Finding Multidimensional Simpson's Paradox

Jay Xu *
Burnaby Mountain Secondary School
8800 Eastlake Drive
Burnaby, BC, Canada V3J 7X5
jayxu2023@gmail.com

Jian Pei
Duke University
308 Research Drive
Durham, NC 27708, USA
j.pei@duke.edu

Zicun Cong
Simon Fraser University
8888 University Drive
Burnaby, BC, Canada V5A 1S6
zicun_cong@sfu.ca

ABSTRACT

Finding and analyzing Simpson's paradox, a well known statistical phenomenon, has found many applications. While the existing literature focuses on only analyzing the causes of identified Simpson's paradox, there is no systematic analysis on Simpson's paradox in multidimensional spaces. In this paper, we develop a simple yet practical approach to automatically identify all Simpson's paradox instances formed by various sub-populations and separator attributes in a multidimensional data set. Moreover, we analyze the distribution of the multidimensional Simpson's paradox instances on three real data sets with respect to dimensionality, size of sub-populations, participation of individual records, redundancy, and more. We obtain a series of interesting observations about a few questions that have never been asked before. The results open doors to a few interesting directions for future study. Moreover, this paper is an outcome from a high-school student summer research internship. It reflects our on-going effort in promoting data science research to youth and high school students.

1. INTRODUCTION

Simpson's paradox [12] is a well known statistical phenomenon where an association between two random variables emerges in a population but disappears or even reverses when the population is divided into subgroups. For example, recently Kügelgen *et al.* [14] report an instance of Simpson's paradox in the COVID-19 case fatality rates about China and Italy. Consider the data in Table 1. The overall fatality rate of Italy (4.4%) was clearly higher than that of China (2.3%), as shown in the last row. However, this association does not appear in any age group. Indeed, in each age group, the fatality rate of Italy is lower than or equal to that in China[1].

As another example, Witmer [15] reports an instance of Simpson's paradox in conviction rate of the white defendants and the minority defendants. Consider the data in Table 2. The overall conviction rate of the white defendants (34.4%) is clearly higher than that of the minority defendants (32.6%), as shown in the last row. However, in each victim race group, the conviction rate of the white defendants is lower than that of the minority defendants.

Table 1: An example of Simpson's paradox. The numerator and the denominator of each ratio are the number of fatalities and the number of confirmed cases, respectively. (Replicated from Table 1, Appendix A in [14], data sources: [16; 4])

Age group	Italy	China
0–9	$\frac{0}{43} = 0\%$	$\frac{0}{0} = 0\%$
10–19	$\frac{0}{85} = 0\%$	$\frac{1}{549} = 0.2\%$
20–29	$\frac{0}{296} = 0\%$	$\frac{7}{3,619} = 0.2\%$
30–39	$\frac{0}{470} = 0\%$	$\frac{18}{7,600} = 0.2\%$
40–49	$\frac{1}{891} = 0.1\%$	$\frac{38}{8,571} = 0.4\%$
50–59	$\frac{3}{1,453} = 0.2\%$	$\frac{130}{10,008} = 1.3\%$
60–69	$\frac{37}{1,471} = 2.5\%$	$\frac{309}{8,583} = 3.6\%$
70–79	$\frac{114}{1,785} = 6.4\%$	$\frac{312}{3,918} = 8.0\%$
≥ 80	$\frac{202}{1,532} = 13.2\%$	$\frac{208}{1,408} = 14.8\%$
Total	$\frac{357}{8,026} = 4.4\%$	$\frac{1,023}{44,672} = 2.3\%$

Simpson's paradox is not rare in real applications. We will show more examples of Simpson's paradox in Example 3 and demonstrate that there exist many instances of Simpson's paradox in real data sets in Section 4.2.

In many applications, instances of Simpson's paradox are interesting for statisticians and data scientists, since they may provide strong hints that potentially lead to discoveries of possible causal effects. For example, Kügelgen *et al.* [14] explain the instance of Simpson's paradox in COVID-19 case fatality rates about China and Italy using mediation analysis and achieve the separation of age-related effects from other effects unrelated to age of the COVID-19 pandemic. Moreover, Witmer [15] explains the instance of Simpson's paradox in conviction rates of the white defendants and the the minority defendants using logistic regression and finds that the race of the victim matters more than that of the defendants.

Most of the existing studies on Simpson's paradox focus on analyzing the hidden causal effect given an observed paradox. However, finding instances of Simpson's paradox is far from easy, particularly on multidimensional data sets

*This work was done when the first author conducted a summer research internship.

[1]In [14], the definition of Simpson's paradox does not require that each sub-population under consideration is not empty. Specifically, the age group "0–9" of China is empty – the sub-population does not contain any cases.

Table 2: An example of Simpson's paradox. The numerator and the denominator of each ratio are the number of convicted and the number of recorded cases, respectively. (Replicated from Tables 1, 2, and 3 in [15], data sources: `http://www.amstat.org/publications/jse/v23n2/Simpson_Stand_Ground_2015.csv`)

Victim Race	White Defendants	Minority Defendants
White	$\frac{40}{107} = 37.4\%$	$\frac{10}{25} = 40\%$
Minority	$\frac{5}{24} = 20.8\%$	$\frac{19}{64} = 29.7\%$
Total	$\frac{45}{131} = 34.4\%$	$\frac{29}{89} = 32.6\%$

with many dimensions. Those dimensions provide an exponential number of ways to partition a population into various sub-populations. Moreover, in a multidimensional data set, instances of Simpson's paradox may appear not only at the whole population level, but may also happen in a sub-population and the further divided sub-sub-populations. For example, as shown in Example 3, in the *Adult* data set in the UCI data repository, we can observe an instance of Simpson's paradox in the sub-population of people in the United States with occupations "Protective-serv" and "Tech-support" and sub-sub-population groups divided by gender.

It is tedious, if not impractical at all, to manually examine an exponential number of possible sub-populations and further divided sub-sub-populations to find multidimensional instances of Simpson's paradox on large data sets. Can we develop a tool to automatically and systematically find all instances of Simpson's paradox in a multidimensional data set? Moreover, although Simpson's paradox has been investigated systematically in literature, the existing studies focus on individual instances of Simpson's paradox, that is, how to explain an instance of Simpson's paradox. To the best of our knowledge, there is no existing study on the collection of all instances of Simpson's paradox in a multidimensional data set. Many interesting questions remain open. For example, in real multidimensional data sets, is Simpson's paradox rare or common? How are instances of Simpson's paradox distributed in subspaces? Are instances of Simpson's paradox caused by a small number of records or by many records? A systematic empirical study on those questions may lead to new insights and inspire further theoretical analysis.

In this study, we systematically investigate Simpson's paradox in multidimensional data sets. We make the following contributions. First, we develop and implement a simple yet practical tool to automatically find the complete set of instances of Simpson's paradox in all possible subspaces in a multidimensional data set. We show that our tool is capable of analyzing many real data sets. Second, using our tool, we empirically study a series of questions about the distribution of instances of Simpson's paradox in several real multidimensional data sets and report the corresponding findings and insights. Last, based on this pilot project, we identify a series of interesting directions for future work.

It is worth mentioning that this paper is the outcome of a high school student summer internship. It also presents our on-going effort and progress in promoting data science to youth and bringing data science research to talented high school students.

The rest of this paper is organized as follows. We formulate the multidimensional Simpson's paradox finding problem in Section 2. We develop our algorithm and the tool in Section 3 and report our findings in the experimental study in Section 4. We review the related work in Section 5. We conclude the paper and discuss future directions in Section 6.

2. PROBLEM FORMULATION

In this section, we define the problem of finding multidimensional Simpson's paradox. For the sake of clarity, we often call an instance of Simpson's paradox simply a Simpson's paradox or a paradox.

2.1 Preliminaries: Multidimensional Data Analysis

Consider a multidimensional table $T(X_1, \ldots, X_n, Y)$, where $X_1, \ldots, X_n$ are n categorical *dimensions* (also known as *attributes*) and Y is a *binary label attribute*. For a record (also known as tuple) $t \in T$, we write $t.X_i$ $(1 \leq i \leq n)$ the value of t on dimension X_i and $t.Y$ the label of t. For each dimension, we assume a meta-symbol $*$ that does not belong to the domain.

A *sub-population* is a subset of tuples in T that share common values on some dimensions. Formally, denote by $c = (x_1, \ldots, x_n) = \{t \mid t.X_i = x_i \vee x_i = *,\ 1 \leq i \leq n\}$ a sub-population, where for $1 \leq i \leq n$, $x_i \in Dom(X_i) \cup \{*\}$. We also call $\{t \mid t.X_i = x_i \vee x_i = *,\ 1 \leq i \leq n\}$ the *coverage* of sub-population $c = (x_1, \ldots, x_n)$, denoted by $cov(c) = cov(x_1, \ldots, x_n) = \{t \mid t.X_i = x_i \vee x_i = *,\ 1 \leq i \leq n\}$.

A sub-population $c = (x_1, \ldots, x_n)$ is *k-dimensional* if $k = |\{x_i \neq *, 1 \leq i \leq n\}|$. That is, in total there are k attributes where x_i takes a value not $*$.

We can compute the frequencies of the label variable in a sub-population. Assuming that the frequencies are unbiased estimates of the underlying probabilities, we can write the frequencies as conditional probabilities, that is $Pr(Y \mid c)$.

Example 1. *Let us use the* Adult *data set in the UCI data repository*[2] *to demonstrate the concepts. The data set contains* 32,561 *records of census data. Let us consider the following categorical dimensions in the table: workclass (W), education (E), marital-status (M), occupation (O), relationship (P), race (R), sex (S), and native-country (N). The label attribute is income-over-50K (I), which indicates whether an adult recorded in the table has annual income over 50K. In the rest of this paper, we write the table as*

$$Adult(W, E, M, O, P, R, S, N, I)$$

Sub-population c = (*Federal-gov, Bachelors, *, *, *, *, Male, *) contains those males working for the federal government who have a bachelors degree. The sub-population is* 3*-dimensional, since the sub-population takes non-$*$ values on three attributes in total, namely* W, E, *and* S. $Pr(Y = 1 \mid c)$ *is the probability of having annual income over 50K in the sub-population.*

For sub-populations $c = (x_1, \ldots, x_n)$ and $c' = (x'_1, \ldots, x'_n)$, c is called a *parent (population)* of c' and c' a *child (population)* of c, if (1) there exists i_0 $(1 \leq i_0 \leq n)$ such that $x_{i_0} = *$ and $x'_{i_0} \neq *$; and (2) for all other i $(1 \leq i \leq n, i \neq i_0)$, $x_i = x'_i$.

[2] `https://archive.ics.uci.edu/ml/datasets/adult`

We call X_{i_0} the *differential dimension*. Apparently, if c is a parent population of c', then $cov(c) \supseteq cov(c')$.

If sub-populations c_1 and c_2 share a same parent on the same differential attribute, then c_1 and c_2 are called two *siblings*. Clearly, for sibling populations c_1 and c_2, $cov(c_1) \cap cov(c_2) = \emptyset$.

For sub-populations $c = (x_1, \ldots, x_n)$ and $c' = (x'_1, \ldots, x'_n)$, c is called an *ancestor (population)* of c' and c' a *descendant (population)* of c, if $c \neq c'$ and $x_i = x'_i$ or $x_i = *$ for every $1 \leq i \leq n$. In such a case, $cov(c) \supseteq cov(c')$.

EXAMPLE 2. *In the* Adult *data set, sub-populations (Federal-gov, Bachelors, *, *, *, *, Male, *) and (Federal-gov, Bachelors, *, *, *, *, Female, *) are siblings, since they share a same parent (Federal-gov, Bachelors, *, *, *, *, *, *) on the same differential attribute sex.*

*Please note that two sub-populations sharing a same parent may not always be siblings. For example, (Federal-gov, Bachelors, *, *, *, *, Male, *) and (Federal-gov, Bachelors, *, *, *, *, *, US) share a same parent (Federal-gov, Bachelors, *, *, *, *, *, *) but not on the same differential attribute. Thus, they are not siblings.*

2.2 Simpson's Paradox

DEFINITION 1. *For two sibling sub-populations* $c = (x_1, \ldots, x_n)$ *and* $c' = (x'_1, \ldots, x'_n)$ *and one separator attribute* X_{i_0} *such that* $x_{i_0} = x'_{i_0} = *$, *a* Simpson's paradox *appears if*

1. $Pr(Y = 1 \mid c) \geq Pr(Y = 1 \mid c')$;
2. *for each value* $v \in Dom(X_{i_0})$, *the sub-populations*

   $cov(c[X_{i_0} = v]) = cov(x_1, \ldots, x_{i_0-1}, v, x_{i_0+1}, \ldots, x_n) \neq \emptyset$

   and

   $cov(c'[X_{i_0} = v]) = cov(x'_1, \ldots, x'_{i_0-1}, v, x'_{i_0+1}, \ldots, x'_n) \neq \emptyset$,

   that is, they are not empty, and $Pr(Y = 1 \mid c[X_{i_0} = v]) \leq Pr(Y = 1 \mid c'[X_{i_0} = v])$; *and*
3. *the strict inequality holds in at least either (1) or (2).*

We call (c, c', X_{i_0}) *a* k-dimensional Simpson's paradox, *where* k *is the dimensionality of sub-populations* c *and* c'.

EXAMPLE 3. *In the* Adult *data set,* c_1*=(*, *, *, *, *, *, *, India) and* c'_1*=(*, *, *, *, *, *, *, Taiwan) are two sibling sub-populations. We observe* $Pr(I = 1 \mid c_1) = 40.00\% > Pr(I = 1 \mid c'_1) = 39.22\%$. *However, for each value in the domain of sex, that is, Male and Female, we have* $Pr(I = 1 \mid c_1[sex = Male]) = 42.70\% < Pr(I = 1 \mid c'_1[sex = Male]) = 44.44\%$ *and* $Pr(I = 1 \mid c_1[sex = Female]) = 18.18\% < Pr(I = 1 \mid c'_1[sex = Female]) = 26.67\%$. *This is a 1-dimensional Simpson's paradox.*

As another example of Simpson's paradox, let us look at sibling sub-populations c_2*=(*, *, *, Protective-serv, *, *, *, United-States) and* c'_2*=(*, *, *, Tech-support, *, *, *, United-States). Since* $Pr(I = 1 \mid c_2) = 33.50\% > Pr(I = 1 \mid c'_2) = 30.24\%$, *but for each value in the domain of sex, male and female,* $Pr(I = 1 \mid c_2[sex = Male]) = 36.26\% < Pr(I = 1 \mid c'_2[sex = Male]) = 40.60\%$ *and* $Pr(I = 1 \mid c_2[sex = Female]) = 12.68\% < Pr(I = 1 \mid c'_2[sex = Female]) = 12.89\%$. *This is a 2-dimensional Simpson's paradox.*

It is also easy to verify that Tables 1 and 2 present two 1-dimensional Simpson's paradoxes, as discussed in Section 1.

In some applications, one may want to distinguish between whether the association in a population simply disappears or indeed reverses in the sub-populations divided by an attribute. Thus, we call a Simpson's paradox *strong* if the inequality strictly holds in both conditions (1) and (2) in Definition 1.

In some situations, to ensure the statistical significance of the Simpson's paradox, one may want to constrain that the association is significant. To accommodate this demand, we define δ-Simpson's paradox as follows.

DEFINITION 2. *For two sibling sub-populations* $c = (x_1, \ldots, x_n)$ *and* $c' = (x'_1, \ldots, x'_n)$, *one separator attribute* X_{i_0} *such that* $x_{i_0} = x'_{i_0} = *$, *and a parameter* $\delta \geq 0$, *a* δ-Simpson's paradox *appears if*

1. $Pr(Y = 1 \mid c) - Pr(Y = 1 \mid c') \geq \delta$;
2. *for each value* $v \in Dom(X_{i_0})$, *the sub-populations*

   $cov(c[X_{i_0} = v]) = cov(x_1, \ldots, x_{i_0-1}, v, x_{i_0+1}, \ldots, x_n)) \neq \emptyset$

   and

   $cov(c'[X_{i_0} = v]) = cov(x'_1, \ldots, x'_{i_0-1}, v, x'_{i_0+1}, \ldots, x'_n) \neq \emptyset$,

   that is, they are not empty, and $Pr(Y = 1 \mid c'[X_{i_0} = v]) - Pr(Y = 1 \mid c[X_{i_0} = v]) \geq \delta$; *and*
3. *if* $\delta = 0$, *then the strict inequality holds in at least either (1) or (2).*

We have the following immediately.

PROPOSITION 1. *A 0-Simpson's paradox is a Simpson's paradox (Definition 1) and vice versa. A Simpson's paradox is strong if and only if it is a* δ*-Simpson's paradox with* $\delta > 0$.

2.3 Problem Definition

Example 3 shows that multiple Simpson's paradoxes may appear in a multidimensional table. Moreover, Simpson's paradoxes may involve multidimensional sub-populations. Therefore, it is important to find all multidimensional Simpson's paradoxes in a given multidimensional table.

Given a table T, the *problem of finding multidimensional Simpson's Paradoxes* is to find all multidimensional Simpson's paradoxes. The task can be extended to finding multidimensional strong and δ-Simpson's paradoxes according to the corresponding definitions.

3. THE METHOD

In this section, we describe our computational algorithms for finding multidimensional Simpson's paradoxes. We first present the general framework. Then, we discuss the two phases of our approach step by step.

3.1 The Framework

In order to find all possible Simpson's paradoxes in various sub-populations, we need the statistics of all possible sub-populations. Then, using the statistics, we can apply the definition of Simpson's paradox to examine those sub-populations and identify possible paradoxes. Accordingly, the framework of our approach consists of two phases, as shown in Algorithm 1.

Algorithm 1: The framework of our approach.

Input: a multidimensional data set
Output: all instances of Simpson's paradox

1 Phase 1: materialize the statistics of all multidimensional sub-populations (Section 3.2);
2 Phase 2: find Simpson's paradoxes using the statistics (Section 3.3);

Algorithm 2: The algorithm materializing the statistics of all sub-populations.

Input: a multidimensional data set T
Output: statistics of all multidimensional sub-populations

1 *Merge records having the same dimension values;*
2 **for** *each record* $t = (x_1, \ldots, x_n, y)$ *in* T *after merging* **do**
3 **for** *each ancestor* $(x'_1, \ldots, x'_n)$ *of* $(x_1, \ldots, x_n)$ **do**
4 *Update the counters of the statistics in cell* $P(x'_1, \ldots, x'_n)$;
5 **end**
6 **end**

3.2 Materializing Statistics of Multidimensional Sub-populations

Conceptually, we use a multidimensional array to store the statistics of all multidimensional sub-populations. Let us use a simple example to illustrate the idea.

EXAMPLE 4. *Suppose the multidimensional data set is* $T = (A, B, C, Y)$*, where* A*,* B*, and* C *are three categorical attributes and* Y *is a binary label attribute. Let the domains of the attributes* $Dom(A) = \{a_1, a_2, a_3\}$*,* $Dom(B) = \{b_1, b_2\}$*, and* $Dom(C) = \{c_1, c_2\}$*. We use a 3-dimensional array* P *to record the statistics of sub-populations. For every sub-population* (a_i, b_j, c_k) *where* $a_i \in Dom(A) \cup \{*\}$*,* $b_j \in Dom(B) \cup \{*\}$*, and* $c_k \in Dom(C) \cup \{*\}$*, the statistics are accumulated and stored in cell* $P[a_i, b_j, c_k]$ *in the array. Two statistics are computed: the total number of records falling in this sub-population and the total number of records having* $Y = 1$ *in this sub-population. Clearly, we can use two counters to maintain those two statistics.*

There are many feasible methods to materialize the statistics of multidimensional sub-populations. We notice that the problem of materializing the statistics of multidimensional sub-populations can be reduced to computing a data cube [6]. That is, we can compute a data cube where each aggregate is the two statistics. There is a rich body of literature on data cube computation, such as the array-based approach [18] and the bottom-up approach [2]. Please see the data cube and data warehousing chapter in textbook [7] for a comprehensive introduction.

To keep our implementation simple, we maintain the multidimensional array for statistics in main memory and adopt a one-scan method as shown in Algorithm 2. We use an example to demonstrate the procedure.

EXAMPLE 5. *Consider the setting in Example 4 again. Suppose we read a record* $t = (a_3, b_1, c_2, 1)$ *from the multidimensional data set* T*. In total,* t *has* $2^3 - 1 = 7$ *ancestors, namely* $(*, b_1, c_2)$*,* $(a_3, *, c_2)$*,* $(a_3, b_1, *)$*,* $(*, *, c_2)$*,* $(*, b_1, *)$*,* $(a_3, *, *)$*, and* $(*, *, *)$*. We increase the population size counters of those ancestor cells by* 1*, indicating that the sub-populations in those cells are increased by* 1*. Since* $t.Y = 1$*, we also increase the counters for the number of records having* $Y = 1$ *in those sub-populations by* 1*. We update the statistics for* (a_3, b_1, c_2) *as well. The statistics are used in the multidimensional paradox search later.*

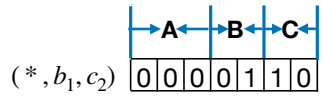


Figure 1: An example of the encoding dimension values of a sub-population $e = (*, b_1, c_2)$. The highest 3 bits $\langle 0, 0, 0 \rangle$ encode the value $*$ on attribute A, the next 2 bits $\langle 0, 1 \rangle$ encode the value b_1 on attribute B, and the lowest 2 bits $\langle 1, 0 \rangle$ encode the value c_2 on attribute C.

On a data set of n dimensions and m distinct records, in total we need to update the statistics of $2^n * m$ times. This phase is a heavy cost of finding multidimensional Simpson's paradox. As computing the statistics of all sub-populations is equivalent to materializing a data cube, the output size is exponential with respect to the input size in the worst case. Thus, it can be shown that computing the statistics of all sub-populations is #P-hard – there does not exist a polynomial time algorithm.

In a multidimensional data set, particularly when the dimensionality is high, data is sparse. That is, many sub-populations may be empty and do not cover any record in the data set. We may take the advantage of data sparsity. In implementation we keep the statistics of sub-populations using a key-value store. In this way, only those sub-populations that are not empty are assigned storage space and the statistics are stored. To further reduce the space cost, we encode a dimension value whose domain has l categorical values using $\lceil \log_2 l+1 \rceil$ bits and pack the dimension values of a sub-population into one or multiple bytes as a key.

EXAMPLE 6. *Take the data set* T *in Example 4 as an example, Figure 1 illustrates how the dimension values of sub-population* $e = (*, b_1, c_2)$ *are encoded and packed.*

For each record t that has n dimensions, we have to update the statistics of 2^n sub-populations. If two records have the same dimension values, they update the same set of sub-populations. To reduce the update workload, at the beginning of Algorithm 2 (Line 1), we first merge all records sharing the same dimension values into one and maintain the counts. We only need to update the statistics of sub-populations using the records and the associated counts after merging.

3.3 Finding Simpson's Paradoxes

Once the statistics of all non-empty sub-populations are computed, we can find Simpson's paradoxes according to the definition. The algorithm is shown Algorithm 3.

For each non-empty k-dimensional sub-population c, that is, c has non-$*$ values on k dimensions, where $1 \leq k \leq n$, Algorithm 3 takes $O(\sum_{x_i = x_j = *, i \neq j} |Dom(D_i)|^2 \cdot |Dom(D_j)|)$ time.

Algorithm 3: The algorithm of finding all instances of Simpson's paradox.

Input: the statistics of all non-empty sub-populations
Output: all instances of Simpson's paradox

1 **for** *each non-empty sub-population* $c = (x_1, \ldots, x_n)$ **do**
2 **for** *each dimension* D_i $(1 \le i \le n)$ *such that* $x_i = *$ **do**
3 **for** *each pair of values* $y, y' \in Dom(D_i)$ *such that* $c_1 = (x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n)$ *and* $c_2 = (x_1, \ldots, x_{i-1}, y', x_{i+1}, \ldots, x_n)$ *are non-empty sub-populations* **do**
4 **for** *each dimension* D_j $(1 \le j \le n)$ *such that* $j \ne i$ *and* $x_j = *$ **do**
5 *Check whether* (c_1, c_2, D_j) *is a Simpson's paradox according to Definition 1 (or Definition 2 if* δ*-Simpson's paradox is to be found)*
6 **end**
7 **end**
8 **end**
9 **end**

To speed up Algorithm 3, we apply the following techniques. First, in the most outside loop, we only need to consider those non-empty sub-populations that have at least two dimensions taking value $*$. In other words, in Line 1 of the algorithm, we only need to consider the sub-populations up to $(n-2)$-dimensional. Since those $(n-1)$- or n-dimensional sub-populations do not allow siblings and separators in the rest of the algorithm.

Second, we notice that, in order to form a Simpson's paradox, a sub-population has to have a minimum number of records. Therefore, in Line 1 of Algorithm 3, if $|c|$ is too small, we do not need to conduct any search within the sub-population. To be specific, we have the following interesting results. We start with the case of strong Simpson's paradox.

THEOREM 1. *If* (c, c', D) *is a strong Simpson's paradox then* $|c| + |c'| > 4 \times |Dom(D)|$.

PROOF. *According to the definition of strong Simpson's paradox, since* (c, c', D) *is a strong Simpson's paradox,* $Pr(Y = 1|c) > Pr(Y = 1|c')$ *and, for any* $x_i \in Dom(D)$, $Pr(Y = 1|c[D = x_i]) < Pr(Y = 1|c'[D = x_i])$, *where* $c[D = x_i]$ *is the sub-sub-population of* c *such that* $D = x_i$. *Moreover, each sub-population* $c[D = x_i]$ *contains at least one record, and so does* $c'[D = x_i]$. *In other words,* $|c| \ge |Dom(D)|$ *and* $|c'| \ge |Dom(D)|$.

Since for every $x_i \in Dom(D)$, $Pr(Y = 1|c[D = x_i]) < Pr(Y = 1|c'[D = x_i])$, $c'[D = x_i]$ *must contain at least one record* t' *such that* $t'.Y = 1$. *Otherwise,* $Pr(c'[D = x_i]) = 0$ *and thus* $Pr(Y = 1|c[D = x_i]) < Pr(Y = 1|c'[D = x_i])$ *cannot hold. Similarly,* $c[D = x_i]$ *must contain at least one record* t *such that* $t.Y = 0$. *Otherwise,* $Pr(c[D = x_i]) = 1$ *and thus* $Pr(Y = 1|c[D = x_i]) < Pr(Y = 1|c'[D = x_i])$ *cannot hold. In other words,* $|c[Y = 0]| \ge |Dom(D)|$ *and* $c'[Y = 1]| \ge |Dom(D)|$.

At the same time, c *must contain at least one record* t *such that* $t.Y = 1$, *otherwise* $Pr(Y = 1|c) = 0$ *and* $Pr(Y = 1|c') > 0$ *and thus* $Pr(Y = 1|c) > Pr(Y = 1|c')$ *does not hold. Similarly,* c' *must contain at least one record* t' *such that* $t'.Y = 0$, *otherwise* $Pr(Y = 1|c') = 1$ *and* $Pr(Y = 1|c) < 1$ *and thus* $Pr(Y = 1|c) > Pr(Y = 1|c')$ *does not hold. Therefore, we have* $|c| > |Dom(D)|$ *and* $|c'| > |Dom(D)|$.

Notice $Pr(Y = 1|c) \le 1 - \frac{|Dom(D)|}{|c|}$ *and* $Pr(Y = 1|c') \ge \frac{|Dom(D)|}{|c'|}$. *Due to* $Pr(Y = 1|c) > Pr(Y = 1|c')$, *we have* $1 - \frac{|Dom(D)|}{|c|} > \frac{|Dom(D)|}{|c'|}$. *That is,* $|c| + |c'| < \frac{|c| \cdot |c'|}{|Dom(D)|}$. *Consider curves* $f = \frac{|c| \cdot |c'|}{|Dom(D)|}$ *and* $g = |c| + |c'|$ *in the two-dimensional space of* $(|c|, |c'|)$. *The tangent is achieved when* $|c| = |c'| = 2 \times |Dom(D)|$. *Therefore, when* $|c| + |c'| < \frac{|c| \cdot |c'|}{|Dom(D)|}$, $|c| + |c'| > 4 \times |Dom(D)|$.

Table 3: An example data set containing a strong Simpson's paradox. Each row of the table is a record t with two attributes A and B, and a binary label Y.

A	B	C
0	0	1
0	0	1
0	0	1
0	0	0
0	1	0
1	0	1
1	1	1
1	1	0
1	1	0

Indeed, the upper bound is tight, as shown in the following example.

EXAMPLE 7. *Consider the following nine records in Table 3. We have*

$$Pr(Y = 1|(0, *)) = 0.6 > Pr(Y = 1|(1, *) = 0.5.$$

However,

$$Pr(Y = 1|(0, 0)) = 0.75 < Pr(Y = 1|(1, 0)) = 1$$

and

$$Pr(Y = 1|(0, 1)) = 0 < Pr(Y = 1|(1, 1)) = 0.33.$$

$((0, *), (1, *), B)$ *is a strong Simpson's paradox and* $|(0, *)| + |(1, *)| = 9 = 4 \times |B| + 1$.

In general, we have the following bound on the size of Simpson's paradox.

COROLLARY 1. *If* (c, c', D) *is a Simpson's paradox then* $|c| + |c'| > 2 \times |Dom(D)|$.

PROOF. *Apparently, each sub-population* $c[D = x_i]$ *contains at least one record, and so does* $c'[D = x_i]$. *In other words,* $|c| \ge |Dom(D)|$ *and* $|c'| \ge |Dom(D)|$. *That is,* $|c| + |c'| \ge 2 \times |Dom(D)|$.

If $|c| + |c'| = 2 \times |Dom(D)|$, *then for each* $x_i \in Dom(D)$, $Pr(Y = 1|c[D = x_i])$ *and* $Pr(Y = 1|c'[D = x_i])$ *are either* 0 *or* 1. *Therefore, either for any* $x_i \in Dom(D)$, $Pr(Y = 1|c[D = x_i]) = Pr(Y = 1|c'[D = x_i])$ *or there exists at least one* x_i *such that* $Pr(Y = 1|c[D = x_i]) > Pr(Y = 1|c'[D = x_i])$. *In the latter case, it is not a Simpson's paradox. In the former case,* $Pr(Y = 1|c) = Pr(Y = 1|c')$, *that is, it is not a Simpson's paradox, either. Contradiction.*

In summary, we have $|c| + |c'| > 2 \times |Dom(D)|$.

Table 4: An example data set containing a Simpson's paradox. The table has the same structure as Table 3.

A	B	Y
0	0	1
0	1	0
1	0	1
1	1	0
1	1	0

Indeed, this bound is also tight.

EXAMPLE 8. *Consider the* 5 *records in Table 4. We have* $Pr(Y = 1|(0, *)) = 0.5 > Pr(Y = 1|(1, *) = 0.33$. *However,* $Pr(Y = 1|(0, 0)) = Pr(Y = 1|(1, 0)) = 1$ *and* $Pr(Y = 1|(0, 1)) = Pr(Y = 1|(1, 1)) = 0$. $((0, *), (1, *), B)$ *is a (trivial) Simpson's paradox and* $|(0, *)|+|(1, *)| = 5 = 2\times|B|+1$.

4. EXPERIMENTAL RESULTS

In this section, we empirically explore multidimensional Simpson's paradox on three real-world data sets using our proposed approach and investigate a series of interesting questions.

4.1 Data Sets

We use three real-world data sets in our experiments, namely, the Adult data set[3], the Mushroom data set[4], and the Loan data set[5]. Table 5 shows some statistics of the three data sets. The row "# of Merged Records" shows the number of records after the merging step (Line 1) in Algorithm 2.

The **Adult** data set contains census data extracted from the U.S. Census Bureau database [6]. Each record in the data set records the personal information of an adult with 6 numeric and 8 categorical attributes. In our experiments, we only use the 8 categorical attributes, namely, work class, education, marital status, occupation, relationship, race, sex, and native country. The cardinalities of the eight attributes are 9, 16, 7, 15, 6, 5, 2, and 42, respectively. The binary label Y is income-over-50K, which indicates whether an adult recorded in the data set has an annual income over 50K.

The **Mushroom** data set consists of the descriptions of $8,124$ hypothetical samples of different species of gilled mushrooms. Each record in the data set is the description of a mushroom sample with 22 categorical attributes. To ensure our experiments on the data set can complete within a reasonable amount of time, we randomly pick 12 categorical attributes of the data set in our experiments. Those attributes picked are cap-shape, cap-surface, cap-color, odor, gill-spacing, gill-size, gill-color, stalk-shape, stalk-root, spore-print-color, population, and habitat. The cardinalities of the 12 attributes are 6, 4, 10, 9, 2, 2, 12, 2, 5, 9, 6, and 7, respectively. The binary label Y indicates whether a recorded mushroom sample is edible. Please note that our current implementation uses only one computer and only main memory. We believe that our approach can be sped up using parallel programming and thus can handle data sets with more attributes. We leave this to the future work.

The **Loan** data set consists of a large set of hypothetical loan applications, which is constructed to demonstrate the approval of loans by a bank. Each record in the data set has 19 numeric and 8 categorical attributes. In our experiments, we use the 8 categorical attributes, namely, year, home ownership, income category, term, purpose, interest payments, grade, and region. The cardinalities of the 8 categorical attributes are 9, 6, 3, 2, 14, 2, 7, and 5, respectively. The binary label Y indicates whether a recorded loan is good.

We choose these three data sets as examples in our empirical study since they are popularly used real data sets and from different domains. At the same time, we acknowledge that there exist many more real data sets where multidimensional Simpson's paradox analysis is interesting. We hope that the results from the three data sets can inspire more interest in finding and analyzing multidimensional Simpson's paradoxes in more data sets and applications.

Our proposed approach is implemented in Java. All experiments are conducted on a PC with Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz, 16GB main memory, 138GB SSD, and a 931 GB HDD running Windows 10. Our source code is published at GitHub `https://github.com/JXu2023/Multidimensional-Simpson-s-Paradox`.

In the rest of this section, we present our empirical study answering a series of research questions.

Table 5: Some statistics of the datasets.

Data Set	Adult	Mushroom	Loan
# Records	32,561	8,124	1,048,575
# of Merged Records	8,688	2,494	31,231
Dimensionality	8	12	8
# Sub-populations	436,414	1,389,900	470,566

Table 6: The total number of Simpson's paradoxes and strong Simpson's paradoxes in the three data sets.

Data Set	# Simpson's paradox	# Strong Simpson's paradox
Adult	3,905	1,109
Mushroom	9,898	149
Loan	15,032	11,014

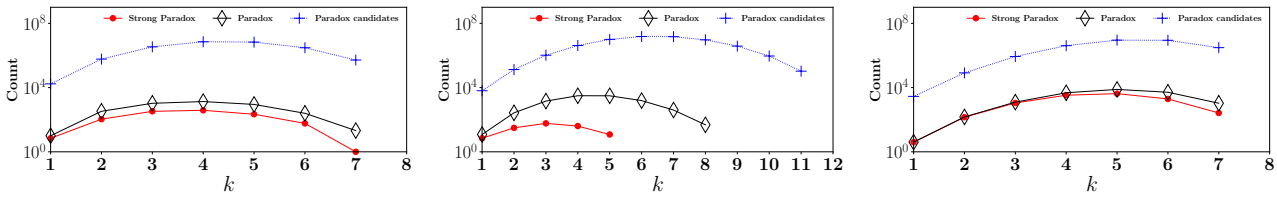
4.2 Are Simpson's Paradoxes Rare?

The first question is whether Simpson's paradoxes are rare in real data sets. Table 6 shows the numbers of paradoxes and strong paradoxes in the three data sets.

Surprisingly, the three data sets allow many Simpson's paradoxes. Even we only consider strong Simpson's paradoxes, still hundreds or even thousands of instances can be observed. To the best of our knowledge, we are the first to report the number Simpson's paradoxes in multidimensional subspaces on multiple real data sets. These results strongly suggest that Simpson's paradoxes may be far from rare in many applications.

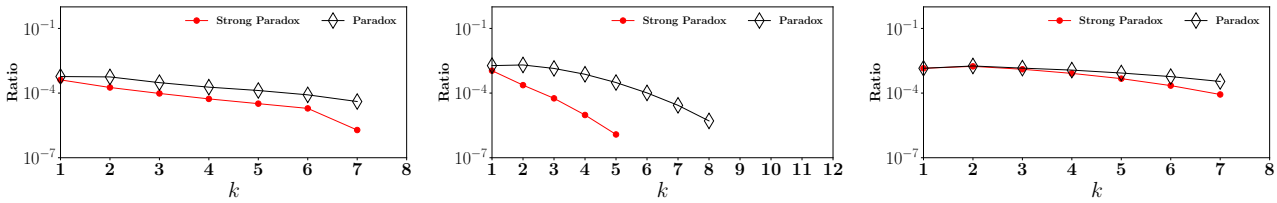
The Loan and the Adult data sets have the same dimension-

[3] `https://archive.ics.uci.edu/ml/datasets/adult`
[4] `https://archive.ics.uci.edu/ml/datasets/mushroom`
[5] `https://www.kaggle.com/datasets/ikpeleambrose/irish-loan-data`
[6] `https://www.census.gov/data.html`



(a) Adult (b) Mushroom (c) Loan

Figure 2: The number of multidimensional Simpson's paradoxes with respect to subspace dimensionality.



(a) Adult (b) Mushroom (c) Loan

Figure 3: The paradox likelihood with respect to dimensionality.

ality but contain very different numbers of paradoxes. Obviously, the reason is that the two data sets have different distributions. Particularly, the relations among attributes are very different on these data sets.

4.3 What Is the Effect of Dimensionality?

How are the Simpson's paradoxes (c, c', D) distributed with respect to the dimensionality of c and c'? Again, to the best of our knowledge, this question has not been asked or addressed in literature.

Figure 2 shows, on the three real data sets, the number of multidimensional Simpson's paradoxes with respect to dimensionality, that is, the number of k-dimensional Simpson's paradoxes with respect to k. The figure shows the numbers of both the Simpson's paradoxes and strong Simpson's paradoxes. Moreover, to provide a reference base for comparison, we also plot in the figure the number of non-empty k-dimensional tuples (c, c', D), that is, two k-dimensional siblings c and c' and a dimension D where both c and c' take value $*$. We call such tuples *paradox candidates*. Obviously, a k-dimensional (strong) Simpson's paradox must be a paradox candidate. Thus, for a dimensionality k, the ratio of the number of (strong) Simpson's paradoxes versus the number of paradox candidates is the likelihood that a paradox candidate is indeed a paradox. We call this ratio the *paradox likelihood*.

We observe several interesting results from Figure 2. First, it is intuitive that the number of strong Simpson's paradoxes is smaller than that of Simpson's paradoxes. The gap between these two depends on the individual data sets. Please note that the Y-axis is in logarithmic scale. On the *Adult* and *Mushroom* data sets, the gap is large. On the *Loan* data set, most of the Simpson's paradoxes are strong.

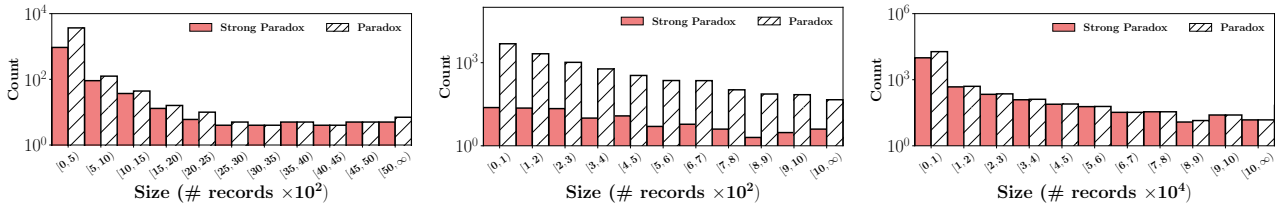
Second, as the dimensionality k increases, the number of paradoxes increases first and then decreases. There are two conflicting factors. On the one hand, as the dimensionality increases, the number of possible sibling pairs increases exponentially. The increase potentially allows more paradoxes. On the other hand, in a high dimensional space, the data becomes sparse. Many sub-populations are empty. When the dimensionality is high, the sparsity leads to the decrease of number of paradox candidates and also the number of paradoxes. In fact, there is no strong Simpson's paradox when $k \geq 8$ on all three data sets. The dimensionalities of the Adult and the Loan data sets are 8. By definition, the largest dimensionality of a paradox cannot exceed 7. However, even on the Mushroom data set whose dimensionality is 12, there is not strong Simpsons' paradox over 5 dimensions and no Simpson's paradox over 8 dimensions.

Last, while the three curves, that is, the numbers of paradox candidates, paradoxes, and strong paradoxes, have similar trends, they do not peak in a synchronized pace. In fact, the curve of strong paradoxes always peaks first and then drops first, followed by the curve of paradoxes. The reason is that strong paradoxes have a stronger and thus harder to satisfy requirement on the distinction of data associations. It also needs more data records as shown by our theoretical analysis (Theorem 1).

To better understand the paradox likelihood, Figure 3 plots the paradox likelihood with respect to dimensionality k. Interestingly, in all three data sets, the likelihoods for both strong Simpson's paradoxes and Simpson's paradoxes are strictly decreasing exponentially. Simpson's paradoxes are often caused by unobserved causal factors. The increase of dimensionality leads to noisy high dimensional spaces, which do not necessarily have embedded strong causal factors. Therefore, the likelihood of Simpson's paradoxes does not sustain as the dimensionality increases.

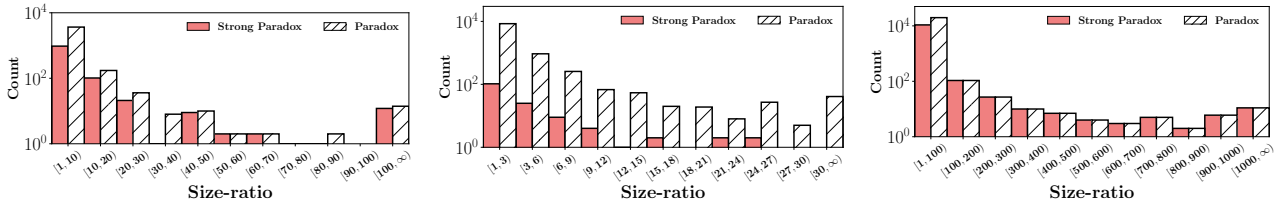
4.4 What Are the Sizes of Simpson's Paradoxes?

An interesting and important question is about the size of Simpson's paradoxes. That is, how many records are involved in a Simpson's paradox?



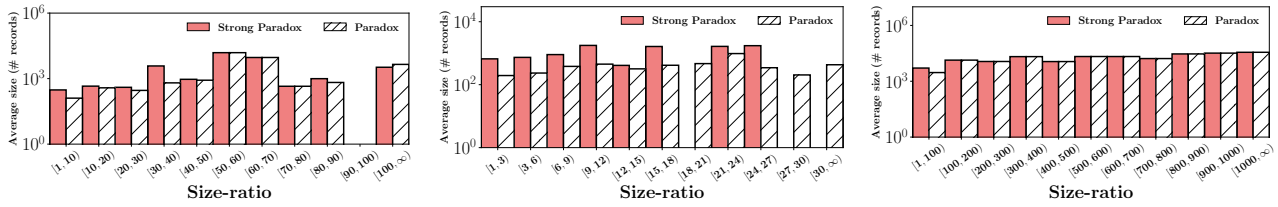
(a) Adult (b) Mushroom (c) Loan

Figure 4: The number of multidimensional paradoxes with respect to the total number of records $|c| + |c'|$.



(a) Adult (b) Mushroom (c) Loan

Figure 5: The number of multidimensional paradoxes with respect to the size-ratio $\frac{\max\{|c|,|c'|\}}{\min\{|c|,|c'|\}}$.



(a) Adult (b) Mushroom (c) Loan

Figure 6: The average size of multidimensional paradoxes with respect to size-ratios.

Indeed, this question involves two sub-questions. First, for a Simpson's paradox (c, c', D), what is $|c|+|c'|$, the total number of records in the sub-populations c and c'? Second, we are also interested in the ratio between the sizes of c and c'. We define the size-ratio of paradox (c, c', D) as $\frac{\max\{|c|,|c'|\}}{\min\{|c|,|c'|\}}$. The larger the ratio, the more different the relative sizes of the two sub-populations. Third, we explore the size of the smaller sibling sub-population $\min\{|c|, |c'|\}$ of a paradox.

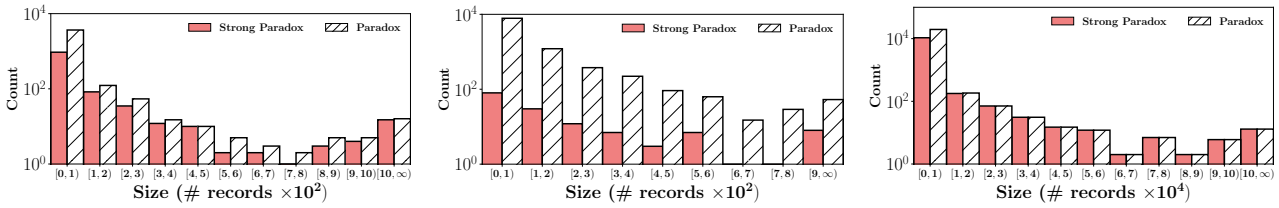
It happens that the largest Simpson's paradoxes in the three real data sets, respectively, are also the largest strong Simpson's paradoxes with sizes 29,753, 5,720, and 759,618. We believe that those are just coincidences.

Figure 4 shows the histograms of the total number of records in the paradoxes in the three data sets. We partition the size of paradoxes into several intervals and count the number of paradoxes whose sizes fall into each interval. The Y-axis represents the count in logarithmic scale. In general, as the size increases, the number of (strong) paradoxes decreases. There are more sub-populations of smaller sizes than those of larger sizes, which lead to more sibling pairs and potentially more paradoxes. The differences between the numbers of paradoxes and strong paradoxes in the intervals vary with respect to data sets. For example, the differences on the Adult and Loan data sets are small and that on the Mushroom data set is large.

The "smallest" strong paradoxes that involve the least number of records on the three data sets have 10 (Adult), 42 (Mushroom), and 10 (Loan) records, respectively. All those three strong paradoxes are statistical significant in the tests reported in Section 4.10.

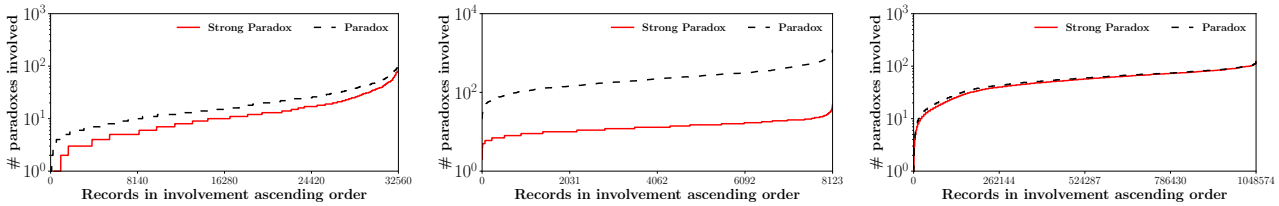
Figure 5 shows the histograms of the size-ratios of paradoxes in the three data sets. The Y-axis of the figure is in logarithmic scale. Again, we partition the size-ratio into intervals and count the number of paradoxes whose size-ratios falling into each interval. The numbers of (strong) paradoxes decrease as the size-ratio increases from small to medium (the first 50% of the intervals). The sizes of may sub-populations tend to close to the average size. Thus, the chance that the sizes of two siblings are dramatically different is low. However, we do observe paradoxes whose size ratios are huge, for example, over 1,000 times in the Loan data set. As future work, it is interesting to investigate the statistical significance of such paradoxes of extremely large size ratios.

Figure 6 plots the average size of paradoxes in each interval of size-ratio on the real data sets. Interestingly but without any surprise, the average sizes in most size-ratio intervals are not far away. This clearly shows that the sizes of most sub-populations tend to be close to the overall average. In some size-ration intervals, the number of paradoxes is small and thus the average size in the interval may be substantially



(a) Adult (b) Mushroom (c) Loan

Figure 7: The number of multidimensional paradoxes with respect to the size of the smaller sibling sub-population $\min\{|c|, |c'|\}$.



(a) Adult (b) Mushroom (c) Loan

Figure 8: The numbers of multidimensional paradoxes that an individual record is involved.

deviate from the average.

Figure 7 shows the histograms of the sizes of the smaller sibling sub-population in the paradoxes in the three data sets. The Y-axis of the figure is in logarithmic scale. We partition the size of the smaller sibling sub-population into several intervals and count the number of paradoxes whose size of the smaller sibling sub-population fall into each interval. We observe that a large number of (strong) paradoxes include at least 100 records in their sibling sub-populations. The observation suggests that many of the multidimensional paradoxes may include a non-trivial number of records in their sibling sub-populations.

4.5 Are Many Records Involved in Simpson's Paradoxes?

While the previous research questions analyze how Simpson's paradoxes are distributed, now we ask a question from the individual record perspective. In a real data set, are many records involved in some Simpson's paradoxes or only very few of them involved in many paradoxes? This problem has not been touched in literature.

To be specific, a record t is involved in a (strong) Simpson's paradox (c, c', D), if $t \in cov(c) \cup cov(c')$. For each record, we count the number of (strong) Simpson's paradoxes in which the record is involved.

Figure 8 shows the number of paradoxes that an individual record is involved. We sort all records in a data set in the ascending order on the number of paradoxes involved. The Y-axis is in logarithmic scale. Some statistics of the number of paradoxes that an individual record is involved are reported in Table 7. We have three observations.

First, surprisingly almost all records are involved in some paradoxes. This is because, as reported in Section 4.4, a few paradoxes in the data sets involve a large portion of the records in the data sets. Second, a record can be involved in multiple paradoxes. Due to the sparsity of data, multiple sub-populations may contain the same set of records in a data set, that is, their coverages are the same. Thus, multiple paradoxes may contain the same group of records (more details in Section 4.6). Last, as shown in Table 7, on the three data set, the max values are multiple standard deviations larger than the corresponding averages. Some records are involved in significantly more paradoxes than the other records. As future work, it is interesting to further investigate the effects of those records on the paradoxes.

Table 7: Some statistics of the number of paradoxes that an individual record is involved.

Data Set	Paradox Type	Mean and Std dev.	Max	Min
Adult	Strong paradox	13.7 ± 12.6	88	0
	Paradox	20.5 ± 16.0	127	0
Mushroom	Strong paradox	13.9 ± 5.0	50	2
	Paradox	254.7 ± 156.7	1237	21
Loan	Strong paradox	55.7 ± 23.3	123	0
	Paradox	58.1 ± 22.7	124	0

4.6 Redundancy in Simpson's Paradoxes

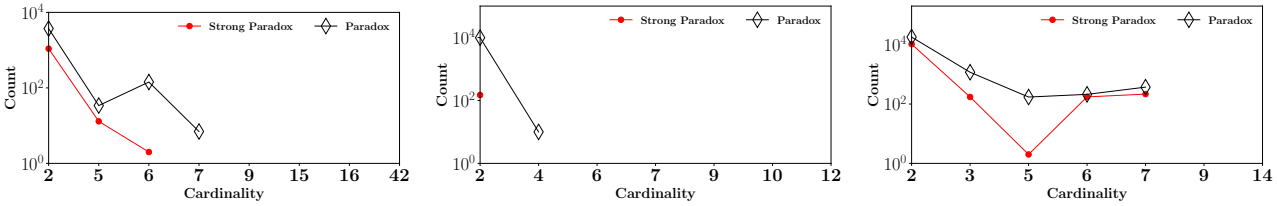
Some observations in Section 4.5 motivates the question whether there exists redundancy in Simpson's paradoxes. Specifically, two Simpson's paradoxes (c_1, c_1', D) and (c_2, c_2', D) are said to be *redundant* if $cov(c_1) = cov(c_2)$ and $cov(c_1') = cov(c_2')$, that is, sub-populations c_1 and c_2 contain the exactly same set of records, and so do c_1' and c_2'.

Table 8 shows three strong Simpson's paradoxes in the Adult data set, which are formed by the same set of thirteen records. The three paradoxes have workclass and sex as the differential dimension and the separator attribute, respectively, where $Dom(\text{sex}) = \{\text{Male}, \text{Female}\}$.

The reason for the redundancy in Simpson's paradoxes is due to the closure of sub-populations, which is found by closed frequent patterns [9] and quotient cube [8]. In the context of multidimensional Simpson's paradox analysis, a sub-population c_1 is not closed if there exists a parent sub-population c_2 of c_1 such that c_1 and c_2 contain the same set of records. Apparently, for siblings c_1 and c_1' and siblings

Table 8: Three strong paradoxes formed by the same group of 13 records. "Self-emp-not-inc" and 'Local-gov' are short for "Unincorporated Self Employment" and "Local Goverment", respectively. The numerator and the denominator of each ratio are the number of recorded adults with income above 50K and the number of recorded adults in the corresponding sub-population.

Sibling Sub-population							Male	Female	Total
Education	**Marital Status**	**Occupation**	**Relationship**	**Race**	**Native Country**	**Workclass**			
Doctorate	Never-married	Prof-specialty	Not-in-family	*	*	Self-emp-not-inc	$\frac{6}{7} = 0.86$	$\frac{1}{3} = 0.33$	$\frac{7}{10} = 0.7$
Doctorate	Never-married	Prof-specialty	Not-in-family	*	*	Local-gov	$\frac{1}{1} = 1$	$\frac{1}{2} = 0.5$	$\frac{2}{3} = 0.67$
Sibling Sub-population							**Male**	**Female**	**Total**
Education	**Marital Status**	**Occupation**	**Relationship**	**Race**	**Native Country**	**Workclass**			
Doctorate	Never-married	*	Not-in-family	White	*	Self-emp-not-inc	$\frac{6}{7} = 0.86$	$\frac{1}{3} = 0.33$	$\frac{7}{10} = 0.7$
Doctorate	Never-married	*	Not-in-family	White	*	Local-gov	$\frac{1}{1} = 1$	$\frac{1}{2} = 0.5$	$\frac{2}{3} = 0.67$
Sibling Sub-population							**Male**	**Female**	**Total**
Education	**Marital Status**	**Occupation**	**Relationship**	**Race**	**Native Country**	**Workclass**			
Doctorate	Never-married	Prof-specialty	Not-in-family	White	*	Self-emp-not-inc	$\frac{6}{7} = 0.86$	$\frac{1}{3} = 0.33$	$\frac{7}{10} = 0.7$
Doctorate	Never-married	Prof-specialty	Not-in-family	White	*	Local-gov	$\frac{1}{1} = 1$	$\frac{1}{2} = 0.5$	$\frac{2}{3} = 0.67$



(a) Adult (b) Mushroom (c) Loan

Figure 9: The numbers of paradoxes with respect to the cardinality of separator attribute.

c_2 and c_2', if c_1 is a parent of c_2 and $cov(c_1) = cov(c_2)$, c_1' is a parent of c_2' and $cov(c_1') = cov(c_2')$, then (c_1, c_1', D) is a Simpson's paradox (c_1, c_1', D) if and only if (c_2, c_2', D) is also a Simpson's paradox.

This finding leads to an immediate task for future work: how can we compute a concise (i.e., non-redundant) representation of multidimensional Simpson's paradoxes.

4.7 What Is the Effect of Separator Cardinality?

The separator attribute also plays an important role in Simpson's paradox. Intuitively, the fewer possible values on the separator attribute, the easier a Simpson's paradox can be formed. Is this the case on real data sets?

Figure 9 shows the numbers of (strong) Simpson's paradoxes with respect to the cardinality of the separator attribute on the three data sets. In general, our intuition holds in most of the cases. There are some exceptions. In the *Adult* data set, the dimension "relationship" acting as the separator attribute, which has cardinality 6, leads to an unusually large number of Simpson's paradoxes. Moreover, in the *Loan* data set, the dimension "region" acting as the separator attribute, which has cardinality 5, leads to an unusually small number of strong Simpson's paradoxes.

We have to admit that our study here is not thorough enough to answer this question conclusively, since the cardinality is highly coupled with the specific dimensions in the real data sets. Using these real data sets, we cannot control the cardinality and compare the numbers of Simpson's paradoxes in a fair manner. We leave the thorough investigation about the effect of separator attribute cardinality for future work.

Table 9: The runtime of our approach (in seconds).

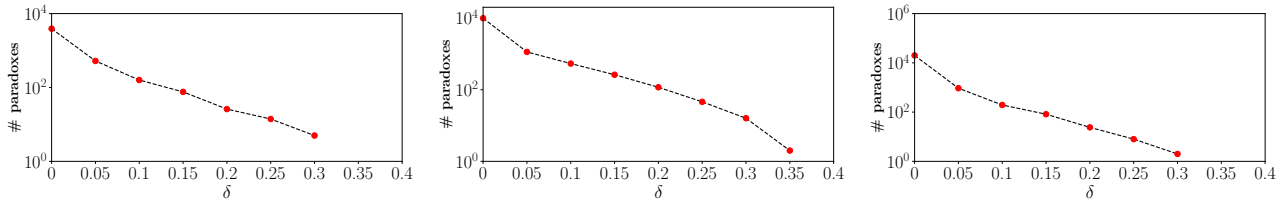
Data Set	Phase 1	Phase 2	Phase 2 w/o Pruning	Speedup (%)
Adult	0.63	1.53	3.00	49%
Mushroom	3.22	7.82	11.93	34.5%
Loan	3.55	1.95	2.05	5%

4.8 Runtime

In this subsection, we report the runtime of our proposed approach. In addition, to evaluate the effectiveness of our pruning strategy, we implement an ablation by disabling the pruning strategy using Theorem 1.

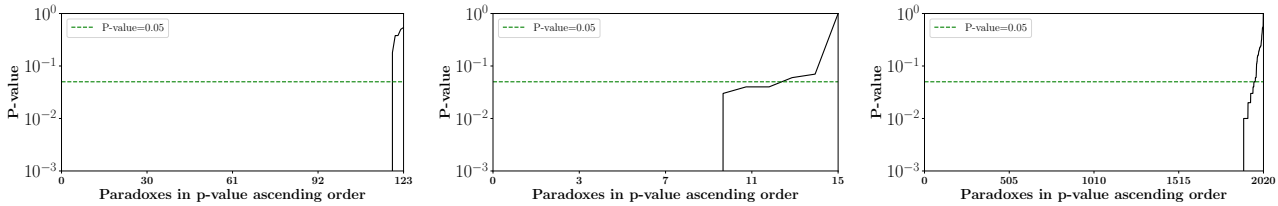
The runtime of our method is reported in Table 9. Our method can always finish within 12 seconds. The runtime of Phase 1 of our method on the Loan data set is larger than that on the other two data sets. This is because the Loan data set contains much more records than the other two data sets, which take a longer time to process. As shown in Table 5, the Mushroom data set has the largest number of sub-populations. Therefore, it takes our method the longest time to search paradoxes in the Mushroom data set.

Based on Theorem 1, we do not need to conduct any search within the sibling sub-populations c and c' if their sizes are too small. Comparing the Phase 2 time of our method with that of the ablation, we can see that our pruning strategy can speed up the Phase 2 runtime by 5-49%. The pruning



(a) Adult (b) Mushroom (c) Loan

Figure 10: The numbers of δ-paradoxes.



(a) Adult (b) Mushroom (c) Loan

Figure 11: The p-values of the paradoxes in p-value ascending order. The Y-axis is in logarithmic scale.

strategy is less effective on the Loan data set. Due to the large population and low cardinality of the Loan data set, it tends to generate large sub-populations, which have more records than the threshold identified in Theorem 1.

4.9 What Is the Effect of δ-Simpson's Paradox?

Last, we study Simpson's paradoxes that are statistical significant, that is, δ-Simpson's paradoxes (Definition 2). In this experiment, we vary δ from 0 to 0.4 and report the number of δ-paradoxes in the three data sets in Figure 10. The Y-axis of the figure is in logarithmic scale. There is no δ-paradox in the Adult, Mushroom, and Loan data sets when $\delta = 0.4$. In all three data sets, the numbers of δ-paradoxes decrease exponentially. This observation is intuitive: the more significance requirement, the less paradoxes can be formed.

4.10 Are Small Simpson's Paradoxes Statistically Significant?

Some paradoxes may only include a small number of records. Are those paradoxes statistically significant? In this section, we conduct statistical hypothesis tests on the small paradoxes in the three data sets to study their statistical significance.

In each data set, we focus on the strong paradoxes whose size is smaller than or equal to a threshold. The thresholds in the Adult, Mushroom, and Loan data sets are set to 0.1%, 1%, and 0.01% of the total number of records, that is, 33, 82, and 105 records, respectively. Given the thresholds, 124, 16, and 2021 strong paradoxes are found in the Adult, Mushroom, and Loan data sets, respectively.

The null hypothesis is that the strong paradoxes are caused by chance. The p-value threshold is set to 0.05. For each paradox (c, c', D), We perturb the data set $c \cup c'$ as follows. We uniformly randomly select $\lceil |c \cup c'| \times 5\% \rceil = \lceil (|c| + |c'|) \times 5\% \rceil$ pairs of records (t, t') such that $t \in c$, $t' \in c'$, and $t.Y \neq t'.Y$, and then switch the class labels of the records. That is, we randomly choose a small number of pairs of records from c and c' whose class labels are different, and swap the class labels. This perturbation method maintains the class distribution in subpopulation $c \cup c'$ unchanged. We obtain 100 perturbed sample data sets for each paradox. The p-value is computed as the probability that the paradox is still observed in the perturbed samples.

Please note that the perturbation is very small. For example, in the Adult data set, the largest size of $|c \cup c'|$ is up to 33 under consideration in this experiment. In our experiment setting, in each perturbed sample set, we only change no more than $\lceil 33 \times 5\% \rceil = 2$ pairs of records.

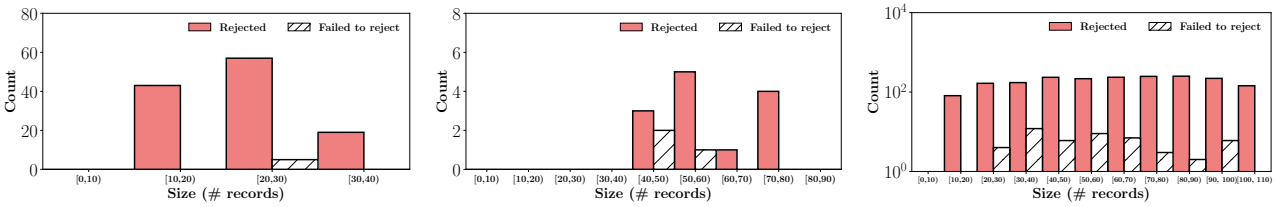
Figure 11 shows the p-values of the strong paradoxes. We sort the strong paradoxes in the p-value ascending order. The results clearly show that most of the paradoxes are statistically significant, that is p-value < 0.05. Even in the Adult data set, only one or two pairs of records are perturbed, still over 95% of the paradoxes very likely disappear in the perturbed sample sets and thus the null hypothesis is strongly rejected.

Figure 12 groups the paradoxes into the intervals of the size $|c \cup c'|$ and, for each interval, plots the number of statistically significant paradoxes (i.e., # rejected) and that of statistically insignificant ones (i.e., # failed to reject). The results show that, even for the paradoxes involving a very small sub-population, there are still statistically significant ones.

The experiments on statistical significance demonstrate that finding multidimensional Simpson's paradoxes is statistically meaningful.

5. RELATED WORK

Simpson [12] describes the paradoxical phenomenon that is named after him later, though the similar effects are mentioned before by Pearson [11] and Yule [17]. Simpson's



(a) Adult (b) Mushroom (c) Loan

Figure 12: The statistical significance of paradoxes with respect to the total number of records in $c \cup c'$.

paradox has been well investigated in literature. Particularly, a rich body of literature explores how to explain Simpson's paradox using causal inference. Sprenger and Weinberger [13] provide a thorough discussion and literature summary about Simpson's paradox and applications. Pearl [10] presents an insightful review. Clearly, a thorough review of the literature on Simpson's paradox is far beyond the capacity of this paper. We refer interested readers to the comprehensive and thoughtful reviews [13; 10].

Almost all existing studies focus on analyzing individual Simpson's paradox. To the best of our knowledge, except for two recent studies [1; 5], the problem of automatically finding Simpson's paradoxes in multidimensional data sets has never been touched.

Alipourfard *et al.* [1] present a statistical method that automatically find Simpson's paradox in data, which is the work most related to our study here. Although both Alipourfard *et al.*'s work [1] and ours here focus on finding Simpson's paradox, there are two fundamental differences between the two. First, Alipourfard *et al.*'s method [1] considers only two dimensional projected subspaces of the *whole* data set instead of all possible subspaces and sub-population pairs. To be specific, for a data set $T(X_1, \ldots, X_n, Y)$, Alipourfard *et al.*'s method [1] considers up to $n(n-1)$ dimension pairs (X_i, X_j) $(i \neq j)$ such that X_i is the differential attribute and X_j is the separator attribute. In our study, we consider an exponential number of possible Simpson's paradox candidates (c, c', X_i) where c and c' are sibling sub-populations. The search space in our study and the number of Simpson's paradoxes are much larger than those considered by Alipourfard *et al.* [1].

Second, Alipourfard *et al.* [1] use the Simpson's paradoxes found from a data set collected from Stack Exchange to analyze the factors affecting question-answerering performance, the likelihood that an answer provided by a user is accepted by the asker as the best answer to the question. The data analysis focuses on an application. Our study here is concerned with the general questions about distributions of Simpson's paradoxes in subspaces and sub-populations. We use three different real data sets with different applications to conduct the experiments. The data analysis tasks are completely different.

Therefore, although Alipourfard *et al.*'s work [1] is related to ours, they are still substantially different. Their method cannot be used to accomplish the objectives in our study and our study does not touch the target application explored by Alipourfard *et al.* [1]. Therefore, we cannot use the method by Alipourfard *et al.* [1] as a baseline.

Fabris and Freitas [5] consider Simpson's paradox that may formed using hierarchies of different dimensions, such as claimants may be divided by gender, state/country/city, claimant type, or age-group. They identify Simpson's paradox in two dimensional projected subspaces of the whole data set and use the hierarchy information of dimensions to compute the surprisingness of the discovered paradoxes. However, their method does not automatically explore all possible sub-populations. Therefore, their approach is fundamentally different from ours and cannot serve the purpose of this study.

6. CONCLUSIONS AND DISCUSSION

In this paper, we explore Simpson's paradoxes hidden in a multidimensional data set. We present a simple yet practical method to automatically compute all Simpson's paradoxes formed by various sub-populations and separator attributes. Applying our algorithm on three real data sets, we obtain interesting observations on a series of questions that have never been investigated systematically in literature.

Our vision is to develop a systematic, end-to-end, human-in-the-loop approach and toolbox for automatic Simpson's paradox detection and analysis. This study is just the very beginning step and opens the door to a series of interesting future work. First, in terms of computation, how to scale up the automatic finding of Simpson's paradoxes on high dimensional and large data sets remains a challenge. On such high dimensional data sets, likely the main memory is insufficient to hold the statistics of all sub-populations. Moreover, there are an exponential number of paradox candidates that need to be checked. Developing scalable and possibly distributed methods for fast multidimensional Simpson's paradox finding is an interesting direction.

Second, as pointed out in Section 4.6, we observe redundancy among multidimensional Simpson's paradoxes. Therefore, it is important to systematically investigate concise and non-redundant representation of multidimensional Simpson's paradoxes and how to compute multidimensional Simpson's paradoxes in such a representation.

Third, as a Simpson's paradox may be explained by some causal factors, it is interesting to develop systematic methods and tools to explain multiple related Simpson's paradoxes in multiple overlapping multidimensional subspaces. For example, can we identify the correlation and causality among Simpson's paradoxes in different subspaces? This may involve causal inference crossing multiple subspaces.

Fourth, there are many variant definitions of Simpson's paradox, such as those on continuous data [1], using association [3], and constraining sub-populations in different ways [14]. It is interesting to extend our approach and anal-

ysis to those variants and explore the general principles.
Fifth, in various applications, a user may want to apply various constraints to direct the search of Simpson's paradox, such as the size on sub-population size, the characteristics of attributes, and even the relations among the found Simpson's paradoxes. How to utilize those constraints to speed up the search and build an interactive, human-in-the-loop search system is an interesting engineering project.
In addition to the scientific future work, this project also gains us experience and confidence in promoting data science to youth and talented high-school students. The first author has gained highly rewarding experience in basic research methodology. The second author has learned how to devise a research project that carries the mission of research on data science frontier and, at the same time, fits the capability and capacity of a talented high school student for a summer internship. It is extremely encouraging to see that the first author's wonderful potential has been realized through a few months of hands-on research and development.

7. REFERENCES

[1] Nazanin Alipourfard, Peter G. Fennell, and Kristina Lerman. Can you trust the trend? discovering simpson's paradoxes in social data. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 19–27, New York, NY, USA, 2018. Association for Computing Machinery.

[2] Kevin Beyer and Raghu Ramakrishnan. Bottom-up computation of sparse and iceberg cube. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD '99, pages 359–370, New York, NY, USA, 1999. Association for Computing Machinery.

[3] Peter J Bickel, Eugene A Hammel, and J William O'Connell. Sex bias in graduate admissions: Data from berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175):398–404, 1975.

[4] Istituto Superiore di Sanità. Epidemia COVID-19: Aggiornamento nazionale, 09 Marzo 2020 – ore 16:00. `https://www.epicentro.iss.it/coronavirus/bollettino/Bollettino-sorveglianza-integrata-COVID-19_09-marzo-2020.pdf`.

[5] Carem C. Fabris and Alex A. Freitas. Discovering surprising instances of simpson's paradox in hierarchical multidimensional data. *International Journal of Data Warehousing and Mining*, 2(1):27–49, 2006.

[6] J. Gray, A. Bosworth, A. Lyaman, and H. Pirahesh. Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals. In *Proceedings of the Twelfth International Conference on Data Engineering*, pages 152–159, 1996.

[7] Jiawei Han, Jian Pei, and Hanghang Tong. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 4th edition, 2022.

[8] Laks V. S. Lakshmanan, Jian Pei, and Jiawei Han. Quotient cube: How to summarize the semantics of a data cube. In *Proceedings of the 28th International Conference on Very Large Data Bases*, VLDB'02, pages 778–789. VLDB Endowment, 2002.

[9] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In *Proceedings of the 7th International Conference on Database Theory*, ICDT 99, pages 398–416, Berlin, Heidelberg, 1999. Springer-Verlag.

[10] Judea Pearl. Comment: Understanding simpsons paradox. *The American Statistician*, 68(1):8–13, 2014.

[11] Karl Pearson, Alice Lee, and Leslie Bramley-Moore. Vi. mathematical contributions to the theory of evolution. —vi. genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred racehorses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 192:257–330, 1899.

[12] E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

[13] Jan Sprenger and Naftali Weinberger. Simpson's Paradox. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

[14] Julius von Kügelgen, Luigi Gresele, and Bernhard Schölkopf. Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects. *arXiv e-prints*, page arXiv:2005.07180, May 2020.

[15] Jeff Witmer. How much do minority lives matter? *Journal of Statistics Education*, 23(2), 2015.

[16] Zunyou Wu and Jennifer M. McGoogan. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72,314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA*, 323(13):1239–1242, 04 2020.

[17] G. Udny Yule. Notes on the theory of association of attributes in statistics. *Biometrika*, 2(2):121–134, 1903.

[18] Yihong Zhao, Prasad M. Deshpande, and Jeffrey F. Naughton. An array-based algorithm for simultaneous multidimensional aggregates. *SIGMOD Rec.*, 26(2):159–170, jun 1997.