A tool framework for tweaking features in synthetic datasets

J.W. Zhang #1, Y.C. Tay #2

# School of Computing, National University of Singapore

1 jiangwei@u.nus.edu

2 dcstayyc@nus.edu.sg

Abstract—Researchers and developers use benchmarks to compare their algorithms and products. A database benchmark must have a dataset. To be application-specific, this dataset D should be empirical. However, D may be too small, or too large, for the benchmarking experiments. D must, therefore, be scaled to the desired size.

To ensure the scaled D is similar to D, previous work typically specifies or extracts a fixed set of features \( F \) \( \{ F_1, F_2, \ldots, F_n \} \) from D, then uses \( F \) to generate synthetic data for \( \tilde{D} \). However, this approach \( D \rightarrow F \rightarrow \tilde{D} \) becomes increasingly intractable as \( F \) gets larger, so a new solution is necessary.

Different from existing approaches, this paper proposes ASPECT to scale D to enforce similarity. ASPECT first uses a tweaking tool \( T \) to make sure \( D \) has the required feature \( \tilde{F}_1 \). ASPECT coordinates the tweaking of \( T_{1, \ldots, n} \) to \( D \), so \( \tilde{T}_k(\cdot)(\tilde{T}(D)) \ldots \) has the required features \( \tilde{F}_1, \ldots, \tilde{F}_n \).

By shifting from \( D \rightarrow F \rightarrow \tilde{D} \rightarrow \tilde{F} \), data scaling becomes flexible. The user can customise the scaled dataset with their own interested features. Extensive experiments on real datasets show that ASPECT can enforce similarity in the dataset effectively and efficiently.

I. INTRODUCTION

We have two motivations for introducing ASPECT:

Motivation 1: Benchmarks are ubiquitous in the computing industry and academia. Developers use benchmarks to compare products and algorithms, while researchers use them similarly in research.

For 20 odd years, the popular benchmarks for database management systems were the ones defined by the Transaction Processing Council (TPC) 1. However, the small number of TPC benchmarks are increasingly irrelevant to the myriad of diverse applications, and the TPC standardization process is too slow. This led to a proposal for a paradigm shift, from a top-down design of domain-specific benchmarks by committee consensus, to a bottom-up collaboration to develop tools for application-specific benchmarking.

A database benchmark must have a dataset. For the benchmark to be application-specific, it must start with an empirical dataset \( D \). This \( D \) may be too small or too large for the benchmarking experiment, so the first tool to develop would be for scaling \( D \) to a desired size.

Motivation 2: Apart from benchmarking, dataset scaling plays important roles in other fields as well. A start-up company with a small dataset may want a larger dataset for testing the scalability of their system architecture. On the other hand, an enterprise with a large dataset may want a scaled down version to provide quick answers to aggregation queries (averages, count, etc.).

Given this outlook, a tool that scales an empirical dataset \( D \) to a synthetic and similar \( \tilde{D} \) will be very appealing. This generation of artificial data is necessary if \( D \) is larger, and helpful if \( D \) is smaller or equal in size. For all cases: \( \tilde{D} \) must be similar to \( D \). Moreover, the similarity definition should be application-specific. It can be measured by graph properties, query results, etc., depends on the application.

A. Existing approach and the limitations

To ensure \( \tilde{D} \) is similar to \( D \), previous work typically follows the framework in Fig. 1. Each algorithm extracts a fixed set of features \( F = \{ F_1, F_2, \ldots, F_n \} \) from \( D \), then scales \( F \) to \( \tilde{F} \) as a predicting feature for the scaled dataset \( \tilde{D} \). \( D \) is finally synthesized based on \( \tilde{F} \). \( \tilde{F} \) here defines the similarity between \( D \) and \( \tilde{D} \); the more features in \( F \), the greater the similarity between \( D \) and \( \tilde{D} \).

For example, if \( D \) is a graph and \( F = \{ F_1, F_2 \} \), where \( F_1 \) is density and \( F_2 \) is number of triangles, then we would expect \( D \) and \( \tilde{D} \) are similar in terms of density and triangles. However, there are some limitations from the perspective of a developer and a user:

1) The developer faces the implementation reusability and scalability issue: Algorithm Implementation Reusability: Consider the scenario where one application developer implements an algorithm \( A_1 \) using the feature set \( \{ F_1, F_2 \} \). Later, another developer may find it more important for her application to preserve \( \{ F_2, F_3 \} \), where \( F_3 \) is the number of rectangles. So she implements another algorithm \( A_2 \) to preserve \( \{ F_2, F_3 \} \). However, a third developer might want to preserve \( \{ F_1, F_2, F_3 \} \); what should he do? In this case,
Algorithm Implementation Scalability: As mentioned previously, the more features in \( F \), the greater the similarity between \( D \) and \( \tilde{D} \). However, a large feature set dramatically increases the difficulty of designing an algorithm that maintains the features simultaneously. For example, if \( F = \{ F_1, F_2, F_3, F_4 \} \), where \( F_4 \) is the fraction of nodes with degree 1, then it is less likely one can design a single algorithm which preserves all 4 features. If we only consider degree distributions as features, then it is already NP-hard to decide whether there exists a graph satisfying certain degree distributions [5].

2) The user does not have a choice of the features: In the current framework, once an algorithm is implemented, the features are fixed. Consider the same example used above, \( A_1 \) is implemented to preserve \( F = \{ F_1, F_2 \} \), and \( A_2 \) is implemented to preserve \( F = \{ F_3, F_4 \} \). The user can only choose to preserve \( \{ F_1, F_2 \} \) or \( \{ F_3, F_4 \} \), but not the union.

B. Overcoming the limitations

In this paper, we propose ASPECT, a flexible framework for synthetic data scaling. Unlike existing approaches, ASPECT takes the following two steps as illustrated in Fig.2.

Step1: Use a size-scaler \( S_0 \) to scale \( D \) to \( \tilde{D}_0 \) of desired size.

Step2: For the desired feature set \( \{ \tilde{F}_1, \tilde{F}_2, \ldots, \tilde{F}_k \} \), apply independently developed tools \( T_1, T_2, \ldots, T_n \) on \( \tilde{D}_0 \) in order. Each tool \( T_i \) generates a dataset \( \tilde{D}_i \) by adjusting \( D_i-1 \). After the adjustment of \( T_i \), \( \tilde{D}_i \) satisfies \( \{ \tilde{F}_1, \tilde{F}_2, \ldots, \tilde{F}_i \} \). Note that Step2 does not depend on the size-scaler in Step1. We call this tweaking \( \tilde{D}_i-1 \) by tool \( T_i \). The final dataset is \( \tilde{D} \).

For the above-mentioned limitations in Sec. I-A, ASPECT resolves them with ease: For implementation reusability, each feature tweaking tool is independently developed. Once a tweaking tool \( T_i \) for feature \( \tilde{F}_i \) is implemented, the user can apply \( T_i \) together with other tweaking tools whenever it is needed. No re-coding! For implementation scalability, to preserve the feature set with \( n \) features, the developer just needs to implement \( n \) tweaking tools, instead of hardcoding all \( n \) features into a single piece of software. And each tweaking tool \( T_i \) tweaks the feature \( \tilde{F}_i \). For the issue of feature choice, once tweaking tools \( T_1, T_2, T_3 \) and \( T_4 \) are implemented for \( \tilde{F}_1, \tilde{F}_2, \tilde{F}_3 \) and \( \tilde{F}_4 \) respectively, the user can choose \( \{ T_2, T_3 \} \) to get \( \{ \tilde{F}_2, \tilde{F}_3 \} \), or \( \{ T_1, T_3, T_4 \} \) to get \( \{ \tilde{F}_1, \tilde{F}_3, \tilde{F}_4 \} \), etc.

Hence, to enforce greater similarity in the scaled dataset \( \tilde{D} \), we just need to apply more tweaking tools. We envision having developers from the database community contributing tools \( T_i \) to a repository for tweaking synthetic datasets. Then ASPECT will have more tools for the user to customise the scaled datasets. This would go some way towards realising the suggested paradigm shift to a bottom-up collaboration for application-specific benchmarking.

However, the tools require some coordination, since some changes to \( D_i-1 \) by one tool \( T_i \) may be undone by another tool \( T_j \). Moreover, tools are developed independently by different developers. Different developers might have different tweaking techniques. To ensure a tool is compatible with ASPECT, ASPECT must explicitly standardise types of modifications that could be made on a dataset while tweaking.

C. Overview

To summarize, our contribution in this paper are fourfold:

1. We propose ASPECT, a framework for flexible application of tweaking tools to enforce target features in synthetic dataset.
2. We present results from extensive experiments on real datasets, to verify that ASPECT can enforce similarity in the dataset effectively and efficiently.
3. We present necessary and sufficient conditions, and tweaking algorithms, for three new complex features.
4. We state Feature Tweaking Bound and Order Problems that offer a rewarding challenge for research on dataset tweaking.

We first introduce ASPECT architecture in Sec.II followed by three new complex features that serve to illustrate the ASPECT framework in Sec.III. One of them concerns inter-column and inter-row correlation induced by implicit relationships in a social network dataset; we thus provide here a solution to a problem highlighted previously [32]. See IV describes the datasets and similarity measures used in the experiments, and the results are presented in Sec.V Sec.VI points out some limitations and insights of ASPECT. Related work is surveyed in Sec.VII before Sec. VII concludes with a summary.

II. ASPECT ARCHITECTURE

As shown in Fig 2 an input dataset is first scaled by the size-scaler \( S_0 \) which returns a scaled dataset \( D_0 \) of the desired size. Note that \( S_0 \) could be any tool which guarantees the number of tuples in each table generated is as expected and there are no invalid foreign key values. For example, \( S_0 \) may be DSCALER [30], or it could be ReX [8]; we will show that ASPECT is able to preserve the features well for both DSCALER and ReX in Sec.V The choice of \( S_0 \) is outside the scope of this paper. After the dataset is resized, ASPECT then coordinates the application of tools \( T_i \) on \( D_i-1 \), to make sure the feature \( \tilde{F}_i \) is reflected in the tweaked dataset \( \tilde{D}_i \). In the tweaking process, there are a few issues:

1. How do we get the target feature \( \tilde{F}_i \)?
12. Given an target feature $\tilde{F}_t$, how can we tweak $\tilde{D}_{t-1}$ to ensure that the tweaked dataset $\tilde{D}_t$ contains $\tilde{F}_t$?
13. Given $\tilde{D}_{n-1}$ already contains $\tilde{F}_1, \tilde{F}_2, \ldots, \tilde{F}_{n-1}$, how can we maintain $\tilde{F}_1, \tilde{F}_2, \ldots, \tilde{F}_{n-1}$ while tweaking $\tilde{F}_n$?
14. Tools are developed independently by different developers. How can we make sure these independently developed tools are compatible with ASPECT?

A. ASPECT flow

To address the above 4 issues, we illustrate the tweaking process for, say, $\mathcal{T}_4$.

Step1. ASPECT first calls the tool $\mathcal{T}_4$ to start tweaking, then calls previously applied tools $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$ to start preparation.

Step2.1. $\mathcal{T}_4$ then finds the target feature $\tilde{F}_4$ by calling its Feature Generator.

Step2.2. $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$ call their respective Feature Calculators to calculate the corresponding features $\tilde{F}_1, \tilde{F}_2, \tilde{F}_3$. This step is concurrent with Step 2.1.

Step3. $\mathcal{T}_4$ starts the Tweaking Algorithm. Every time $\mathcal{T}_4$ needs to modify $\tilde{D}_3$, $\mathcal{T}_4$ sends the intended modification to ASPECT for validation.

Step4. ASPECT calls $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$ to confirm the modification with their own Feature Validators. ASPECT summarizes the feedback from $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$ and replies “yes/no” to $\mathcal{T}_4$.

Step5.1. If the reply is “yes”, ASPECT modifies $\tilde{D}_3$ and tells $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$ to update the feature statistics by using their Feature Updaters.

Step5.2. If the reply is “no”, ASPECT tells $\mathcal{T}_4$ to find an alternative modification.

Step6. Repeat from Step3 until $\mathcal{T}_4$ halts.

In the flow above, I1 is addressed by Step2.1 (Feature Generator), I2 is addressed by Step3 (Tweaking Algorithm) and I3 is handled in Step2.2, Step4 and Step5 (Feature Calculator, Feature Validator, Feature Updater). We will explain I1, I2, and I3 in Sec[II-B] and I4 in Sec[II-C] in detail.

B. Tweaking tool component

We now explain how each individual tool $\mathcal{T}_i$ should be implemented. Each $\mathcal{T}_i$ must have at least 5 components.

Feature Generator: This module generates the target feature statistics for the tweaked dataset. Such a generation can be done in 3 ways: (i) User input: The user might have their own target feature statistics for the scaled dataset. For example, the user might want to specify the number of males in the population. Hence, the user can manually input target features. (ii) Developer generation: When a developer implements the tweaking tool for a specific feature, the developer has a better understanding of how the feature changes while the dataset scales. Therefore, the developer can provide the feature generation tool for his/her own developed feature. (iii) Generate through historical data: Apart from the previous two methods, statistical tools can be developed for certain features for general purposes, e.g. frequency distribution for attribute values. One can first take chronological snapshots of the dataset (if applicable), $\tilde{D}_1, \tilde{D}_2, \ldots, \tilde{D}_k$, then extract the feature $\tilde{F}_i$ from each snapshot dataset. Next, apply data fitting methods on $\tilde{F}_i$ to fit $\tilde{F}_i$ into different statistical models, e.g. Poisson distribution. Once the best matching model is learned, say, Poisson distribution, we can learn how the shape parameter $\lambda$ varies as the dataset grows. Hence, we can get the target feature $\tilde{F}_i$. Such an approach is orthogonal to this paper, and will be elaborated in a separate paper.

Tweaking Algorithm: It tweaks the dataset $\tilde{D}_{t-1}$ to make sure that $\tilde{D}_t$ has the target feature $\tilde{F}_t$ at the end of tweaking. Note that it is not trivial to provide a tweaking algorithm for a complex feature, e.g. linear feature as presented in Sec. III. The developer has to code the tweaking algorithm. Moreover, the tweaking algorithm can only modify the dataset through similar operations presented in Fig. 3.

Feature Calculator: It calculates the feature statistics for $\tilde{F}_i$ from a given dataset.

Feature Validator: It checks whether a proposed tuple insertion/deletion/replacement affects some existing feature. Assuming 1 modification on a tuple is needed when tweaking feature $\tilde{F}_n$ on dataset $\tilde{D}_{n-1}$. Modifying either $t_1$ or $t_2$ will satisfy $\tilde{F}_n$, and modifying $t_1$ changes a previously tweaked feature $\tilde{F}_1$, but modifying $t_2$ does not. Then, modify $t_2$ instead of $t_1$.

At times, it is too strict when a tuple modification is allowed only if no previously tweaked feature affected. For example, $\tilde{F}_1 = \{\text{more than half of the customers are men}\}$ and $\tilde{F}_2 = \{\text{more than half of the customers are women}\}$. These 2 features are contradictory, so one of it has to be violated. Hence, the validation needs to be relaxed. In this paper, a tuple modification is allowed if the resulting errors for all previously tweaked features are less than 5%, which is the threshold $\epsilon_{\text{threshold}}$. Consider the example used previously: if error $< \epsilon_{\text{threshold}}$, then modification on $t_1$ is allowed as well; however, if error $\geq \epsilon_{\text{threshold}}$, then only $t_2$ can be chosen.

In the worst case, if no tuple modification can satisfy previously tweaked feature’s $\epsilon_{\text{threshold}}$, ASPECT allows more tuple modification by relaxing the validation on fewer features. So, some of the feature’s error might be larger than $\epsilon_{\text{threshold}}$.

Feature Updater: After each modification of the tuples, the Feature Updater updates the tweaked features’ statistics.

Fig. 3: Feature Updater Interface

Under ASPECT, it is the tool developers’ responsibility to ensure that the above requirements are correctly implemented and adhere to ASPECT’s structure. If the developer does not, say, properly validate the modifications for Feature Validator, then it is highly likely the corresponding feature will be affected by subsequent tweaks.

C. ASPECT compatibility guarantee

As mentioned in I4, we need to guarantee each individually developed tool can be used in ASPECT. So, we standardise
Fig. 4: Demonstration for Updator Interface: The update is on table with tableID 5. Step1 demonstrates `deleteValues`; Step2 demonstrates `insertValues`; Step3 demonstrates `replaceValues`. The affected values are highlighted using red colour, and the operation for each step is listed at the bottom.

### III. Example Features For Demonstration

In ASPECT, the more tools we apply, the more features we can preserve. As pointed out previously, for some feature $F_j$, it is inevitable that some previously tweaked feature $F_i$ may be affected when tweaking $F_j$. The concern is how much $F_i$ is affected while tweaking $F_j$ under ASPECT.

In this paper, we run experiments to demonstrate how $F_i$ is affected while tweaking $F_j$ empirically. We will apply 3 tools $T_1, T_2, T_3$ sequentially on $D_0$ to fix the features $F_1, F_2, F_3$. At the end of the tweaking, we will examine how well the three features are preserved. We propose three important and complex features: **linear**, **coappear**, **pairwise** as example features. These features are selected based on two criteria:

**Popularity:** Our ultimate goal is to build an application-specific system for dataset scaling. Hence, it only makes sense if the features are widely used. The features we consider are used widely in the literature [1], [14], [18], [29].

**Complexity:** Since we want to check how well ASPECT can maintain features if tools can undo previously applied tools. Hence, the features we use for demonstration should be complex and affect each other. Simple features such as “# of null values in each table”, “# of tuples in each table” are easy to tweak. To avoid presenting a strawman test, we skip such simple features in this paper. Nevertheless, these simple feature tweaking tools are already implemented in ASPECT. The user can apply such simple features if needed.

To the best of our knowledge, we are the first to publish these 3 features within one dataset.

In this section, we make the following assumption to avoid confusion. If $T'$ references $T$ (denoted $T' \rightarrow T$), it does so via one foreign key constraint only. This assumption can be easily relaxed. Also, $t_1 \rightarrow t_2, t_1 \in T_1, t_2 \in T_2$ means tuple $t_1$ references $t_2$.

### A. Linear feature

Applications are often interested in computing $T_{k[i]} \cdots \cdots \rightarrow T_1$ for some reference chain $T_k \rightarrow \cdots \rightarrow T_1$. For example, to count the number of (distinct) movies with reviews that are commented on by users, one may need to take the join of a reference chain from comments to reviews to movies. This is what we call a **linear** feature.

Fig. 5 illustrates the concept of a **linear** feature. For any reference chain $T_k \rightarrow \cdots \rightarrow T_1$, the linear feature describes how one tuple $t_i \in T_i$ is transitively referenced by other tuples $t_j \in T_j$, for any $j > i$. In Fig. 5, $a_2 \in T_A$ is directly referenced by $b_2, b_3 \in T_B$, and indirectly referenced by $c_1, c_2, c_3 \in T_C$. However, $a_2$ is not indirectly referenced by any tuple in $T_D$.

Variants of linear joins are widely used to generate query result approximations, such as database sampling [14] and positive/false-negative cases for these types of validations, but we believe the effect is minor. In Sec. [III-B] we need to modify $k$ tables simultaneously. To fit into ASPECT, we validate/modify the tables one by one. Experiments in Sec. [V] show this approach is effective.

| Table ID | PK | C1 | C2 | C3 |
|----------|----|----|----|----|
| 1        | 1  | 1  | 1  | 1  |
| 2        | 1  | 1  | 1  | 1  |
| 3        | 1  | 1  | 1  | 1  |

| Table ID | PK | C1 | C2 | C3 |
|----------|----|----|----|----|
| 1        | 1  | 1  | 1  | 1  |
| 2        | 1  | 1  | 1  | 1  |
| 3        | 1  | 1  | 1  | 1  |

| Table ID | PK | C1 | C2 | C3 |
|----------|----|----|----|----|
| 1        | 1  | 1  | 1  | 1  |
| 2        | 1  | 1  | 1  | 1  |
| 3        | 1  | 1  | 1  | 1  |

| Table ID | PK | C1 | C2 | C3 |
|----------|----|----|----|----|
| 1        | 1  | 1  | 1  | 1  |
| 2        | 1  | 1  | 1  | 1  |
| 3        | 1  | 1  | 1  | 1  |

| Table ID | PK | C1 | C2 | C3 |
|----------|----|----|----|----|
| 1        | 1  | 1  | 1  | 1  |
| 2        | 1  | 1  | 1  | 1  |
| 3        | 1  | 1  | 1  | 1  |
Fig. 5: **Linear feature.** There are 4 tables $T_D \rightarrow T_C \rightarrow T_B \rightarrow T_A$. Each node in the tree is a tuple in the table. $a_i$ represents the $ith$ tuple in $T_A$. $b_1, c_1, d_1$ are defined similarly. $a_3$ is referenced by $T_B, T_C, T_B$ and $a_2$ is only referenced by $T_B, T_C$. $H_{eg}$ is the corresponding linear join matrix.

The database generation process. This paper presents $T_{linear}$, an algorithm to tweak a dataset so it accurately scales the size of linear joins.

**Definition 1:** A tuple $t_1 \in T_1$ is a root of $T_k \rightarrow \cdots \rightarrow T_1$ if there are tuples $t_2 \in T_2, \ldots, t_k \in T_k$ such that $t_k \rightarrow \cdots \rightarrow t_1$. Let $S_{i,j}$ be the set of roots of $T_j \rightarrow \cdots \rightarrow T_i$, and $h_{j,i} = |S_{j,i}|$.

In Fig. 5, $a_2$ is a root of $T_C \rightarrow T_B \rightarrow T_A$, but not a root of $T_D \rightarrow T_C \rightarrow T_B \rightarrow T_A$. If $T_1 = T_A, T_2 = T_B, T_3 = T_C$ and $T_4 = T_D$, then $S_{4,2} = \{b_4, b_3\}$, so $h_{4,2} = 2$. The $h_{j,i}$ values form a matrix, as follows:

**Definition 2:** For a maximal chain $T_k \rightarrow \cdots \rightarrow T_1$, define its **linear join matrix** as a lower triangular matrix

$$H = \begin{bmatrix} 0 & & & 0 \\ h_{2,1} & 0 & & \\ h_{3,1} & h_{3,2} & \ddots & \\ \vdots & \vdots & \ddots & \ddots \\ h_{k,1} & h_{k,2} & \cdots & h_{k,k-1} & 0 \end{bmatrix}$$

$T_k \rightarrow \cdots \rightarrow T_1$ is maximal if there is no $T_{k+1}$ such that $T_{k+1} \rightarrow T_k \rightarrow \cdots \rightarrow T_1$ and $T_k \rightarrow \cdots \rightarrow T_1$ for all $k + 1$. In Fig. 5, there are 2 roots $a_2, a_3$ for $T_C \rightarrow T_B \rightarrow T_A$, and 3 roots $a_1, a_2, a_3$ for $T_B \rightarrow T_A$, so $h_{3,1} = 2, h_{2,1} = 3$.

Let $H$ be a linear join matrix in some $D_i$, before tweaked by $T_{linear}$ and $\tilde{H}$ the target linear join matrix. $T_{linear}$ tweak $H$ to become $\tilde{H}$. There are two concerns:

1) Is it possible to tweak $H$ to $\tilde{H}$? (necessary conditions)
2) How to tweak $H$ to $\tilde{H}$? (sufficient conditions)

We first address concern 1 using the following theorem.

**Theorem 1:** [necessity] Let $H$ be the linear join matrix of $T_k \rightarrow \cdots \rightarrow T_1$ before tweaked by $T_{linear}$, and $\tilde{H}$ be the target linear join matrix. $H$ can be tweaked to $\tilde{H}$ only if

- (L1) $\tilde{h}_{j,i} \leq \min_{1 \leq i < j \leq k} |T_i|$ for all $1 \leq i < j \leq k$.
- (L2) $\tilde{h}_{i+1,i} \geq h_{i+1,i} \geq \cdots \geq h_{k,i}$ for all $1 \leq i \leq k - 1$.
- (L3) $\tilde{h}_{j,1} \leq h_{j,2} \leq \cdots \leq h_{j,j-1}$ for all $2 \leq j \leq k$.
- (L4) $\tilde{h}_{j,i+1} - h_{j,i+1,i} \geq \tilde{h}_{j,i+1,i} - h_{j,i+1,i}$ for all $1 \leq i < j - 1 < k - 1$.

Next, we describe how $T_{linear}$ tweaks $H$ to $\tilde{H}$. $T_{linear}$ tweaks $H$ row by row. For $ith$ row, $T_{linear}$ then tweaks the row entry by entry. $T_{linear}$ first does **leadingAdjust**, tweaking $h_{i,1}, h_{i,2}, \ldots, h_{i,i-1}$ to $h_{i,1}, \ldots$. $T_{linear}$ then does **nonLeadingAdjust:** it tweaks $h_{i,1}, \ldots$ to $h_{i,2}, h_{i,3}, \ldots$ to $h_{i,1}, \ldots$. Instead of providing a formal proof, we present an example of tweaking from $H_{eg}$ to $H_{exp}$ in Fig. 6 and attach the proofs in the appendix.

**Second Row:** We are expecting one less root for $T_B \rightarrow T_A$. $T_{linear}$ chooses an existing root, say $a_1$, and plucks all its descendants ($b_1$) and attach them to some other root, say $a_2$. After such modification, we will have 2 roots $a_2, a_3$ for $T_B \rightarrow T_A$. This is reflected in step 1.

**Third Row:** No modifications are needed for the first entry. For the second entry, one more root for $T_C \rightarrow T_B$ is expected. Hence, we pluck $c_1$ from $b_2$ and attach $c_1$ to $b_1$. This completes the tweaking for the second row and it is reflected in step 2.

Fig. 6: **Tweaking demonstration for $T_{linear}$.** The red dotted lines are the modifications made on the dataset. The linear join matrix after each step is presented at the bottom.
Fig. 7: **Coappear feature.** There are 3 tables $T_A, T_B, T_C$ referencing to the same tables $T_K, T_H$. $(k_1, h_2)$ appeared together in $T_A, T_B$ for 3 times, in $T_C$ for 1 time.

**Fourth Row:** For the first entry, we expect 1 more root for $T_D \rightarrow T_C \rightarrow T_B \rightarrow T_A$, say $a_2$. Hence, we pluck $d_1$ from $c_4$ and attach $d_1$ to $c_1$. Now we have 2 roots $a_2, a_3$ for $T_D \rightarrow T_C \rightarrow T_B \rightarrow T_A$, and the last row becomes $(2, 3, 3, 0)$. This is reflected in step 3. For the second entry, we expect 2 more roots for $T_D \rightarrow T_C \rightarrow T_B$. Hence, we pluck $d_2$ from $c_4$ and attach it to $c_2$. This is reflected in step 4, and the last row is $(2, 4, 4, 0)$ now. Lastly, we pluck $d_4$ from $c_5$ and attach it to $c_3$ which ends the tweaking. This is reflected in step 5.

When tweaking the $i$th row, $T_{i \text{linear}}$ always plucks the tuples in $i$th table and attaches them to the $(i - 1)$th table. $T_{i \text{linear}}$ never re-modify the entries in previously tweaked rows, which gives some intuition that the tweaking is always possible.

So far, we only considered tweaking one linear join matrix. In general, a dataset can have multiple overlapping reference chains. Suppose we have already tweaked the matrix for $T_1 \rightarrow T_2 \rightarrow T_3 \rightarrow T_1$, then tweak the matrix for an overlapping $T_1 \rightarrow T_3 \rightarrow T_2$. This can undo the tweaking for $T_1 \rightarrow T_3 \rightarrow T_1$.

The issue is not just for overlapping linear joins but, in general, applies to any pair of tweaking algorithms. For example, running $T''$ (e.g. $T_{\text{pair-wise}}$) after $T'$ (e.g. $T_{\text{coappear}}$) can undo the work done by $T'$. We do not have a solution yet. Instead, we adopt the heuristic as presented in Sec. II-B.

**B. Coappear feature**

Fig. 7 illustrates the concept of a **coappear** feature. The tables $T_A$, $T_B$ and $T_C$ may be for comment, share and like in a social network service, referencing tables $T_K$ and $T_H$ for post and users. Thus, the same $(\text{postID, userID})$ may appear multiple times in the same table and in multiple tables. This **coappear** feature can be used for user profiling: e.g. if Alice comments, shares and likes a post about volunteerism many times, it is more likely that Alice is interested in volunteer work [29]. Other examples include group theme prediction [10] and on-line recommendation [13].

Tweaking is done via tweaking a frequency distribution that captures the correlation in foreign key appearances:

**Definition 3:** Suppose $T_1, \ldots, T_k$ reference the same tables $T'_1, \ldots, T'_m$ and $b_1, \ldots, b_m$ coappear as foreign keys $v_i$ times in $T'_i, \ldots, v_k$ times in $T'_k$. If there are $n$ such $(b_1, \ldots, b_m)$, then $\xi_{T_1, \ldots, T_k}(v_1, \ldots, v_k) = n$. We call $(v_1, \ldots, v_k)$ a **coappear vector** and $\xi_{T_1, \ldots, T_k}$ the **coappear distribution**. To simplify notation, we refer to $\xi_{T_1, \ldots, T_k}$ as $\xi$ if there is no ambiguity.

In Fig.7, $(k_1, h_2)$ appears 3 times in $T_A$, 3 times in $T_B$, and 1 time in $T_C$, so $\xi(3, 3, 1) = 1$. Further, $(k_2, h_3)$ and $(k_3, h_1)$ each appears 1 time in $T_A$, 1 time in $T_B$ and 2 times in $T_C$, so $\xi(1, 1, 2) = 2$.

Like for linear joins, we present necessary and sufficient conditions for tweaking the coappear distribution:

**Theorem 2:** [**necessity**] Suppose tables $T_1, \ldots, T_k$ reference the same tables $T'_1, \ldots, T'_m$ in some $D_i$. $\xi$ is the coappear distribution before being tweaked by $T_{\text{coappear}}$, and $\tilde{\xi}$ is the target coappear distribution. $\xi$ can be tweaked to $\tilde{\xi}$ only if:

\[
(C1) \sum_{v_i} v_i \xi(v) = |T_i| \quad \text{for } 1 \leq i \leq k
\]

\[
(C2) \sum_{v} \tilde{\xi}(v) = \prod_{i=1}^{m} |T'_i|.
\]

**Proof:** (C1) $\xi(v)$ is the number of different foreign key tuples $(b_1, \ldots, b_m)$ with coappear vector $v$. Hence, each $(b_1, \ldots, b_m)$ appears $v_i$ times in $T_i$, so $\sum_{v} v_i \xi(v) = |T_i|$.\n\
\[
(C2) \quad \sum_{v} \tilde{\xi}(v) = \prod_{i=1}^{m} |T'_i|.
\]

Next, we explain how $T_{\text{coappear}}$ tweaks $\xi$ to $\tilde{\xi}$. Let $\xi = \xi - \tilde{\xi}$, $\Delta^+ = \{ v | \xi^+(v) > 0 \}$, $\Delta^- = \{ v | \xi^-(v) < 0 \}$. $T_{\text{coappear}}$ works as follows:

For each $v = (v_1, \ldots, v_k) \in \Delta^+$, it adds $|\xi^+(v)|$ more foreign key tuples $(b_1, \ldots, b_m)$, each appearing $v_i$ times in $T_i$. It does this by looping $|\xi^+(v)|$ times, and in each iteration:

**CoappearVectorRetrieve:** Pick the closest coappear vector $v' = (v'_1, \ldots, v'_k) \in \Delta^+$, using Manhattan distance.

**TupleRetrieve:** There may be multiple $b = (b_1, \ldots, b_m)$ foreign key tuples with coappear vector $v'$ (e.g. in Fig.7 $(k_2, h_3)$ and $(k_3, h_3)$ both have $v' = (1, 1, 2)$). For each $v'$, choose one such $b$.

**Tuple Modification:** Tuples are tweaked as follows: For $1 \leq i \leq k$, if $v_i' - v_i > 0$, remove $v_i - v_i'$ tuples with foreign key values $b$ from $T_i$; if $v_i' - v_i < 0$, add $v_i - v_i'$ tuples with foreign key values $b$ into $T_i$.

**StatsUpdate:** Update $\xi^+(v) \times 1$ and $\xi^-(v') \times -1$.

The job is done when the loop terminates. We can prove that the necessary conditions are sufficient for the tweaking.

**Theorem 3:** [**sufficiency**] Suppose tables $T_1, \ldots, T_k$ reference the same tables $T'_1, \ldots, T'_m$. Let $\xi$ be the coappear distribution in some $D_i$ before tweaking and $\tilde{\xi}$ the target coappear distribution. If $\xi$ satisfies the necessary conditions in Theorem 2 then $T_{\text{coappear}}$ tweaks $\xi$ to become $\tilde{\xi}$.

The formal proof is provided in the appendix. In a dataset, we might have multiple coappear distributions. Suppose $T_D$ and $T_E$ reference $T_A$ and $T_B$, while $T_G$ and $T_H$ reference $T_B$ and $T_C$, so there are two coappear distributions: $\xi_{T_D, T_E}$ and $\xi_{T_G, T_H}$.
For each type of user have multiple instantiations: tables: involves both inter-column and inter-row correlation. 

A user-to-user tie is captured by the following distribution:

\[ (x, y) \in \Theta^+ \text{ such that } x \neq y \]  

where \( x, y \) tuples also represent \( k \) \( \{v_i, v_j\} \) pairs, so \( \tilde{\rho}_R(x, y) = k \).

(P2) As above, for each \( \tilde{\rho}_R(x, y) = k \), there are \( k \) \( \{u_i, v_i\} \) pairs, and each pair has \( x+y \) tuples in \( \text{response2post} \). These \( k(x+y) \) tuples are double-counted by \( \tilde{\rho}_R(x, y) \), so we get the equality in (P2).

(P3) Similarly, there are \( |U|(|U| - 1) \) user pairs, and each is counted once by \( \tilde{\rho}_R(x, y) \), so (P3) follows.

Next, we explain how \( \mathcal{T}_{\text{pairwise}} \) tweaks \( \rho \) to \( \tilde{\rho} \). Let \( \rho^*_R = \rho_R - \tilde{\rho}_R \), \( \Theta^+ = \{(x, y) | \rho^*_R(x, y) > 0\} \) and \( \Theta^- = \{(x, y) | \rho^*_R(x, y) < 0\} \). \( \mathcal{T}_{\text{pairwise}} \) loops through each \( \text{response2post} \) table \( R \). For each \( (x, y) \in \Theta^- \), it adds \( |\rho^*_R(x, y)| \) pairs \( \{u_i, v_i\} \), where user \( u_i \) has \( x+y \) response2post tuples in \( R \) referencing \( v_i \)’s post. It does this by looping \(|\rho^*_R(x, y)|\) times, and in each iteration:

**PairwiseVectorRetrieve:** Pick \( v' = (x', y') \in \Theta^+ \) that is closest to \( (x, y) \) by Manhattan distance.

**TupleModification:** Choose users \( u_i \) and \( v_i \) with pairwise vector \((x', y')\) and tweak \( u_i \)’s responses to \( v_i \)’s post, as follows: If \( x < x' \), then \( u_i \) has \( x' - x \) more responses to \( v_i \)’s post than desired, so \( \mathcal{T}_{\text{pairwise}} \) randomly chooses and removes \( x' - x \) such responses. If \( x > x' \), we add \( x - x' \) responses from \( u_i \) on \( v_i \)’s post. If \( v_i \) has no post, we artificially create a post for \( v_i \). To do this, we pick another user \( w_i \) who has more than 1 post and pick a post \( p_{w_i} \) with minimum responses among \( w_i \)’s posts; we make \( p_{w_i} \) a post by \( v_i \), and shift the responses to \( p_{w_i} \) to other posts by \( w_i \). If (rare case) all users have at most 1 post, we will make a new post \( p \) for \( v_i \), and add \( x - x' \) responses to \( p \). We similarly tweak \( v_i \)’s responses to \( u_i \)’s post.

**StatsUpdate:** Increase \( \rho^*_R(x, y) \) and \( \rho^*_R(y, x) \) by 1 and decrease \( \rho^*_R(x', y') \) and \( \rho^*_R(y', x') \) by 1.

We can prove the above mentioned conditions in Theorem 4 are sufficient for \( \mathcal{T}_{\text{pairwise}} \) tweaks \( \rho_R \) to \( \tilde{\rho}_R \) by Theorem 5.

**Theorem 5:** [Sufficiency] For each \( \text{response2post} \) table \( R \) in some \( \mathcal{D}_R \), \( \rho_R \) is the pairwise distribution before tweaking and \( \tilde{\rho}_R \) is the target pairwise distribution. If \( \tilde{\rho}_R \) satisfies the necessary conditions in Theorem 4 then \( \mathcal{T}_{\text{pairwise}} \) tweaks \( \rho_R \) to \( \tilde{\rho}_R \). Moreover, the extra tuples added to the \( \text{post} \) table \( P \) is at most \( |U| - |P| \), where \( U \) is the \( \text{user} \) table.

The formal proof is provided in the appendix. Since \( \text{response2post} \) can have several instantiations (e.g. share, like, etc.), a social network dataset can have multiple pairwise distributions, but they can be tweaked independently. For example, suppose a \( \text{post} \) table \( P \) has two \( \text{response2post} \) tables \( R_1 \) and \( R_2 \), and \( \rho_{R_1} \) is tweaked to \( \tilde{\rho}_{R_1} \) first. When tweaking \( \rho_{R_2} \), we only modify the tuples in \( R_2 \), so it does not affect the tweaked \( \tilde{\rho}_{R_1} \). Moreover, adding tuples in \( P \) does not affect \( \tilde{\rho}_{R_1} \) as well.

For the above mentioned tweaking tools, they modify the
dataset by calling the functions in Sec. II-C. \( T_{\text{linear}} \) modifies the dataset through the operation replaceValues. \( T_{\text{coappear}} \) and \( T_{\text{pairwise}} \) modify the dataset through the operations deleteValues, insertValues.

IV. EXPERIMENT SETUP

In the following, each experiment is run on a Linux machine with 64GB memory and an Intel Xeon 2.4GHz processor. We now describe the datasets and similarity measures used in our experiments. ASPECT is implemented in Java.

A. Datasets

In this paper, due to the space constraint, we only present experiments on Xiami\(^2\) Xiami contains music-related data with 28 tables and more than 90M tuples. Reader can refer to the appendix for experiments on three more datasets, DoubanBook, DoubanMusic and DoubanMovie. Each dataset is larger than 10GB originally. However, there are columns, e.g. song_name, movie_name, that are irrelevant to the experiments. We do not want to exaggerate ASPECT's capability of handling big datasets. We hence purposely filter out those irrelevant columns and only conduct experiment on the relevant columns.

We take 6 snapshots of each dataset, \( \mathcal{D}_1 \subset \mathcal{D}_2 \subset \mathcal{D}_3 \subset \mathcal{D}_4 \subset \mathcal{D}_5 \subset \mathcal{D}_6 \). For each \( \mathcal{D}_i \), ASPECT takes \( \mathcal{D}_1 \) as input, and first uses a size-scaler to scale \( \mathcal{D}_1 \) to \( \mathcal{D}_0 \), where \( \mathcal{D}_0 \) and \( \mathcal{D}_1 \) are of the same size. We then apply tweaking tools on \( \mathcal{D}_0 \) to achieve the target features. After the tweaking process is done, ASPECT outputs \( \tilde{\mathcal{D}}_1 \) which is similar to \( \mathcal{D}_1 \). For our experiments, we use \( \mathcal{D}_1 \) as the ground-truth, and compare the similarity between \( \tilde{\mathcal{D}}_1 \) and \( \mathcal{D}_1 \).

B. Size-scaler

Size-scalers are orthogonal to enforcing features in the final dataset \( \mathcal{D} \). Our experiments show that ASPECT is able to generate datasets with small errors for three different size-scalers, Dscaler [36], ReX [8] and Rand, described below:

DSCALER is the first solution to scale relational tables by different ratios. It uses a correlation database which captures fine-grained, per-table correlations to scale the original dataset.

ReX is an automated representative extrapolation technique [8]. It scales all tables by the same ratio. Since tables in the ground-truth dataset do not scale uniformly, the targeted features do not satisfy the necessary conditions in Sec. III for datasets generated by ReX. We, therefore, modify the targeted features to enforce the necessary conditions before tweaking.

Rand is a randomised size-scaler. The tuples are generated randomly. However, it satisfies two requirements: (i) the number of tuples are generated as expected and (ii) the tuples generated satisfy the foreign key constraints.

C. Similarity measure

As stated previously, the similarity between tweaked dataset \( \tilde{\mathcal{D}} \) and ground truth dataset \( \mathcal{D} \) are defined through the feature set \( \mathcal{F} \). Hence, we measure how well ASPECT preserves the features. In the experiment, we only apply 3 tweaking tools presented in Sec. III to preserve the corresponding features. Hence, we measure similarity based on these 3 features.

1) Feature Accuracy: we individually measure the similarity of the 3 features.

\textbf{Linear Feature:} For a target linear join matrix \( H \) (ground truth) and the corresponding \( \tilde{H} \) in the final tweaked dataset, let \( \epsilon_H \) be the mean relative error among the entries. For example, \[ \tilde{H} = \begin{bmatrix} 0 & 0 & 0 \\ 5 & 0 & 0 \\ 2 & 3 & 0 \end{bmatrix}, \quad H = \begin{bmatrix} 0 & 0 & 0 \\ 4 & 0 & 0 \\ 3 & 4 & 0 \end{bmatrix} \]

then \( \epsilon_H = \frac{1}{3} (|\frac{0-0}{0}| + |\frac{5-4}{5}| + |\frac{3-3}{3}|) = \frac{5}{18} \). The linear feature error of \( \tilde{\mathcal{D}} \) is the mean of all \( \epsilon_H \), thus unbounded.

\textbf{Coappear Feature:} For each coappear distribution, let \( \xi \) be the target (ground truth) and \( \tilde{\xi} \) the tweaked distribution. The coappear distribution error \( \epsilon_\xi \) is \[ \epsilon_\xi = \frac{1}{N_{FK}} \sum_v |\xi(v) - \tilde{\xi}(v)|, \]

where \( N_{FK} \) is the number of foreign key vectors. \( \epsilon_\xi \) is bounded by \( \frac{1}{N_{user-pair}} (\sum_v |\xi(v)| + |\tilde{\xi}(v)|) = 2 \). The coappear feature error of \( \tilde{\mathcal{D}} \) is the mean of all \( \epsilon_\xi \).

\textbf{Pairwise Feature:} Similarly, for a pairwise distribution, let \( \rho \) be the target (ground truth) and \( \tilde{\rho} \) the tweaked distribution. The pairwise distribution error \( \epsilon_\rho \) is \[ \epsilon_\rho = \frac{1}{N_{user-pair}} \sum_v |\rho(v) - \tilde{\rho}(v)|, \]

where \( N_{user-pair} \) is the number of user pairs; \( \epsilon_\rho \) is at most 2. The pairwise distribution error of \( \tilde{\mathcal{D}} \) is the mean of all \( \epsilon_\rho \).

2) Query Accuracy: we also measure similarity by the result of an aggregate queries (COUNT, AVERAGE) that are related to the 3 features. The query error is measured by \( \epsilon_q = \frac{|q(\tilde{\mathcal{D}}) - q(\mathcal{D})|}{q(\mathcal{D})} \).

V. RESULTS AND ANALYSIS

In our experiment, ASPECT coordinates \( T_{\text{linear}}, T_{\text{coappear}}, T_{\text{pairwise}} \) on the scaled dataset generated by a size-scaler to realize the corresponding features. There are 3! = 6 ways of ordering these tweaking tools. We use P-L-C, say, to denote the permutation where \( T_{\text{pairwise}}, T_{\text{linear}}, T_{\text{coappear}} \) are applied in that order.

We first compare the feature similarity in Sec. V-A followed by query similarity experiments in Sec. V-B. Later, we discuss the possible improvements in Sec. V-B. Lastly, we present the execution time of ASPECT in Sec. V-D.

A. Feature similarity

For each feature, the plots are organized as follows: x-axis represents the dataset snapshots; y-axis is the feature error. In each plot, we compare how the 6 permutations perform against the baseline (without tweaking).
1) **Linear feature**: In Xiami, there are in total 38 linear join matrices. Fig 9 plots the average error of these linear join matrices. In general, the later $T_{\text{linear}}$ is applied, the smaller the linear feature error, i.e. C-L-P and P-L-C have smaller errors than L-C-P and L-P-C, and C-P-L and P-C-L have 0 error. All permutations reduce the error tremendously for all size-scalers on all datasets.

Different size-scalers generate a scaled dataset with different errors. In Fig 9 take D6 for example, Dscaler generates a dataset with error around 1.0, while ReX generates a dataset with error around 25. Regardless of the initial error difference, ASPECT is able to reduce the error tremendously after applying the tweaking tools.

2) **Coappear feature**: There are 12 coappear distributions for Xiami. Fig 10 plots the average error of these coappear distributions. It shows that, like for $T_{\text{linear}}$, the later $T_{\text{coappear}}$ is applied in the tweaking order, the smaller the coappear error. In general, we find that permutations where $T_{\text{coappear}}$ is after $T_{\text{linear}}$ reduces the errors more than if $T_{\text{coappear}}$ is before $T_{\text{linear}}$. This is expected, since $T_{\text{linear}}$ modifies the coappearing tables massively after $T_{\text{coappear}}$ is done.

Similar to linear feature, most tweaking permutations significantly reduce the coappear errors. However, for the plot Dscaler-Xiami, we observe that the tweaking permutation C-L-P and C-P-L have a smaller error reduction. One possible reason may be small original error, that gives limited room for improvement. The other possible reason could be the highly overlapping structure: the coappear distribution involves many tables. Take $\xi_T$ for example, where $T$ is \{Listen_Artist, Lib_Artist, Artist_Fan, Artist_Comment\}. This $\xi_T$ overlaps with 8 linear joins, so it is modified by 8 linear tweaking tools if $T_{\text{linear}}$ is applied after $T_{\text{coappear}}$. This increases the difficulty of getting a validated modification as described in Sec II-B. We will discuss how to improve the similarity for such highly overlapping features in Sec IV-C.

Nevertheless, ReX-Xiami and Rand-Xiami still have small errors despite such a highly overlapping features.

3) **Pairwise feature**: Xiami has 4 pairwise distributions. Fig 11 plots the error of these pairwise distributions. It again shows that, the later $T_{\text{pairwise}}$ is applied in a tweaking order, the smaller the pairwise feature error in the tweaked dataset. Moreover, all tweaking permutations reduce the errors tremendously for all size-scalers.

In summary, the later a tool $T_i$ is applied, the smaller the error for the feature $F_j$. Moreover, all tweaking permutations reduce the errors tremendously for most of the cases. If the features are highly overlapping, it is possible that the error reduction is not very significant. In the next section, we will discuss how to improve this.

### B. Query similarity

For each query $q$, we compare the query results on the ground-truth dataset $q(D)$ and the scaled dataset $q(D')$. As mentioned in Sec V-B, ReX cannot scale the dataset to arbitrary sizes, so it cannot be used for query similarity experiments. The 4 queries used are: $Q_1$ computes the number of users who have uploaded a photo with commenters; $Q_2$ computes the number of Music Videos that have been commented on by at most 10 different users; $Q_3$ computes the average number of listeners per song; $Q_4$ computes the number of user pairs having interactions through profile page.

Fig 12 presents results for the 4 queries. The first row uses Dscaler as a size-scaler, the second row uses Rand as a size-scaler. The x-axis represents the dataset snapshots, and y-axis represents query error.

As we can see from Fig 12 all tweaking permutations reduce the query error significantly on both size-scalers. The errors are reduced to $< 0.05$ for most of the tweaking
permutations. For Q2, even though the initial error after the size-scaler is relatively low for D2, ASPECT is still able to reduce the error further.

C. Similarity improvement over iterations

Even though ASPECT significantly reduces the errors for most of the cases, there are some rare exceptions, e.g. Dscaler-Xiami in Fig[10] where ASPECT generates a dataset with coappear error > 0.1. Such cases happen when tools modify previously tweaked features. To improve the performance, we run ASPECT for multiple iterations. Previously, we applied tools Tcoappear, Tlinear, Tpairwise sequentially for the permutation C-L-P which results in D. But now, we apply tools Tcoappear, Tlinear, Tpairwise on D in the same order with another few iterations. We find that by having more iterations of tweaking, the error is further reduced tremendously.

In Dscaler-Xiami, C-L-P, C-P-L have larger errors. Fig[13] presents the results of C-L-P and C-P-L with more iterations. The x-axis is the features; the y-axis represents the errors; the bar represents the iterations.

For C-L-P, the coappear error is 0.08 for the first iteration, and reduced to 0.04 in the second iteration and further reduced to 0.02 in the third iteration. For linear feature, the error reduction is greater, from 0.05 to 10^{-3} from third iteration onwards. For C-P-L, we observe similar phenomenon. Moreover, the error reduction is faster (the error stabilises from second iteration onwards).

In summary, the errors are reduced significantly as the number of iterations increases. From the second or third iteration onwards, the resulting error will be really small ~0.02. Hence, the room for improvement will be limited. The reader can find significant error reduction for other datasets in the appendix as well.

D. ASPECT execution time

So far, we have verified that ASPECT is effective in tweaking the features. Next, we will show that ASPECT is efficient as well. Fig[14] presents the running time of each tweaking permutation. Similar to the previous plots, the x-axis represents the dataset snapshot; y-axis represents the running time (minutes).

In Fig[14] the execution time increases linearly with the dataset size for most of the experiments. All tweaking permutations finish within 100 minutes. Moreover, different size-scalers result in different execution time. This is expected, as different size-scalers have different feature errors. Hence, the amount of tweaking is different.

For the same size-scalers and the same dataset, different tweaking permutations have different execution times. In general, L-C-P and L-P-C are more efficient than other tweaking permutations.

VI. LIMITATIONS AND OBSERVATIONS

ASPECT aims to tweak the features F1, F2, ..., Fn so that the tweaked dataset has the corresponding target features ˜F1, ˜F2, ..., ˜Fn. Extensive experiments above show that ASPECT has the capability of tweaking complex features with reasonably small errors within reasonable running time.

A. Limitations

While tweaking a feature, we might modify some already tweaked features. These already tweaked features may take various forms, which significantly increases the difficulty of proving some error bound of previously tweaked features. We state this issue as the Feature Tweaking Bound Problem:

\textbf{Assuming a dataset }\hat{D}_n \textbf{has features }F_1, ..., F_n. \textbf{If a tweaking tool }T_{n+1} \textbf{is applied on }\hat{D}_n, \textbf{how much does it affect the previous features }F_1, ..., F_n? \textbf{?}

Solving the Feature Tweaking Bound Problem for general features might not be possible. Proving error bounds should be easier if the features satisfy certain properties. Consider the following trivial example: if each ˜F_i represents the attribute distribution of a distinct column, then one can easily see that tweaking tool T_{n+1} will never affect ˜F_1, ..., ˜F_n. An example of a non-trivial restriction would be limiting the features to just 1 join.

Besides the Feature Tweaking Bound Problem, there is also the Feature Tweaking Order Problem:

\textbf{When tweaking a dataset for }n \textbf{features, which tweaking order results in the least error?}

We believe the Feature Tweaking Bound Problem and Feature Tweaking Order Problem are issues that offer a rewarding challenge for research on dataset tweaking.

B. Observations

In developing ASPECT, we arrive at the following observations:

\textbf{(O1) Non-overlapping features.} If the features are not overlapping, then regardless of the modification of T_{n+1}, all the previous modified features will be preserved.

\textbf{(O2) Determination of non-overlapping features.} Given (O1), we would want to determine which features among

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig13.png}
\caption{Tweaking error over iterations (vertical axis is error)}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig14.png}
\caption{Execution time for Xiami (minutes)}
\end{figure}
The Dataset Scaling Problem (DSP) was first advocated by Tay [32]. There have been several solutions to this problem in the field of relational database. UpSizeR [33] is the first solution to DSP, which uses attribute correlation extracted from an empirical dataset to generate a synthetic dataset. ReX [8] is a later work that scales up the original dataset by an integer factor \( s \), using an automated representative extrapolation technique. Chronos [17] scales the streaming data by focusing on capturing and simulating streaming data with both column correlation and temporal correlation. Recently, DSP was extended to non-uniform DSP (nuDSP) [55]. As a solution to nuDSP, DSCALER uses a correlation database which captures fine-grained, per-tuple correlations for scaling.

Data scaling is extended into other fields as well. In [25], the authors propose a dataset scaling problem for RDF data and provide a solution RBench that scales the original input dataset by preserving 4 features: resource identity (resource name, resource type, resource degree), relationship patterns (subgraphs with only relationship edges), predicate dictionary (frequency counts of the words) and attribute stars (frequency counts of the star structure). In [21], the authors lift the scaling approach from the pure database level to the OBDA level, where the domain information of ontologies and mappings are also taken into account as well. VIG [21] maintains the similarity for OBDA data by preserving the following features: size of columns clusters and disjointness, schema dependencies and column-based duplicates and NULL Ratios. However, VIG only supports dataset where each table has at most one foreign key only. In most storage systems, compression time and compression ratio are important issues. Hence, these two criteria should also be used for similarity measurement. In [15], SDGen is proposed to scale an input dataset to an arbitrary size which preserves these two similarities. In [37], the authors extend DSP to the Graph Scaling Problem for directed graphs.

DSCALER is proposed as a solution that maintains not only local graph properties, e.g. degree distribution, but also global graph properties, e.g. effective diameter. In the broader field, much work [7], [16], [20], [30] have been done on query-independent application-specific database generation (not scaling). The query-independent database generation refers to generating the database in terms of a given real data set (data-driven) or a set of its character descriptions (character-driven), which is initiated by Jim Gray [16].

For character-driven data generation, a parallel algorithm is first proposed to generate a dataset with a predefined schema and a predefined distribution [16]. However, no correlations between attributes are preserved. Later, Bruno and Chaudhuri introduce DGL [7], a simple specification language to generate datasets with complex synthetic distributions and inter-table correlations. In [20], Houkjaer proposes a table-wise graph model which holds various statistical information about the foreign key and column content. By using such a graph model, a more realistic dataset can be generated. PSDG [19] is another parallel solution to generate “industrial sized” dataset. PSDG supports easy parallelism, using a construct for specifying foreign keys. However, PSDG does not allow independent generation of dependent tables. Referenced tables have to be created for generating dependent tables. This is very inefficient if the referenced tables are not needed, and are huge in size. To overcome this, PDGF [27] is proposed.

Character-driven data generation also appears in graphs as well [2], [12], [22], [5]. For example, the Erdős-Rényi model generates a graph of any size \( n \) with a specified edge probability \( p \). gMark [5] is another tool that generates a graph database based on users’ specifications, e.g. degree distribution and node type occurrence. Recently, TrillionG [23] is proposed to generate large scale graphs in a short time. It can generate trillion-node graphs within two hours by using 10PCs.

For data-driven synthetic database generation, the input is always a real dataset. Mudd [30] is the first tool which uses a real dataset for database generation. However, Mudd uses very little information (name and address) from the real dataset. Similarly, TEXTURE [11] is another microbenchmark for text query workloads. However, it only extracts simple properties, e.g. word distribution, document length. As an extension to PDGF, DBSynth [3] scales the database by utilizing the meta-data (e.g. min/max constraints) and the statistics from the input dataset. Strictly speaking, DSP is a sub-problem of query-independent synthetic database scaling. The problem is that two tweaking tools may access the same tuple more than once. In ASPECT, each tweak has only access to the dataset via the functions similar to the ones provided in Fig. [3]. Hence, ASPECT knows if any two tweaking tools have accessed the same tuples. Then the problem is reduced to finding independent sets in graph theory, where the nodes are the tweaking tools. If two tweaking tools access the same tuples, there will be an edge linking the two nodes. Even though, it is NP-Hard to find a maximum independent set, Robson [28] has proven that it can done in \( O(1.226^n) \) time, which is a reasonable complexity for a small number of tools \( n \).

(03) **Conflicting overlapping features**. Overlapping features are called conflicting if no dataset can satisfy all of them. A simple example of conflicting features of a social network dataset is: \( F_1 = \) more than half of the customers are men and \( F_2 = \) more than half of the customers are women. Such features must be modified to resolve the conflict, and ASPECT always modifies the features that are applied earlier.

(04) **Non-conflicting overlapping features**. For non-conflicting features, it is not always feasible to synthesize a dataset that satisfies all of them, even if one exists. For example, it is already NP-Hard to decide whether there exists a graph that satisfies certain degree distributions [5], so there is no polynomial algorithm that generates a graph for such distributions. For the sake of efficiency, we may have to sacrifice some feature accuracy. Even so, for the features in this paper, ASPECT maintains the features accurately. As we can see from Sec. V-C, the error is reduced to 0.02 after 2 to 3 iterations in the experiments.

**VII. RELATED WORK**

For character-driven data generation, a parallel algorithm is first proposed to generate a dataset with a predefined schema and a predefined distribution [16]. However, no correlations between attributes are preserved. Later, Bruno and Chaudhuri introduce DGL [7], a simple specification language to generate datasets with complex synthetic distributions and inter-table correlations. In [20], Houkjaer proposes a table-wise graph model which holds various statistical information about the foreign key and column content. By using such a graph model, a more realistic dataset can be generated. PSDG [19] is another parallel solution to generate “industrial sized” dataset. PSDG supports easy parallelism, using a construct for specifying foreign keys. However, PSDG does not allow independent generation of dependent tables. Referenced tables have to be created for generating dependent tables. This is very inefficient if the referenced tables are not needed, and are huge in size. To overcome this, PDGF [27] is proposed.

Character-driven data generation also appears in graphs as well [2], [12], [22], [5]. For example, the Erdős-Rényi model generates a graph of any size \( n \) with a specified edge probability \( p \). gMark [5] is another tool that generates a graph database based on users’ specifications, e.g. degree distribution and node type occurrence. Recently, TrillionG [23] is proposed to generate large scale graphs in a short time. It can generate trillion-node graphs within two hours by using 10PCs.

For data-driven synthetic database generation, the input is always a real dataset. Mudd [30] is the first tool which uses a real dataset for database generation. However, Mudd uses very little information (name and address) from the real dataset. Similarly, TEXTURE [11] is another microbenchmark for text query workloads. However, it only extracts simple properties, e.g. word distribution, document length. As an extension to PDGF, DBSynth [3] scales the database by utilizing the meta-data (e.g. min/max constraints) and the statistics from the input dataset. Strictly speaking, DSP is a sub-problem of query-independent synthetic database scaling.
There are also domain-specific benchmarks that generate datasets for specific domains. They include TPC generation (data-driven).

For example, MWGen uses road and floor plans as input and generates a set of real world infrastructure together with moving objects in different transportation modes. Such domain-specific data generation usually generates only one fixed dataset with fixed schema for all applications under the same domain. Hence, domain-specific data generation is different from application-specific data scaling.

Nevertheless, previous works generate datasets with predefined features, instead of flexible features. To the best of our knowledge, ASPECT is the first framework which allows dataset scaling with flexible features through coordinating the tweaking tools.

VIII. Conclusion and Future Work

This paper introduces ASPECT, a framework for flexible application of tweaking tools to enforce target features in synthetic dataset. To generate a scaled dataset with greater similarity comparing to the original dataset, one just needs to apply more tweaking tools. We demonstrate ASPECT by coordinating 3 highly overlapping and complex tweaking tools on real datasets to realize the target features. Extensive experiments show that ASPECT effectively reduces the errors by orders of magnitudes in the synthetic data without sacrificing efficiency.

ASPECT is a step towards the vision for application-specific benchmark data generation. It facilitates bottom-up collaboration among developers in contributing tools for tweaking synthetic datasets to enforce similarity with empirical data.

Our current work is on the Feature Tweaking Bound and Order Problems. We hope to make some progress by restricting the feature types (e.g. single joins).

REFERENCES

[1] S. Acharya, P. B. Gibbons, V. Poosala, and S. Ramaswamy. Join synopses for approximate query answering. In SIGMOD, pages 275–286, 1999.
[2] W. Aiello, F. Chung, and L. Lu. A random graph model for power law graphs. Experimental Mathematics, 10(1):53–66, 2001.
[3] A. Arasu, R. Kaushik, and J. Li. Data generation using declarative constraints. In SIGMOD, pages 685–696, 2011.
[4] T. G. Armstrong, V. Ponnekkanti, D. Borthakur, and M. Callaghan. LinkBench: A database benchmark based on the facebook social graph. In SIGMOD, pages 1185–1193, 2013.
[5] G. Bagan, A. Bonifati, R. Ciucaan, G. H. Fletcher, A. Lemay, and N. Advokaat. gMark: schema-driven generation of graphs and queries. IEEE TKE, 2016.
[6] Z. Bao, Y. C. Tay, and J. Zhou. sonSchema: A conceptual schema for social networks. In Int. Conf. Conceptual Modeling (ER), pages 197–211, 2013.
[7] N. Bruno and S. Chaudhuri. Flexible database generators. In VLDB, pages 1097–1107, 2005.
[8] T. Buda, T. Cerqueus, et al. ReX: Extrapolating relational data in a representative way. In Data Science, LNCS 9147, pages 95–107. Springer, 2015.
[9] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears. Benchmarking cloud serving systems with YCSB. In ACM Symp. Cloud Computing, pages 143–154, 2010.
[10] P. Cui, T. Zhang, F. Wang, and P. He. Perceiving group themes from collective social and behavioral information. In AAAI, pages 65–71, 2015.
[11] V. Ercegovac, D. J. DeWitt, and R. Ramakrishnan. The TEXTURE benchmark: measuring performance of text queries on a relational DBMS. In VLDB, pages 313–324, 2005.
[12] P. Erdős and A. Rényi. On the evolution of random graphs. In Publication of the Mathematical Institute of the Hungarian Academy of Science, pages 17–61, 1960.
[13] O. Erling, A. Averbuch, J. Larriba-Pey, et al. The LDBC social network benchmark: interactive workload. In SIGMOD, pages 619–630, 2015.
[14] R. Gemulla, P. Rösch, and W. Lehner. Linked Bernoulli synopses: Sampling along foreign keys. In Scientific and Statistical Database Management, pages 6–23, 2008.
[15] R. Gracia-Tinedo, D. Harnik, D. Naor, et al. SDGen: Mimicking datasets for content generation in storage benchmarks. In USENIX Conf. File and Storage Technologies (FAST), pages 317–330, Santa Clara, CA, 2015.
[16] J. Gray, P. Sundaresan, et al. Quickly generating billion-record synthetic databases. In SIGMOD, pages 243–252, 1994.
[17] L. Gu, M. Zhou, Z. Zhang, et al. Chronos: An elastic parallel framework for stream benchmark generation and simulation. In ICDE, pages 101–112, 2015.
[18] X. He, H. Zhang, M.-Y. Kan, and T.-S. Chua. Fast matrix factorization for online recommendation with implicit feedback. In SIGIR, volume 16, 2016.
[19] J. E. Hoag and C. W. Thompson. A parallel general-purpose synthetic data generator. SIGMOD Rec., 36(1):19–24, Mar. 2007.
[20] K. Houkjer, K. Torp, and R. Wind. Simple and realistic data generation. In VLDB, pages 1243–1246, 2006.
[21] D. Lanti, G. Xiao, and D. Calvanes. Fast and simple data scaling for OBDA benchmarks. Proc. BLINK, 2016.
[22] S. Mussmann, J. Moore, J. J. Pfeiffer, and J. Neville. Incorporating assortativity and degree dependence into scalable network models. In AAAI, 2015.
[23] H. Park and M.-S. Kim. TrillionG: A trillion-scale synthetic graph generator using a recursive vector model. In SIGMOD, pages 913–928, 2017.
[24] N. Patki, R. Wedge, and K. Veeramachaneni. The synthetic data vault. In DSAA, pages 399–410, Oct 2016.
[25] S. Qiao and Z. M. Özsoyoglu. RBench: Application-specific RDF benchmarking. In SIGMOD, pages 1825–1838, 2015.
[26] T. Rabl, M. Danisch, et al. Just can’t get enough: Synthesizing big data. In SIGMOD, pages 1457–1462, 2015.
[27] T. Rabl, M. Frank, H. M. Sergioh, and H. Kosch. A data generator for cloud-scale benchmarking. In TPC Tech. Conf. (TPC(TC)), pages 41–56, 2010.
[28] J. M. Robson. Algorithms for maximum independent sets. Journal of Computer Science, 7(3):425–440, 1986.
[29] X. Song, Z.-Y. Ming, L. Nie, Y.-L. Zhao, and T.-S. Chua. Volunteer tendency prediction via harvesting multiple social networks. ACM Trans. Inf. Syst., 34(2):10–1:27, Feb. 2016.
[30] J. M. Stephens and M. Poess. M UDD: a multi-dimensional data generator. In SIGSOFT Software Engineering Notes, pages 104–109, 2004.
[31] M. Stonebraker. A new direction for TPC? In TPCC, pages 11–19, 2009.
[32] Y. C. Tay. Data generation for application-specific benchmarking. PVLDB, 4(12):1470–1473, 2011.
[33] Y. C. Tay, B. T. Dai, et al. UpSizeR: Synthetically scaling an empirical relational database. Inf. Syst., 38(8):1168–1183, 2013.
[34] L. Wang, J. Zhan, C. Luo, et al. BigDataBench: A big data benchmark suite from internet services. In High Performance Computer Architecture (HPCA), pages 488–499, 2014.
[35] J. Xu and R. H. Gütting. MWGen: A mini world generator. In Mobile Data Management (MDM), pages 258–267, July 2012.
[36] J. W. Zhang and Y. C. Tay. Dscaler: Synthetically scaling a given relational database. In VLDB, pages 1671–1682, 2016.
[37] J. W. Zhang and Y. C. Tay. GSCALER: Synthetically scaling a given graph. In EDBT, pages 53–64, 2016.
[38] J. W. Zhang and Y. C. Tay. Synthetic dataset scaling with flexible features. http://www.comp.nus.edu.sg/~upsizer/, 2017.
IX. APPENDIX

A. Algorithm pseudocode

In this section, we present the pseudocode for three tweaking tools: $\mathcal{T}_{\text{linear}}, \mathcal{T}_{\text{coappear}}, \mathcal{T}_{\text{pairwise}}$.

Algorithm 1: Linear Feature Tweaking

1. for each row of $H$
   2. leadingElementAdjust() //Lemma 1
   3. for 2 to row_length
      4. nonLeadingElementAdjust() //Lemma 3

Algorithm 2: Coappear Feature Tweaking

1. for each $(v_1, \ldots, v_k) \in \Delta^-$ do
   2. while $\xi^-(v_1, \ldots, v_k) \neq 0$ do
      3. $(v'_1, \ldots, v'_k) \leftarrow \text{closest}(\Delta^+, (v_1, \ldots, v_k))$
      4. for $i \in [1, k]$ do
         5. $t_i \leftarrow \text{tupleRetrieve}(T_i, (v'_1, \ldots, v'_k))$
         6. tupleModification($T_i, v'_i, v_i$)
         7. statsUpdate($\xi^-(v_1, \ldots, v_k), \xi^+(v'_1, \ldots, v'_k)$)

Algorithm 3: Pairwise Feature Tweaking

1. for each $\rho_R^*$ do
   2. for each $(x, y) \in \Theta^+$ do
      3. while $\rho_R^*(x, y) \neq 0$ do
         4. $(x', y') \leftarrow \text{closest}(\Theta^+, (x, y))$
         5. $u, v \leftarrow \text{tupleRetrieve}(R, (x', y'))$
         6. tupleModification($R, x' - x, y' - y, u, v$
         7. statsUpdate($\rho_R^*(x', y')$
         8. statsUpdate($\rho_R^*(x', y')$

B. Theorems and proofs for linear feature

In this section, we present the the formal proofs that are related to the linear feature. Define $H^* = H - \bar{H}$ and $h_{j,i}^* = h_{j,i} - \bar{h}_{j,i}$ for all $1 \leq i, j \leq k$.

Lemma I (leadingElementAdjust): If the first $n$ rows of $H$ and $H$ are the same, then $(h_{n+1,1}^*, h_{n+1,2}^*, \ldots, h_{n+1,n}^*)$ can be tweaked to $(0, h_{n+1,2}^*, \ldots, h_{n+1,n}^*)$, where $h_{n+1,i}^* \geq 2 \leq i \leq n$.

Proof: There are two cases: $h_{n+1,1}^* > 0$ and $h_{n+1,1}^* < 0$.

Case $h_{n+1,1}^* > 0$: There are $h_{n+1,1}^*$ more tuples in $T_1$ having descendants in $T_{n+1}$. Hence, tweaking is needed to make these $h_{n+1,1}^*$ tuples have no descendants in $T_{n+1}$. It takes two steps: Leaf Tuple Plucking and Leaf Tuple Attaching.

Leaf Tuple Plucking: Consider $S_{n+1,1}$, the tuples in $T_1$ which have descendants in $T_{n+1}$. Let $R_{n+1,1}$ be the $h_{n+1,1}^*$ tuples from $S_{n+1,1}$ with least number of descendants in $T_{n+1}$.

Let $Q_{n+1,1}$ be the tuples in $T_{n+1}$ which are descendants of tuples in $R_{n+1,1}$.

Pluck all tuples in $Q_{n+1,1}$ by removing their foreign key reference to $T_n$. Then, all tuples in $R_{n+1,1}$ have no descendants in $T_{n+1}$, so, $h_{n+1,1}^* = 0$ after the tuple plucking. Next, we will do Leaf Tuple Attaching.

Leaf Tuple Attaching: Let $V_{n+1,1}$ be the tuples in $T_n$ which are descendants of $S_{n+1,1} - R_{n+1,1}$. All tuples in $S_{n+1,1} - R_{n+1,1}$ already have descendants in $T_{n+1}$. Hence, attaching more tuples to $V_{n+1,1}$ will not change $h_{n+1,1}^*$. Therefore, attach back all the tuples in $Q_{n+1,1}$ by setting their foreign key reference randomly to tuples in $V_{n+1,1}$.

Case $h_{n+1,1}^* < 0$: There are $h_{n+1,1}^*$ more tuples in $T_1$ which do not have descendants in $T_{n+1}$. The tweaking takes two steps by doing tuple plucking first, and then tuple attaching.

Leaf Tuple Plucking: $|h_{n+1,1}^*|$ tuples in $T_{n+1}$ need to be found first. Once these $|h_{n+1,1}^*|$ tuples are plucked, we can attach them to the tuples in $T_n$ to increase $h_{n+1,1}^*$. For each tuple in $S_{n+1,1}$ (the tuples in $T_1$ which have descendants in $T_{n+1}$), pick 1 descendant in $T_{n+1}$ to form a leaf set $Leaf_{n+1,1}$. It is obvious that plucking any tuples from $T_{n+1} - Leaf_{n+1,1}$ never modify $h_{n+1,1}^*$. Then,

$$|T_{n+1} - Leaf_{n+1,1}| - |h_{n+1,1}^*| = |T_{n+1}| - h_{n+1,1} - (h_{n+1,1}^* - h_{n+1,1}) = |T_{n+1}| - (h_{n+1,1}^* - h_{n+1,1}) \geq 0$$

(by L1)

Hence, $|h_{n+1,1}^*|$ tuples in $T_{n+1} - Leaf_{n+1,1}$ can be randomly plucked. Next, we will attach these tuples back.

Leaf Tuple Attaching: Consider $S_{n+1,1} - S_{n+1,1}$, the set of tuples in $T_1$ having descendants in $T_n$ but no descendants in $T_{n+1}$. Since

$$|S_{n+1,1} - S_{n+1,1}| - |h_{n+1,1}^*| = |S_{n+1,1} - S_{n+1,1}| - |h_{n+1,1}^*| = |S_{n+1,1} - S_{n+1,1}| - |h_{n+1,1} - h_{n+1,1}^*| = |S_{n+1,1} - S_{n+1,1}| - (h_{n+1,1} - h_{n+1,1}^*) = h_{n+1,1} - h_{n+1,1} = \bar{h}_{n+1,1} = \bar{h}_{n+1,1} \geq 0$$

(by L2)

Then, there are $|h_{n+1,1}^*|$ tuples from $S_{n+1,1} - S_{n+1,1}$, denoted as $Sub_{n+1,1}$. For each tuple in $Sub_{n+1,1}$, it must have a descendant in $T_n$. Hence, randomly attach the $|h_{n+1,1}^*|$ tuples from Leaf Tuple Plucking: to the decent in $T_n$. Then all tuples in $Sub_{n+1,1}$ have descendants in $T_{n+1}$ now. Hence, $h_{n+1,1}^* = 0$ after attachment.

We are done with the leading element tweaking; next, we will prove the correctness non-leading element tweaking. Some care is needed to tweak $h_{n+1,1}^*$. Suppose $h_{n+1,1}^* = 1$ for Fig 8, so we want to remove descendants in $T_D$ for 1 tuple in $T_B$. If we do this by plucking $d_5$ from $c_5$, then $h_{n+1,1}^*$ is decreased by 1 as well, so we should instead pluck $d_6$ or $d_7$. Therefore, when tweaking $h_{n+1,1}$, we should avoid affecting $h_{n+1,1}^*$. If $h_{n+1} = -1$, we need to add descendants in $T_D$ for 1 more tuple in $T_B$. We can pluck a leaf, say $d_1$, and attach
Fig. 15: Non-Leading Element Tweaking Demonstration

it to $c_6$; this will increase $\bar{h}_{4,2}$, but it will also increase $\bar{h}_{4,1}$. However, if we first pluck the subtree rooted at $b_6$ and attach it to $a_2$, then pluck $d_1$ and attach it to $c_6$, $\bar{h}_{4,2}$ will increase without affecting other $h_{j,i}$ values. This leads us to the following definition:

**Definition 5:** For a reference chain $T_k \rightarrow \cdots \rightarrow T_1$, suppose we pluck $t \in T_i$ from $t' \in T_{i-1}$ and attach $t$ to some other $t'' \in T_{i-1}$. We call this an **isomorphic adjustment** if the linear join matrix is unchanged.

**Lemma 2:** For a reference chain $T_k \rightarrow \cdots \rightarrow T_1$, we can make $|S_{k-1,i} - S_{k,i}| - |S_{k-1,i-1} - S_{k,i-1}|$ isomorphic adjustments to $T_i$.

**Proof:** ($S_{k-1,i-1} - S_{k,i}$) are the tuples in $T_{i-1}$ having descendants in $T_{k-1}$ but no descendants in $T_k$. Similarly, $(S_{k-1,i} - S_{k,i})$ are the tuples in $T_i$ having descendants in $T_{k-1}$ but no descendants in $T_k$. For each tuple in $S_{k-1,i-1} - S_{k,i}$, pick 1 descendant in $T_1$ with descendants in $T_{k-1}$ to form a set $S_y$. Then, $S_y \subseteq S_{k-1,i} - S_{k,i}$. Let’s consider $S_{k-1,i} - S_{k,i} - S_y$. For any tuple in $S_{k-1,i} - S_{k,i} - S_y$, we can pluck it and attach it back randomly to tuples in $S_{k,i-1}$. These adjustments are isomorphic. Hence, the maximum number is $|S_{k-1,i} - S_{k,i} - S_y| = |S_{k-1,i} - S_{k,i}| - |S_{k-1,i-1} - S_{k,i-1}|$.

**Lemma 3 (nonLeadingElementAdjust):** Suppose the first $n$ rows of $H$ and $\bar{H}$ are the same. Then $(0, \ldots, 0, h_{n+1,1}^*, \ldots, h_{n+1,n}^*)$ can be tweaked to $(0, \ldots, 0, h_{n+1,1+1}^*, \ldots, h_{n+1,n}^*)$.

**Proof:** The tweaking steps are similar to **Lemma 2**. There are two cases:

**Case $h_{n+1,i}^* > 0$:** Similar to **Lemma 2**, there are two steps:

**Leaf Tuple Plucking:** For each tuple in $S_{n+1,i-1}$, pick one descendant $t_y$ in $T_i$, where $t_y$ has descendants in $T_{n+1}$. Use $R_y$ to denote the set of all such $t_y$. Therefore, $|R_y| = |S_{n+1,i-1}|$. For any tuple in $S_{n+1,i-1} - R_y$, $h_{n+1,i}$ will be decreased if all its descendants in $T_{n+1}$ are detached. Moreover, such detachment will not affect $h_{n+1,i-1}$. Thus,$|S_{n+1,i} - R_y| = h_{n+1,i}^* - h_{n+1,i-1}$

$$= h_{n+1,i}^* - \bar{h}_{n+1,i-1}$$

$$\geq h_{n+1,i}^* - \bar{h}_{n+1,i}$$

(By L3)

$$= h_{n+1,i}^*$$

Hence, $h_{n+1,i}^*$ tuples in $S_{n+1,i} - R_y$ can be randomly plucked.

**Leaf Tuple Attaching:** Similar to **Lemma 1**

**Case $h_{n+1,i}^* < 0$:** Similar to **Lemma 1**, there are two steps: Leaf Tuple Plucking and Leaf Tuple Attaching.

**Leaf Tuple Plucking:** Similar to **Lemma 1**

**Leaf Tuple Attaching:** If no isomorphic adjustment is needed, then it is the same as **Lemma 1**. Otherwise, based on **Lemma 2**, the maximum isomorphic adjustment is

$$|S_{n,i}| - |S_{n+1,i}| - (|S_{n,i-1}| - |S_{n+1,i-1}|)$$

Moreover,

$$|S_{n,i}| - |S_{n+1,i}| - (|S_{n,i-1}| - |S_{n+1,i-1}|) - |h_{n+1,i}^*|$$

$$= h_{n,i} - h_{n+1,i} + h_{n+1,i-1} - h_{n,i-1} - \bar{h}_{n+1,i} + h_{n+1,i}$$

$$= h_{n,i} + h_{n+1,i-1} - h_{n,i-1} - \bar{h}_{n+1,i}$$

$$= \bar{h}_{n,i} + \bar{h}_{n+1,i-1} - \bar{h}_{n,i-1} - \bar{h}_{n+1,i} \geq 0$$ (By LA)

Therefore, at least $h_{n+1,i}^*$ isomorphic adjustments in $T_i$ can be made. Let $Sub_{n+1,i}$ be the set of tuples that undergo isomorphic adjustments. For each tuple in the subset $Sub_{n+1,i}$, pick 1 descendant in $T_n$ and randomly attach a tuple plucked from the previous step. Hence, the leaf tuple attaching can be done.

**Theorem 6:** For a reference chain $T_k \rightarrow \cdots \rightarrow T_1$ in some $\bar{D}_i$, let $H$ be the linear join matrix before tweaking and $\bar{H}$ the target linear join matrix. If $\bar{H}$ satisfies the necessary conditions in **Theorem 1**, then Algorithm 1 tweaks $H$ to give $\bar{H}$.

**Proof:** Algorithm 1 iterates over the rows of $H^* = H - \bar{H}$; for each row, the first entry is tweaked to 0 with **Lemma 2** and the following entries are tweaked to 0 with **Lemma 3**.

**C. Theorem and proofs for coappear feature**

In this section, we present formal proofs that are related to coappear feature.

**Theorem 3 (sufficiency):** Suppose tables $T_1, \ldots, T_k$ reference the same tables $T_1', \ldots, T_m'$. Let $\xi$ be the coappear distribution in some $\bar{D}_i$ before tweaking and $\xi$ the target coappear distribution. If $\xi$ satisfies the necessary conditions in **Theorem 2**, then $T_{coappear}$ tweaks $\xi$ to become $\xi$.

**Proof:** Both $\xi$ and $\xi$ satisfy $C_1$, so $|T_i|$ is unaffected by the tweaking. Similarly, $C_2$ ensures $\sum_{v} \xi = \sum_{v} \xi$, so $\sum_{v} \xi = 0$. Therefore

$$\sum_{v \in \Delta^+} \xi(v) + \sum_{v \in \Delta^0} \xi(v) + \sum_{v \in \Delta^-} \xi(v) = 0,$$

so $\sum_{v \in \Delta^+} \xi(v) = \sum_{v \in \Delta^-} -\xi(v)$, i.e. $\sum_{v \in \Delta^+} \xi(v') = \sum_{v \in \Delta^-} -\xi(v')$. Each tweak decreases $\xi(v')$ by 1 and increases $\xi(v)$ by $1$ for some $v' \in \Delta^+$ and $v \in \Delta^-$. After $\sum_{v \in \Delta^+} \xi(v')$ iterations, we get $\sum_{v \in \Delta^-} -\xi(v') = 0 = \sum_{v \in \Delta^-} -\xi(v')$, so $\xi = \xi$. 


D. Theorem and proofs for pairwise feature

In this section, we present formal proofs that are related to linear feature. We first prove Theorem 5.

Theorem 5 [sufficiency] For each response2post table R, let \( \rho_R \) be the pairwise distribution in \( \mathcal{D} \) and \( \tilde{\rho}_R \) the target pairwise distribution. If \( \rho_R \) satisfies (P3) in Theorem 4 then Algorithm 3 tweaks \( \rho_R \) to become \( \tilde{\rho}_R \). Moreover, the extra tuples added to the post table \( P \) is at most \( |U| - |P| \), where \( U \) is the user table.

Proof: Since \( \rho_R \) and \( \tilde{\rho}_R \) both satisfy (P3), and \( \mathcal{T}_\text{pairwise} \) does not affect \( |U| \), we have \( \sum_{x,y} \rho_R(x,y) = \sum_{x,y} \tilde{\rho}_R(x,y) \), and so \( \sum_{x,y} \rho_R^*(x,y) = 0 \). As in the proof of Theorem 3 this implies

\[
\sum_{(x,y) \in \Theta^+} \rho^*_R(x,y) - \sum_{(x,y) \in \Theta^-} \rho^*_R(x,y) = 0.
\]

Each tweak by Algorithm 3 increases \( \rho^*_R(x,y) \) and \( \rho^*_R(y,x) \) by 1, and decreases \( \rho^*_R(x',y) \) and \( \rho^*_R(y',x') \) by 1 for \( (x,y) \in \Theta^- \) and \( (x',y') \in \Theta^+ \). After \( 1/2 \) number of loops, we get \( \sum_{(x,y) \in \Theta^+} \rho^*_R(x,y) = \sum_{(x,y) \in \Theta^-} |\rho^*_R(x,y)| \), so \( \rho_R = 0 \), i.e. \( \rho = \tilde{\rho} \).

Moreover, if a new post needs to be added to \( P \), then each user has at most 1 post. Thus, there are \( |U| - |P| \) users who do not have posts, so that many posts need to be added to ensure each user has 1 post.

Previously, we assume a user does not respond to his/her own post. Now, we remove the assumption. However, we separate the distribution \( \rho_R \) into 2 distributions: \( \rho_S \) and \( \rho_N \), where \( \rho_S \) is the distribution generated by user self-responding behavior, and \( \rho_N \) does not contain any pairwise vector generated by self-responding. Section III-C has discussed the case for \( \rho_N \), so we now discuss tweaking for \( \rho_S \).

Theorem 7: For a response2post table \( R \) in some \( \mathcal{D} \), let \( \rho_S \) be the pairwise distribution generated by user self-responding before tweaking, and \( \tilde{\rho}_S \) the target pairwise distribution. If \( \rho_S \) can be tweaked to become \( \tilde{\rho}_S \), then \( \tilde{\rho}_S \) satisfies the following conditions:

1. \( (SP1) \sum_x 2x \tilde{\rho}_S(x,x) + \sum_{x,y} (x+y) \tilde{\rho}_S(x,y) = 2|\mathcal{T}_R| \)
2. \( (SP2) \sum_x \tilde{\rho}_S(x,x) = |U| \)

Proof: (SP1) For each response tuple \( t \), made from \( u_i \) to \( v_i \). If \( u_i \neq v_i \), assuming their pairwise vector is \( (x,y) \), then it is double-counted by \( \rho_R^*(x,y) \) and \( \rho_R^*(y,x) \). If \( u_i = v_i \), then there are only \( x \) tuples, which are double-counted by \( (x+y) \). So we get the equality in \( (SP1) \).

(SP2) There are \( |U| \) users, each user can respond to himself. Hence counted only once.

Let \( \rho_S^* = \rho_S - \rho_N \), \( \Theta^+_S = \{(x,x) | \rho^*_S(x,x) > 0 \} \) and \( \Theta^-_S = \{(x,x) | \rho^*_S(x,x) < 0 \} \). For each \( (x,x) \in \Theta^-_S \), it adds \( |\rho^*_S(x,x)| \) pairs \( (u_i, u_i) \), where user \( u_i \) has \( x \) response2post tuples in \( R \) referencing \( u_i \)’s post. It does this by looping \( |\rho^*_S(x,x)| \) times, and in each iteration:

- **PairwiseVectorRetrieve**: Pick \( v' = (x',x') \in \Theta^+_S \) that is closest to \( (x,x) \) by Manhattan distance.
- **TupleModification**: Choose users \( u \) with self-respond pairwise vector \( (x',x') \) and tweak \( u \)'s responses to \( u \)'s post, as follows: If \( x < x' \): this means \( u \) has \( x' - x \) more responses to \( u \)'s post than desired, so \( \mathcal{T}_\text{pairwise} \) randomly chooses and removes \( x' - x \) such responses. If \( x > x' \): \( \mathcal{T}_\text{pairwise} \) adds \( x - x' \) responses from \( u \) on \( u \)'s post. If \( u \) has no post, we artificially create a post for \( u \). To do this, we pick another user \( w \) who has more than 1 post and, among \( w \)'s posts, pick a post \( p_w \) with minimum responses; we make \( p_w \) a post by \( u \), and shift the responses to \( p_w \) to other posts by \( u \). If \( w \) has no response, we create a new post \( p \) for \( u \), and add \( x - x' \) responses to \( p \).
- **StatsUpdate**: Increase \( \rho^*_S(x,x) \) by 1 and decrease \( \rho^*_S(x',x') \) by 1.

The following theorem says that conditions in Theorem 7 suffices to ensure that \( \mathcal{T}_\text{pairwise} \) tweaks \( \rho_S \) to become \( \tilde{\rho}_S \).

Theorem 8: For each response2post table \( R \) in some \( \mathcal{D} \), let \( \rho_S \) be the self-responded pairwise distribution before tweaking and \( \tilde{\rho}_S \) the target pairwise distribution. If \( \rho_S \) satisfies the conditions in Theorem 7 then \( \rho_S \) can be tweaked to \( \tilde{\rho}_S \). Moreover, the extra tuples added to the post table \( P \) is at most \( |U| - |P| \), where \( U \) is the user table.

Proof: The proof is similar to the proof in Sec III-C.

E. Dataset summary

In this section, we summarize the datasets used in the experiments. We used 4 datasets from Douban\footnote{https://www.douban.com} and Xiami\footnote{https://www.xiami.com}.

Douban is a Chinese social network website that allows the creation and sharing of content related to movies, books, music, recent events and activities in Chinese cities. Xiami is a Chinese online music website that provides recommendations of music services, offline music activities, and other interactive content. The short summary of the 4 datasets are the following:

1. DoubanMovie contains movie-related data in 17 tables, with table sizes ranging from 10856 to 36747342 tuples.
2. DoubanBook contains book-related data in 12 tables, with table sizes ranging from 686605 to 12891598 tuples.
3. DoubanMusic contains music-related data in 10 tables, with table sizes ranging from 52078 to 7086936 tuples.
4. Xiami also contains music-related data, but is larger. It has 70 tables and more than 90millions tuples.

Fig 16 presents the dataset size for each partition. For example, the 6th partition of DoubanMovie \( \mathcal{D}_6 \) is 2.5 Gigabytes.

The schema of each dataset is presented as follows.
Fig. 16: Dataset Size Summary

Fig. 17: Schema For DoubanMusic: There are 11 tables; The tables with the same color (except grey color) share the same coappear distribution. For example, Album_Comment, Album_Listening, Album_Heard, Album_Wish reference to both Album and User tables; Review is the post table; Review_Comment is the response2post table.

Fig. 18: Schema For DoubanBook: There are 12 tables; The tables with the same color (except grey color) share the same coappear distribution. For example, Book_Comment, Book_Reading, Book_Read, Book_Wish, Diary reference to both Book and User tables; Diary and Review are the post tables; Diary_Comment Review_Comment are the response2post tables.

Fig. 19: Schema For DoubanMovie: There are 17 tables; The tables with the same color (except grey color) share the same coappear distribution. For example, Movie_Actor, Movie_Script, Movie_Director reference to both Star and Movie tables; Movie_Review and Movie_Photo are the post tables; Review_Comment and Photo_Comment are the response2post tables.

Fig. 20: Schema For Xiami: There are 28 tables; The tables with the same color (except grey color) share the same coappear distribution. For example, Listen_Song, Lib_Song reference to both Song and User tables; Collection, Photo, Space and Thread are the post tables; Photo_Comment, Space_Comment, Collect_Like and Thread_Comment are the response2post tables.

F: Feature similarity for DoubanMovie, DoubanMusic, DoubanBook

In this section, we presents the feature similarity for DoubanMovie, DoubanMusic, DoubanBook.
1) Linear feature similarity: Fig.21 presents the linear feature similarity results. All tables are involved in at least one linear join matrix for all datasets. For DoubanMovie, each of the 17 tables is involved in one of 24 linear join matrices. For example, Movie.Comment → Movie and Trailer.Comment → Trailer → Movie are maximal linear joins. Similarly, the 12 tables in DoubanBook have 15 linear join matrices, the 11 tables in DoubanMusic have 14 linear join matrices.

In general, the later $T_{\text{linear}}$ is applied, the smaller the linear feature error, i.e. C-L-P and P-L-C have smaller errors than L-C-P and L-P-C, and C-P-L and P-C-L have 0 error. All permutations reduce the error tremendously for all size-scalers on all datasets.

Even though the error reduction is huge, there are still some cases that the error is > 0.1. For example, Rand-DoubanBook for L-P-C. It reduces linear feature error from 2 to 0.2. We further investigate this issue, the largest error occurs on the join Book.Comment → User. For L-P-C, while tweaking the coappear distribution for $\xi_T$, where $T$ is (Book.Comment, Book.Read, Book.Reading, Book.Wish, Book.Review), it overlaps with 1 pairwise distribution — Book.Review as a post table, and Review.Comment as a response2post table. Moreover, it overlaps with 12 linear joins (e.g. Book.Comment → User, Book.Comment → Book). As stated in Sec V-A such highly overlapped features increase the difficulty of getting a validated modification as described in Section IV. Hence, this could be a potential reason that error is > 0.1.

2) Coappear feature similarity: Fig.22 presents the coappear feature similarity results. There are 6 coappear distributions for DoubanMovie. For example, The 6 tables (Movie.Seen, Movie.Watching, Movie.Wish, Movie.Photo, Movie.Review, Movie.Comment) reference Movie and User. Similarly, DoubanMusic has 4 coappear distributions, DoubanBook has 4 and Xiami has 12. In each case, each table is involved in one or more coappear distributions.

Fig.22 shows that, like for $T_{\text{linear}}$, the later $T_{\text{coappear}}$ is applied in the tweaking order, the smaller the coappear error. In general, we find that permutations where $T_{\text{coappear}}$ is after $T_{\text{linear}}$ reduces the errors more than if $T_{\text{coappear}}$ is before $T_{\text{linear}}$. This is expected, since $T_{\text{linear}}$ modifies the coappearing tables massively after $T_{\text{coappear}}$ is done.

For average error, all permutations of tweaking significantly reduce the error for all datasets for all size-scalers. It is below 0.1 for all most tweaking.

For the plot Dscaler-DoubanMovie, we observe that the tweaking permutations ($T_{\text{linear}}$ applied after $T_{\text{coappear}}$) have an error around 0.2. By looking at the details, we find that this happens for the coappear distribution involving many tables. Take $\xi_T$ for example, where $T$ is (Movie.Comment, Movie.Seen, Movie.Watching, Movie.Wish, Movie.Review, Movie.Photo). This coappear distribution overlaps with 12 linear join matrices and 2 pairwise distribution. This coappear distribution will be modified by 12 linear tweaking tools if $T_{\text{linear}}$ applied after $T_{\text{coappear}}$. Hence, increase the difficulty of getting a validated modification as described in Section II. Nevertheless, we still have a small error for ReX-DoubanMovie, Rand-DoubanMovie for such a highly overlapped structure.

For ReX-DoubanMovie, even through the error without tweaking is as low as 0.01. All tweaking permutations are still able to reduce the error.

3) Pairwise feature similarity: Fig.23 presents the pairwise feature similarity results. DoubanMovie has 2
pairwise distributions: (i) Review as post table and Review_Comment as response2post table; and (ii) Photo as post and Photo_Comment as response2post. DoubanMusic has 1 pairwise distribution, and DoubanBook has 2 pairwise distributions.

Fig. 23 again shows that, the later $T_{\text{pairwise}}$ is applied in a tweaking order, the smaller the pairwise feature error in the tweaked dataset. For DoubanMusic and Xiami, all tweaking permutations reduce the errors tremendously for all size-scalers. For DoubanMovie, all tweaking permutations on data generated by DSCALER significantly reduce the pairwise feature error; most tweaking permutations on data generated by Rand significantly reduce the error, except L-P-C and P-L-C. For Dscaler-DoubanMovie, the error without tweaking is small ($< 0.05$), some tweaking permutations increase the errors. For DoubanBook, all tweaking permutations reduce the errors tremendously for all size-scalers except three tweaking permutations on Dscaler-DoubanBook.

G. Query similarity for DoubanMovie, DoubanMusic, DoubanBook

In this section, we similarly run queries on DoubanMovie, DoubanMusic, DoubanBook, and compare the query results on ground-truth dataset and scaled dataset.

1) Query similarity for DoubanMovie: Fig. 24 presents the query results on DoubanMovie. The 4 queries used are: $Q_1$ computes the number of users that have written a album-view with commenters; $Q_2$ computes the number of stars that have at most 10 different fans; $Q_3$ computes the average number of interested listeners of a album; $Q_4$ computes the number of user pairs having interactions through a album review.

As we can see from Fig. 24 all tweaking permutations reduce the query error significantly on both size-scalers. The errors are reduced to $< 0.05$ for most of the tweaking permutations.

2) Query similarity for DoubanMusic: Fig. 25 presents the query results on DoubanMusic. The 4 queries used are: $Q_1$ computes the number of users that have written a album-view with commenters; $Q_2$ computes the number of stars that have at most 10 different fans; $Q_3$ computes the average number of interested listeners of a album; $Q_4$ computes the number of user pairs having interactions through a album review.

Similar to DoubanMovie, all permutations reduce the errors tremendously.

3) Query similarity for DoubanBook: Fig. 26 presents the query results on DoubanBook. The 4 queries used are: $Q_1$ computes the number of users that have written a book-view with commenters; $Q_2$ computes the number of diaries that have at most 10 different commenters; $Q_3$ computes the average number of interested readers of a book; $Q_4$ computes the number of user pairs having interactions through a book review.

As we can see from Fig. 26 most of the tweaking permutations reduce the errors tremendously except for few rare cases, e.g. Dscaler-DoubanBook-$Q_1$. For the L-C-P permutation, we can see that it has a larger error than the baseline. This is expected, since $Q_1$ is a linear feature related query, and the linear feature that were tweaked by $T_{\text{linear}}$ is subsequently modified by $T_{\text{coappear}}$ and $T_{\text{pairwise}}$. Such a
scenario can be improved by having more iterations. In Fig 27 we run L-C-P on Dscaler-DoubanBook with more iterations. We can see that from second iteration onwards, the Q1 error is reduced to less than 0.001.

### H. Similarity improvement over iterations

In this section, we present feature similarity results for different iterations. Fig 28 Fig 29 and Fig 30 present the results of running 6 tweaking permutations for up to 4 iterations on the dataset generated by DSCALER, ReX and Rand. Take Fig 28 for example, for coappear feature, 4th column for C-L-P is 0.031. It means that after running C-L-P permutation on the data generated by DScaler for 4 times, the coappear feature error is 0.031. It is a 10-fold decrease from 0.306 (the No-Tweak baseline).

For all the three figures, we can see that the more iterations of tweaking, the less error we will have. On average, ASPECT can achieve an error of around 0.02 after 2 or 3 iterations.

### I. Execution time for DoubanMovie, DoubanMusic, DoubanBook

We can see that, the execution time increases linearly with the dataset size for most of the experiments. DoubanMovie is the largest dataset, it takes more time. Nevertheless, most experiments finishes with 60 minutes for the largest snapshot of DoubanMovie. DoubanMusic and DoubanBook are the smaller datasets, hence, it takes less time, within 60 minutes, for the worst tweaking permutation.

For the same dataset, different size-scaler will result in different execution time. This is understandable, the data generated by the size-scalers have different feature errors. Hence, the amount of tweaking is different. Take DoubanMovie for example, the execution time for each permutation varies among the different size-scaler. Moreover, for the same size-scaler and the same dataset, different tweaking permutation has different execution time. In general we find that L-C-P and L-P-C are more efficient than other tweaking permutations.