Learning to Discover Various Simpson's Paradoxes

Jingwei Wang
Ant Group
Shanghai, China
wangjingwei.wjw@antgroup.com

Jianshan He
Ant Group
Beijing, China
yebai.hjs@antgroup.com

Weidi Xu
Ant Group
Shanghai, China
weidi.xwd@antgroup.com

Ruopeng Li
Ant Group
Beijing, China
ruopeng.lrp@antgroup.com

Wei Chu
Ant Group
Hangzhou, China
weichu.cw@antgroup.com

ABSTRACT

Simpson's paradox is a well-known statistical phenomenon that has captured the attention of statisticians, mathematicians, and philosophers for more than a century. The paradox often confuses people when it appears in data, and ignoring it may lead to incorrect decisions. Recent studies have found many examples of Simpson's paradox in social data and proposed a few methods to detect the paradox automatically. However, these methods suffer from many limitations, such as being only suitable for categorical variables or one specific paradox. To address these problems, we develop a learning-based approach to discover various Simpson's paradoxes. Firstly, we propose a framework from a statistical perspective that unifies multiple variants of Simpson's paradox currently known. Secondly, we present a novel loss function, *Multi-group Pearson Correlation Coefficient* (**MPCC**), to calculate the association strength of two variables of multiple subgroups. Then, we design a neural network model, coined **SimNet**, to automatically disaggregate data into multiple subgroups by optimizing the MPCC loss. Experiments on various datasets demonstrate that SimNet can discover various Simpson's paradoxes caused by discrete and continuous variables, even hidden variables. The code is available at https://github.com/ant-research/Learning-to-Discover-Various-Simpson-Paradoxes.

CCS CONCEPTS

• **Mathematics of computing** → **Statistical paradigms**; • **Information systems** → **Data mining**; • **Computing methodologies** → *Neural networks*.

KEYWORDS

Simpson's paradox, neural networks, data mining

ACM Reference Format:
Jingwei Wang, Jianshan He, Weidi Xu, Ruopeng Li, and Wei Chu. 2023. Learning to Discover Various Simpson's Paradoxes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3580305.3599859

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599859

1 INTRODUCTION

Simpson's paradox is a famous statistical phenomenon first described by Karl Pearson et al. in 1899 [26] and then named by Edward Simpson in 1951 [33]. It originally referred to the fact that the association between a pair of variables (T, Y) reverses sign conditioning on a third variable, Z. A well-known example of Simpson's paradox comes from a study of gender bias in UC Berkeley graduate admissions [7]. There appears to be a statistically significant bias against women in the aggregate admission data. However, when data is disaggregated by department, women have a slight advantage over men in most departments (see Table 3). This association reversal in Simpson's paradox is surprising and therefore has attracted a great deal of attention from research scholars over the past century [5, 12, 23, 34].

Simpson's paradox is not an uncommon statistical phenomenon and occurs in many real-life contexts [20]. For instance, Judea Pearl [24] recently pointed out the presence of Simpson's paradox in the Centers for Disease Control (CDC) data on coronavirus. The CDC data shows that whites have a lower case fatality rate in every age category (except ages 0-4) than non-whites. However, when aggregating all of the ages, whites have a higher fatality rate. The paradox arises because more whites are over 75 than non-whites, and older people are at high risk of dying from COVID. A similar paradox has been presented in [35] that COVID-19 case fatality rates are lower in Italy for every age group but higher overall when compared with China. This phenomenon can be explained by a stark difference in case demographic between the two countries. Besides, a recent study showed that the typical offline evaluation of recommendation systems suffers from Simpson's paradox that Model A outperforms Model B in general. Still, the situation reverses when stratifying the data [14]. The above instances of Simpson's paradox show that opposite conclusions can be reached depending on whether the data are aggregated or disaggregated. Ignoring Simpson's paradox may lead to spurious associations being accepted or even mistaken for causality, which puts decision-makers at great risk of taking improper strategies [28]. For example, Lerman [17] found Simpson's paradox in a large amount of social and behavioral data and suggested the trends discovered in aggregated data result in wrong conclusions about the underlying individual behavior. Therefore, detecting the presence of Simpson's paradox in data becomes critical for data-driven decision-making.

However, automated tools for discovering Simpson's paradox have not received much attention until recent years. The long-term

Table 1: Characteristics of SimNet and existing methods. * Variable refers to the condition variable Z.

Methods	Learning -based	Paradox type	Variable type*	Variable number*
[31, 32, 36]	×	AR	Discrete	Single
[1, 2]	×	AAR	Discrete Continuous	Single
SimNet (ours)	√	AR, AAR, AMP, YAP	Discrete Continuous	Single Multiple

neglect may be due to two reasons. First, most of the reported data with Simpson's paradox are from medical or sociological studies [15, 33]. The data in these studies generally have fewer variables, and the dataset sizes are very small. One can quickly find Simpson's paradox by manual computation without developing automated tools. Second, early studies focused on binary or categorical variables with a few values [7, 37]. For the two types of variables, it is easy to detect the presence of Simpson's paradox in data. The approach is to divide the data into subgroups based on all values taken by a variable and then observe if the trend of each group is opposite to the overall trend. Yet, there are many limitations to this simple approach. For example, grouping by value cannot be applied to continuous variables. In addition, this manual detection approach is far from competent in situations where there are many variables and a large amount of data.

Recently, a few works have proposed methods for the automatic detection of Simpson's paradox. For instance, Xu et al. [36] proposed an algorithm to find Simpson's reversal using categorical variables to partition the whole data set into subgroups. Alipourfard et al. [1] designed a trend-based algorithm to detect Simpson's paradox by binning elements according to the values of variables, which can be applied to discrete variables with large range. They also devised more sophisticated binning techniques to disaggregate data for continuous variables [2]. The above studies have promoted the automatic discovery of Simpson's paradox but still suffer from some limitations. First, there are multiple variants of Simpson's paradox, but these methods can only detect one specific paradox. Second, methods that rely on complicated binning techniques require manually setting the number and size of subgroups, which cannot be learned automatically from data. Third, all existing methods only detect Simpson's paradox caused by a single variable and fail to discover those caused by multiple variables.

To address the above problems, we develop a learning-based approach to discover Simpson's paradox. Firstly, we propose a framework from a statistical perspective that unifies multiple variants of Simpson's paradox currently known. Secondly, we design a novel function, i.e., ***M**ulti-group **P**earson **C**orrelation **C**oefficient* (**MPCC**), to calculate the association strength of two variables of multiple subgroups, which supports differential operations and can be used as a loss function for neural networks. Then, we formulate the discovery of **Sim**pson's Paradox as an optimization problem and design a simple neural **net**work to solve it, coined **SimNet**. SimNet can automatically disaggregate data into multiple subgroups by optimizing the proposed MPCC loss. To the best of our knowledge, this is the first work using a neural network for Simpson's paradox discovery. The advantages of our method over the existing methods are shown in Table 1. Experiments on various datasets demonstrate that our method has superior performance in discovering Simpson's paradox caused by single and multiple variables, regardless of the type of variables.

The main contributions of this paper are summarized below:

- We present a unified framework that incorporates multiple variants of Simpson's paradox currently known.
- We design a neural network model (**SimNet**) to discover Simpson's paradox, handling both discrete and continuous variables, as well as single or multiple variables.
- Extensive experimental results demonstrate that our model **SimNet** can discover various Simpson's paradoxes.

2 RELATED WORK

Simpson's Paradox Detection. Simpson's paradox has been studied over a long period of time, and many versions have been proposed by scholars [1, 11, 30, 33, 37]. We have found four different definitions in the literature [1, 34], which will be described in the next section. The common version of Simpson's paradox is called association reversal. Specifically, if we partition the data into multiple subgroups, each representing a specific value of the third variable Z, the phenomenon appears as a sign reversal of the association between (T, Y) measured in the disaggregated subgroups relative to the aggregated data, which describes the population as a whole. The aforementioned methods [1, 2, 31, 32, 36] all rely on this definition to detect the Simpson's paradox. However, the shortcomings of these methods are also obvious. For example, for unordered categorical variables such as college and species, it is unable to bin the elements of a categorical variable by comparing the magnitude of the values. Moreover, the number of subgroup combinations increases dramatically as the number of subgroups increases. When a variable takes many values, it is clearly infeasible to traverse all combinations. Besides, Simpson's paradox can be caused by a hidden variable that is unobserved in the data. To our knowledge, there is no research that has focused on the discovery of Simpson's paradox due to hidden variables.

Simpson's Paradox and Causal Inference. Data exhibiting Simpson's paradox are unproblematic from the perspective of mathematics and probability theory but nevertheless strike many people as surprising. Understanding the paradox is essential for drawing the correct conclusions from the data. Many researchers have widely discussed its role in causal inference and its implications for probabilistic theories of causality [4, 8, 13, 30]. For example, Pearl has studied the relation between Simpson's paradox and causal inference many times in the framework of graphical causal models [21–24]. Recently, he pointed the causal resolution of Simpson's paradox [23] to be one of the crown achievements of causal inference [25]. In his view, the paradox lays bare the core differences between causal and statistical thinking, and its resolution brings an end to a century of debates and controversies by the best philosophers of our time. Although the causal graph model provides a rational explanation for Simpson's paradox, detecting methods of the paradox still remains to be developed.

3 METHODS

3.1 Simpson's Paradox: A Unified Framework

Various versions of Simpson's paradox have been presented during the last century, and we found four different definitions in the literature. The first three were summarised in [34], which are Association Reversal (AR) [30], Yule's Association Paradox (YAP) [37], and Amalgamation Paradox (AMP) [11]. The three paradoxes are defined by k, a measure of the strength of the association between two variables. There is a logical order inferred by the generality of their definitions, i.e., **YAP** $\Rightarrow$ **AR** $\Rightarrow$ **AMP**. In other words, YAP is a special case of AR, and AR is a special case of AMP. However, whether the association between a pair of variables is significant or not is not taken into account in the above three versions. Therefore, Alipourfard et al. [1] proposed a new definition incorporating the strength of the association and a significance test (p-value). We call their version Averaged Association Reversal (AAR) since it takes into account the average trend across subgroups.

Previous studies employ different measures to represent the association strength k between two variables. For instance, Xu et al. [36] use the correlation coefficient to measure a trend between two variables. Alipourfard et al. [1] employ a linear regression model to compute the relationship between two variables and then use the fitted trend parameter to represent the correlation. To incorporate the four versions of Simpson's paradox into one framework and figure out their logical strength, we specify the definition of k by Pearson correlation coefficient (PCC) with its p-value. PCC is defined as the ratio between the covariance of two variables T and Y and the product of their standard deviations, that is,

$$\rho_{T,Y} := \frac{cov(T,Y)}{\sigma_T \sigma_Y}.$$

$\rho_{T,Y}$ can also be written as

$$\rho_{T,Y} = \frac{\mathbb{E}[TY] - \mathbb{E}[T]\,\mathbb{E}[Y]}{\sqrt{\mathbb{E}[T^2] - (\mathbb{E}[T])^2}\sqrt{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}}. \tag{1}$$

Definition 3.1 (*Association Strength*)**.** Let ρ be the PCC between a pair of variables, and p be the p-value. With a significance level α (e.g., 0.05), the association strength k between the two variables is defined as

$$k := \begin{cases} \rho & \text{if } p < \alpha, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Suppose a dataset $\mathcal{D}$ with n samples is divided into m subgroups $\mathcal{D}_i (i = 1, ..., m)$, with the corresponding strength of association denoted by k_i. Then the four variants of Simpson's paradox are given as follows.

Definition 3.2 (*Association Reversal* (AR))**.** AR occurs if and only if one of the following two conditions holds,

$$\begin{array}{ll} \text{(AR1)} & k \le 0 \text{ and } k_i \ge 0 \quad \forall\, i = 1, ..., m; \\ \text{(AR2)} & k \ge 0 \text{ and } k_i \le 0 \quad \forall\, i = 1, ..., m, \end{array} \tag{3}$$

where either side has to be strict.

AR is the most widely concerned Simpson's paradox, while YAP is the first to receive attention. As a special case of AR, YAP [37] allows the right-hand side in Eq. (3) to be equal.

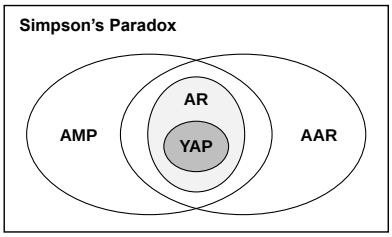

Figure 1: Relation of four variants of Simpson's paradox.

Definition 3.3 (*Yule's Association Paradox* (YAP))**.** YAP occurs when there is no association in all subgroups, but an association emerges in the overall dataset, that is,

$$k_i = 0 \quad \forall\, i = 1, ..., m \quad \text{but} \quad k \neq 0. \tag{4}$$

A more general version of Simpson's paradox is the AMP [11]. AMP occurs when the overall strength of association is greater (or less) than that of each subgroup.

Definition 3.4 (*Amalgamation Paradox* (AMP))**.** AMP occurs when one of the following two conditions holds,

$$\begin{array}{ll} \text{(AMP1)} & k > \max\ k_i \quad \forall\, i = 1, ..., m; \\ \text{(AMP2)} & k < \min\ k_i \quad \forall\, i = 1, ..., m. \end{array} \tag{5}$$

Next, we define the fourth variant AAR which takes the sign of k into consideration and compare the averaged sign of all subgroups with the sign of the overall data.

Definition 3.5 (*Averaged Association Reversal* (AAR))**.** Let $\text{sgn}(x) \in \{1, -1, 0\}, \forall x \in \mathbb{R}$ be the sign function. AAR occurs when

$$\text{sgn}(k) \neq \text{sgn}\Big(\frac{1}{m}\sum_{i=1}^{m} \text{sgn}(k_i)\Big). \tag{6}$$

It is clear that AR is a special case of AAR, but there is no logical relation between AMP and AAR. Therefore, the four variants of Simpson's paradox follow the logical relation as (shown in Fig. 1):

$$\textbf{YAP} \Rightarrow \textbf{AR} \Rightarrow \textbf{AMP}, \text{ and } \textbf{YAP} \Rightarrow \textbf{AR} \Rightarrow \textbf{AAR}. \tag{7}$$

3.2 Multi-group Pearson Correlation Coefficient

The goal of Simpson's paradox discovery is to determine whether one or more paradoxes described above will occur after disaggregating data by one variable Z. If the data is divided into m subgroups according to the value taken by Z, the common technique is to calculate the PCC of each subgroup separately and then compare it with the PCC of the overall data to discover the paradox [1, 36].

To make it possible to compute the PCC of multiple subgroups simultaneously, we propose a new function coined **MPCC**. Let $C \in \{0, 1\}^{n \times m}$ be a group assignment matrix derived from Z, where $C_{i,j} = 1$ if sample i belongs to group j, and 0 otherwise.

Definition 3.6 (*Multi-group Pearson Correlation Coefficient* (MPCC))**.** MPCC is defined as

$$MPCC(T, Y, C) := \frac{\mathbb{E}_C[TY] - \mathbb{E}_C[T] \odot \mathbb{E}_C[Y]}{\sqrt{D_C(T)} \odot \sqrt{D_C(Y)}}, \tag{8}$$

where $\mathbb{E}_C[TY]$, $\mathbb{E}_C[T]$, and $\mathbb{E}_C[Y]$ denote the expectation of TY, T, and Y of multiple subgroups given by C, respectively; $D_C(T)$ and $D_C(Y)$ denote the standard deviation of multiple groups given by C, respectively. In particular, $\mathbb{E}_C[TY] = (C^\top \mathbf{1}_n) \odot (C^\top (T \odot Y))$, where $\mathbf{1}_n$ is a $1 \times n$ column vector having all n elements equal to one. $\mathbb{E}_C[T] = C^\top T$, and $\mathbb{E}_C[Y] = C^\top Y$. Then,

$$\begin{aligned} D_C(T) &= \mathbb{E}_C[T^2] - \mathbb{E}_C[T] \odot \mathbb{E}_C[T] \\ &= (C^\top \mathbf{1}_n) \odot (C^\top (T \odot T)) - (C^\top T) \odot (C^\top T). \end{aligned} \tag{9}$$

$$\begin{aligned} D_C(Y) &= \mathbb{E}_C[Y^2] - \mathbb{E}_C[Y] \odot \mathbb{E}_C[Y] \\ &= (C^\top \mathbf{1}_n) \odot (C^\top (Y \odot Y)) - (C^\top Y) \odot (C^\top Y). \end{aligned} \tag{10}$$

The output of the MPCC function is the PCC of the m subgroups. With this result, we can determine whether Simpson's paradox occurs by grouping (T, Y) according to the group assignment C given by Z. In particular, MPCC supports differential operations and thus can be used as a loss function for neural networks.

3.3 Simpson's Paradox Discovery

As mentioned above, a naive approach to discovering Simpson's paradox is that one firstly groups (T, Y) by Z, then calculates the PCC of all subgroups, and finally compares them with the PCC of the overall data [1, 36]. However, this approach is only applicable when Z is a binary variable or a categorical variable with a few categories. It cannot be applied to a discrete variable with many elements as well as a continuous variable. For the latter two types of variables, combining some elements into one subgroup is necessary to reduce the number of subgroups. Otherwise, the number of subgroups is too large, leading to poor interpretability; and the number of samples within each subgroup is too small, resulting in insignificant correlations.

However, it is not an easy task to group data by conditioning on continuous or discrete variables with a large range. Using a discrete variable Z_d with n different values as an instance, when one attempts to combine the elements by the values of Z_d, they immediately confront the combinatorial explosion problem. The larger n is, the more possible ways of partitioning the data can be.

Theorem 3.1. *When a discrete variable with n different values is selected as a conditional variable to disaggregate data for Simpson's paradox discovery, the number of ways to partition the data is $B_n - 1$, where B_n is the n-th Bell number.*

Proof. Assuming that the given number of subgroups is m ($2 \le m \le n$), the number of ways to partition a set of n distinct elements into the m nonempty subgroups are known as the Stirling number of the second kind, denoted by $S(n, m)$. The Bell number can be formulated as a sum of Stirling numbers of the second kind [6], i.e., $B_n = \sum_{m=0}^{n} S(n, m)$. Since $2 \le m \le n$, the number of ways to partition the data is $\sum_{m=2}^{n} S(n, m) = B_n - 1$. □

It is known that $B_5 = 52, B_{10} = 115975, B_{15} \approx 1.38 \times 10^9$, and $B_{20} \approx 5.17 \times 10^{13}$. This fact and Theorem 3.1 tell us that for discrete variables with large ranges, it is infeasible to discover Simpson's paradoxes by testing all combinations of values. Thus, Alipourfard et al. [1] employed fixed-size bins to partition the data and avert the issue of combinatorial explosion. Yet, this approach is susceptible to bin size and unsuitable for unordered discrete variables.

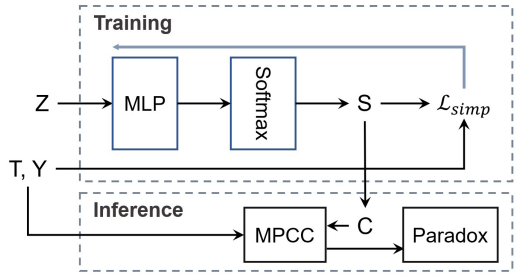

Figure 2: Schema of the SimNet model.

To overcome the aforementioned limitations, we develop a new learning-based method to discover Simpson's paradox. First, we consider Simpson's paradox discovery as an optimization problem. Assuming that the strength of association between two variables T and Y in the aggregated data is greater than zero (i.e., $k > 0$), we can minimize the association strength of T and Y in each subgroup to discover various paradoxes except for AMP2. Then, this problem can be formulated as

$$\min \ \frac{1}{m} \sum_{i=1}^{m} k_i. \tag{11}$$

It is clear that $\frac{1}{m} \sum_{i=1}^{m} k_i = \text{avg}(MPCC(T, Y, C))$, where $\text{avg}(x)$ represents the averaging function and C corresponds to the group assignment of each sample. Thus, the optimization objective can be expressed as

$$\underset{C \in \{0,1\}^{n \times m}}{\arg\min} \ \text{avg}(MPCC(T, Y, C)). \tag{12}$$

Now, the goal turns to finding a group assignment C that minimizes the mean of PCC for all subgroups.

Here, we design a simple neural network (**SimNet**) for learning the group assignment C that may lead to Simpson's paradox, as shown in Fig. 2. In particular, we employ a multi-layer perceptron (MLP) with a softmax operation on the output layer to obtain a soft group assignment S of samples using the conditional variable Z:

$$S = \text{softmax}(\text{MLP}(Z; \Theta_s)), \tag{13}$$

where Θ_s represents the trainable parameter. This network maps the value $Z_i = z$ of each sample into the i-th row of a soft group assignment matrix $S \in [0, 1]^{n \times m}$. The softmax operation guarantees $S_{ij} \in [0, 1]$ and enforces $S\mathbf{1}_m = \mathbf{1}_n$. Here, S is the relaxed continuous formulation of C that can be solved by gradient descent algorithms.

The parameter Θ_s of the network is optimized by minimizing an unsupervised loss function (referred to *Simpson loss*) composed of two terms:

$$\mathcal{L}_{simp} = \mathcal{L}_{mpcc} + \beta \mathcal{L}_{reg}, \tag{14}$$

where β is a balancing hyper-parameter.

The *mpcc loss* term, $\mathcal{L}_{mpcc} = \text{sign}(\rho_{T,Y}) \cdot \text{avg}(MPCC(T, Y, S))$, evaluates the MPCC of T and Y given by the soft group assignment S and is bounded by $-1 \le \mathcal{L}_{mpcc} \le 1$. Throughout the rest of this work, MPCC is calculated by the mean of PCC for all subgroups and represents the average trend of all subgroups.

The sign of $\rho_{T,Y}$ drives the optimization direction opposite to the trend of the aggregated data. Minimizing $\mathcal{L}_{mpcc}$ encourages samples with opposite trends to the aggregated data and consistent internal trends to be clustered together. When $\rho_{T,Y} > 0$, minimizing MPCC can induce the neural network to identify paradoxes such as AAR, YAP, AR, and AMP1. To discover AMP2, it is necessary to maximize MPCC, which can be done by merely setting $\text{sign}(\rho_{T,Y}) = -1$. The same applies to the case where $\rho_{T,Y} < 0$.

Besides, we introduce a *regularization loss* term, $\mathcal{L}_{reg}$, to balance the subgroup size. $\mathcal{L}_{reg} = \frac{\sqrt{m}}{n}\|\sum_{i=1}^{n} S_i^\top\|_2 - 1$, where $\|\cdot\|_2$ is the l_2 norm and $S_i^\top$ indicates the i-th column of $S^\top$ (i.e., the i-th row of S). It is easy to see that $0 \le \mathcal{L}_{reg} \le \sqrt{m} - 1$. $\mathcal{L}_{reg}$ reaches its maximum $\sqrt{m} - 1$, when all samples are assigned to one subgroup. $\mathcal{L}_{reg}$ reaches its minimum 0, when exactly $\frac{n}{m}$ samples are assigned to each subgroup.

With the unsupervised *Simpson loss*, the soft group assignment S is trained to segment data in a way that disaggregated data has significant differences from aggregated data. After training, the final group assignment is inferred by the *argmax* function, that is, $C = \text{argmax}(S)$. Then each sample is assigned to one subgroup, and the strength of association is calculated within each subgroup to determine if any variant of Simpson's paradox occurs. Note that this model can not only detect Simpson's paradox caused by a single variable (i.e., $Z \in \mathbb{R}^n$) but also detect one caused by multiple variables (i.e., $Z \in \mathbb{R}^{n\times m}$), which can be viewed as the proxy of a hidden variable (see Section 4.3.3).

4 EXPERIMENTS

4.1 Dataset

We consider a wide range of real-world datasets in which Simpson's paradox has been reported. The statistics of these datasets are shown in Table 2, including the number of discrete and continuous variables and their size. Here, the size refers to the number of complete records in data since we remove those incomplete records. The details of each dataset are provided as the followings.

Table 2: Statistics of these datasets.

Dataset	No. Discrete	No. Continuous	Size
Iris	1	4	150
Auto MPG	3	5	392
Titanic	4	0	2201
UC Berkeley	3	0	4526
Atlanta CES	4	1	8244
Synthetic	12	2	44000

Iris. This dataset consists of 50 samples from each of 3 species of Iris (setosa, versicolor, and virginica) [3, 10]. Each sample has 4 continuous attributes: sepal length, sepal width, petal length, and petal width. Here, the values of the only discrete variable (specie class) are encoded as numbers; that is, 0, 1, and 2 stand for setosa, versicolor, and virginica, respectively.

Auto MPG. The data concerns city-cycle fuel consumption in miles per gallon (MPG), containing 3 multi-valued discrete and 5 continuous variables [27]. The 3 multi-valued discrete variables are cylinders, model year, and origin. The 5 continuous variables are MPG, displacement, horsepower, weight, and acceleration. We remove 6 instances with unknown values for the *horsepower* variable and retain 392 complete records.

Titanic. This dataset provides information on the fate of 2201 passengers on the fatal maiden voyage of the ocean liner *Titanic*, summarized according to class, sex, age, and survival [9]. Class is a multi-valued discrete variable with 4 values: First, Second, Third, and Crew. The other 3 variables are dichotomous, such as sex (male or female), age (child or adult), and survived (yes or no).

UC Berkeley. This dataset comes from a study of gender bias among graduate school admissions to the University of California, Berkeley [7], involving 4526 applicants from the six largest departments. It has only three discrete variables: gender (male or female), department (A-F), and admission status (admitted or rejected).

Atlanta CES. This dataset[1] collects Atlanta city employee salary (CES) information for the year 2015 and contains the following fields: name of an employee, age, gender, ethnicity, job title, department, and annual salary. The 8246 employees involved in the dataset belong to 770 positions (job titles), and many job titles appear only once. We remove two fields (i.e., name of employee and job title) because neither is suitable as a conditional variable to disaggregate data. We also delete records with the ethnicity "Native Hawaiian or Other Pacific" because there are only 2 records. The remaining data retain four discrete and one continuous variable. Here, we treat age as a multi-valued discrete variable.

Synthetic. We generate the synthetic data by the following process:

$$T \sim \mathcal{N}(0, \sigma_t) + Z; \quad Y \sim \mathcal{N}(0, \sigma_y) + Z + 1 - T,$$

where Z is a variable of randomly sampled values $(1, 2, 3, 4)$ indicating four subgroups, $\sigma_t = 0.5$, and $\sigma_y = 0.3$. This generation process introduces hidden confounding between T and Y since they depend on Z. Specifically, T and Y are positively correlated overall ($PCC \approx 0.65$) while negatively correlated in each subgroup ($MPCC \approx -0.86$) conditioning on Z, as shown in Fig. 8. Following [19], we create the noisy proxy X (i.e., multiple observed variables) for the hidden variable Z. We first code the Z with one-hot encoding and replicate this three times. We then randomly and independently flipped each of these 12 bits (i.e., 12 observed variables). We vary the probability of flipping from 0.05 to 0.5, the latter indicating there is no direct information about the hidden variable Z. Under each noise level, we generate the synthetic data with 4000 samples (1000 in each subgroup).

4.2 Experimental Setup

4.2.1 Baselines. To investigate the effectiveness of our method, we compare it with the detection algorithms for Simpson's paradoxes proposed in [1, 31, 32, 36]. We refer to these algorithms as naive methods, and their results are shown in Appendix A.

[1] https://github.com/bbrewington/atlanta-salary-data

4.2.2 Setup. We use PyTorch to implement the SimNet model where the MLP consists of 3 fully-connected layers and SELU activation function [16]. SimNet is trained using AdamW [18] with mini-batch optimization. Unless otherwise specified, we set the hyper-parameters as $\beta = 10$, batch size=64. If Z is a discrete variable, we encode it as a one-hot vector before feeding it into SimNet. In all experiments, we set the significance level $\alpha = 0.05$. The computational complexity of SimNet is discussed in Appendix B.

4.3 Discovering Simpson's Paradox

4.3.1 Conditioning on a discrete variable. We use three benchmark datasets (UC Berkeley, Iris, and Atlanta CES) to test the performance of the proposed method in discovering Simpson's paradox by conditioning on a discrete variable. In particular, we focus on the ability of SimNet to disaggregate data by merging the elements of one discrete variable.

Table 3: UC Berkeley gender bias. The percentage indicates the acceptance rate.

Department	Male	Female	Size	PCC	p	k
A	62.1%	**82.4%**	933	0.14	0.00	0.14
B	63.0%	**68.0%**	585	0.02	0.62	0
C	**36.9%**	34.1%	918	-0.03	0.39	0
D	33.1%	**34.9%**	792	0.02	0.59	0
E	**27.7%**	23.9%	584	-0.04	0.32	0
F	5.9%	**7.0%**	714	0.02	0.54	0
Total	**44.5%**	30.4%	4526	-0.14	0.00	-0.14

Table 4: AR discovered by SimNet in UC Berkeley.

Subgroup		Department	Men	Women	PCC	p
$m = 2$	1	A,B	62.5%	**79.7%**	0.10	0.00
	2	C,D,E,F	25.5%	**26.5%**	0.01	0.54
$m = 3$	1	A,B	62.5%	**79.7%**	0.10	0.00
	2	D	33.1%	**34.9%**	0.02	0.59
	3	C, E, F	21.9%	**24.1%**	0.03	0.23
Overall		A~F	**44.5%**	30.4%	-0.14	0.00

UC Berkeley. The most famous example of Simpson's paradox is the gender bias among graduate school admissions to the University of California, Berkeley. As shown in Table 3, the overall admissions of six departments (including 2691 male applicants and 1835 female applicants) in 1973 showed that men were more likely to be admitted than women. However, when looking at most departments, the trend actually reverted. For data with categorical variables, existing algorithms [1, 36] disaggregate data by conditioning on all values of a categorical variable (e.g., department) to discover Simpson's paradox, failing to combine some values into one subgroup. Their results are consistent with those in Table 3, where four of the six departments show Simpson's reversal. Table 3 shows an AR paradox since $k_i \geq 0$ for all departments while $k < 0$ overall.

Our model benefits from its nonlinear mapping capability to generate subgroups in which Simpson's reversal appears in each subgroup. Table 4 shows the results of our method found in the UC Berkeley data when the number of subgroups is set to 2 and 3, respectively. As can be seen, our method combines the two departments, A and B, into one subgroup, producing a significant reversal. In contrast, the previous reversal in department B was insignificant (see p in Table 3). Moreover, when the number of subgroups is set to 2 ($m = 2$), departments C, D, E, and F are merged into one group by our model to create a new Simpson's reversal that is impossible with previous methods. A similar result is obtained when $m = 3$. A new reversal appears when departments C, E, and F are merged into one group.

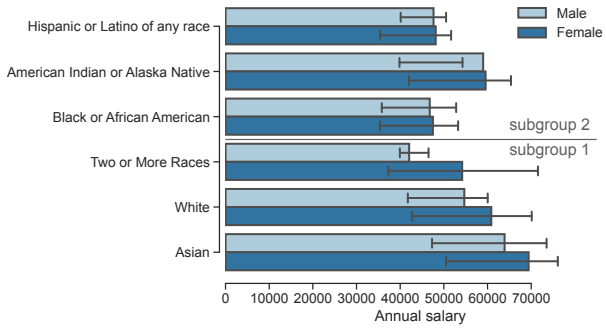

Figure 3: Simpson's paradox (AAR) discovered in Atlanta CES by conditioning on ethnicity.

Table 5: Simpson's paradoxes (AR) discovered in Atlanta CES by conditioning on the department.

Subgroup		Department	Size	PCC	p
$m = 2$	1	AUD, AWD, CCN, COR, CRB, DHR, DIT, DOA, DOF, DOP, DWM, ETH, EXE, JDA, LAW, PCD, PDA, SOL	3587	0.05	0
	2	AFR, APD, DPW, PRC	4657	0.09	0
$m = 3$	1	AFR, DPW, PRC	2308	0.06	0
	2	APD, COR	2669	0.19	0
	3	AUD, AWD, CCN, CRB, DHR, DIT, DOA, DOF, DOP, DWM, ETH, EXE, JDA, LAW, PCD, PDA, SOL	3267	0.04	0.01
Overall		All	8244	0	0.95

Atlanta CES. Previous works showed no connection between gender and the annual salary of city employees in the aggregated data. However, when the data were stratified by ethnicity, women's median salary was actually higher than men's by an obvious margin in some ethnic groups such as White and Asian (see Fig. 3). Existing

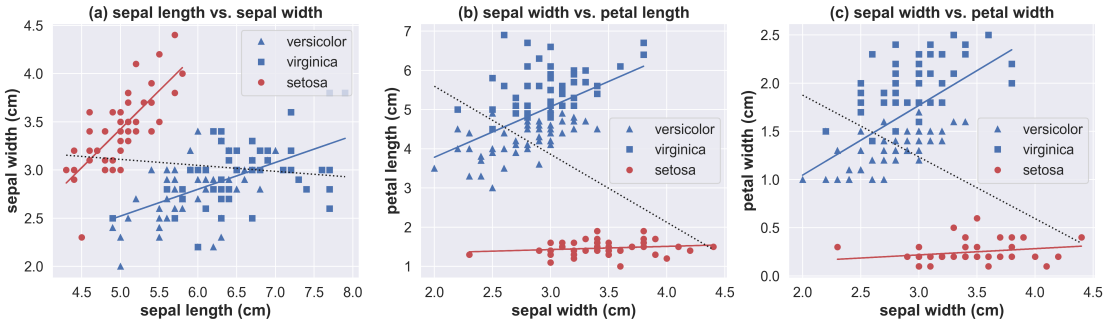

Figure 4: Simpson's paradoxes (all AR) discovered in Iris by conditioning on the class. Black dotted lines show the overall trends and colored solid lines show the subgroup trends. Each subgroup is represented by the same color, and each class is represented by the same shape.

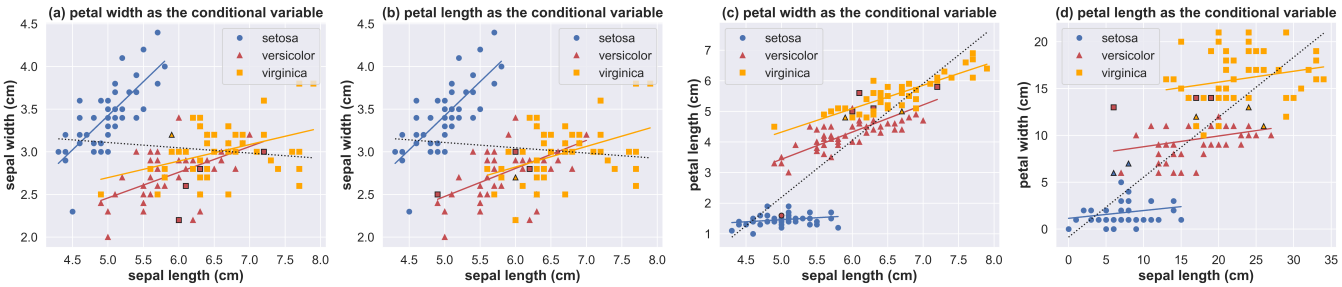

Figure 5: Simpson's paradoxes (a&b: AR, c: AMP, d: YAP) discovered in Iris by conditioning on continuous variables. Black dotted lines show the overall trends and colored solid lines show the subgroup trends. Each subgroup is represented by the same color, and each class is represented by the same shape. Shapes with black boxes represent samples of misclassification.

algorithms can find the AAR paradox by conditioning on all values of ethnicity, and the numerical results are shown in Table 8. Table 9 shows another AAR paradox found by SimNet where the six ethnic groups are divided into two subgroups with significant differences. The first subgroup contains White, Asian, and "Two or More Races", in which females' median salary is higher than males. The second subgroup includes the other ethnic groups in which there is no significant difference in income between men and women. This result can also be seen in Fig. 3, where the grey line indicates the demarcation.

In addition, we discover a new AAR paradox when the data are stratified by the department, and the result given by naive methods can be seen in Table 10. Note that there are 22 departments, and their abbreviations are shown in Table 7. As mentioned above, it is infeasible for naive methods to discover Simpson's paradox by combining the elements of a discrete variable with a large range. However, SimNet has this capability and is able to discover more rigorous paradoxes, such as AR. Table 5 shows two AR paradoxes found by our method in Atlanta CES when the number of subgroups is set to 2 and 3, respectively. When $m = 2$, our method divides 22 departments into two non-overlapping subgroups, each of which shows a significant positive correlation between gender and income. A similar result is obtained when $m = 3$, and the second subgroup exhibits a stronger correlation than either department of APD and COR. The above results again demonstrate the excellent performance of SimNet in discovering Simpson's paradox when conditioned on discrete variables.

Iris. The Iris data contains three classes of 50 instances each, where each class refers to a type of iris plant. Following [36], we use the categorical attribute (i.e., class) of Iris as the conditioning variable to disaggregate data. Our model not only found the nine trend reversals of three pairs of continuous attributes that were reported in [36] but also discovered another six trend reversals with two classes merged into one subgroup. As shown in Figure 4, versicolor and virginica are much closer to each other, and there is a large gap between them and setosa. Because *"virginica and versicolor were from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus"* [3]. Our model captured this feature and grouped them into one subset in the three cases. All three paradoxes in Fig. 4 belong to the AR type, and the numerical results are summarized in Tables 12-14.

4.3.2 Conditioning on a continuous variable. Our method can disaggregate data by conditioning on not only discrete but also continuous variables without any specified operation, going beyond the capability of previous methods [1, 2, 36].

Iris. In Fig. 5, we show four interesting cases of Simpson's paradox where the Iris data is disaggregated by continuous variables. The two cases in Fig. 5(a) and 5(b) are AR, which is the standard

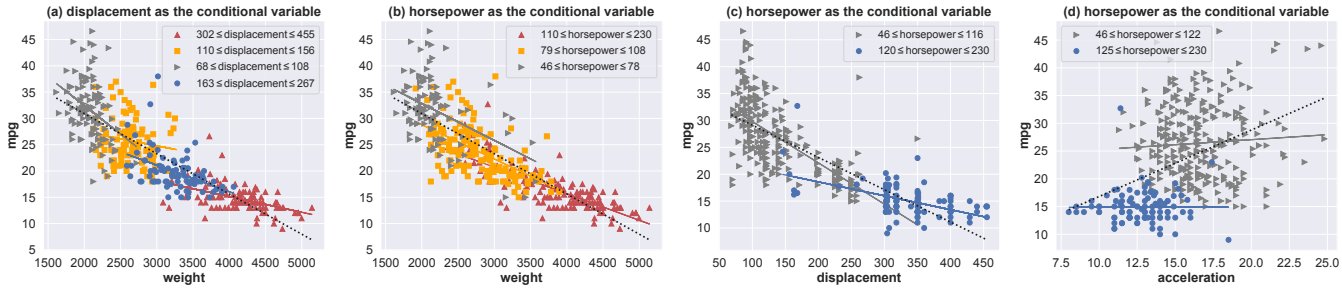

Figure 6: Simpson's paradoxes (a,b,c: AMP, d: YAP) discovered in Auto MPG by conditioning on continuous variables. Black dotted lines show the overall trends and colored solid lines show the subgroup trends. Samples from the same subgroup have the same color and shape.

Simpson's reversal. Overall, sepal length and sepal width are negatively correlated, but the correlation in each subgroup becomes significantly positive when data are grouped by petal width or petal length. The third case in Fig. 5(c) shows an AMP between sepal length and petal length conditioning on petal width. Overall, sepal length and petal length are highly positively correlated, but the correlation in each subgroup becomes weak when samples are grouped according to petal width. It can be seen in Table 17 that the PCC of the three subgroups is lower than that of the whole. The fourth case in Fig. 5(d) shows a YAP between sepal length and petal width conditioning on petal length. Overall, sepal length and petal width are highly positively correlated. Still, none of the correlations are statistically significant when grouped according to petal length since $p > 0.05$ in all three subgroups (see Table 18).

Surprisingly, the subgroups detected by our model are highly consistent with the real classes of Iris data. We use Marco-F1 to evaluate the matching degree of detected subgroups and real classes, and it exceeds 0.92 in all four cases. As shown in Fig. 5, these few samples of misclassification are denoted by black boxes. Note that our model does not use any category information, and the results are learned in an unsupervised manner. In particular, the threshold for partitioning continuous variables in these cases is found automatically by our model without any data preprocessing. Details about the four cases are summarized in Tables 15-18.

Auto MPG. The Auto MPG dataset contains 3 discrete and 5 continuous variables, with the main focus being on which variables are associated with MPG. Previous work [36] found six Simpson's reversals of three continuous variables by conditioning on discrete variables. As shown in the light grey row in Tables 19-21, however, five of the six reversals are not significant. According to our definitions, only two AMP paradoxes were found by the naive method.

Figure 6 shows four new paradoxes discovered by SimNet with continuous variables as the conditional variable. Overall, vehicle weight and MPG are negatively correlated, but the correlation within each group becomes slightly weakened, regardless of whether conditioning is done on displacement or horsepower. The correlation between displacement and MPG is similar. On the other hand, a positive correlation between acceleration and MPG is evident, but this correlation disappears after a horsepower-based segmentation. The thresholds for grouping shown in Fig. 6 suggest SimNet is able to segment data by conditioning on continuous variables automatically. Note that the trend line is obtained by linear fitting with the least squares method, whose slope is different from PCC. The numerical results of the four paradoxes are summarized in Tables 22-25.

4.3.3 Conditioning on multiple variables. In particular, our method can discover Simpson's paradox by conditioning on multiple variables, going beyond the capabilities of all known methods.

Atlanta CES. We continue to focus on the correlation between gender and annual salary but are conditional on two discrete variables (i.e., ethnicity and department). Both of the discrete variables are one-hot encoded. SimNet discovers four new paradoxes of two types (AR and AAR) when the number of subgroups is set from 2 to 5. Overall, there is no association between gender and annual salary; however, a positive correlation is observed in many of the subgroups, as shown in Table 11. Particularly, we display the feature distribution of different subgroups of the four cases in Fig. 7. It can be seen that the feature distribution of each subgroup is distinct from that of the others in each case. These results suggest that our method can segment the data into more homogeneous subgroups and uncover surprising subgroups that have patterns different from the overall.

Synthetic. To further show the power of our model in discovering Simpson's paradox by conditioning on multiple variables, we conduct experiments on the synthetic data with Simpson's paradox.

Table 6: MPCC and NMI values of SimNet for Simpson's paradox detection on synthetic data.

NL	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
MPCC	-0.86	-0.84	-0.79	-0.70	-0.54	-0.44	-0.27	-0.19	-0.10	0.03	0.06
NMI	1.00	1.00	0.96	0.92	0.83	0.76	0.66	0.60	0.56	0.52	0.50

The aim of the experiment is to discover Simpson's paradox from the noisy data. Note that it is impossible to find the original Simpson's paradox by conditioning on any one observed variable when the noise level is higher than 0.05. Table 6 shows the MPCC and NMI [29] values of SimNet on the synthetic data with different noise levels. Since the overall PCC of T and Y is positive, a negative

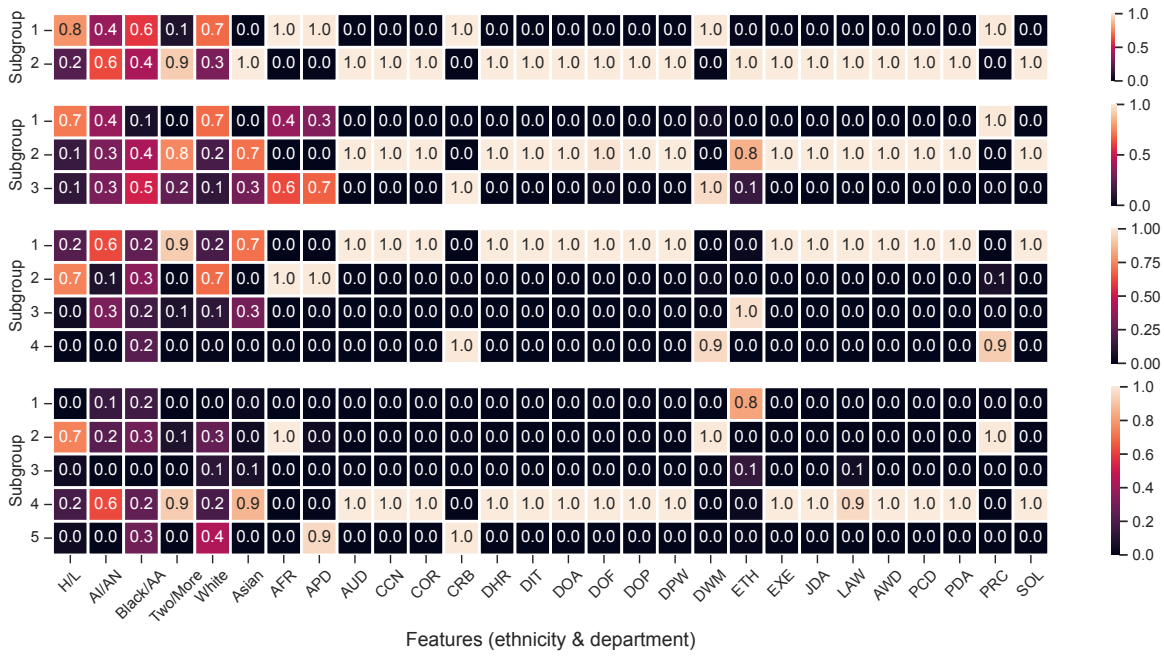

Figure 7: Feature distribution of different subgroups of Atlanta CES data when disaggregated by SimNet.

MPCC means that the subgroups detected by our model are likely to lead to Simpson's paradox, and lower is better. NMI values measure the agreement of the detected subgroups and the hidden variable Z, and a high NMI value indicates that the detected subgroups are close to Z. As can be seen from Table 6, our model achieves excellent performance when the noise level is below 0.3. Besides, Fig. 9 visualizes the inferred hidden variable Z at the noise level of 0.05, 0.15, and 0.25, respectively. The four subgroups in different colors represent the inferred hidden variable, which is highly consistent with the ground truth.

5 DISCUSSION

The above results show that our proposed method SimNet has an excellent performance in discovering various Simpson's paradoxes. In particular, SimNet is not restricted to the type of condition variables, handling both discrete and continuous variables, as well as single or multiple variables.

One limitation of SimNet is that it may not be suitable to measure the association between an unordered discrete variable and other variables. When T is an unordered discrete variable with a large range, the PCC value for T and Y is affected by the order of elements of T. Taking the Titanic data as an example, Class is an unordered discrete variable due to the inclusion of the Crew class. Crew class can not be placed higher or lower than the first class, second class, and third class since the cabins assigned to crew members were mixed up with these three. If the Crew class is placed above the first class, then an AR paradox can be observed when the data is disaggregated by gender (see Table 27). However, if the Crew class is placed below the third class, the paradox would be avoided (see Table 28). One solution is to remove the unordered elements in T and keep only the ordered elements. For instance, when the Crew class is not considered, a stable AMP paradox can be found in Titanic (see Table 29). Yet, this approach is not always applicable.

Another issue may be the choice of m when using SimNet. It is recommended to set m between 2 and 5, which is sufficient to discover Simpson's paradox if it exists in the data. A large m is also feasible for large datasets and variables with a large range.

6 CONCLUSION

Although Simpson's paradox is a famous phenomenon, it has not been exhaustively studied, still suffering from various definitions and limited detection ability. In this paper, we present a unified framework for studying Simpson's paradox from a statistical perspective and figure out the logical relationships between multiple variants. Further, we propose a learning-based method to discover various Simpson's paradoxes caused by single and multiple variables on datasets, regardless of variable types. Our results on various datasets demonstrate that **SimNet** can discover various Simpson's paradoxes and can be a powerful tool to analyze heterogeneous data, e.g., segmenting heterogeneous data into more homogeneous subgroups of similar elements. Moreover, discovering Simpson's paradox in data can serve as a stimulus for researchers to conduct comprehensive analyses of causality, explore plausible explanations, and suggest effective solutions.

ACKNOWLEDGMENTS

We thank Dr. Yu Wu for countless insightful discussions, technical support, and manuscript editing; Prof. Judea Pearl for helpful feedback; and Taifeng Wang, Xiaodong Yan, Lei Zhang, and Hongting Zhou for their insightful comments and support.

REFERENCES

[1] Nazanin Alipourfard, Peter G. Fennell, and Kristina Lerman. 2018. 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 2018)*. 19–27.

[2] Nazanin Alipourfard, Peter G. Fennell, and Kristina Lerman. 2018. Using Simpson's Paradox to Discover Interesting Patterns in Behavioral Data. In *Proceedings of the Twelfth International Conference on Web and Social Media (ICWSM 2018)*. 2–11.

[3] Edgar Anderson. 1936. The species problem in Iris. *Annals of the Missouri Botanical Garden* 23, 3 (1936), 457–509.

[4] Prasanta S. Bandyoapdhyay, Davin Nelson, Mark C. Greenwood, Gordon G. Brittan, and Jesse Berwald. 2010. The logic of Simpson's paradox. *Synthese* 181 (2010), 185–208.

[5] Prasanta S. Bandyopadhyay, R. Venkata Raghavan, Don Wallace Dcruz, and Gordon Brittan Jr. 2015. Truths about Simpson's Paradox: Saving the Paradox from Falsity. In *Logic and Its Applications - 6th Indian Conference (ICLA 2015)*, Vol. 8923. 58–73.

[6] HW Becker and John Riordan. 1948. The arithmetic of Bell and Stirling numbers. *American Journal of Mathematics* 70, 2 (1948), 385–394.

[7] Peter J. Bickel, E A Hammel, and J W O'connell. 1975. Sex Bias in Graduate Admissions: Data from Berkeley. *Science* 187 (1975), 398 – 404.

[8] Colin Ross Blyth. 1972. On Simpson's Paradox and the Sure-Thing Principle. *J. Amer. Statist. Assoc.* 67 (1972), 364–366.

[9] Robert J MacG Dawson. 1995. The "unusual episode" data revisited. *Journal of Statistics Education* 3, 3 (1995).

[10] Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 2 (1936), 179–188.

[11] I. J. Good and Y. Mittal. 1987. The Amalgamation and Geometry of Two-by-Two Contingency Tables. *The Annals of Statistics* 15, 2 (1987), 694–711.

[12] Yue Guo, Carsten Binnig, and Tim Kraska. 2017. What you see is not what you get!: Detecting Simpson's Paradoxes during Data Exploration. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*. 2:1–2:5.

[13] Miguel A Hernán, David Clayton, and Niels Keiding. 2011. The Simpson's paradox unraveled. *International journal of epidemiology* 40, 3 (2011), 780–785.

[14] Amir H Jadidinejad, Craig Macdonald, and Iadh Ounis. 2021. The Simpson's Paradox in the Offline Evaluation of Recommendation Systems. *ACM Transactions on Information Systems* 40, 1 (2021), 1–22.

[15] Rogier Andrew Kievit, Willem E. Frankenhuis, Lourens J. Waldorp, and Denny Borsboom. 2013. Simpson's paradox in psychological science: a practical guide. *Frontiers in Psychology* 4 (2013).

[16] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems (NeurIPS 2017)*. 972–981.

[17] Kristina Lerman. 2017. Computational social scientist beware: Simpson's paradox in behavioral data. *Journal of Computational Social Science* 1 (2017), 49–58.

[18] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[19] Christos Louizos, Uri Shalit, Joris M. Mooij, David A. Sontag, Richard S. Zemel, and Max Welling. 2017. Causal Effect Inference with Deep Latent-Variable Models. In *Advances in Neural Information Processing Systems (NeurIPS 2017)*. 6446–6456.

[20] Eric Neufeld. 1995. Simpson's Paradox in Artificial Intelligence and in Real Life. *Computational Intelligence* 11 (1995), 1–10.

[21] Judea Pearl. 2009. Simpson's Paradox, Confounding, and Collapsibility.

[22] Judea Pearl. 2011. Simpson's Paradox: An Anatomy.

[23] Judea Pearl. 2013. Understanding Simpson's Paradox. *ERN: Other Econometrics: Econometric & Statistical Methods (Topic)* (2013).

[24] Judea Pearl. 2020. *Race, COVID Mortality, and Simpson's Paradox (by Dana Mackenzie)*. http://causality.cs.ucla.edu/blog/index.php/2020/07/06/race-covid-mortality-and-simpsons-paradox-by-dana-mackenzie/

[25] Judea Pearl. 2023. *Causal Inference (CI) - A year in review*. http://causality.cs.ucla.edu/blog/index.php/2023/01/04/causal-inference-ci-a-year-in-review/

[26] Karl Pearson, Alice Lee, and Leslie Bramley-Moore. 1899. Genetic (Reproductive) Selection: Inheritance of Fertility in Man, and of Fecundity in Thoroughbred Racehorses. *Philosophical Transactions of the Royal Society A* 192 (1899), 257–330.

[27] J Ross Quinlan. 1993. Combining instance-based and model-based learning. In *Proceedings of the Tenth International Conference on International Conference on Machine Learning*. 236–243.

[28] Chanapong Rojanaworarit. 2020. Misleading Epidemiological and Statistical Evidence in the Presence of Simpson's Paradox: An Illustrative Study Using Simulated Scenarios of Observational Study Designs. *Journal of Medicine and Life* 13 (2020), 37 – 44.

[29] Simone Romano, James Bailey, Vinh Nguyen, and Karin Verspoor. 2014. Standardized mutual information for clustering comparisons: one step further in adjustment for chance. In *International Conference on Machine Learning*. PMLR, 1143–1151.

[30] Myra L. Samuels. 1993. Simpson's Paradox and Related Phenomena. *J. Amer. Statist. Assoc.* 88 (1993), 81–88.

[31] Rahul Sharma, Huseyn Garayev, Minakshi Kaushik, Sijo Arakkal Peious, Prayag Tiwari, and Dirk Draheim. 2022. Detecting Simpson's Paradox: A Machine Learning Perspective. In *Database and Expert Systems Applications: 33rd International Conference, DEXA 2022, Vienna, Austria, August 22–24, 2022, Proceedings, Part I*. Springer, 323–335.

[32] Rahul Sharma, Minakshi Kaushik, Sijo Arakkal Peious, Markus Bertl, Ankit Vidyarthi, Ashwani Kumar, and Dirk Draheim. 2022. Detecting Simpson's Paradox: A Step Towards Fairness in Machine Learning. In *New Trends in Database and Information Systems: ADBIS 2022 Short Papers, Doctoral Consortium and Workshops: DOING, K-GALS, MADEISD, MegaData, SWODCH, Turin, Italy, September 5–8, 2022, Proceedings*. Springer, 67–76.

[33] Edward Simpson. 1951. The Interpretation of Interaction in Contingency Tables. *Journal of the royal statistical society series b-methodological* 13 (1951), 238–241.

[34] Jan Sprenger and Naftali Weinberger. 2021. Simpson's Paradox. In *The Stanford Encyclopedia of Philosophy*.

[35] Julius von Kügelgen, Luigi Gresele, and Bernhard Schölkopf. 2021. Simpson's Paradox in COVID-19 Case Fatality Rates: A Mediation Analysis of Age-Related Causal Effects. *IEEE Transactions on Artificial Intelligence* 2, 1 (2021), 18–27.

[36] Chenguang Xu, Sarah M. Brown, and Christan Grant. 2018. Detecting Simpson's Paradox. In *Proceedings of the Thirty-First International Florida Artificial Intelligence Research Society Conference (FLAIRS 2018)*. 221–224.

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

A ADDITIONAL RESULTS

A.1 Atlanta CES

This section presents additional information and the detailed results of naive methods and SimNet in the Atlanta CES data. Table 7 provides the abbreviations of 22 departments. Tables 8-10 show the results of naive methods and SimNet when a single variable (ethnicity or department) is used as the conditional variable to disaggregate data. Here, AA, AI, and AN stand for African American, American Indian, and Alaska Native, respectively. Table 11 shows the results of SimNet when two variables (ethnicity and department) are used as the conditional variable to disaggregate data.

Table 7: Department and its abbreviation in Atlanta CES.

Abbr.	Department
AFR	AFR Atlanta Fire & Recuse
APD	APD Atlanta Police Department
AUD	AUD Audit Administration
AWD	PCD Atlanta Workforce Development Agency
CCN	CCN City Council
COR	COR Department of Corrections
CRB	CRB Administration
DHR	DHR Department of Human Resources
DIT	DIT Department of Information Technology
DOA	DOA Department of Aviation
DOF	DOF Department of Finance
DOP	DOP Department of Procurement
DPW	DPW Department of Public Works
DWM	DWM Department of Watershed Management
ETH	ETH Ethics Administration
EXE	EXE Executive Offices
JDA	JDA Municipal Court Operations
LAW	LAW Law Department
PCD	PCD Planning & Community Development
PDA	PDA Public Defender Administration
PRC	PRC Parks, Recreation, & Cultural Affairs
SOL	SOL Solicitor Office

Table 8: AAR discovered by naive methods in Atlanta CES. Ethnicity as Z. Gender as T. Annual salary as Y.

Subgroup	Ethnicity	Size	PCC	p	Sign
1	White	1642	-0.11	0	-1
2	Asian	103	-0.08	0.42	0
3	Two or More Races	12	-0.36	0.25	0
4	Black/AA	6304	-0.02	0.14	0
5	Hispanic/Latino	170	-0.01	0.88	0
6	AI/AN	13	-0.01	0.98	0
Overall	All	8244	0	0.95	0

Table 9: AAR discovered by SimNet in Atlanta CES. Ethnicity as Z. Gender as T. Annual salary as Y.

Subgroup	Ethnicity	Size	PCC	p	Sign
1	White Asian Two or More Races	1757	-0.11	0	-1
2	Black/AA Hispanic/Latino AI/AN	6487	-0.02	0.14	0
Overall	All	8244	0	0.95	0

Table 10: AAR discovered by naive methods in Atlanta CES. Department as Z. Gender as T. Annual salary as Y.

Subgroup	Department	Size	PCC	p	Sign
1	AFR	1063	0.02	0.54	0
2	APD	2349	0.18	0.0	1
3	AUD	13	-0.24	0.43	0
4	AWD	50	0.04	0.8	0
5	CCN	99	0.05	0.62	0
6	COR	320	0.07	0.22	0
7	CRB	6	0.20	0.7	0
8	DHR	133	0.15	0.09	0
9	DIT	124	0.19	0.04	1
10	DOA	544	0.12	0.01	1
11	DOF	145	0.26	0.0	1
12	DOP	40	0.43	0.01	1
13	DPW	799	-0.05	0.14	0
14	DWM	1325	0.04	0.11	0
15	ETH	4	-0.55	0.45	0
16	EXE	294	-0.03	0.55	0
17	JDA	147	0.32	0.0	1
18	LAW	82	0.11	0.33	0
19	PCD	153	0.11	0.17	0
20	PDA	34	0.14	0.43	0
21	PRC	446	-0.03	0.52	0
22	SOL	74	0	0.97	0
Overall	All	8244	0	0.95	0

Table 11: AR ($m = 2$) and AAR ($m = 3, 4, 5$) discovered by SimNet in Atlanta CES. Ethnicity and department as Z. Gender as T. Annual salary as Y.

Subgroup		Size	PCC	p	Sign
$m = 2$	1	4951	0.10	0	1
	2	3293	0.04	0.03	1
$m = 3$	1	1665	0.14	0	1
	2	3052	0.03	0.09	0
	3	3527	0.09	0	1
$m = 4$	1	1990	0.10	0	1
	2	3447	0.13	0	1
	3	1315	0.04	0.16	0
	4	1492	-0.03	0.32	0
$m = 5$	1	1086	-0.02	0.45	0
	2	2393	0.07	0	1
	3	205	0.17	0.02	1
	4	2000	0.10	0	1
	5	2560	0.19	0	1
Overall		8244	0	0.96	0

A.2 Iris

This section presents the detailed results of SimNet in the Iris data. Tables 12-14 show the results of SimNet when *Class* is used as the conditional variable. Tables 15-18 show the results of SimNet when continuous variables are used as the conditional variable.

Table 12: AR discovered by SimNet in Iris. Class as Z. Sepal length as T. Sepal width as Y.

Subgroup		Class	Size	PCC	p
$m = 2$	1	0	50	0.74	0.00
	2	1,2	100	0.55	0.00
$m = 3$	1	0	50	0.74	0.00
	2	1	50	0.52	0.00
	3	2	50	0.46	0.00
Overall		0,1,2	150	-0.12	0.15

Table 13: AR discovered by SimNet in Iris. Class as Z. Sepal width as T. Petal width as Y.

Subgroup		Class	Size	PCC	p
$m = 2$	1	0	50	0.23	0.10
	2	1,2	100	0.57	0.00
$m = 3$	1	0	50	0.23	0.10
	2	1	50	0.66	0.00
	3	2	50	0.54	0.00
Overall		0,1,2	150	-0.37	0.00

Table 14: AR discovered by SimNet in Iris. Class as Z. Sepal width as T. Petal length as Y.

Subgroup		Class	Size	PCC	p
$m = 2$	1	0	50	0.18	0.22
	2	1,2	100	0.52	0.00
$m = 3$	1	0	50	0.18	0.22
	2	1	50	0.56	0.00
	3	2	50	0.40	0.00
Overall		0,1,2	150	-0.43	0.00

Table 15: AR discovered by SimNet in Iris. Petal width as Z. Sepal length as T. Sepal width as Y.

Subgroup	Petal width	Size	PCC	p	Class	F1
1	[0.1, 0.6]	50	0.74	0.00	0	1.00
2	[1.0, 1.6]	52	0.51	0.00	1,2	0.94
3	[1.7, 2.5]	48	0.41	0.00	1,2	0.94
Overall	[0.1, 2.5]	150	-0.12	0.15	0,1,2	0.96

Table 16: AR discovered by SimNet in Iris. Petal length as Z. Sepal length as T. Sepal width as Y.

Subgroup	Petal length	Size	PCC	p	Class	F1
1	[1.0, 1.9]	50	0.74	0.00	0	1.00
2	[3.0, 4.8]	49	0.53	0.00	1,2	0.93
3	[4.9, 6.9]	51	0.44	0.00	1,2	0.93
Overall	[1.0, 6.9]	150	-0.12	0.15	0,1,2	0.95

Table 17: AMP discovered by SimNet in Iris. Petal width as Z. Sepal length as T. Petal length as Y.

Subgroup	Petal width	Size	PCC	p	Class	F1
1	[0.1, 0.5]	49	0.27	0.06	0	0.99
2	[0.6, 1.6]	53	0.73	0.00	0,1,2	0.93
3	[1.7, 2.5]	48	0.86	0.00	1,2	0.94
Overall	[0.1, 2.5]	150	0.87	0.00	0,1,2	0.95

Table 18: YAP discovered by SimNet in Iris. Petal length as Z. Sepal length as T. Petal width as Y.

Subgroup	Petal length	Size	PCC	p	Class	F1
1	[1.0, 3.5]	53	0.19	0.18	0,1	0.97
2	[3.5, 4.9]	46	0.26	0.08	1,2	0.90
3	[4.9, 6.9]	51	0.22	0.12	1,2	0.93
Overall	[1.0, 6.9]	150	0.82	0.00	0,1,2	0.93

A.3 Auto MPG

This section presents the detailed results of naive methods and SimNet in the Auto MPG data. Tables 19-21 show the results of naive methods when a discrete variable (cylinder or model year) is used as the conditional variable to disaggregate data. Here, the light grey rows in the three Tables indicate the six Simpson's reversals. Tables 22-25 show the results of SimNet when a continuous variable (displacement or horsepower) is used as the conditional variable to disaggregate data.

Table 19: AMP and AAR discovered by naive methods. Cylinder as Z. Acceleration as T. MPG as Y.

Subgroup	Cylinders	Size	PCC	p	Sign
1	8	103	0.32	0	1
2	4	199	0.08	0.24	0
3	6	83	-0.34	0	-1
4	3	4	-0.82	0.18	0
5	5	3	0.71	0.49	0
Overall	All	392	0.42	0	1

Table 20: No paradox discovered by naive methods. Model year as Z. Horsepower as T. MPG as Y.

Subgroup	Model year	Size	PCC	p
1	70	29	0.46	0.01
2	71	27	0.69	0
3	72	28	0.63	0
4	73	40	0.70	0
5	74	26	0.39	0.05
6	75	30	-0.05	0.79
7	76	34	0.30	0.08
8	77	28	0.42	0.03
9	78	36	0.42	0.01
10	79	29	-0.05	0.79
11	80	27	0.19	0.34
12	81	28	0.17	0.38
13	82	30	0.14	0.47
Overall	All	392	0.42	0

Table 21: AMP discovered by naive methods. Cylinder as Z. Horsepower as T. MPG as Y.

Subgroup	Cylinders	Size	PCC	p
1	8	103	-0.58	0
2	4	199	-0.59	0
3	6	83	0.01	0.91
4	3	4	0.62	0.38
5	5	3	-0.90	0.29
Overall	All	392	-0.78	0

Table 22: AMP discovered by SimNet in Auto MPG. Displacement as Z. Weight as T. MPG as Y.

Subgroup	Displacement	Size	PCC	p
1	[68, 108]	114	-0.36	0
2	[110, 156]	97	-0.16	0.11
3	[163, 267]	83	-0.45	0
4	[302, 455]	98	-0.50	0
Overall	[68, 455]	392	-0.83	0

Table 23: AMP discovered by SimNet in Auto MPG. Horsepower as Z. Weight as T. MPG as Y.

Subgroup	Horsepower	Size	PCC	p
1	[46, 78]	110	-0.44	0
2	[79, 108]	148	-0.58	0
3	[110, 230]	134	-0.79	0
Overall	[46, 230]	392	-0.83	0

Table 24: AMP discovered by SimNet in Auto MPG. Horsepower as Z. Displacement as T. MPG as Y.

Subgroup	Horsepower	Size	PCC	p
1	[46, 116]	286	-0.67	0
2	[120, 230]	106	-0.52	0
Overall	[46, 230]	392	-0.81	0

Table 25: YAP discovered by SimNet in Auto MPG. Horsepower as Z. Acceleration as T. MPG as Y.

Subgroup	Horsepower	Size	PCC	p
1	[46, 122]	291	0.06	0.29
2	[125, 230]	106	0	0.97
Overall	[46, 230]	392	0.42	0

A.4 Synthetic Data

Figure 8 shows the synthetic data with Simpson's paradox. The PCC of all samples is about 0.65, while that of each subgroup is about -0.86. All correlations are significant, so the paradox belongs to the AR type (see Table 26). Figure 9 shows the results of SimNet on synthetic datasets with different noise levels (NL). High NMI values indicate that the detected subgroups are consistent with the ground truth (see Figure 8). Since this is an unsupervised classification task, the order of the colors is not strictly one-to-one.

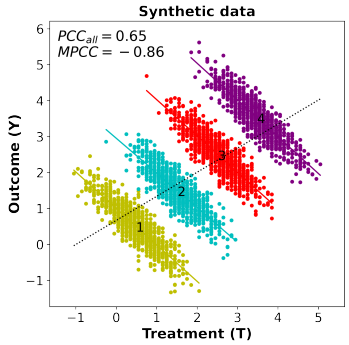

Figure 8: Synthetic data with Simpson's paradox.

Table 26: Simpson's paradox (AR) in synthetic data.

Subgroup	PCC	p	Size
1	-0.87	0	1000
2	-0.84	0	1000
3	-0.87	0	1000
4	-0.87	0	1000
Overall	0.65	0	4000

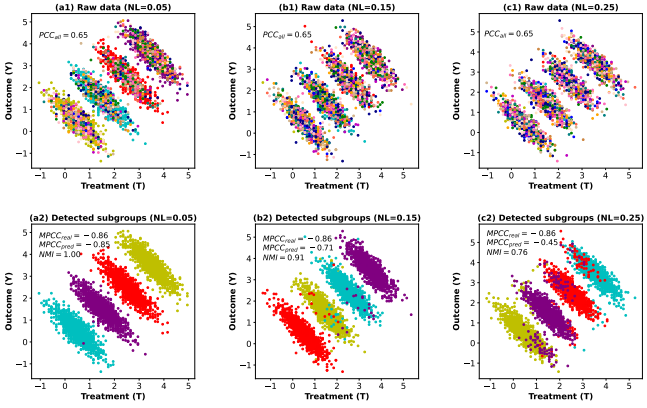

Figure 9: Results of SimNet in synthetic datasets. The first row is the raw data with different NL. The second row is the subgroups (in different colors) given by SimNet.

A.5 Titanic

This section presents the detailed results of SimNet in the Titanic data. Tables 27-29 show the results of SimNet when *Class* is used as the treatment variable.

Table 27: AR discovered in Titanic data. Gender as Z. Class as T. Survive as Y. The Class coding is as follows: Third, 0; Second, 1; First, 2; Crew, 3.

Subgroup	Gender	size	PCC	p
1	Male	1731	0.07	0.01
2	Female	470	0.47	0
Overall	All	2201	0	0.99

Table 28: No paradox discovered in Titanic data. Gender as Z. Class as T. Survive as Y. The Class coding is as follows: Crew, 0; Third, 1; Second, 2; First, 3.

Subgroup	Gender	size	PCC	p
1	Male	1731	0.04	0.13
2	Female	470	0.42	0
Overall	All	2201	0.27	0

Table 29: AMP discovered in Titanic data. Age as Z. Class as T. Survive as Y. There are only 1316 records since the Crew class is removed.

Subgroup	Age	Size	PCC	p
1	Adult	1207	0.322	0
2	Child	109	0.548	0
Overall	All	1316	0.317	0

B COMPUTATIONAL COMPLEXITY

The computational complexity of SimNet is determined by several factors, including the number of samples (n), the number of input features (f), the number of hidden neurons (h), the number of subgroups (m), and the number of hidden layers (l). Since a shallow SimNet model suffices to detect Simpson's paradox, we do not consider the number of hidden layers.

In this study, we trained the SimNet model for 50 epochs on each dataset to ensure model convergence. The SimNet model consists of three fully-connected layers, namely, an input layer, a hidden layer, and an output layer. Table 30 presents the hyper-parameters, parameter counts, floating point operations (FLOPs) per epoch, and average running time per epoch for the SimNet model on each dataset. These figures are calculated according to the experiments corresponding to the results (see brackets) listed in the first column of Table 30. It is worth noting that all experiments were conducted on a MacBook Pro with a 2.6 GHz 6-Core Intel Core i7 processor, and the reported time refers to the CPU time.

From the results in Table 30, it can be seen that our proposed SimNet model is a very small model with high operational efficiency. If users have GPU computing resources, the running speed of the SimNet model will be even faster.

Table 30: The number of parameters and running time of SimNet on five real-world datasets.

Dataset (results)	Hyper Parameters	#Params	#FLOPs	Time (epoch)
Iris (Table 15)	$n = 150, f = 1,$ $h = 64, m = 3.$	4.48k	0.66M	0.007s
Auto MPG (Table 22)	$n = 392, f = 1,$ $h = 64, m = 5.$	4.61k	1.76M	0.017s
Titanic (Table 27)	$n = 2201, f = 2,$ $h = 64, m = 2.$	4.48k	9.44M	0.059s
UC Berkeley (Table 4)	$n = 4526, f = 6,$ $h = 64, m = 2.$	4.74k	20.66M	0.133s
Atlanta CES (Table 11)	$n = 8244, f = 28,$ $h = 256, m = 5.$	74.5k	610.37M	0.404s