A SYNTHETIC GENERATION

In this section, we introduce a few additional findings on coverage redundant Simpson's paradoxes from a data generative perspective. We first discuss the generation of separate instances of Simpson's paradox in Section A.1. We then introduce the process of generating coverage redundant Simpson's paradoxes in Section A.2. Finally, we summarize the overall data synthesization procedure in Section A.3.

A.1 Generating Simpson's Paradoxes

Recall from Definition 2.2 that an association configuration (AC) $p = (s_1, s_2, X, Y)$ is a Simpson's paradox if:

$$\begin{aligned}&(1)\ P(Y|s_1) \leq P(Y|s_2);\ \text{and}\\ &(2)\ P(Y|s_1\langle X \shortrightarrow v\rangle) \geq P(Y|s_2\langle X \shortrightarrow v\rangle),\ \forall v \in \mathrm{Dom}(X).\end{aligned}$$

In particular, the frequency statistics $P(Y|s_j)$ for $j = 1, 2$ in condition (1), is obtained by weighted averaging their sub-population frequency statistics in condition (2), where the weights are determined by the relative coverage sizes of each sub-population. Specifically, we have that:

$$\begin{aligned}P(Y|s_j) &= \sum_{v \in \mathrm{Dom}(X)} \frac{|\,\mathrm{cov}(s_j\langle X \shortrightarrow v\rangle)|}{|\,\mathrm{cov}(s_j)|} \cdot P(Y|s_j\langle X \shortrightarrow v\rangle)\\ &= \boldsymbol{Q}(s_j|X) \cdot \boldsymbol{P}(s_j|Y,X)^\top,\end{aligned}$$

where

$$\boldsymbol{Q}(s_j|X) = \left[\frac{|\,\mathrm{cov}(s_j\langle X \shortrightarrow v_1\rangle)|}{|\,\mathrm{cov}(s_j)|}, \frac{|\,\mathrm{cov}(s_j\langle X \shortrightarrow v_2\rangle)|}{|\,\mathrm{cov}(s_j)|}, \ldots, \frac{|\,\mathrm{cov}(s_j\langle X \shortrightarrow v_{|\,\mathrm{Dom}(X)|}\rangle)|}{|\,\mathrm{cov}(s_j)|}\right]$$

is the sample distribution of s_j partitioned under X, and

$$\boldsymbol{P}(s_j|Y,X) = \left[P(Y|s_j\langle X \shortrightarrow v_1\rangle), P(Y|s_j\langle X \shortrightarrow v_2\rangle), \ldots, P(Y|s_j\langle X \shortrightarrow v_{|\,\mathrm{Dom}(X)|}\rangle)\right]$$

is the frequency statistics of s_j's sub-populations partitioned by X. With this, we can rephrase Definition 2.2 by substituting the terms, that (s_1, s_2, X, Y) is a Simpson's paradox if:

$$\boldsymbol{Q}(s_1|X) \cdot \boldsymbol{P}(s_1|Y,X)^\top < \boldsymbol{Q}(s_2|X) \cdot \boldsymbol{P}(s_2|Y,X)^\top;\ \text{and} \quad (1)$$

$$\boldsymbol{P}(s_1|Y,X)[j] > \boldsymbol{P}(s_2|Y,X)[j],\ 1 \leq j \leq |\,\mathrm{Dom}(X)|. \quad (2)$$

Therefore, the essence of generating an instance of Simpson's paradox is to *(a)* find the set of sub-population frequency statistics $\boldsymbol{P}(s_1|Y,X)$ and $\boldsymbol{P}(s_2|Y,X)$ that satisfy inequality (2); and *(b)* solve for the sample distributions $\boldsymbol{Q}(s_1|X)$ and $\boldsymbol{Q}(s_2|X)$ that satisfy inequality (1). We discuss each in the following paragraphs.

Sub-population Frequency Statistics. The first step of generating an instance of Simpson's paradox is to ensure that each sub-population in s_2, partitioned by X, has an frequency statistics value smaller than its sibling sub-population in s_1 (inequality (2)). A simple pattern that achieves this is:

$$\begin{aligned}\boldsymbol{P}(s_2|Y,X)[1] < \boldsymbol{P}(s_1|Y,X)[1] <\\ \boldsymbol{P}(s_2|Y,X)[2] < \boldsymbol{P}(s_1|Y,X)[2] <\\ \cdots\\ \boldsymbol{P}(s_2|Y,X)[|\,\mathrm{Dom}(X)|] <\\ \boldsymbol{P}(s_1|Y,X)[|\,\mathrm{Dom}(X)|].\end{aligned} \quad (3)$$

This pattern ensures that for any value v_j $(1 \leq j \leq |\,\mathrm{Dom}(X)|)$, we have $\boldsymbol{P}(s_1|Y,X)[j] > \boldsymbol{P}(s_2|Y,X)[j]$, satisfying inequality (2). With this pattern, we can now focus on finding sample distributions that satisfy inequality (1).

Sample Distributions. Given the sub-population frequency statistics pattern established above, we now need to solve for sample distributions $\boldsymbol{Q}(s_1|X)$ and $\boldsymbol{Q}(s_2|X)$ that satisfy inequality (1). To achieve this, we formulate the problem as a quadratic program:

$$\begin{aligned}\text{minimize}\quad & \sum_{j=1}^{2} \left\| \boldsymbol{Q}(s_j|X) - \frac{1}{|\,\mathrm{Dom}(X)|}\mathbf{1} \right\|_2^2\\ \text{subject to}\quad & \text{(i)}\ \sum_{k=1}^{|\,\mathrm{Dom}(X)|} \boldsymbol{Q}(s_j|X)[k] = 1, \quad j = 1, 2;\\ & \text{(ii)}\ \boldsymbol{Q}(s_j|X)[k] > 0, \quad j = 1, 2\\ & \quad \text{and } 1 \leq k \leq |\,\mathrm{Dom}(X)|;\\ & \text{(iii)}\ \boldsymbol{Q}(s_2|X) \cdot \boldsymbol{P}(s_2|Y,X)^\top\\ & \quad > \boldsymbol{Q}(s_1|X) \cdot \boldsymbol{P}(s_1|Y,X)^\top.\end{aligned} \quad (4)$$

We incorporate inequality (1) as a linear constraint specified in condition (iii). Moreover, the objective function of the QP aims to minimize the squared distance between each sample distribution and the uniform distribution $\frac{1}{|\,\mathrm{Dom}(X)|}\mathbf{1}$ to promote uniformity. Our synthetic generator also supports optimizing towards other distribution patterns such as normal or Zipfian distributions to better mimic the statistics of real-world data.

Synthesis of Simpson's Paradox. To summarize, generating an instance of Simpson's paradox requires establishing sub-population frequency statistics satisfying inequality (2) and solving for sample distributions that produce the reversal effect outlined in inequality (1). Algorithm 7 formalizes this generation procedure, taking as input an AC (s_1, s_2, X, Y) and crucially, a size parameter U that controls the number of records covered by the generated Simpson's paradox. The algorithm proceeds in three main steps: first, it generates the sub-population frequency statistics following the pattern in Equation (3); second, it solves the quadratic program in (4) to obtain optimal sample distributions; and third, it populates the output table with $2 \cdot U$ records for the Simpson's paradox (s_1, s_2, X, Y), distributing records across sub-populations according to the sample distributions and assigning label values according to probabilities given by the sub-population frequency statistics.

A.2 Realizing Coverage Redundancies

Having established a method for generating individual instances of Simpson's paradox, we now discuss creating (coverage) redundant Simpson's paradoxes. Building on the discussions in Section 3.1 and

Algorithm 7 Generate non-redundant Simpson's paradox.

Input: An AC (s_1, s_2, X, Y), paradox size U
Output: Data records T for the Simpson's paradox (s_1, s_2, X, Y)

Obtain the sub-population frequency statistics $\boldsymbol{P}(s_1|Y, X)$, $\boldsymbol{P}(s_2|Y, X)$ following Equation (3);
Obtain the sample distributions $\boldsymbol{Q}(s_1|X)$, $\boldsymbol{Q}(s_2|X)$ by solving the quadratic program (QP) in (4);
// Populate data records following the obtained sample distributions and sub-population aggregate statistics
for each $1 \le k \le |\,\mathrm{Dom}(X)|$ and each $j \in [1, 2]$ **do**
    // Find the number of records for each sub-population
    Let $U_{j,k} \leftarrow U \cdot \boldsymbol{Q}(s_j|X)[k]$;
    // Assign labels according to sub-population aggr. stats.
    Add $U_{j,k}$ copies of $s_j\langle X \to v_k\rangle$ as records to T and assign $U_{j,k} \cdot \boldsymbol{P}(s_j|Y, X)[k]$ of them with $(Y = 1)$;
return T.

Definition 3.8, (coverage) redundancies fundamentally arise when distinct populations have identical coverage within the data. To systematically realize (coverage) redundancies, we must understand the conditions under which populations would share the same coverage, as this property serves as the foundation for realizing both sibling child and separator equivalences. To this end, we remark on the following proposition.

Proposition A.1 (Impossibility of coverage identicality). *Let T be a base table with n categorical attributes $\{X_1, \ldots, X_n\}$. If the set of unique records in T corresponds exactly to all possible combinations of attribute values (*i.e., *the complete Cartesian product $\prod_{i=1}^{n} \mathrm{Dom}(X_i)$), then no two distinct populations of T share the same coverage.*

Proof. Assume, by contradiction, that there exist two distinct populations s and s' with $\mathrm{cov}(s) = \mathrm{cov}(s')$. Since $s \neq s'$, let X_{k_0} be an attribute for which $s[k_0] \neq s'[k_0]$. Without loss of generality, assume $s[k_0] = v$ for a fixed value $v \in \mathrm{Dom}(X_{k_0})$ and $s'[k_0] = *$.

Let us consider two records, r_1 and r_2, which are identical on all attributes except X_{k_0}. For r_1, let $r_1.X_{k_0} = v$, and for r_2, let $r_2.X_{k_0} = v'$ where $v' \neq v$ and $v' \in \mathrm{Dom}(X_{k_0})$. For all other attributes X_j where $j \neq k_0$, if $s[j] \neq *$, then $r_1.X_j = r_2.X_j = s[j]$. Observe that both records r_1 and r_2 must exist in T because T contains the complete Cartesian product of all attribute domains.

Now, r_1 is covered by both s and s'. However, record r_2 is covered by s' (because $s'[k_0] = *$) but not by s (because $s[k_0] = v$ but $r_2.X_{k_0} = v', v' \neq v$).

This means $r_2 \in \mathrm{cov}(s')$ but $r_2 \notin \mathrm{cov}(s)$, which contradicts our assumption that $\mathrm{cov}(s) = \mathrm{cov}(s')$. Therefore, no two distinct populations can share the same coverage when T contains the complete Cartesian product of all attribute domains. □

The contrapositive of Proposition A.1 implies that populations sharing identical coverage can only exist when the dataset contains a proper subset of the complete Cartesian product of attribute domains. Therefore, to facilitate the generation of coverage redundant Simpson's paradoxes, we impose a size threshold t that is significantly smaller than $\prod_{i=1}^{n} |\,\mathrm{Dom}(X_i)|$ to constrain the number of unique records in the generated dataset.

Having established the condition for realizing populations with identical coverage, we now proceed to develop methods for realizing each of the three types of coverage equivalences: sibling child equivalence, separator equivalence, and statistic equivalence.

Sibling Child Equivalence. Suppose we have a set of data records T that produces a Simpson's paradox $p_1 = (s_1, s_2, X, Y)$, where $s_1 = s\langle X_0 \to u_1\rangle$ and $s_2 = s\langle X_0 \to u_2\rangle$ are siblings from a common parent s. To realize sibling equivalence, our goal is update the records in T such that they also produce another Simpson's paradox $p_2 = (s_1', s_2', X, Y)$, where $s_1' = s'\langle X_0' \to v_1\rangle$ and $s_2' = s'\langle X_0' \to v_2\rangle$ are siblings from a common parent s', and that p_2 is sibling child equivalent to p_1. According to Definition 3.8, sibling child equivalence requires $\mathrm{cov}(s_1) = cov(s_1')$ and $\mathrm{cov}(s_2) = \mathrm{cov}(s_2')$. Based on the relationship between (s, s_1, s_2) and (s', s_1', s_2'), we have three scenarios:

(1) **Scenario 1**: $s \neq s'$, $X_0 = X_0'$, and $\{u_1, u_2\} = \{v_1, v_2\}$. In this case, sibling child equivalence is achieved by ensuring $\mathrm{cov}(s) = \mathrm{cov}(s')$. To this, for each categorical attribute X_k where $s[k] \neq *$ or $s'[k] \neq *$, we update every record r in $\mathrm{cov}(s)$ (within T), such that $r.X_k = s[k]$ or $r.X_k = s'[k]$. If $s[k] = s'[k] = *$, then no update is needed for $r.X_k$.
(2) **Scenario 2**: $s = s'$, $X_0 \neq X_0'$, and $\{u_1, u_2\} \neq \{v_1, v_2\}$. In this case, to achive sibling child equivalence, we establish a one-to-one mapping $f : \{u_1, u_2\} \mapsto \{v_1, v_2\}$ such that $f(u_1) = v_1$ and $f(u_2) = v_2$. For each record r in T, we set $r.X_0' = f(r.X_0)$ when $r.X_0 \in \{u_1, u_2\}$. This ensures $\mathrm{cov}(s\langle X_0 \to u_k\rangle) = \mathrm{cov}(s\langle X_0' \to f(u_k)\rangle)$ for $k = 1, 2$.
(3) **Scenario 3**: $s \neq s'$, $X_0 \neq X_0'$, and $\{u_1, u_2\} \neq \{v_1, v_2\}$. This combines the previous senarios. To achieve sibling child equivalence, we first ensure $\mathrm{cov}(s) = \mathrm{cov}(s')$ as in Scenario 1, then establish the one-to-one mapping as in Scenario 2.

Example A.2. Consider the data records in a slightly perturbed version of Table 2 where attribute values in D are randomized. Supposed the perturbed Table 2 is populated as a result of generating the Simpson's paradox $p_1 = ((*, b_1, *, *), (*, b_2, *, *), A, Y_1)$. To create a sibling equivalent Simpson's paradox $p_2 = ((*, *, *, d_1), (*, *, *, d_2), A, Y_1)$, we apply both Scenarios 1 and 2.

For Scenario 1, the parent population is $(*, *, *, *)$ for both p_1 and p_2, which are identical, so no adjustment to records in the perturbed table is needed.

For Scenario 2, we define a one-to-one mapping $f : \{b_1, b_2\} \to \{d_1, d_2\}$ where $f(b_1) = d_1$ and $f(b_2) = d_2$. We then update each record in the perturbed table with $B = b_1$ gets $D = d_1$ and each record with $B = b_2$ gets $D = d_2$. This establishes that $cov((*, b_1, *, *)) = cov((*, *, *, d_1))$ and $cov((*, b_2, *, *)) = cov((*, *, *, d_2))$.

In this way, we make p_1 and p_2 to be sibling equivalent, as verified in Example 3.2. □

Algorithm 8 formalizes this process of generating sibling-child-equivalent Simpson's paradoxes.

Separator Equivalence. Recall from Proposition 3.3 and Definition 3.8, Simpson's paradoxes $p_1 = (s_1, s_2, X_1, Y)$ and $p_2 = (s_1, s_2, X_1', Y)$ are separator equivalent if there exists a one-to-one mapping f between $\mathrm{Dom}(X_1)$ and $\mathrm{Dom}(X_1')$ such that for every

Algorithm 8 Realizing sibling child equivalence.

Input: Data records T producing the Simpson's paradox $p_1 = (s_1, s_2, X, Y)$ where $s_1 = s\langle X_0 \shortrightarrow u_1\rangle$ and $s_2 = s\langle X_0 \shortrightarrow u_2\rangle$, sibling populations (s_1, s_2) where $s_1' = s\langle X_0' \shortrightarrow v_1\rangle$ and $s_2' = s\langle X_0' \shortrightarrow v_2\rangle$

Output: Updated data records T producing a sibling-child-equivalent Simpson's paradox $p_2 = (s_1', s_2', X, Y)$

// Scenario 1: Ensure $\mathrm{cov}(s) = \mathrm{cov}(s')$
for each record $r \in T$ s.t. $r \in \mathrm{cov}(s)$ **do**
  for each attribute X_k s.t. $s[k] \neq *$ or $s'[k] \neq *$ **do**
    Set $r.X_k \leftarrow s[k]$ if $s[k] \neq *$, else $r.X_k \leftarrow s'[k]$;
// Scenario 2: Establish the one-to-one mapping
Establish the mapping f where $f(u_j) = v_j$ for $j = 1, 2$;
for each record $r \in T$ **do**
  Set $r.X_0' \leftarrow f(r.X_0)$ if $r.X_0 \in \{u_1, u_2\}$.

$v \in \mathrm{Dom}(X_1)$ and $j = 1, 2$,

$$\mathrm{cov}(s_j\langle X_1 \shortrightarrow v\rangle) = \mathrm{cov}(s_j\langle X_1' \shortrightarrow f(v)\rangle).$$

To achieve this, for every record r in T, we set $r.X_1' = f(r.X_1)$, which we formalize the process in Algorithm 9.

Example A.3. Consider a perturbed version of Table 2 where attribute values in C are initially randomized. Suppose the perturbed Table 2 is populated as a result of generating the Simpson's paradox $p_1 = ((*, b_1, *, *), (*, b_2, *, *), A, Y_1)$. To create a separator equivalent Simpson's paradox $p_2 = ((*, b_1, *, *), (*, b_2, *, *), C, Y_1)$, we establish a one-to-one mapping f between $\mathrm{Dom}(A) = a_1, a_2$ and $\mathrm{Dom}(C) = c_1, c_2$ where $f(a_1) = c_1$ and $f(a_2) = c_2$. We then update each record in the perturbed table so that whenever $A = a_1$, we set $C = c_1$, and whenever $A = a_2$, we set $C = c_2$. This ensures that $\mathrm{cov}((*, b_1, *, *)\langle A \shortrightarrow a_k\rangle) = \mathrm{cov}((*, b_1, *, *)\langle C \shortrightarrow c_k\rangle)$ and $\mathrm{cov}((, b_2, *,)\langle A \shortrightarrow a_k\rangle) = \mathrm{cov}((, b_2, *, *)\langle C \shortrightarrow c_k\rangle)$ for $k \in \{1, 2\}$. As verified in Example 3.4, p_1 and p_2 are separator equivalent. □

Algorithm 9 Realizing separator equivalence.

Input: Set of data records T producing the Simpson's paradox (s_1, s_2, X_1, Y), a separator attribute X_1'

Output: Updated set of data records T producing a separator equivalent Simpson's paradox $p_2 = (s_1, s_2, X_1', Y)$

Let $f : \mathrm{Dom}(X_1) \mapsto \mathrm{Dom}(X_1')$ be the one-to-one map;
for each record $r \in T$ **do**
  Set $r.X_1' \leftarrow f(r.X_1)$.

Statistic Equivalence. Recall from Proposition 3.5 and Definition 3.8, Simpson's paradoxes $p_1 = (s_1, s_2, X, Y_2)$ and $p_2 = (s_1, s_2, X, Y_2')$ are statistic equivalent if for each s_j $(j = 1, 2)$ $P(Y_2|s_j) = P(Y_2'|s_j)$, and for every value $v \in \mathrm{Dom}(X)$, $P(Y_2|s_j\langle X \shortrightarrow v\rangle) = P(Y_2'|s_j\langle X \shortrightarrow v\rangle)$. To achieve this, we simply ensure that each record has identical values for both label attributes Y_2 and Y_2'. We formalize this process in Algorithm 10.

Example A.4. Consider a perturbed version of Table 2 where attribute values in Y_2 are initially randomized. Suppose the perturbed Table 2 is populated as a result of generating the Simpson's paradox $p_1 = ((*, b_1, *, *), (*, b_2, *, *), A, Y_1)$. To create a statistic equivalent Simpson's paradox $p_2 = ((*, b_1, *, *), (*, b_2, *, *), A, Y_2)$, we update each record in the perturbed table so that $Y_2 = Y_1$ for all records. This ensures that $P(Y_1|s_j) = P(Y_2|s_j)$ and $P(Y_1|s_j\langle A \shortrightarrow a_k\rangle) = P(Y_2|s_j\langle A \shortrightarrow a_k\rangle)$ for $j \in \{1, 2\}$ and $k \in \{1, 2\}$. As verified in Example 3.6, p_1 and p_2 are statistic equivalent. □

Algorithm 10 Realizing statistic equivalence.

Input: Data records T producing the Simpson's paradox $p_1 = (s_1, s_2, X, Y_2)$, a label attribute Y_2'

Output: Updated data records T producing a statistic equivalent Simpson's paradox $p_2 = (s_1, s_2, X, Y_2')$

for each record $r \in T$ **do**
  Set $r.Y_2' \leftarrow r.Y_2$.

A.3 Data Generation Workflow

Algorithm 11 Generate redundant Simpson's paradoxes.

Input: Categorical attributes $\{X_i\}_{i=1}^n$, label attributes $\{Y_j\}_{j=1}^m$, size threshold $t \ll \prod_{i=1}^n |\mathrm{Dom}(X_i)|$

Output: Data table T (initially empty)

Let $R \leftarrow \emptyset$ to collect the set of unique data records;
Let $P \leftarrow \emptyset$ to collect the set of generated Simpson's paradoxes;
while $|R| < t$ **do**
  Let $p_1 = (s_1, s_2, X_1, Y_3)$ be an AC not in P;
  // Step 1: Generate distinct Simpson's paradox
  Populate T' for the Simpson's paradox p_1 using Alg. 7;
  // Step 2: Introduce coverage redundancies
  // Sibling child equivalence
  Apply Alg. 8 to T' to create a sibling-child-equivalent Simpson's paradox $p_2 = (s_1', s_2', X_1, Y_2)$;
  // Separator equivalence
  Apply Alg. 9 to T' to create a separator equivalent Simpson's paradox $p_3 = (s_1, s_2, X_1', Y_3)$;
  // Statistic equivalence
  Apply Alg. 10 to T' to create a statistic equivalent Simpson's paradox $p_4 = (s_1, s_2, X_1, Y_2')$;
  Add p_1, p_2, p_3, and p_4 to P;
  Add T' to T and unique records of T' to R;
return T.

Building upon the techniques established in Sections A.1 and A.2, we formulate a systematic approach for synthetic data generation that integrates both individual Simpson's paradox generation and coverage redundancy realization. The process employs a two-phase strategy: first generating distinct instances of Simpson's paradoxes (Section A.1), then systematically introducing coverage redundancies through sibling child, separator, and statistics equivalences (Section A.2). These phases are iterated until reaching a specified threshold $t \ll \prod_{i=1}^n |\mathrm{Dom}(X_i)|$ of unique records populated, which per Proposition A.1 ensures the dataset contains populations with identical coverage necessary for redundancy.

Algorithm 11 formalizes this process, taking categorical attributes $\{X_i\}_{i=1}^n$, label attributes $\{Y_j\}_{j=1}^m$, and the size threshold t as input, and producing a synthetic data table containing groups of (coverage) redundant Simpson's paradoxes.

B MISSING PROOFS

In this section, we present missing proofs of lemmas, propositions, and theorems presented in Section 3 and Section 4.

B.1 Proofs of Redundancy Properties

Lemma 3.1 (Sibling child equivalence) *Consider two association configurations $p = (s_1, s_2, X, Y)$ and $p' = (s_1', s_2', X, Y)$ where $\text{cov}(s_1) = \text{cov}(s_1')$ and $\text{cov}(s_2) = \text{cov}(s_2')$. If p is a Simpson's paradox, then p' is also a Simpson's paradox.*

Proof. Since p is a Simpson's paradox, $P(Y|s_1) > P(Y_|s_2)$. Due to $cov(s_j) = cov(s_j')$ $(j = 1, 2)$, $P(Y|s_j) = P(Y|s_j')$. Therefore, $P(Y|s_1') > P(Y|s_2')$. Furthermore, for every $v \in \text{Dom}(X)$, we have that $\text{cov}(s_j\langle X \shortrightarrow v\rangle) = \text{cov}(s_j'\langle X \shortrightarrow v\rangle)$ $(j = 1, 2)$, implying $P(Y|s_j\langle X \shortrightarrow v\rangle) = P(Y_i'|s_j'\langle X \shortrightarrow v\rangle)$ $(j = 1, 2)$. Hence, for every $v \in \text{Dom}(X)$, $P(Y|s_1'\langle X \shortrightarrow v\rangle) \leq P(Y|s_2'\langle X \shortrightarrow v\rangle)$. It follows, from Def. 2.2, that p' is a Simpson's paradox. □

Lemma 3.3 (Separator equivalence) *Consider two association configurations $p = (s_1, s_2, X, Y)$ and $p' = (s_1, s_2, X', Y)$, where $X \neq X'$ and there exists a one-to-one mapping $f : \text{Dom}(X) \mapsto \text{Dom}(X')$ such that for every $v \in \text{Dom}(X)$ and $s \in \{s_1, s_2\}$, $\text{cov}(s\langle X \shortrightarrow v\rangle) = \text{cov}(s\langle X' \shortrightarrow f(v)\rangle)$. If p is a Simpson's paradox, then p' is also a Simpson's paradox.*

Proof. Since p is a Simpson's paradox, $P(Y|s_1) > P(Y|s_2)$. The populations s_1 and s_2 remain the same in p', so this inequality holds for p' as well. In addition, for every value $v \in \text{Dom}(X)$, $P(Y|s_1\langle X \shortrightarrow v\rangle) \leq P(Y|s_2\langle X \shortrightarrow v\rangle)$. Due to the one-to-one mapping f, for every value $v \in \text{Dom}(X)$, $P(Y|s_j\langle X \shortrightarrow v\rangle) = P(Y|s_j\langle X' \shortrightarrow f(v)\rangle)$ $(j = 1, 2)$. Thus, $P(Y|s_1\langle X' \shortrightarrow f(v)\rangle) \leq P(Y|s_2\langle X' \shortrightarrow f(v)\rangle)$, $\forall v \in \text{Dom}(X)$. It follows from Def. 2.2 that p' is a Simpson's paradox. □

Lemma 3.5 (Statistic equivalence) *Consider two association configurations $p = (s_1, s_2, X, Y)$ and $p' = (s_1, s_2, X, Y')$ such that $Y \neq Y'$. If p is a Simpson's paradox and if any of the following (sufficient, and progressively less restrictive) conditions hold, then p' is also a Simpson's paradox:*

1. *For every $t \in \text{cov}(s_1) \cup \text{cov}(s_2)$, $t.Y = t.Y'$;*
2. *For every $s \in \{s_1, s_2\}$, $P(Y|s) = P(Y'|s)$ and for every $v \in \text{Dom}(X)$, $P(Y|s\langle X \shortrightarrow v\rangle) = P(Y'|s\langle X \shortrightarrow v\rangle)$;*
3. *$\text{sign}(P(Y|s_1) - P(Y|s_2)) = \text{sign}(P(Y'|s_1) - P(Y'|s_2))$, and for every $v \in \text{Dom}(X)$, $\text{sign}(P(Y|s_1\langle X \shortrightarrow v\rangle) - P(Y|s_2\langle X \shortrightarrow v\rangle)) = \text{sign}(P(Y'|s_1\langle X \shortrightarrow v\rangle) - P(Y'|s_2\langle X \shortrightarrow v\rangle))$.*

Proof. Since p is a Simpson's paradox, $P(Y|s_1) > P(Y|s_2)$, and for every value $v \in \text{Dom}(X)$, $P(Y|s_1\langle X \shortrightarrow v\rangle) \leq P(Y|s_2\langle X \shortrightarrow v\rangle)$. We want to show that that p' is also a Simpson's paradox under each case.

Cases (1) and (2): In both cases, we have $P(Y|s_j) = P(Y'|s_j)$ for $j = 1, 2$. This gives that $P(Y'|s_1) > P(Y'|s_2)$. Furthermore, for every $v \in \text{Dom}(X)$, we have $P(Y|s_j\langle X \shortrightarrow v\rangle) = P(Y'|s_j\langle X \shortrightarrow v\rangle)$ for $j = 1, 2$. This gives that $P(Y'|s_1\langle X \shortrightarrow v\rangle) \leq P(Y'|s_2\langle X \shortrightarrow v\rangle)$. It follows, from Def. 2.2, that p' is a Simpson's paradox.

Case (3): Since $P(Y|s_1) > P(Y|s_2)$, we have $P(Y|s_1) - P(Y|s_2) > 0$, thus $\text{sign}(P(Y|s_1) - P(Y|s_2)) = +1$. By the given condition, $\text{sign}(P(Y'|s_1) - P(Y'|s_2)) = +1$, which implies $P(Y'|s_1) > P(Y'|s_2)$. In addition, for every $v \in \text{Dom}(X)$, since $P(Y|s_1\langle X \shortrightarrow v\rangle) \leq P(Y|s_2\langle X \shortrightarrow v\rangle)$, we have $\text{sign}(P(Y|s_1\langle X \shortrightarrow v\rangle) - P(Y|s_2\langle X \shortrightarrow v\rangle)) = -1$. By the given condition, $\text{sign}(P(Y'|s_1\langle X \shortrightarrow v\rangle) - P(Y'|s_2\langle X \shortrightarrow v\rangle)) = -1$, which implies $P(Y'|s_1\langle X \shortrightarrow v\rangle) \leq P(Y'|s_2\langle X \shortrightarrow v\rangle)$. It follows, from Def. 2.2, that p' is a Simpson's paradox.

In all three cases, p' is a Simpson's paradox. □

Theorem 3.9 (Equivalence) *Redundancy of Simpson's paradoxes is an equivalence relation.*

Proof. (Reflexivity) Given any Simpson's paradox $p = (s_1, s_2, X, Y)$. It is trivial that

1. $\text{cov}(s_j) = \text{cov}(s_j)$ $(j = 1, 2)$;
2. $P(Y|s_j) = P(Y|s_j)$ $(j = 1, 2)$; and
3. for every value $v \in \text{Dom}(X)$,
   (a) $\text{cov}(s_j\langle X \shortrightarrow v\rangle) = \text{cov}(s_j\langle X \shortrightarrow v\rangle)$ $(j = 1, 2)$; and
   (b) $P(Y|s_j\langle X \shortrightarrow v\rangle) = P(Y|s_j\langle X \shortrightarrow v\rangle)$ $(j = 1, 2)$.

Hence, coverage redundancy is reflexive.

(Symmetricity) Suppose Simpson's paradoxes p and p' are coverage redundant. It is also straightforward that, for $(j = 1, 2)$,

1. $\text{cov}(s_j) = \text{cov}(s_j') \Leftrightarrow \text{cov}(s_j') = \text{cov}(s_j)$;
2. $P(Y|s_j) = P(Y'|s_j') \Leftrightarrow P(Y'|s_j') = P(Y|s_j)$;
3. suppose a one-to-one mapping f between $\text{Dom}(X)$ and $\text{Dom}(X')$ such that for every value $v \in \text{Dom}(X)$,
   (a) $\text{cov}(s_j\langle X \shortrightarrow v\rangle) = \text{cov}(s_j'\langle X' \shortrightarrow f(v)\rangle) \Leftrightarrow$ $\text{cov}(s_j'\langle X' \shortrightarrow f(v)\rangle) = \text{cov}(s_j\langle X \shortrightarrow v\rangle)$;
   (b) $P(Y|s_j\langle X \shortrightarrow v\rangle) = P(Y'|s_j'\langle X' \shortrightarrow f(v)\rangle) \Leftrightarrow$ $P(Y'|s_j'\langle X' \shortrightarrow f(v)\rangle) = P(Y|s_j\langle X \shortrightarrow v\rangle)$.

Hence, coverage redundancy is symmetric.

(Transitivity) Suppose p, p', p'' are Simpson's paradoxes such that p and p' are coverage redundant, p' and p'' are coverage redundant. It is, again, straightforward that, for $(j = 1, 2)$,:

1. if $\text{cov}(s_j) = \text{cov}(s_j')$ and $\text{cov}(s_j') = \text{cov}(s_j'')$, then $\text{cov}(s_j) = \text{cov}(s_j'')$;
2. if $P(Y|s_j) = P(Y'|s_j')$ and $P(Y'|s_j') = P(Y''|s_j'')$, then $P(Y|s_j) = P(Y''|s_j'')$;
3. suppose one-to-one mappings, f between $\text{Dom}(X)$ and $\text{Dom}(X')$, g between $\text{Dom}(X')$ and $\text{Dom}(X'')$, such that for every value $v \in \text{Dom}(X)$,
   (a) if $\text{cov}(s_j\langle X \shortrightarrow v\rangle) = \text{cov}(s_j'\langle X' \shortrightarrow f(v)\rangle)$ and $\text{cov}(s_j'\langle X' \shortrightarrow f(v)\rangle) = \text{cov}(s_j''\langle X'' \shortrightarrow g(f(v))\rangle)$, then $\text{cov}(s_j\langle X \shortrightarrow v\rangle) = \text{cov}(s_j''\langle X'' \shortrightarrow g(f(v))\rangle)$ note that $g \circ f$ is also a one-to-one mapping;
   (b) if $P(Y|s_j\langle X \shortrightarrow v\rangle) = P(Y'|s_j'\langle X' \shortrightarrow f(v)\rangle)$ and $P(Y'|s_j'\langle X' \shortrightarrow f(v)\rangle) = P(Y''|s_j''\langle X'' \shortrightarrow g(f(v))\rangle)$, then $P(Y|s_j\langle X \shortrightarrow v\rangle) = P(Y''|s_j''\langle X'' \shortrightarrow g(f(v))\rangle)$.

Hence, coverage redundancy is transitive. □

Lemma 3.11 (Product space) *Each redundant paradox group can be characterized by the product of: $\mathcal{E}_1 \times \mathcal{E}_2 \times \mathbf{X} \times \mathbf{Y}$, where $\mathbf{X}$ is a set of separator attributes, $\mathbf{Y}$ is a set of label attributes, and $\mathcal{E}_1, \mathcal{E}_2$ are sets of sibling populations, each containing populations with identical coverage. Any choice of $(s_1, s_2, X, Y) \in \mathcal{E}_1 \times \mathcal{E}_2 \times \mathbf{X} \times \mathbf{Y}$ where s_1, s_2 are siblings is a Simpson's paradox in the redundant paradox group.*

Proof. Let $p = (s_1, s_2, X, Y)$ be a Simpson's paradox in a redundant paradox group $\mathcal{G}$. The following defines the construction of

the product space:

$$\mathcal{E}_1 = \{s' \in \mathcal{P} \mid \text{cov}(s_1) = \text{cov}(s')\},$$
$$\mathcal{E}_2 = \{s' \in \mathcal{P} \mid \text{cov}(s_2) = \text{cov}(s')\},$$
$$\mathbf{X} = \{X' \mid (s_1, s_2, X', Y) \in \mathcal{G}\},$$
$$\mathbf{Y} = \{Y' \mid (s_1, s_2, X, Y') \in \mathcal{G}\}$$

where $\mathcal{P}$ denotes the set of all populations.

We first show that *every paradox in $\mathcal{G}$ belongs to $\mathcal{E}_1 \times \mathcal{E}_2 \times \mathbf{X} \times \mathbf{Y}$*. Let $p' = (s_1', s_2', X', Y')$ be any Simpson's paradox in $\mathcal{G}$. p and p' are redundant and are both in $\mathcal{G}$. By Def. 3.8, redundancy arises from sibling child equivalence (Lemma 3.1), separator equivalence (Lemma 3.3), or statistics equivalence (Lemma 3.5).

- By sibling child equivalence, if $\text{cov}(s_1) = \text{cov}(s_1')$ and $\text{cov}(s_2) = \text{cov}(s_2')$, then (s_1', s_2', X, Y) is a Simpson's paradox redundant with p. Therefore $s_1' \in \mathcal{E}_1$ and $s_2' \in \mathcal{E}_2$.
- By separator equivalence, if there exists a one-to-one mapping f between $\text{Dom}(X)$ and $\text{Dom}(X')$ such that $\text{cov}(s_j\langle X \to v\rangle) = \text{cov}(s_j\langle X' \to f(v)\rangle)$ for every $v \in \text{Dom}(X)$, then (s_1, s_2, X', Y) is redundant with p. Therefore, $X' \in \mathbf{X}$.
- By statistic equivalence, if the frequency statistics under label Y' satisfy any sufficient condition in Lemma 3.5, then (s_1, s_2, X, Y') is redundant with p. Therefore $Y' \in \mathbf{Y}$.

By the transitivity of the equivalence relation (Theorem 3.9), any combination of these equivalences preserves the redundancy. Therefore $(s_1', s_2', X', Y') \in \mathcal{E}_1 \times \mathcal{E}_2 \times \mathbf{X} \times \mathbf{Y}$.

We then show that *every valid element of $\mathcal{E}_1 \times \mathcal{E}_2 \times \mathbf{X} \times \mathbf{Y}$ is a Simpson's paradox in $\mathcal{G}$*. Let $(s_1', s_2', X', Y') \in \mathcal{E}_1 \times \mathcal{E}_2 \times \mathbf{X} \times \mathbf{Y}$ where s_1', s_2' are siblings. We show that $p' = (s_1', s_2', X', Y')$ is a Simpson's paradox redundant with p. Since $s_1' \in \mathcal{E}_1$ and $s_2' \in \mathcal{E}_2$, we have $\text{cov}(s_1') = \text{cov}(s_1)$ and $\text{cov}(s_2') = \text{cov}(s_2)$. By Lemma 3.1, if (s_1, s_2, X, Y) is a Simpson's paradox, then (s_1', s_2', X, Y) is also a Simpson's paradox. Since $X' \in \mathbf{X}$, there exists some paradox in $\mathcal{G}$ with separator X'. By Lemma 3.3 and the construction of $\mathbf{X}$, the AC (s_1', s_2', X', Y) is a Simpson's paradox. Similarly, since $Y' \in \mathbf{Y}$, by Lemma 3.5 and the construction of $\mathbf{Y}$, the AC (s_1', s_2', X', Y') is a Simpson's paradox. Hence, by Theorem 3.9, p' is redundant with p and belongs to $\mathcal{G}$.

Therefore, We have shown that $\mathcal{G} = \{(s_1, s_2, X, Y) \in \mathcal{E}_1 \times \mathcal{E}_2 \times \mathbf{X} \times \mathbf{Y} \mid s_1$ and s_2 are siblings and (s_1, s_2, X, Y) is a Simpson's paradox$\}$. Moreover, any valid choice from $\mathcal{E}_1 \times \mathcal{E}_2 \times \mathbf{X} \times \mathbf{Y}$ (satisfying the sibling constraint) yields a Simpson's paradox in $\mathcal{G}$. □

Property 1 (Convexity of coverage groups) *Let $\mathcal{P}$ be the set of all populations. For each coverage group $\mathcal{E} \in \mathcal{P}/\equiv_{\text{cov}}$, $\mathcal{E}$ is a convex subset of coverage-identical populations. Furthermore, $|\text{up}(\mathcal{E})| = 1$ and the least descendant is the unique upper bound.*

Proof. The proof consists of two parts:

(a) $\mathcal{E}$ is a convex subset;
(b) $\mathcal{E}$'s upper bound is unique.

For part (a), we want to prove that (1) for any pair of populations s and s' in $\mathcal{E}$ such that $s \succ s'$, every intermediate populations s'' where $s \succ s'' \succ s'$ is also in $\mathcal{E}$, and (2) populations in $\mathcal{E}$ are connected.

First, regarding claim (1), let $s, s' \in \mathcal{E}$ where $s \succ s'$, and let s'' be any population such that $s \succ s'' \succ s'$. By definition of ancestor-descendant relation, $\text{cov}(s) \supseteq \text{cov}(s'') \supseteq \text{cov}(s')$. Since $\text{cov}(s) = \text{cov}(s')$, it follows $\text{cov}(s'') = \text{cov}(s) = \text{cov}(s')$. Therefore, $s' \in \mathcal{E}$.

Second, regarding claim (2), let $s_1, s_2 \in \mathcal{E}$ where $s_1 \neq s_2$, there are two possibilities:

(1) $s_1 \succ s_2$ (or $s_2 \succ s_1$ in symmetry). From claim (1), since every intermediate population s'' such that $s_1 \succ s'' \succ s_2$ is in $\mathcal{E}$, s_1 and s_2 are connected ($s_1 \sim s_2$).
(2) $s_1 \not\succ s_2$ (or $s_2 \not\succ s_1$ in symmetry). Then there exists a population $s'' \in \mathcal{E}$ such that s'' is a common descendant (or ancestor) of s_1 and s_2, that is, $s_1 \succ s''$ and $s_2 \succ s''$ (or $s'' \succ s_1$ and $s'' \succ s_2$). From claim (1), we have that $s_1 \sim s''$ and $s'' \sim s_2$. Therefore, $s_1 \sim s_2$.

For part (b), let s_d be the descendant of all populations in $\mathcal{E}$. Specifically, for each attribute X_i ($1 \le i \le n$), we have that:

$$s_d[i] = \begin{cases} v & \text{if there exists } s \in \mathcal{E} \text{ s.t. } s[i] = v \neq *,\ v \in \text{Dom}(X_i) \\ * & \text{otherwise.} \end{cases}$$

In other word, s_d is an upper bound of $\mathcal{E}$. Suppose there exists another upper bound s_d' of $\mathcal{E}$ where $s_d' \neq s_d$. Then there must be an attribute X_i where $s_d'[i] \neq s_d[i]$. This means either:

(1) $s_d'[i] = *$ but $s_d[i] = v$ where $v \in \text{Dom}(X_i)$; or
(2) $s_d'[i] = v'$ but $s_d[i] = v$ where $v' \neq v$ and $v, v' \in \text{Dom}(X_i)$.

In case (1), $s_d' \succ s_d$. Hence, s_d' is not an upper bound of $\mathcal{E}$. In case (2), $\text{cov}(s_d') \neq \text{cov}(s_d)$. Hence, $s_d' \notin \mathcal{E}$. Therefore, s_d is unique. □

Property 2 (Reconstruction from bounds) *Let $\mathcal{E} \subseteq \mathcal{P}$ be a convex subset of populations. Then $s \in \mathcal{E}$ if and only if there exist $s_l \in \text{low}(\mathcal{E})$ and $\{s_u\} = \text{up}(\mathcal{E})$ such that $s_l \preceq s \preceq s_u$.*

Proof. ($\Rightarrow$) Given $s \in \mathcal{E}$, then either $s \in \text{low}(\mathcal{E})$, $s \in \text{up}(\mathcal{E})$, or $s \notin \text{low}(\mathcal{E})$ and $s \notin \text{up}(\mathcal{E})$.

If $s \in \text{low}(\mathcal{E})$, we can set $s_l = s$. Since $\mathcal{E}$ is convex and connected, there must exist an upper bound $s_u \in \text{up}(\mathcal{E})$ such that $s \preceq s_u$.

If $s \in \text{up}(\mathcal{E})$, we can set $s_u = s$. Similarly, there must exist a lower bound $s_l \in \text{low}(\mathcal{E})$ such that $s_l \preceq s$.

If s is neither a lower nor upper bound, then by the convexity of $\mathcal{E}$, there must exist $s_l \in \text{low}(\mathcal{E})$ such that $s_l \prec s$ and $s_u \in \text{up}(\mathcal{E})$ such that $s \prec s_u$. Therefore, we have $s_l \prec s \prec s_u$.

($\Leftarrow$) Suppose there exists $s \in \mathcal{P}$, $s_l \in \text{low}(\mathcal{E})$, and $s_u \in \text{up}(\mathcal{E})$, such that $s_l \preceq s \preceq s_u$. By convexity of $\mathcal{E}$, it follows that $s \in \mathcal{E}$. □

B.2 Proofs of Algorithmic Properties

Theorem 4.1 (#P-Hardness). *Finding all redundant paradox groups in a multidimensional table is #P-hard.*

Proof. We prove #P-hardness via a parsimonious reduction from #SAT. Given a Boolean formula ϕ in CNF with variables $x_1, \ldots, x_n$ and clauses $C_1, \ldots, C_m$, we construct in polynomial time a table $T(\phi)$ such that there exists a bijection between satisfying assignments of ϕ and redundant paradox groups in $T(\phi)$.

Construction. The table $T(\phi)$ contains the following elements:

(1) Categorical Attributes. The table contains $3n + m + 2$ categorical attributes:

- For each variable x_i ($1 \le i \le n$): three attributes A_i, B_i, C_i, each with domain $\{\text{true}, \text{false}\}$. The three copies enable sibling child equivalence.
- For each clause C_j ($1 \le j \le m$): one attribute D_j with domain $\{0, 1\}$, where 1 indicates the clause is satisfied and 0 indicates unsatisfied.
- Two auxiliary attributes U_1 and U_2, each with domain $\{0, 1\}$, which serve as separators and differential attributes.

(2) Label Attributes. We define two binary label attributes Y_1 and Y_2 to create statistic equivalence.

(3) Records. The table contains $2n + 2m + 4$ records.

(3.1) Variable Records: For each variable x_i ($1 \le i \le n$), we create two records:

- r_i^{true}: Set $A_i = B_i = C_i = \text{true}$; for all $\ell \ne i$, set $A_\ell = B_\ell = C_\ell = \text{false}$; set all $D_j = 0$; set $U_1 = 0, U_2 = 0$; set $Y_1 = Y_2 = 0$.
- r_i^{false}: Set $A_i = B_i = C_i = \text{false}$; for all $\ell \ne i$, set $A_\ell = B_\ell = C_\ell = \text{false}$; set all $D_j = 0$; set $U_1 = 0, U_2 = 1$; set $Y_1 = Y_2 = 0$.

Each variable record encodes one possible truth value for its variable. The three attribute copies (A_i, B_i, C_i) taking identical values ensure that multiple distinct populations can have identical coverage.

(3.2) Clause Records: For each clause C_j ($1 \le j \le m$), we create two records:

- r_j^{sat}: For each variable x_i, set $A_i = \text{true}$ if literal x_i appears in C_j, set $A_i = \text{false}$ if literal $\neg x_i$ appears in C_j or x_i does not appear in C_j; set all $B_i = C_i = \text{false}$; set $D_j = 1$ and all $D_\ell = 0$ for $\ell \ne j$; set $U_1 = 1, U_2 = 0$; set $Y_1 = Y_2 = 1$.
- r_j^{unsat}: For each variable x_i, set $A_i = \text{true}$ if literal x_i appears in C_j, set $A_i = \text{false}$ if literal $\neg x_i$ appears in C_j or x_i does not appear in C_j; set all $B_i = C_i = \text{false}$; set $D_j = 0$ and all $D_\ell = 0$ for $\ell \ne j$; set $U_1 = 1, U_2 = 1$; set $Y_1 = Y_2 = 0$.

The clause records encode literal requirements. A population will cover r_j^{sat} if and only if the assignment it encodes satisfies clause C_j.

Padding Records: We add four records to balance frequency statistics:

- $r^{(1)}$: Set all $A_i = B_i = C_i = \text{false}$; set all $D_j = 0$; set $U_1 = 0, U_2 = 0$; set $Y_1 = 0, Y_2 = 0$.
- $r^{(2)}$: Set all $A_i = B_i = C_i = \text{false}$; set all $D_j = 0$; set $U_1 = 0, U_2 = 1$; set $Y_1 = 0, Y_2 = 0$.
- $r^{(3)}$: Set all $A_i = B_i = C_i = \text{false}$; set all $D_j = 0$; set $U_1 = 1, U_2 = 0$; set $Y_1 = 0, Y_2 = 0$.
- $r^{(4)}$: Set all $A_i = B_i = C_i = \text{false}$; set all $D_j = 0$; set $U_1 = 1, U_2 = 1$; set $Y_1 = 0, Y_2 = 0$.

The construction runs in polynomial time: we create $O(n + m)$ attributes and $O(n + m)$ records, with each record constructible in $O(n + m)$ time.

Establishing the Bijection. We now establish the bijection between satisfying assignments and redundant paradox groups.

Claim 1. *For each satisfying assignment $\sigma : \{x_1, \ldots, x_n\} \to \{\text{true}, \text{false}\}$ of ϕ, there exists a unique redundant paradox group $\mathcal{G}_\sigma$ in $T(\phi)$.*

Proof. Given a satisfying assignment σ, we construct two sibling populations s_1^σ and s_2^σ that form the basis of a Simpson's paradox. Define s_1^σ as follows: for each variable attribute A_i, set $s_1^\sigma[A_i] = \text{true}$ if $\sigma(x_i) = \text{true}$ and $s_1^\sigma[A_i] = \text{false}$ if $\sigma(x_i) = \text{false}$; set $s_1^\sigma[B_i] = s_1^\sigma[C_i] = *$ for all i; set $s_1^\sigma[D_j] = *$ for all j; set $s_1^\sigma[U_1] = *$ and $s_1^\sigma[U_2] = 0$. Define s_2^σ identically except $s_2^\sigma[U_2] = 1$.

By construction, s_1^σ and s_2^σ are siblings under differential attribute U_2. Since σ satisfies ϕ, for each clause C_j, the population s_1^σ covers the record r_j^{sat} because the variable attributes of s_1^σ match at least one literal in C_j. The coverage sets are:

$$\text{cov}(s_1^\sigma) = \{r_i^{\sigma(x_i)} : i \in [n]\} \cup \{r_j^{\text{sat}} : j \in [m]\} \cup \{r^{(1)}, r^{(3)}\}$$
$$\text{cov}(s_2^\sigma) = \{r_i^{\sigma(x_i)} : i \in [n]\} \cup \{r_j^{\text{unsat}} : j \in [m]\} \cup \{r^{(2)}, r^{(4)}\}$$

Computing frequency statistics, we have $P(Y_1 = 1|s_1^\sigma) = \frac{m}{n+m+2} > 0 = P(Y_1 = 1|s_2^\sigma)$ since only clause-sat records contribute $Y_1 = 1$ values. When conditioning on separator U_1: for $U_1 = 0$, both populations cover only variable and padding records (all with $Y_1 = 0$), giving equal statistics; for $U_1 = 1$, s_1^σ covers clause-sat records while s_2^σ covers clause-unsat records, and the padding records are constructed to ensure $P(Y_1 = 1|s_1^\sigma\langle U_1 \to 1\rangle) \ge P(Y_1 = 1|s_2^\sigma\langle U_1 \to 1\rangle)$. This establishes that $(s_1^\sigma, s_2^\sigma, U_1, Y_1)$ is a Simpson's paradox according to Definition 2.2.

This paradox belongs to a unique redundant paradox group $\mathcal{G}_\sigma$ exhibiting all three types of redundancy. First, sibling child equivalence arises because we can construct populations using attributes B_i or C_i instead of A_i to encode σ, yielding identical coverage. Second, separator equivalence can be created by introducing additional separator attributes that partition records identically to U_1. Third, statistic equivalence exists because Y_1 and Y_2 take identical values on variable, clause-sat, clause-unsat, and padding records, ensuring equivalent frequency statistics. The group $\mathcal{G}_\sigma$ is unique to σ because populations encoding different variable assignments have different coverage sets (they cover different variable records), and thus cannot be redundant by Definition 3.8. □

Claim 2. *Each redundant paradox group in $T(\phi)$ corresponds to a unique satisfying assignment of ϕ.*

Consider any Simpson's paradox (s_1, s_2, Z, Y) in $T(\phi)$. To achieve the association reversal required by Definition 2.2, population s_1 must cover records with high Y values. In our construction, records with $Y_1 = 1$ are clause-sat records. For s_1 to cover clause-sat records (i.e., r_j^{sat}), the assignment that s_1 represents must satisfy clause C_j.

We extract an assignment σ from s_1: for each variable x_i, if $s_1[A_i] = \text{true}$ (or $s_1[B_i] = \text{true}$ or $s_1[C_i] = \text{true}$), set $\sigma(x_i) = \text{true}$; if $s_1[A_i] = \text{false}$ (or equivalently for B_i, C_i), set $\sigma(x_i) = \text{false}$. For s_1 to cover records with high proportion of $Y = 1$, it must cover r_j^{sat} for all clauses $j \in [m]$. By our construction, this occurs if and only if σ satisfies all clauses in ϕ, making σ a satisfying assignment.

Different satisfying assignments yield distinct redundant paradox groups because they cover different variable records. If $\sigma \ne \sigma'$, then for some variable x_k we have $\sigma(x_k) \ne \sigma'(x_k)$, implying $r_k^{\sigma(x_k)} \ne r_k^{\sigma'(x_k)}$. Populations with different coverage cannot be redundant by Definition 3.8, and thus belong to different redundant paradox groups. This establishes uniqueness. □

Conclusion. The two claims establish a bijection between satisfying assignments of ϕ and redundant paradox groups in $T(\phi)$. Since the construction is polynomial-time and preserves counts exactly, we have a parsimonious reduction from #SAT. As #SAT is #P-complete, counting redundant paradox groups is #P-hard. □

Theorem 4.6 (Completeness). *Algorithm 2 materializes all non-empty populations that satisfy the coverage threshold. Furthermore, after group merging, Algorithm 2 yields maximal convex coverage groups of coverage-identical populations; that is, no population outside a group shares the same coverage as any population within it.*

Proof. We prove by contradiction. Assume there exists a non-empty population s^* that satisfies the coverage threshold but is not materialized by Algorithm 2. Since s^* is non-empty, there exists at least one record $t \in T$ such that $t \in \text{cov}(s^*)$.

Consider the unique path from the root $s_{\text{root}} = (*, *, \ldots, *)$ to $s*$ in the population lattice. This path consists of a sequence of populations $s_0 = s_{\text{root}} \succ s_1 \succ \ldots \succ s_k = s^*$ where each s_{i+1} is the direct child of s_i.

At each step, if $|\text{cov}(s_i)| \geq \theta \cdot |T|$, the DFS continues the traversal to s_{i+1}. If the threshold is not met, all descendants of s_i are pruned.

However, if s^* is pruned due to insufficient coverage, then s^* covers fewer than $\theta \cdot |T|$ records, contradicting our assumption that s^* satisfies the coverage threshold. If s^* is not pruned, then $\text{cov}(s^*) \geq \theta \cdot |T|$. This means for each s_i (where $0 \leq i < k$) in the sequence, $|\text{cov}(s_i)| \geq \text{cov}(s^*) \geq \theta \cdot |T|$ since coverage is monotonic along ancestor-descendant relationships. In other words, the stopping criterion of DFS is not met at s_i and will continue to s_{i+1}. By induction, DFS will not stop at s_{k-1} (the direct parent of s^*) and continues to $s_k = s^*$. This contradicts our assumption that s^* is not reached (or materialized) by the DFS traversal.

Therefore, all non-empty populations (satisfying the coverage threshold) are materialized. □

Proposition 4.7. *Let $p = (s_1, s_2, X, Y)$ be a Simpson's paradox, where s_1 and s_2 belong to coverage groups $\mathcal{E}_1$ and $\mathcal{E}_2$ in $\mathcal{P}/\equiv_{\text{cov}}$, respectively. Then for any $(s'_1, s'_2) \in \mathcal{E}_1 \times \mathcal{E}_2$ such that s'_1 and s'_2 are siblings, the AC $p' = (s'_1, s'_2, X, Y)$ is also a Simpson's paradox and redundant with respect to p.*

Proof. Since $\text{cov}(s'_1) = \text{cov}(s_1)$ and $\text{cov}(s'_2) = \text{cov}(s_2)$, according to Proposition 3.1, p' is also a Simpson's paradox. Since p and p' share identical separator and label attributes, according to Definition 3.8, p and p' are coverage redundant. □

Proposition 4.9. *Let $\mathbf{P}$ be a set of sibling-child-equivalent Simpson's paradoxes with separator X and label Y. Suppose (s'_1, s'_2, X', Y'), where $X' \neq X$ or $Y' \neq Y$, is a Simpson's paradox redundant with respect to some paradox in $\mathbf{P}$. Then for every $p = (s_1, s_2, X, Y) \in \mathbf{P}$, the AC (s_1, s_2, X', Y') is also a redundant Simpson's paradox with respect to p.*

Proof. Let $p = (s_1, s_2, X, Y) \in \mathbf{P}$. Since p' is (coverage) redundant to p, by Definition 3.8, we have:

(1) $\text{cov}(s_j) = \text{cov}(s'_j)$ $(j = 1, 2)$;
(2) $P(Y|s_j) = P(Y'|s'_j)$ $(j = 1, 2)$; and
(3) there exists a one-to-one mapping f between $\text{Dom}(X)$ and $\text{Dom}(X')$ such that for every $v \in \text{Dom}(X)$ and $j \in \{1, 2\}$:
   (a) $\text{cov}(s_j\langle X \shortrightarrow v\rangle) = \text{cov}(s'_j\langle X' \shortrightarrow f(v)\rangle)$;
   (b) $P(Y|s_j\langle X \shortrightarrow v\rangle) = P(Y'|s'_j\langle X' \shortrightarrow f(v)\rangle)$.

For the AC $p'' = (s_1, s_2, X', Y')$, we need to show it's a Simpson's paradox. First, since p is a Simpson's paradox, we know $P(Y|s_1) > P(Y|s_2)$. From sibling child and statistic equivalences between p and p', we have $P(Y'|s_1) > P(Y'|s_2)$. Second, from separator equivalence between p and p', we have $P(Y'|s_1\langle X' \shortrightarrow f(v)\rangle) \leq P(Y'|s_2\langle X' \shortrightarrow f(v)\rangle)$ for every $v \in \text{Dom}(X)$. This shows that p'' satisfies Definition 2.2 and is a Simpson's paradox.

We then show that p'' is (coverage) redundant to p. First, the same one-to-one mapping f that established separator equivalence between p and p' also establishes separator equivalence between p and p''. Second, from statistic equivalence between p and p', p and p'' are also statistic equivalent. Therefore, by Definition 3.8, p'' is (coverage) redundant to p. □

Lemma 4.12. *Two Simpson's paradoxes p and p' are redundant if and only if $\textsc{Sig}(p) = \textsc{Sig}(p')$.*

Proof. ($\Rightarrow$) If $p = (s_1, s_2, X, Y)$ and $p' = (s'_1, s'_2, X', Y')$ are redundant, then by Definition 3.8:

- $\text{cov}(s_j) = \text{cov}(s'_j)$ for $j = (1, 2)$;
- $P(Y|s_j) = P(Y'|s'_j)$ for $j = (1, 2)$;
- There exists an one-to-one mapping $f : \text{Dom}(X) \rightarrow \text{Dom}(X')$ where for every $v \in \text{Dom}(X)$:
  - $\text{cov}(s_j\langle X \shortrightarrow v\rangle) = \text{cov}(s'_j\langle X' \shortrightarrow f(v)\rangle)$; and
  - $P(Y|s_j\langle X \shortrightarrow v\rangle) = P(Y'|s'_j\langle X' \shortrightarrow f(v)\rangle)$.

Therefore, $\textsc{Sig}(p) = \textsc{Sig}(p')$.

($\Leftarrow$) If $\textsc{Sig}(p) = \textsc{Sig}(p')$, then p is sibling child, separator, and statistic equivalent to p'. Hence, p and p' are redundant. □