e^*(\phi, N, \delta) = O(\sqrt{d^3 \log^3(Nd)})$. And this deviation term vanishes if $N = \omega(d^2 \log^3(Nd))$. As a more concrete example, if $N = \frac{1}{2}d^3$, then the deviation term is $e^*(\phi, N, \delta) = O(\frac{\log^3(\sqrt{d})}{d})$ which vanishes at a rather fast rate with increasing $d$. Moreover, the dependency of the deviation term in $\delta$ is mild, and scales as $\log^{3/2}(1/\delta)$ which is close to a sub-exponential dependence.\footnote{For sub-exponential random variables, the dependence on $\delta$ is $\log(1/\delta)^{\frac{3}{2}}$.}  

**5.1 Proof of Theorem 5.1: A Confidence Interval for the Mutual Information**

In this section, we discuss in detail the proof of Theorem 5.1. We focus on the case in which $C$ is a degenerate random variable ($d_C = 1$) since the main components of the proof are already present in this
simple case. When \( d_C = 1 \), the conditional mutual information is reduced to the standard mutual information \( I(A_S; B_S) \). In turn, the random set \( S \) which determines the random relation is chosen uniformly at random from all subsets of \([d_A] \times [d_B]\) of a given size. To avoid confusion with the non-degenerate model, we denote this size by \( \eta \) (rather than by \( N \)) whenever \( d_C = 1 \). The mutual information can then be decomposed as \( I(A_S; B_S) = H(A_S) + H(B_S) - H(A_S, B_S) \) [8, Section 2.4], and by the definition of the random model, \( R_S = (A_S, B_S) \) is distributed uniformly over the possible sets of size \( \eta \). Thus \( H(A_S, B_S) = \log \eta \) with probability 1 (over the choice of \( S \)), and \( I(A_S; B_S) = H(A_S) + H(B_S) - \log \eta \). Due to symmetry, the analysis of \( H(A_S) \) and \( H(B_S) \) is analogous, and so we next focus on the former. The main ingredient of the proof of Theorem 5.1 is a confidence interval for the random entropy \( H(A_S) \), when \( A_S \) is chosen from a random relation model similar to the one of Definition 5.2, albeit with a degenerated \( C \), that is, \( d_C = 1 \). At a later stage, we discuss the generalization of this result to \( d_C > 1 \). The confidence bound on \( H(A_S) \) is as follows:

**Theorem 5.2.** Let \( A_S \) be drawn according to the random relation model of Definition 5.2 with \( d_C = 1 \) and \( N = \eta \). Assume w.l.o.g. that \( d_A \geq d_B \) and that

\[
\eta \geq 128 d_A \log \left( \frac{128 d_A}{\delta} \right).
\]

Then, for any probability \( \delta \in (0, 1) \) it holds that

\[
\log d_A \geq H(A_S) \geq \log d_A - 20 \sqrt{\frac{d_A \log \left( \frac{d_A}{\eta} \right)}{\eta}}
\]

with probability \( 1 - \delta \), over the random choice of the set \( S \).

The proof of Theorem 5.2 comprises most of the proof of Theorem 5.1, and requires a diverse set of mathematical techniques, discussed in Section 5.2. For now, taking the result of Theorem 5.2 as given, a high probability bound on the value of \( I(A_S; B_S) \) for the random relation model with degenerated \( C \) can be obtained as a simple corollary to Theorem 5.2, as follows:

**Corollary 5.2.1.** Let \( \overline{\rho} = \frac{d_A d_B}{\eta} - 1 \). Then, under the same assumptions of Theorem 5.2,

\[
I(A_S; B_S) \geq \log (1 + \overline{\rho}) - 40 \sqrt{\frac{d_A \log \left( \frac{d_A}{\eta} \right)}{\eta}},
\]

with probability \( 1 - \delta \), over the random choice of the set \( S \).

Corollary 5.2.1 will be used in the proof of Theorem 5.1. Beyond that, it also reveals the tightness of the mutual information bound in the simpler setting of a degenerated MVD (\( d_C = 1 \)). Indeed, since \( A_S \subseteq [d_A] \) and \( B_S \subseteq [d_B] \) for all realizations of \( S \), then \( \rho(R_S, \phi) \leq \frac{d_A d_B}{\eta} - 1 = \overline{\rho} \). Thus, Corollary 5.2.1 implies

\[
I(A_S; B_S) \geq \log (1 + \rho(R_S, \phi)) - 40 \sqrt{\frac{d_A \log \left( \frac{d_A}{\eta} \right)}{\eta}},
\]

but actually shows the stronger bound (42).

**Proof outline of Theorem 5.1.** At this point, let us take the results of Theorem 5.2 and Corollary 5.2.1 as granted. Then, the proof of Theorem 5.1 is essentially a generalization of the result of Theorem 5.2 to the case in which \( d_C > 1 \). Let us define \( R_S = [S] = \sigma_S(\mathcal{R}) \). Then, in the random relation model \( N_S(t) = |S| \) is a random variable, and beyond the randomness in the joint distribution of \( (A_S, B_S) \) when conditioned on any \( C = t \), there is also randomness in the number of tuples in the random relation, whenever \( C = t \). Hence, the mutual information conditioned on the specific value of \( C = t \), to wit, \( I(A_S; B_S \mid C_S = t) \), is drawn from the random model in Definition 5.2 with \( N \) replaced by \( N_S(t) \) (the latter being a random variable due to the random choice of \( S \)). The result of Corollary 5.2.1, regarding the mutual information of a pair of random variables \( A_S, B_S \), can then be used conditionally on \( C_S = t \), where \( \eta \) is being replaced by \( N_S(t) \). In order for this result to hold, the qualifying condition of Corollary 5.2.1, to wit \( N_S(t) \geq 128 d_A \log \left( \frac{128 d_A}{\delta} \right) \), should hold for all \( t \in [d_C] \). The proof begins by showing that this condition indeed holds for all \( t \in [d_C] \) with high probability. This is proved in [13], and is based on the fact that \( N_S(t) \) is a hypergeometric random variable, and on a concentration result by Serfling [22] for such random variables. The proof then assumes that all the following holds: (I) For all \( t \in [d_C] \), \( N_S(t) \) is sufficiently large so that the qualifying condition of Corollary 5.2.1 holds. (II) For each \( t \in [d_C] \), the confidence bound in Corollary 5.2.1 holds. (III) \( H(C_S) \) is close to \( \log d_C \).

Specifically, the proof in [13] assures that the first condition holds with high probability; Corollary 5.2.1 assures that the second condition holds with high probability; a simple modification of Theorem 5.2 shows that the third condition holds with high probability. By the union bound, the event in which the set \( S \subseteq [d_A] \times [d_B] \times [d_C] \) simultaneously satisfies properties (I), (II) and (III) has high probability. The proof is completed by considering a set \( S \subseteq [d_A] \times [d_B] \times [d_C] \) which satisfies properties (I), (II) and (III), and relating \( \log(1 + \rho(R_S, \phi)) \) to the mutual information. Concretely, an application of the \( \log \sum \) inequality [8, Thm. 2.7.1] shows that

\[
\log \left( 1 + \rho(R_S, \phi) \right) \leq \log d_C - H(C_S)
\]

\[
+ \sum_{t \in [d_C]} P[C_S = t] \log \left( 1 + \overline{\rho}_t(t) \right),
\]

where \( \overline{\rho}_t(t) = \frac{d_A d_B}{\eta} - 1 \).
which can be bounded by conditional mutual information, and an additional additive deviation term, utilizing the aforementioned assumption that \( s \) satisfies properties (I), (II) and (III).

5.2 Confidence Bound of the Conditional Entropy

In this section, we describe the proof of the confidence interval in Theorem 5.2, which is comprised of three main steps on its own: (I) A bound on the expected value of \( H(\mathcal{A}_S) \), which is shown to be asymptotically close to \( \log \mathcal{d}_A \) under the random relation model. (II) A concentration result of \( H(\mathcal{A}_S) \) to its expected value. (III) A combination of these bounds. In the next two subsections we provide a formal statement of the first two steps, and outline their proof. The full proof is rather long, and appears in [13].

5.2.1 The expected value of the entropy. In this section, we state a bound on the average mutual information \( \mathbb{E}[H(\mathcal{A}_S)] \) and outline its proof. Let us denote, for notational brevity,

\[
C(d) \overset{\text{def}}{=} \frac{2 \log(d)}{\sqrt{d}}.
\]

Proposition 5.4 (Bounds on the expected entropy). Assume that \( \mathcal{d}_A \geq \mathcal{d}_B \) and that \( \eta \geq 60d_A \). If \( S \) is chosen uniformly at random from one of the possible subsets of \([d_A] \times [d_B]\) of size \( \eta \) then

\[
0 \leq \log \mathcal{d}_A - \mathbb{E}[H(\mathcal{A}_S)] \leq C(d_B),
\]

where \( C(d) \) is as defined in (45). An analogous result hold for \( H(\mathcal{B}_S) \):

\[
0 \leq \log \mathcal{d}_B - \mathbb{E}[H(\mathcal{B}_S)] \leq C(d_A).
\]

We next present the main ideas of the proof of Proposition 5.4. As a first step, we identify that the expected value \( \mathbb{E}[H(\mathcal{A}_S)] \) is, in fact, a conditional entropy \( H(A \mid S) \), to wit,

\[
\mathbb{E}[H(\mathcal{A}_S)] = \sum_s \mathbb{P}[S = s] \cdot H(\mathcal{A}_S) = H(A \mid S).
\]

The crux of the proof of Proposition 5.4 requires lower bounding \( H(A \mid S) \). Nonetheless, to illuminate the challenge in the proof, it is insightful to first note that as conditioning reduces entropy [8, Theorem 2.6.5], and so

\[
H(A \mid S) \leq H(A).
\]

Using the symmetry of the distribution of the set \( S \), it follows that \( H(A) = \log \mathcal{d}_A \) (see [13] for a rigorous proof). So, Proposition 5.4 states that \( H(A \mid S) \) is close to its unconditional value \( H(A) \), up to \( C(d_B) \). In other words, we need to show that the conditioning (over the random variable \( S \)) only slightly reduces entropy in (49).

To further delve into the proof of this property, we closely inspect \( H(A \mid S) \). For any \((i, j) \in [d_A] \times [d_B] \), we define the random variable \( \mathcal{U}_S(i, j) \overset{\text{def}}{=} 1 \{ (i, j) \in S \} \), which indicates if the tuple \( (i, j) \) is in the random relation \( \mathcal{R}_S \) (see Definition 5.2). By symmetry, \( \mathbb{P}(\mathcal{U}_S(i, j) = 1) = \frac{\eta}{\mathcal{d}_A \mathcal{d}_B} \) for all \((i, j) \in [d_A] \times [d_B] \), and hence \( \{ \mathcal{U}_S(i, j) \}_{(i, j) \in [d_A] \times [d_B]} \) are identically distributed. The values \( \{ \mathcal{U}_S(i, j) \}_{(i, j) \in [d_A] \times [d_B]} \) uniquely determine \( S \), and so also the entropy \( H(\mathcal{A}_S) \). However, \( \{ \mathcal{U}_S(i, j) \}_{(i, j) \in [d_A] \times [d_B]} \) are dependent random variables, and such random variables are typically more difficult to handle than independent ones. Letting \( Y_S \equiv \mathcal{Y}_S(1) \overset{\text{def}}{=} \frac{1}{\mathcal{d}_B} \sum_{j \in [d_B]} \mathcal{U}_S(1, j) \), it can be shown that

\[
H(\mathcal{A} \mid S) = -\frac{\mathcal{d}_A \mathcal{d}_B}{\eta} \mathbb{E}[Y_S \cdot \log(Y_S)] + \log \frac{\eta}{\mathcal{d}_B}.
\]

Noting that \( f(t) \overset{\text{def}}{=} t \log(t) \) is a convex function on \( \mathbb{R}_+ \), one obtains from Jensen’s inequality that

\[
-\mathbb{E}[Y_S \cdot \log(Y_S)] \leq -\mathbb{E}[Y_S] \log \mathbb{E}[Y_S],
\]

and since \( \mathbb{E}(Y_S) = \frac{\eta}{\mathcal{d}_B} \), it immediately follows that

\[
H(A \mid S) \leq \log \mathcal{d}_A = H(A),
\]

as is already known from the conditioning reduces entropy property (49). From the above discussion, we deduce that in order to obtain a lower bound on \( H(A \mid S) \), which is close to \( H(A) = \log \mathcal{d}_A \), it is required to show that the Jensen-based bound in (51) is close to an equality. Trivially, if \( Y_S \) had been a deterministic quantity, then any Jensen-based inequality is satisfied with equality, and specifically (51). Continuing this line of thought, one expects that if \( Y_S \) is tightly concentrated around its expected value (i.e., “close” to being deterministic), then (51) approximately holds with equality. Indeed, such relations have been extensively explored via the functional entropy of a non-negative random variable \( X \), defined as

\[
\text{Ent}(X) \overset{\text{def}}{=} \mathbb{E}[\log(X)] - \mathbb{E}[(\log(X))^+].
\]

The functional entropy\(^3\) is non-negative, and is conveniently upper bounded via logarithmic Sobolev inequalities (LSIs) [5, Chapter 5]. Specifically, these inequalities bound \( \text{Ent}(X) \) by the Efron-Stein variance of \( X \) [5, Chapter 5], which in turn quantifies the concentration of \( Y \) around its expected value – low Efron-Stein variance implies tight concentration around the expected value, and thus low functional entropy by LSIs. Therefore, the proof addresses the bounding of \( \text{Ent}(Y_S) \). Nonetheless, LSIs are typically derived for functions of independent random variables, whereas here, as discussed, \( Y_S \equiv \frac{1}{\mathcal{d}_B} \sum_{j \in [d_B]} \mathcal{U}_S(1, j) \) is an average of dependent random variables. To address this matter, we define a new set of random variables \( \{ \mathcal{V}(j) \}_{j \in [d_B]} \), so that each \( \mathcal{V}(j) \) has the same marginal distribution as \( \mathcal{U}_S(1, j) \), but where the \( \{ \mathcal{V}(j) \}_{j \in [d_B]} \) are independent. In other words, \( \{ \mathcal{V}(j) \}_{j \in [d_B]} \) is a set of Bernoulli random variables for which \( \mathbb{P}(\mathcal{V}(j) = 1) = \frac{\eta}{\mathcal{d}_A \mathcal{d}_B} \), thus possibly asymmetric. We then define \( \mathcal{Y} \overset{\text{def}}{=} \frac{1}{\mathcal{d}_B} \sum_{j \in [d_B]} \mathcal{V}(j) \), and instead of directly bounding \( \text{Ent}(Y_S) \) as is required for the proof, we bound \( \text{Ent}(\mathcal{Y}) \) and the difference between the two functional entropies, to wit, we write

\[
\text{Ent}(Y_S) = \text{Ent}(\mathcal{Y}) + [\text{Ent}(Y_S) - \text{Ent}(\mathcal{Y})],
\]

and then separately bound each of the terms. Denoting \( \rho \overset{\text{def}}{=} \frac{d_A d_B}{\eta} - 1 \) (which is an upper bound on the relative number of spurious tuples), the proof shows that

\[
\text{Ent}(\mathcal{Y}) \leq \frac{2 \log(1/\rho)}{1 - \rho} \cdot \frac{1}{\mathcal{d}_B}.
\]

The proof of this property is based on a LSI for the asymmetric Bernoulli random variables \( \{ \mathcal{V}(j) \}_{j \in [d_B]} \) [5, Chapter 5], along with a careful bounding of the Efron-Stein variance of \( \mathcal{Y} \). The next term in the decomposition of \( \text{Ent}(Y_S) \) in (54) is absolutely bounded in the proof as

\[
\left| \text{Ent}(Y_S) - \text{Ent}(\mathcal{Y}) \right| \leq \frac{2 \log^2(\mathcal{d}_B)}{\mathcal{d}_B}.
\]

\(^3\)Not to be confused with the Shannon entropy of a random variable \( H(Y) \), see [5].
Summing the bounds on $\operatorname{Ent}(\tilde{Y})$ and $|\operatorname{Ent}(Y_S) - \operatorname{Ent}(\tilde{Y})|$ leads to a bound on $\operatorname{Ent}(Y_S)$, which in turn shows that the Jensen-bound in (51) is close to equality. This shows that $H(A | S) \leq \log d_A$ in fact approximately achieved, up to the defined vanishing term $C(d_B)$.

### 5.2.2 The concentration to the expected value of the entropy.

We next discuss the second step of the proof of Theorem 5.2. We state a concentration bound on

$$H(A_S) = \mathbb{E}[H(A_S)]$$

The next step is to replace the hypergeometric random variable $Z_S$ with a Poisson random variable $W$. And a larger multiplicative pre-factor $(21d_A^2)$ instead of just $d_B$. The next matter to address is that $g(t) = -t \log t$ is not a Lipschitz function since its derivative is unbounded for $t \uparrow 0$ as well as $t \downarrow 0$ (note that while $Z_S \leq d_B$ with probability 1, $W$ is unbounded). We first add the $t \downarrow 0$ case. Since $W$ is supported on integers, the minimal non-zero argument possible for $g(\frac{W}{\eta})$ is $1/\eta$. So, if we restrict $t \in [\frac{1}{\eta}, 1]$ then $g(t)$ is a Lipschitz function with semi-norm $\frac{1}{\eta} \log \eta$. Based on this observation we propose a function $\tilde{g}_\eta(t)$ which well approximates $g(t)$ on one hand, and is Lipschitz on the other hand. By an application of the triangle inequality, the term in (61) is upper bounded as

$$\frac{W}{\eta} - \mathbb{E}[H(A_S)] \leq \frac{W}{\eta} - \tilde{g}_\eta \left( \frac{W}{\eta} \right) +$$

$$\left| \mathbb{E}[Z_S] \right| \leq \frac{W}{\eta} - \tilde{g}_\eta \left( \frac{W}{\eta} \right) +$$

(62)

The first term in (62) is bounded as $1/\eta$ with probability 1 directly from the construction of the function $\tilde{g}_\eta(t)$. The last term in (62) is difficult to bound, and the proof mainly uses a Poisson LSI [5, Thm. 6.17] along with various approximation steps. The proof continues by bounding the probability that the middle term in (62) is larger than some value. The main tool for this bound in a concentration bound for Lipschitz functions of Poisson random variables. This bound is not used directly, since $\tilde{g}_\eta(t)$ is not Lipschitz over the entire real line. However, the argument $W/\eta$ is small enough with high probability, and thus belong to the Lipschitz continuous part of this function. Additional approximation arguments show that this suffices to obtain tight upper bound. The combination of the bounds for all three terms then establishes the proof of Theorem 5.2.

### 6 Conclusion

We show that the KL-Divergence is a useful measure for capturing the loss of an AJD with respect to the number of redundant tuples generated by the acyclic join. Our proposed random database model has allowed us to establish a high probability upper-bound on the percentage of redundant tuples, which coincides with the deterministic lower bound for large databases. Overall, our findings provide insights into the information-theoretic nature of AJD loss.

### Acknowledgments

The work of B.K. was supported by the US-Israel Binaational Science Foundation (BSF) Grant No. 2030983, and the work of N.W. was supported in part by the Israel Science Foundation (ISF), Grant No. 11782/22. The work was also supported by the Technion MLIS-TDSI Grant No. 86703064. The authors thanks Or Glassman for various numerical computations relatd to this research. N.W. thanks Nadav Merlin for a discussion on tail bounds for hypergeometric random variables and sampling without replacement.
REFERENCES

[1] Foto N. Afrati and Phokion G. Kolaitis. 2009. Repair checking in inconsistent databases: algorithms and complexity. In Database Theory - ICDT 2009, 12th International Conference, St. Petersburg, Russia, March 23-25, 2009. Proceedings. 31–41. https://doi.org/10.1007/978-3-642-00076-82489

[2] Catriel Beeri, Ronald Fagin, David Maier, Alberto O. Mendelzon, Jeffrey D. Ullman, and Mihalis Yannakakis. 1981. Properties of Acyclic Database Schemes. In Proceedings of the 13th Annual ACM Symposium on Theory of Computing, May 11-13, 1981, Milwaukee, Wisconsin, USA. 355–362. https://doi.org/10.1145/800076.802289

[3] Catriel Beeri, Ronald Fagin, David Maier, and Mihalis Yannakakis. 1983. On the Desirability of Acyclic Database Schemes. J. ACM 30, 3 (July 1983), 479–513. https://doi.org/10.1145/2402.322389

[4] Leopoldo E. Bertossi. 2011. Elements of Information Theory. Wiley-Interscience, Hoboken, NJ, USA.

[5] E. F. Codd. 1975. Recent Investigations in Relational Data Base Systems. In ACM Transactions on Database Systems, 3, 1 (Mar. 1978), 262–278. https://doi.org/10.1145/320557.320571

[6] E. F. Codd. 1971. Further Normalization of the Data Base Relational Model. IBM Research Report, San Jose, California RJ909 (1971).

[7] T. M. Cover and J. A. Thomas. 2006. Elements of Information Theory. Wiley-Interscience, Hoboken, NJ, USA.

[8] Tony T. Lee. 1987. An Information-Theoretic Analysis of Relational Databases - Part I: Data Dependencies and Information Metric. IEEE Trans. Software Eng. 13, 10 (1987), 1049–1061. https://doi.org/10.1109/TSE.1987.232847

[9] Tony T. Lee. 1987. An Information-Theoretic Analysis of Relational Databases - Part II: Information Structures of Database Schemas. IEEE Trans. Software Eng. 13, 10 (1987), 1061–1072.

[10] Robert J Serfling. 1974. Probability inequalities for the sum in sampling without replacement. The Annals of Statistics (1974), 39–48.

[11] Batya Kenig. 2019. Mining Approximate Acyclic Schemes from Relations. In Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14–19, 2020. ACM, 297–312. https://doi.org/10.1145/3318464.3380573

[12] Batya Kenig, Pranay Mundra, Guna Prasad, Babak Salimi, and Dan Suciu. 2020. Computing Machinery, New York, NY, USA, 114–126. https://doi.org/10.1145/2882903.2882939

[13] Batya Kenig and Nir Weinberger. 2022. Quantifying the Loss of Acyclic Join Dependencies. arXiv preprint arXiv:2210.14572 (2022).

[14] Mahmoud Abo Khamis, Hung Q. Ngo, XuanLong Nguyen, Dan Olteanu, and Maximilian Schleich. 2018. AC/DC: In-Database Learning Thunderstruck. In Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning, DEEM@SIGMOD 2018, Houston, TX, USA, June 15, 2018. 8:1–8:10. https://doi.org/10.1145/3209889.3209896

[15] Tony T. Lee. 1987. An Information-Theoretic Analysis of Relational Databases - Part I: Data Dependencies and Information Metric. IEEE Trans. Software Eng. 13, 10 (1987), 1049–1061. https://doi.org/10.1109/TSE.1987.232847

[16] Tony T. Lee. 1987. An Information-Theoretic Analysis of Relational Databases - Part II: Information Structures of Database Schemas. IEEE Trans. Software Eng. 13, 10 (1987), 1061–1072.

[17] Mark Levene and George Loizou. 2003. Why is the snowflake schema a good data warehouse design? Inf. Syst. 28, 3 (2003), 225–240. https://doi.org/10.1016/S0304-4379(02)00021-2

[18] M. Mitzenmacher and E. Upfal. 2017. Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis. Cambridge University Press.

[19] Dan Olteanu and Jakub Zavodny. 2012. Factorised representations of query results: size bounds and readability. In ICDT. ACM, 285–298.

[20] Maximilian Schleich, Dan Olteanu, and Radu Ciucanu. 2016. Learning Linear Regression Models over Factorized Joins. In Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016. 3–18. https://doi.org/10.1145/2882903.2882939

[21] Maximilian Schleich, Dan Olteanu, Mahmoud Abo Khamis, Hung Q. Ngo, and XuanLong Nguyen. 2019. A Layered Aggregate Engine for Analytics Workloads. In Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 – July 5, 2019. 1642–1659. https://doi.org/10.1145/3299869.3324961

[22] Robert J Serfling. 1974. Probability inequalities for the sum in sampling without replacement. The Annals of Statistics (1974), 39–48.

[23] Mihalis Yannakakis. 1981. Algorithms for Acyclic Database Schemes. In Proceedings of the Seventh International Conference on Very Large Data Bases - Volume 7 (Cannes, France) (VLDB ’81), VLDB Endowment, 82–94. http://dl.acm.org/citation.cfm?id=1260831.1260840