A simple criterion for controlling selection bias

Eunice Yuh-Jie Chen
Cognitive Systems Laboratory
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
eyjchen@cs.ucla.edu

Judea Pearl
Cognitive Systems Laboratory
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
judea@cs.ucla.edu

Abstract

Controlling selection bias, a statistical error caused by preferential sampling of data, is a fundamental problem in machine learning and statistical inference. This paper presents a simple criterion for controlling selection bias in the odds ratio, a widely used measure for association between variables, that connects the nature of selection bias with the graph modeling the selection mechanism. If the graph contains certain paths, we show that the odds ratio cannot be expressed using data with selection bias. Otherwise, we show that a d-separability test can determine whether the odds ratio can be recovered, and when the answer is affirmative, output an unbiased estimand of the odds ratio. The criterion can be test in linear time and enhances the power of the estimand.

1 INTRODUCTION

Controlling selection bias is one of the most critical issues in machine learning and statistical inference. Selection bias occurs when data is sampled preferentially under a hidden selection mechanism. As a result, certain associations between variables are preferred in the data, and the data fails to represent the distribution of interest. Much research therefore has been devoted into recovering unbiased estimates from selection-biased data.

The odds ratio (OR) is a wildly used measure for association between two binary variables. Bareinboim and Pearl (2012) defined the OR as G-recoverable when an unbiased estimate of the OR can be expressed in terms of selection-biased data using the assumptions embedded in a graph representing the selection mechanism (see Geneletti et al., 2009; Hernán et al., 2004; Spirtes et al., 1995). For example, the OR between the outcome (Y) and the treatment (X) is G-recoverable when the selection mechanism is modeled by the graph in Figure 1, where S denotes the selection and the selection mechanism is active when $S = 1$. Bareinboim and Pearl (2012) have also developed a recursive graphical criterion to determine if there exists a set of variables $\mathbf{Z}$ such that conditioning on $\mathbf{Z}$ renders the OR G-recoverable; and find $\mathbf{Z}$ if it exists. In addition, they implemented the criterion using a polynomial-time algorithm.

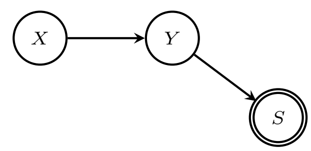


Figure 1: A graph for a selection mechanism where the odds ratio between X and Y is G-recoverable.

In this paper, we show that the criterion proposed by Bareinboim and Pearl (2012) is equivalent to a non-recursive criterion. We first show that the OR is not G-recoverable when the graph G representing the selection mechinism contains *resilient paths* (defined in Subsection 3.1). Next we show that when G does not contain resilient paths, the OR is G-recoverable if and only if there exists a set of observable *non-resilient* (defined in Subsection 3.2) variables $\mathbf{Z}$ that d-separates X and Y in a subgraph of G, and conditioning on any such $\mathbf{Z}$ renders the OR G-recoverable. The criterion can be tested in linear time, and can deliver a minimum cardinality $\mathbf{Z}$, which improves the power of the estimand, in polynomial time.

We assume that G is a directed acyclic graph (DAG)

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

and S is a sink. The assumptions hold for typical selection mechanisms. We also assume that G contains an edge $X \to Y$, reflecting the association of interest. In addition, we denote a singleton set $\{W\}$ as W; an ancestral set of W as $\mathrm{An}(W)$; and when clarity demands, denote an ancestral set of W in G as $\mathrm{An}(W)_G$.

2 BACKGROUND

This section provides a background on the OR and discusses previous work on OR recovery.

2.1 The Odds Ratio

We will be considering the OR that measures the association between two binary variables X and Y given a set of variables $\mathbf{C}$, namely, the $\mathbf{c}$-specific odds ratio. It is defined as the ratio between the odds of y when $X = x'$ and when $X = x$ given $\mathbf{C} = \mathbf{c}$.

Definition 1. *(Odds Ratio)*
Given two binary variables X, Y and a set of variables $\mathbf{C}$, the $\mathbf{c}$-specific odds ratio *between X and Y, denoted as $OR(Y, X|\mathbf{C})$, is*

$$\frac{\Pr(y|x', \mathbf{c}) / \Pr(y'|x', \mathbf{c})}{\Pr(y|x, \mathbf{c}) / \Pr(y'|x, \mathbf{c})}.$$

Following Bareinboim and Pearl (2012), we will confine our attention to covariate sets $\mathbf{C}$ consisting of X-independent variables or otherwise the $\mathbf{c}$-specific OR does not have a causal interpretation.

2.2 Previous Work

The OR may be recovered from selection-biased data by applying Lemma 1, which is stated by Didelez et al. (2010) and follows Cornfield (1951); Geng (1992); Whittemore (1978).

Lemma 1. *(OR Collapsibility)*
Given two binary variables X, Y and two sets of variables $\mathbf{T}, \mathbf{U}$, then $OR(Y, X|\mathbf{T}, \mathbf{U})$ is collapsible over $\mathbf{T}$, that is, $OR(Y, X|\mathbf{T}, \mathbf{U}) = OR(Y, X|\mathbf{U})$, if either $(X \perp\!\!\!\perp \mathbf{T}|Y, \mathbf{U})$ or $(Y \perp\!\!\!\perp \mathbf{T}|X, \mathbf{U})$.

Consider the selection mechanism modeled by the DAG G in Figure 1, which shows $(X \perp\!\!\!\perp S|Y)_G$. Let $\mathbf{T} = S$ and $\mathbf{U} = \{\}$. Then by Lemma 1, $OR(Y, X|S = 1) = OR(Y, X)$. Since $OR(Y, X|S = 1)$ can be estimated from selection-biased data, $OR(Y, X)$ is recoverable from the data.

Note that Lemma 1 involves statistical independence: $(X \perp\!\!\!\perp \mathbf{T}|Y, \mathbf{U})$ or $(Y \perp\!\!\!\perp \mathbf{T}|X, \mathbf{U})$; while the example above uses graphical independence: $(X \perp\!\!\!\perp S|Y)_G$. Statistical independence is implied by graphical independence, yet graphical independence is not always implied by statistical independence (Pearl, 1988).

Therefore we follow Bareinboim and Pearl (2012) and define a weaker notion of OR recoverability, that is, the OR recoverability over the graph representing the selection mechanism.

Definition 2. *(OR G-Recoverability)*
*Given a graph G representing a selection mechanism and a set of variables $\mathbf{C}$, then $OR(X, Y|\mathbf{C})$ is G-*recoverable *from selection-biased data when the assumptions embedded in G render it expressible in terms of observable distribution $P(\mathbf{V}_{-S}|S = 1)$, where $\mathbf{V}_{-S}$ denotes $\mathbf{V} \setminus S$. Formally, for every two probability distributions $P_1(.)$ and $P_2(.)$ compatible with G, $P_1(\mathbf{V}_{-S}|S = 1) = P_2(\mathbf{V}_{-S}|S = 1)$ implies $OR_1(X, Y|\mathbf{C}) = OR_2(X, Y|\mathbf{C})$.*

Following this definition, $OR(Y, X)$ is G-recoverable from the DAG in Figure 1. Now consider the task of recovering $OR(Y, X)$ from the DAG G in Figure 2. Since $(Y \perp\!\!\!\perp S|X, W)_G$, apply Lemma 1 with $\mathbf{T} = S, \mathbf{U} = W$ and obtain $OR(Y, X|W, S = 1) = OR(Y, X|W)$. Then since $(X \perp\!\!\!\perp W|Y)_G$, reapply Lemma 1 with $\mathbf{T} = W, \mathbf{U} = \{\}$ and obtain $OR(Y, X|W) = OR(Y, X)$. Consequently $OR(Y, X) = OR(Y, X|W, S = 1)$, and $OR(Y, X)$ is G-recoverable.

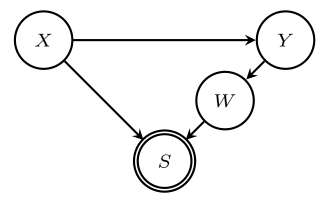


Figure 2: A DAG for a selection mechanism where $OR(Y, X)$ is G-recoverable.

In the examples above, the OR is recovered by repeatedly applying Lemma 1 till it can be expressed with selection-biased data. This approach exemplifies the OR G-recoverability criterion proposed by Bareinboim and Pearl (2012), which states that $OR(Y, X|\mathbf{C})$ is G-recoverable if and only if there exists a sequence of variables $\mathbf{Z} = (Z_1, \ldots, Z_n)$ such that by recursively applying Lemma 1 to G, $OR(Y, X|\mathbf{C}, \mathbf{Z}, S = 1) = OR(Y, X|\mathbf{C}, \mathbf{Z}) = OR(Y, X|\mathbf{C}, Z_1, \ldots, Z_{n-1}) = \ldots = OR(Y, X|\mathbf{C}, Z_1) = OR(Y, X|\mathbf{C})$. In addition, they implemented the recursive criterion using a polynomial-time algorithm.

3 A SIMPLE CRITERION FOR ODDS RATIO RECOVERY

In this section, we first define resilient paths, and provide an intuition on why they render the OR not G-recoverable. Next we present the OR G-recovery test.

In the following, $\mathbf{C}$ denotes a set of X-independent variables, which is the $\mathbf{C}$ in $OR(X,Y|\mathbf{C})$.

3.1 Non-Recoverability and Resilient Paths

Now we define resilient paths and present the theorem stating that resilient paths renders the OR not G-recoverable.

Definition 3. *(Resilient Paths)*
Given a DAG G, a resilient path *for S is*

1. *A path between Y and S passing X as a collider and is not blocked when conditioning on $\mathbf{C}\bigcup X$.*
2. *A path between X and S passing Y as a collider and is not blocked when conditioning on $\mathbf{C}\bigcup Y$.*
3. *A path p between X and Y along which there exists a non-collider T such that there is a path p' between T and S, and both p and p' are not blocked when conditioning on $\mathbf{C}$.*

Theorem 1. *Given a DAG G, then $OR(Y,X|\mathbf{C})$ is not G-recoverable if G contains resilient paths.*

Proof. See Appendix. ☐

Below, we provide examples for Theorem 1, and discuss the intuition behind.

3.1.1 Type 1 Resilient Path

Consider the task of recovering $OR(Y,X)$ from the DAG G in Figure 3. In G there exists a Type 1 resilient path $S \leftarrow W_2 \rightarrow X \leftarrow W_1 \rightarrow Y$. Consider conditioning on W_1 to recover $OR(Y,X)$. Since $(Y \perp\!\!\!\perp S|X,W_1)_G$, by Lemma 1, $OR(Y,X|W_1,S=1) = OR(Y,X|W_1)$. However because of the resilient path, there does not exist a $\mathbf{U}$ such that at least one of $(X \perp\!\!\!\perp W_1|Y,\mathbf{U})_G$ and $(Y \perp\!\!\!\perp W_1|X,\mathbf{U})_G$ holds. Consequently Lemma 1 cannot be applied to remove W_1 from the conditioning set in $OR(Y,X|W_1)$. The best we can do is to recover $OR(Y,X|W_1)$. An attempt to recover OR(Y, X) by conditioning on W_2 would fail for similar reasons.

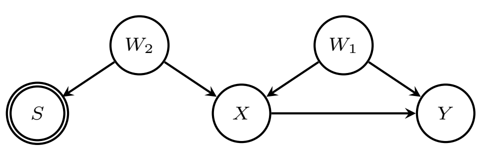


Figure 3: A DAG for a selection mechanism where given $\mathbf{C} = \{\}$, there exists a Type 1 resilient path $S \leftarrow W_2 \rightarrow X \leftarrow W_1 \rightarrow Y$.

One might surmise, at this point, that we may be able to recover $OR(Y,X)$ by averaging $OR(Y,X|W_1)$ over all W_1 (Pearl, 2012). Unfortunately even when the prior probability $P(W_1)$ is known, this cannot be accomplished because $OR(Y,X)$ is non-linear in $OR(Y,X|W_1)$ and $P(W_1)$.

3.1.2 Type 2 Resilient Path

The case for a Type 2 resilient path is similar to that of Type 1, with Y being the collider. For example, consider the task of recovering $OR(Y,X)$ from the DAG G in Figure 4. In G there exists a Type 2 resilient path $X \rightarrow Y \leftarrow W \rightarrow S$. Since $(Y \perp\!\!\!\perp S|X,W)_G$, by Lemma 1, $OR(Y,X|W,S=1) = OR(Y,X|W)$. However because of the resilient path, there does not exist a $\mathbf{U}$ such that at least one of $(X \perp\!\!\!\perp W|Y,\mathbf{U})_G$ and $(Y \perp\!\!\!\perp W|X,\mathbf{U})_G$ holds. Consequently Lemma 1 cannot be applied to remove W from the conditioning set in $OR(Y,X|W)$, and $OR(Y,X)$ is not G-recoverable.

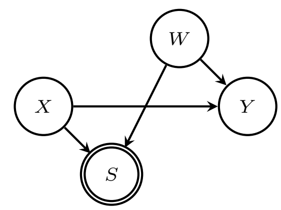


Figure 4: A DAG for a selection mechanism where given $\mathbf{C} = \{\}$, there exists a Type 2 resilient path $X \rightarrow Y \leftarrow W \rightarrow S$.

3.1.3 Type 3 Resilient Path

Consider the task of recovering $OR(Y,X)$ from the DAG G in Figure 5. In G there exists a Type 3 resilient paths $X \leftarrow W \rightarrow Y$. Since $(Y \perp\!\!\!\perp S|X,W)_G$, by Lemma 1, $OR(Y,X|W,S=1) = OR(Y,X|W)$. However because of the resilient path, there does not exist a $\mathbf{U}$ such that at least one of $(X \perp\!\!\!\perp W|Y,\mathbf{U})_G$ and $(Y \perp\!\!\!\perp W|X,\mathbf{U})_G$ holds. Consequently Lemma 1 cannot be applied to remove W from the conditioning set in $OR(Y,X|W)$, and $OR(Y,X)$ is not G-recoverable.

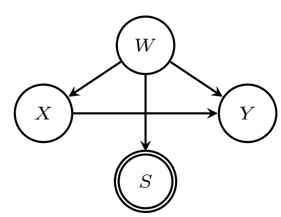


Figure 5: A DAG for a selection mechanism where given $\mathbf{C} = \{\}$, there exists a Type 3 resilient path $X \leftarrow W \rightarrow Y$.

3.2 Recovery Test

Recall that the OR G-recovery criterion proposed by Bareinboim and Pearl (2012) finds a $\mathbf{Z} = (Z_1, \ldots, Z_n)$ such that by recursively applying Lemma 1 to G, $OR(Y, X|\mathbf{C}, \mathbf{Z}, S = 1) = OR(Y, X|\mathbf{C}, \mathbf{Z}) = OR(Y, X|\mathbf{C}, Z_1, \ldots, Z_{n-1}) = \ldots = OR(Y, X|\mathbf{C}, Z_1) = OR(Y, X|\mathbf{C})$. The discussions on resilient paths suggest that if there exist generalized resilient paths for a $Z \in \mathbf{Z}$, as defined below, Z cannot be removed from the conditioning set of the OR, and consequently $OR(Y, X|\mathbf{C}, \mathbf{Z}, S = 1) \neq OR(Y, X|\mathbf{C})$.

Definition 4. *(Generalized Resilient Paths)*
Given a DAG G and a node Z, a generalized resilient path *for Z is defined the same as a resilient path in Definition 3, expect for replacing the S with Z.*

Hence when $OR(Y, X|\mathbf{C}, \mathbf{Z}, S = 1) = OR(Y, X|\mathbf{C})$, $\mathbf{Z}$ must be *non-resilient*, defined below.

Definition 5. *(Non-Resilient Variables)*
Given a DAG G, a set of variables $\mathbf{Z}$ is non-resilient *if no $Z \in \mathbf{Z}$ induces generalized resilient paths.*

Now we present the OR G-recovery theorem.

Definition 6. *(Drainage Graph)*
Given a DAG G, a drain *is a path between X and Y with colliders such that the colliders are in* $\text{An}(\mathbf{C} \bigcup S)$ *and the non-colliders are not in $\mathbf{C}$; and the* drainage graph *$\hat{G}$ is the subgraph of G consisting of the drains and the directed paths from the colliders on the drains to $\mathbf{C} \bigcup S$.*

Theorem 2. *Given a DAG G with no resilient paths, then $OR(Y, X|\mathbf{C})$ is G-recoverable if and only if there exists a set of observable non-resilient variables $\mathbf{Z} \subseteq \text{An}(\mathbf{C} \bigcup S)_G$ such that $(X \perp\!\!\!\perp Y|S, \mathbf{C}, \mathbf{Z})_{\hat{G}}$, where $\hat{G}$ is the drainage graph of G. Moreover $OR(Y, X|\mathbf{C}) = OR(Y, X|\mathbf{C}, \mathbf{Z}, S = 1)$.*

Proof. See Appendix. □

Recall the task of recovering $OR(Y, X)$ from the DAG G in Figure 2. DAG G has no resilient paths and its drainage graph $\hat{G}$ is depicted in Figure 6. Since $(X \perp\!\!\!\perp Y|S, W)_{\hat{G}}$, by Theorem 2, $OR(Y, X) = OR(Y, X|W, S = 1)$, as concluded. In this example, to see the intuition behind Theorem 2, note that since $(X \perp\!\!\!\perp Y|S, W)_{\hat{G}}$ and X, Y are non-colliders in G, at least one of $(X \perp\!\!\!\perp S|W, Y)_G$ and $(Y \perp\!\!\!\perp S|W, X)_G$ holds. Consequently by Lemma 1, $OR(Y, X|W) = OR(Y, X|W, S = 1)$. Then since in G a collider S lies between X and W, and Y is a non-collider, $(X \perp\!\!\!\perp W|Y, S)_G$. Consequently by Lemma 1, $OR(Y, X) = OR(Y, X|W)$.

Then recall the task of recovering $OR(Y, X)$ from the DAG G in Figure 1. DAG G has no resilient paths and its drainage graph $\hat{G}$ is the empty graph. Therefore $\mathbf{Z} = \{\}$ satisfies Theorem 2 and $OR(Y, X) = OR(Y, X|S = 1)$, as concluded.

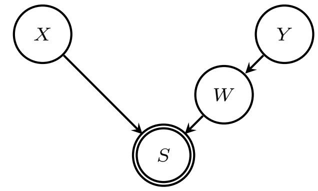
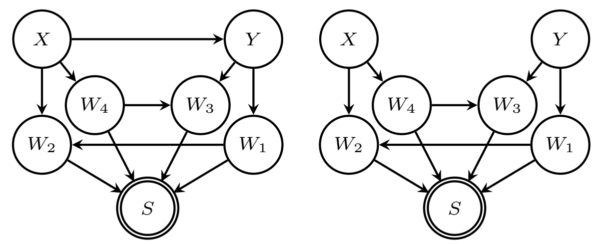


Figure 6: The drainage graph of Figure 2.

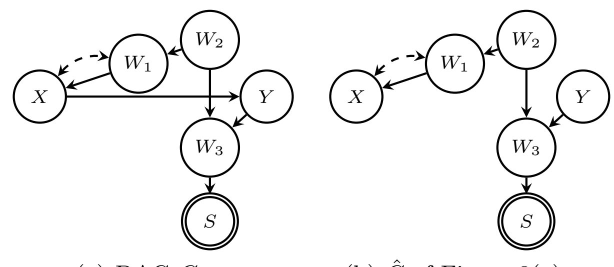
Now consider the task of recovering $OR(Y, X)$ from the DAG G in Figure 7(a) (Bareinboim and Pearl, 2012). DAG G has no resilient paths and its drainage graph $\hat{G}$ is shown in Figure 7(b). By Theorem 2, $\mathbf{Z}$ may be $\{W_1, W_2, W_4\}, \{W_1, W_3, W_4\}$ and $\{W_1, W_2, W_3, W_4\}$. Note that it is easy to find the $\mathbf{Z}$'s with minimum cardinality to enhance the power of the estimand (see Subsection 4.2).



(a) DAG G. (b) $\hat{G}$ of Figure 7(a).

Figure 7: A DAG for a selection mechanism where $OR(Y, X)$ is G-recoverable, and its drainage graph.

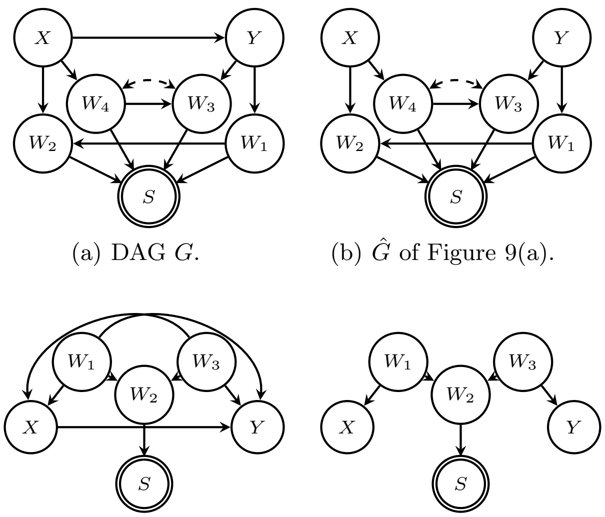
Another example is the task of recovering $OR(Y, X)$ from the DAG G in Figure 8(a). Again G has no resilient paths. Note that W_1 is an ancestor of S. Consequently the drainage graph $\hat{G}$ is the DAG in Figure 8(b). By Theorem 2, $\mathbf{Z}$ can be $\{W_2\}, \{W_1, W_2\}$, $\{W_2, W_3\}$ and $\{W_1, W_2, W_3\}$.



(a) DAG G. (b) $\hat{G}$ of Figure 8(a).

Figure 8: A DAG for a selection mechanism where $OR(Y, X)$ is G-recoverable, and its drainage graph.

Then consider the tasks of recovering $OR(Y, X)$ from the DAGs in Figure 9(a) (Bareinboim and Pearl, 2012) and 9(c). The DAGs do not contain resilient paths. Note that in the DAG in Figure 9(c), both W_1 and W_3 are not non-resilient. For both DAGs, there does not exist a set of observable non-resilient variables $\mathbf{Z} \subseteq \mathrm{An}(S)_G$ such that $(X \perp\!\!\!\perp Y|S, \mathbf{C}, \mathbf{Z})_{\hat{G}}$. Consequently by Theorem 2, $OR(Y, X)$ is not G-recoverable.



(a) DAG G. (b) $\hat{G}$ of Figure 9(a).

(c) DAG G. (d) $\hat{G}$ of Figure 9(c).

Figure 9: DAGs for selection mechanisms where $OR(Y, X)$ is not G-recoverable, and their drainage graphs.

4 ALGORITHMS

In this section, we present algorithms implementing the criterion we proposed. We use G^m to denote the moral graph of G, and $G_{\mathbf{W}}$ to denote the subgraph of G induced by a set of nodes $\mathbf{W}$.

4.1 Resilient Paths

Recall that the definition of resilient paths involves d-separation, which we would test with Theorem 3 (Lauritzen et al., 1990).

Theorem 3. *Given a DAG G and a set of variables $\mathbf{W}$, $(X \perp\!\!\!\perp Y|\mathbf{W})_G$ if and only if $\mathbf{W}$ separates X and Y in $\big(G_{\mathrm{An}(\mathrm{X} \bigcup \mathrm{Y} \bigcup \mathbf{W})}\big)^m$.*

Consider testing if there exist Type 1 resilient paths in a DAG G. Let G' denote $\big(G_{\mathrm{An}(\mathbf{C} \bigcup \mathrm{X} \bigcup \mathrm{Y} \bigcup \mathrm{S})}\big)^m$. DAG G' can be constructed in $O\big(|E(G')|\big)$ time. Then on G' two breadth-first searches, one starting from Y and the other from S, can together enumerate all the paths between Y and S and check whether any passes X as a collider in $O\big(|E(G')|\big)$ time. The other two types of resilient paths can be tested similarly.

4.2 Odds Ratio Recovery

Recall Theorem 2. Algorithms similar to the ones in Subsection 4.1 can find the non-resilient variables in $\mathrm{An}(\mathbf{C} \bigcup S)_G$ and construct the drainage graph $\hat{G}$. Then using the algorithms proposed by Acid and De Campos (1996) and Tian et al. (1998), an observable non-resilient $\mathbf{Z} \subseteq \mathrm{An}(\mathbf{C} \bigcup S)_G$ such that $(X \perp\!\!\!\perp Y|S, \mathbf{C}, \mathbf{Z})_{\hat{G}}$ can be found in $O\big(|E(\hat{G}^m)|\big)$ time. Moreover a minimum $\mathbf{Z}$ can be obtained in $O\big(|\mathbf{N}| \cdot |E(\hat{G}^m)|\big)$ time, where $\mathbf{N}$ is the set of observable non-resilient variables in $\mathrm{An}(\mathbf{C} \bigcup S)_G$ in $\hat{G}$.

5 CONCLUSION

This paper presents a simple criterion for determining if the OR can be recovered from selection-biased data, and when the answer is affirmative, it outputs an unbiased estimand of the OR. The criterion can be implemented using linear-time algorithms and enhances the power of the estimand.

Appendix

Our proofs for Theorem 1 and 2 involve the following theorem (Bareinboim and Pearl, 2012).

Theorem 4. *Given a DAG G, then $OR(Y, X|\mathbf{C})$ is G-recoverable if and only if there exists an ordered set of observable variables $\mathbf{Z} = (Z_1, \ldots, Z_n)$ such that*

$$(X \perp\!\!\!\perp Z_i^S|\mathbf{C}, Y, Z_1^S, \ldots, Z_{i-1}^S)_G \text{ or}$$
$$(Y \perp\!\!\!\perp Z_i^S|\mathbf{C}, X, Z_1^S, \ldots, Z_{i-1}^S)_G \quad (1)$$

for $1 \leq i \leq n+1$, where $Z^S = (\mathbf{Z}, S)$. Moreover if $\mathbf{Z}$ exists, $OR(Y, X|\mathbf{C}) = OR(Y, X|\mathbf{C}, \mathbf{Z}, S = 1)$.

In the following, we use $p(X, Y)$ to denote a path p between X and Y; $p^{\rightsquigarrow}|\mathbf{W}$ to denote that p is not blocked when conditioning on $\mathbf{W}$; and $\mathbf{Z}_{1,i}$ to denote $\{Z_1, \ldots Z_i\}$. In addition, we refer to a generalized resilient path for Z simply as a resilient path for Z.

Theorem 1

Lemma 2. *Given a Type 3 resilient path p, recall that p' denotes the path between T and S. Then the path p_x formed by the subpath of p between X and T and p' is $p_x^{\rightsquigarrow}|\mathbf{C} \bigcup Y$; and the path p_y formed by the subpath of p between T and Y and p' is $p_y^{\rightsquigarrow}|\mathbf{C} \bigcup X$.*

Proof. First consider the case where p is of the form $X - \ldots - \leftarrow T \leftarrow - \ldots - Y$ and p' is of the form $T \leftarrow - \ldots - S$. Since $p^{\rightsquigarrow}|\mathbf{C}$ and $p'^{\rightsquigarrow}|\mathbf{C}$, $p_x^{\rightsquigarrow}|\mathbf{C} \bigcup Y$. Then since $p^{\rightsquigarrow}|\mathbf{C}$ and $\mathbf{C}$ is X-independent, T must be is an ancestor of X on p. Consequently since $p^{\rightsquigarrow}|\mathbf{C}$ and $p'^{\rightsquigarrow}|\mathbf{C}$, $p_y^{\rightsquigarrow}|\mathbf{C} \bigcup X$. The remaining cases can be reasoned similarly. ∎

Now we prove Theorem 1.

Proof. For each type of resilient paths, we show that Equation 1 fails for some i, and consequently by Theorem 4 the OR is not G-recoverable.

Consider a Type 1 resilient path p. Since $p^{\leadsto}|\mathbf{C}\bigcup X$, to satisfy Equation 1 for $i = n+1$, on p there must exist a non-collider $Z_{i'}^S \in \mathbf{Z}^S$, where $i' < i$. Repeat this reasoning till possible $Z_{i'}^S$ are exhausted and Equation 1 fails for some i. For a Type 2 resilient path, similar arguments apply. Then consider a Type 3 resilient path. By Lemma 2, $p_x^{\leadsto}|\mathbf{C}\bigcup Y$ and $p_y^{\leadsto}|\mathbf{C}\bigcup X$. Then similar arguments apply. ☐

Theorem 2

To prove Theorem 2, we first prove Theorem 5.

Definition 7. *(Generalized Drain and Generalized Drainage Graph)*
Given a DAG G and a set of variables $\mathbf{Z}$, then a generalized drain *and the* generalized drainage graph *$\tilde{G}$ are defined the same as a drain and the drainage graph in Definition 6, expect for replacing $\mathbf{C}\bigcup S$ with $\mathbf{C}\bigcup\mathbf{Z}\bigcup S$.*

Theorem 5. *Given a DAG G with no resilient paths, then $OR(Y, X|\mathbf{C})$ is G-recoverable if and only if there exists a set of observable non-resilient variables $\mathbf{Z}$ such that $(X \perp\!\!\!\perp Y|S, \mathbf{C}, \mathbf{Z})_{\tilde{G}}$, where $\tilde{G}$ is the generalized drainage graph of G. Moreover $OR(X, Y|\mathbf{C}) = OR(X, Y|\mathbf{C}, \mathbf{Z}, S = 1)$.*

Given Theorem 1 and 4, to prove Theorem 5, it suffices to show the following lemma to be true.

Lemma 3. *Given a DAG G with no resilient paths, an ordered set of variables $\mathbf{Z}$ satisfies Equation 1 if and only if it is non-resilient and $(X \perp\!\!\!\perp Y|S, \mathbf{C}, \mathbf{Z})_{\tilde{G}}$.*

($\Rightarrow$)

Lemma 4. *Given a DAG G with no resilient paths, and a set of non-resilient variables $\mathbf{Z}$, then in $\tilde{G}$, every path $p(X, Y)$ passes a collider not in $\mathrm{An}(\mathbf{C})$.*

Proof. When p passes at least one collider, since $\mathbf{C}$ is X-independent, the collider that is the closest to X is not in $\mathrm{An}(\mathbf{C})$. Then consider when p only passes non-colliders. Since p is in $\tilde{G}$, it consists of generalized drains. As a result, $p^{\leadsto}|\mathbf{C}$. Moreover since $\mathbf{C}$ is X-independent, on p there exists a T such that there is a path between T and a $Z \in \mathbf{Z}^S$ that is not blocked when conditioning on $\mathbf{C}$. Consequently p is a Type 3 resilient path for Z, a contradiction. ☐

Now we prove the ($\Rightarrow$) part of Lemma 3.

Proof. Arguments similar to those used for Theorem 1 show that $\mathbf{Z}$ must be non-resilient. Then we show that $(X \perp\!\!\!\perp Y|S, \mathbf{C}, \mathbf{Z})_{\tilde{G}}$. Consider a $p(X, Y)$ in $\tilde{G}$. First note that $p^{\leadsto}|\mathbf{C}\bigcup\mathbf{Z}^S$. Then note that p does not pass nodes in $\mathbf{C}$ as non-colliders, and that by Lemma 4, p passes a collider not in $\mathrm{An}(\mathbf{C})$. Consequently if $(X \not\perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}^S)_{\tilde{G}}$, in $\tilde{G}$, Equation 1 must fail for some i. ☐

($\Leftarrow$)

Lemma 5. *Given a DAG G with no resilient paths, and a set of non-resilient variables $\mathbf{Z}$ such that $(X \perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}^S)_{\tilde{G}}$, then $(X \perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}_{1,i}^S)_{\tilde{G}}$ for $1 \leq i \leq n+1$.*

Proof. By Lemma 4, $(X \perp\!\!\!\perp Y|\mathbf{C})_{\tilde{G}}$. Since G is a DAG, from $(X \perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}^S)_{\tilde{G}}$ to $(X \perp\!\!\!\perp Y|\mathbf{C})_{\tilde{G}}$, there must exist a sequence of Z_i such that the removal of each Z_i from the conditioning set does not render X d-connect to Y, that is, $(X \perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}_{1,i})_{\tilde{G}}$ for $1 \leq i \leq n$. ☐

Lemma 6. *Given a DAG G with no resilient paths, and a set of non-resilient variables $\mathbf{Z}$ such that $(X \perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}^S)_{\tilde{G}}$, then in G there do not exist paths $p(Y, Z_i^S)^{\leadsto}|\mathbf{C}\bigcup X\bigcup\mathbf{Z}_{1,i-1}^S$ that pass exactly one collider in $\mathrm{An}(X) \setminus \mathrm{An}(\mathbf{C}\bigcup\mathbf{Z}_{1,i-1}^S)$.*

Proof. Assume the contrary, and let T denote the collider in $\mathrm{An}(X) \setminus \mathrm{An}(\mathbf{C}\bigcup\mathbf{Z}_{1,i-1}^S)$ on p. Since $p^{\leadsto}|\mathbf{C}\bigcup X\bigcup\mathbf{Z}_{1,i-1}^S$ and there are no Type 1 or 3 resilient paths for $\mathbf{Z}^S$, on p between Y and T there must exist colliders T' in $\mathrm{An}(\mathbf{Z}_{1,i-1}^S) \setminus \mathrm{An}(\mathbf{C})$. Then since $p^{\leadsto}|\mathbf{C}\bigcup X\bigcup\mathbf{Z}_{1,i-1}^S$, there must exist a $Z_{i'}^S \in \mathbf{Z}_{1,i-1}^S$ such that $p_x(X, Z_{i'}^S)^{\leadsto}|\mathbf{C}\bigcup\mathbf{Z}_{1,i'-1}^S$ and $p_y(Y, Z_{i'}^S)^{\leadsto}|\mathbf{C}\bigcup\mathbf{Z}_{1,i'-1}^S$, where $p_x : X \leftarrow \ldots \leftarrow T \leftarrow - \ldots - \rightarrow T' \rightarrow \ldots \rightarrow Z_{i'}^S$ and $p_y : Y - \ldots - \rightarrow T' \rightarrow \ldots \rightarrow Z_{i'}^S$. By Definition 7, p_x and p_y are in $\tilde{G}$, and consequently $(X \not\perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}_{1,i'}^S)_{\tilde{G}}$, a contradiction by Lemma 5. ☐

Lemma 7. *Given a DAG G with no resilient paths, and a set of non-resilient variables $\mathbf{Z}$ such that $(X \perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}^S)_{\tilde{G}}$, then in G there do not exist paths $p(Y, Z_i^S)^{\leadsto}|\mathbf{C}\bigcup X\bigcup\mathbf{Z}_{1,i-1}^S$ that pass more than one collider in $\mathrm{An}(X) \setminus \mathrm{An}(\mathbf{C}\bigcup\mathbf{Z}_{1,i-1}^S)$.*

Proof. Assume the contrary, and consider the colliders in $\mathrm{An}(X) \setminus \mathrm{An}(\mathbf{C}\bigcup\mathbf{Z}_{1,i-1}^S)$ on p . Let T denote the one closest to Y, and T' denote the one closest to Z_i^S. Then consider path $p' : Y - \ldots - \rightarrow T \rightarrow \ldots \rightarrow X \leftarrow \ldots \leftarrow T' \leftarrow - \ldots - Z_i^S$. Since $p^{\leadsto}|\mathbf{C}\bigcup X\bigcup\mathbf{Z}_{1,i-1}^S$, for p' not to be a Type 1 resilient path for $\mathbf{Z}^S$, there must exist colliders T'' in $\mathrm{An}(\mathbf{Z}_{1,i-1}^S) \setminus \mathrm{An}(\mathbf{C})$ between Y and T. Then since $p^{\leadsto}|\mathbf{C}\bigcup X\bigcup\mathbf{Z}_{1,i-1}^S$, there must exist a $Z_{i'}^S \in \mathbf{Z}_{1,i-1}^S$ such that $p_x(X, Z_{i'}^S)^{\leadsto}|\mathbf{C}\bigcup\mathbf{Z}_{1,i'-1}^S$ and $p_y(Y, Z_{i'}^S)^{\leadsto}|\mathbf{C}\bigcup\mathbf{Z}_{1,i'-1}^S$, where $p_x : X \leftarrow \ldots \leftarrow$

$T \leftarrow - \ldots - \rightarrow T'' \rightarrow \ldots \rightarrow Z^S_{i'}$ and $p_y : Y - \ldots - \rightarrow T'' \rightarrow \ldots \rightarrow Z^S_{i'}$. By Definition 7, p_x and p_y are in $\tilde{G}$, and consequently $(X \not\perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}^S_{1,i'})_{\tilde{G}}$, a contradiction by Lemma 5. □

Lemma 8. *Given a DAG G with no resilient paths, and a set of non-resilient variables* $\mathbf{Z}$ *such that* $(X \perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}^S)_{\tilde{G}}$*, then in G there do not exist paths* $p(X, Z^S_i)^{\rightsquigarrow}|\mathbf{C} \bigcup Y \bigcup \mathbf{Z}^S_{1,i-1}$ *that pass exactly one collider in* $\mathrm{An}(Y) \setminus \mathrm{An}(\mathbf{C} \bigcup \mathbf{Z}^S_{1,i-1})$.

Proof. Similar to that of Lemma 6. □

Lemma 9. *Given a DAG G with no resilient paths, and a set of non-resilient variables* $\mathbf{Z}$ *such that* $(X \perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}^S)_{\tilde{G}}$*, then in G there do not exist paths* $p(X, Z^S_i)^{\rightsquigarrow}|\mathbf{C} \bigcup Y \bigcup \mathbf{Z}^S_{1,i-1}$ *that pass more than one collider in* $\mathrm{An}(Y) \setminus \mathrm{An}(\mathbf{C} \bigcup \mathbf{Z}^S_{1,i-1})$.

Proof. Similar to that of Lemma 7. □

Lemma 10. *Given a DAG G with no resilient paths, and a set of non-resilient variables* $\mathbf{Z}$ *such that* $(X \perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}^S)_{\tilde{G}}$*, then in G there do not both exist a path* $p_x(X, Z^S_i)^{\rightsquigarrow}|\mathbf{C} \bigcup Y \bigcup \mathbf{Z}^S_{1,i-1}$ *that does not pass colliders in* $\mathrm{An}(Y) \setminus \mathrm{An}(\mathbf{C} \bigcup \mathbf{Z}^S_{1,i-1})$*, and a path* $p_y(Y, Z^S_i)^{\rightsquigarrow}|\mathbf{C} \bigcup X \bigcup \mathbf{Z}^S_{1,i-1}$ *that does not pass colliders in* $\mathrm{An}(X) \setminus \mathrm{An}(\mathbf{C} \bigcup \mathbf{Z}^S_{1,i-1})$.

Proof. Assume the contrary. Since p_x does not pass colliders in $\mathrm{An}(Y) \setminus \mathrm{An}(\mathbf{C} \bigcup \mathbf{Z}^S_{1,i-1})$, $p_x^{\rightsquigarrow}|\mathbf{C} \bigcup \mathbf{Z}^S_{1,i-1}$. Similarly $p_y^{\rightsquigarrow}|\mathbf{C} \bigcup \mathbf{Z}^S_{1,i-1}$. Then let p denote path $X - \ldots - T - \ldots - Y$, where T is the intersection of p_x and p_y, and let p' denote the subpath of p_x between T and Z^S_i.

We first show that on p there exist colliders T' in $\mathrm{An}(\mathbf{Z}^S_{1,i-1}) \setminus \mathrm{An}(\mathbf{C})$. Consider when T is a non-collider on p. Since $p_x^{\rightsquigarrow}|\mathbf{C} \bigcup \mathbf{Z}^S_{1,i-1}$, $p_y^{\rightsquigarrow}|\mathbf{C} \bigcup \mathbf{Z}^S_{1,i-1}$, and $\mathbf{C}$ is X-independent, for p not to be a Type 3 resilient path for $\mathbf{Z}^S$, on p there must exist colliders T' in $\mathrm{An}(\mathbf{Z}^S_{1,i-1}) \setminus \mathrm{An}(\mathbf{C})$. Then consider when T is a collider on p. Since $p_x^{\rightsquigarrow}|\mathbf{C} \bigcup \mathbf{Z}^S_{1,i-1}$ and $\mathbf{C}$ is X-independent, either there exists colliders in $\mathrm{An}(\mathbf{Z}^S_{1,i-1}) \setminus \mathrm{An}(\mathbf{C})$ on p between X and T or T is in $\mathrm{An}(\mathbf{Z}^S_{1,i-1}) \setminus \mathrm{An}(\mathbf{C})$.

Then since $p_x^{\rightsquigarrow}|\mathbf{C} \bigcup \mathbf{Z}^S_{1,i-1}$ and $p_y^{\rightsquigarrow}|\mathbf{C} \bigcup \mathbf{Z}^S_{1,i-1}$, there must exist a $Z^S_{i'} \in \mathbf{Z}^S_{1,i-1}$ such that ${p'_x}^{\rightsquigarrow}|\mathbf{C} \bigcup \mathbf{Z}^S_{1,i'-1}$ and ${p'_y}^{\rightsquigarrow}|\mathbf{C} \bigcup \mathbf{Z}^S_{1,i'-1}$, where $p'_x : X - \ldots - \rightarrow T' \rightarrow \ldots \rightarrow Z^S_{i'}$ and $p'_y : Y - \ldots - \rightarrow T' \rightarrow \ldots \rightarrow Z^S_{i'}$. By Definition 7, p_x and p_y are in $\tilde{G}$, and consequently $(X \not\perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}^S_{1,i'})_{\tilde{G}}$, a contradiction by Lemma 5. □

Now we prove the ($\Leftarrow$) part of Lemma 3.

Proof. Assume Equation 1 fails for some i. By Lemma 6, 7, 8, 9, and 10, a contradiction occurs. □

Before we prove Theorem 2, we first prove the following lemmas.

Lemma 11. *(Acid and De Campos, 1996) Given a DAG G, nodes* V, V'*, and a set of nodes* $\mathbf{U}$*, if there exists a set of nodes* $\mathbf{U}'$ *such that* $(V \perp\!\!\!\perp V'|\mathbf{U}, \mathbf{U}')_G$*, then the minimal sets* $\mathbf{U}^* \subseteq \mathbf{U}'$ *such that* $(V \perp\!\!\!\perp V'|\mathbf{U}, \mathbf{U}^*)_G$ *are in* $\mathrm{An}(V \bigcup V' \bigcup \mathbf{U})$.

Lemma 12. *If there exists a set of observable non-resilient variables* $\mathbf{Z}$ *satisfying Equation 1, then there exists a set of observable non-resilient variables* $\mathbf{Z}' \subseteq \mathrm{An}(\mathbf{C} \bigcup X \bigcup Y \bigcup S)_G$ *satisfying Equation 1.*

Proof. Follows directly from Lemma 11. □

Lemma 13. *Given a DAG G with no resilient paths, and a set of observable non-resilient variables* $\mathbf{Z} \subseteq \mathrm{An}(\mathbf{C} \bigcup X \bigcup Y \bigcup S)_G$ *such that* $(X \perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}^S)_{\tilde{G}}$*, then when* $\tilde{G}$ *is not the empty graph, Y is in* $\mathrm{An}(S)_G \setminus \mathrm{An}(\mathbf{C} \bigcup X \bigcup Y)_G$.

Proof. Consider paths $p(X, Y)$ in $\tilde{G}$. First note that $p^{\rightsquigarrow}|\mathbf{C} \bigcup \mathbf{Z}^S$. Then note that by Lemma 4, p passes colliders. Among these colliders, let T be the one closest to Y on p. Then recall that G always contains $X \rightarrow Y$, and consider the path p' in G that consists of $X \rightarrow Y$ and subpath of p between Y and T. Since $p^{\rightsquigarrow}|\mathbf{C} \bigcup \mathbf{Z}^S$ and $\mathbf{C}$ is X-independent, T is in $\mathrm{An}(\mathbf{Z}^S)_G \setminus \mathrm{An}(\mathbf{C})_G$. Now we show that p' is of the form $X \rightarrow Y \rightarrow \ldots \rightarrow T$. Assume the contrary, that is, p' is of the form $X \rightarrow Y \leftarrow \ldots \leftarrow\rightarrow \ldots \rightarrow T$. Since $T \in \mathrm{An}(\mathbf{Z}^S)_G \setminus \mathrm{An}(\mathbf{C})_G$, there exists a Type 2 resilient path for $\mathbf{Z}^S$, a contradiction. Then recall that $\mathbf{Z} \subseteq \mathrm{An}(\mathbf{C} \bigcup X \bigcup Y \bigcup S)_G$. Since G is a DAG, T is in $\mathrm{An}(S)_G \setminus \mathrm{An}(\mathbf{C} \bigcup X \bigcup Y)_G$. Consequently Y is in $\mathrm{An}(S)_G \setminus \mathrm{An}(\mathbf{C} \bigcup X \bigcup Y)_G$. □

Lemma 14. *Given a DAG G with no resilient paths, and a set of observable non-resilient variables* $\mathbf{Z}' \subseteq \mathrm{An}(\mathbf{C} \bigcup X \bigcup Y \bigcup S)_G$ *such that* $(X \perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}'^S)_{\tilde{G}}$*, then if* $\tilde{G}$ *is the empty graph, Equation 1 is satisfied when* $\mathbf{Z} = \{\}$.

Proof. Assume the contrary, that is, Equation 1 is not satisfied when $\mathbf{Z} = \{\}$. Since G contains no resilient paths, there must exist drains. Consequently $\tilde{G}$ is not empty, a contradiction. □

Now we prove Theorem 2.

Proof. When $\tilde{G}$ is the empty graph, by Lemma 14, Theorem 2 is true. Then consider when $\tilde{G}$ is not empty.

By Lemma 12, if there exists an observable non-resilient $\mathbf{Z}$ such that $(X \perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}^S)_{\tilde{G}}$, there exists an observable non-resilient $\mathbf{Z}' \subseteq \mathrm{An}(\mathbf{C} \bigcup X \bigcup Y \bigcup S)_G$ such that $(X \perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}'^S)_{\tilde{G}}$. Recall that G always contains $X \to Y$. Then by Lemma 13, X, Y are in $\mathrm{An}(S)_G$. Consequently $\mathbf{Z}' \subseteq \mathrm{An}(\mathbf{C} \bigcup S)_G$ and $\tilde{G} = \hat{G}$. As a result, $(X \perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}'^S)_{\tilde{G}}$ if and only if $(X \perp\!\!\!\perp Y|\mathbf{C}, \mathbf{Z}'^S)_{\hat{G}}$. Then by Theorem 5, Theorem 2 is true. □

Acknowledgement

The authors would like to thank the reviewers for their helpful comments. This research was supported in parts by grants from NSF #IIS-1249822 and ONR #N00014-13-1-0153 and #N00014-10-1-0933.

References

S. Acid and L. M. De Campos. An algorithm for finding minimum d-separating sets in belief networks. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pages 3–10, 1996.

E. Bareinboim and J. Pearl. Controlling selection bias in causal inference. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pages 100–108, 2012.

J. Cornfield. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*, 11(6):1269, 1951.

V. Didelez, S. Kreiner, and N. Keiding. Graphical models for inference under outcome-dependent sampling. *Statistical Science*, 25(3):368–387, 2010.

S. Geneletti, S. Richardson, and N. Best. Adjusting for selection bias in retrospective, case–control studies. *Biostatistics*, 10(1):17–31, 2009.

Z. Geng. Collapsibility of relative risk in contingency tables with a response variable. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 585–593, 1992.

M.A. Hernán, S. Hernández-Díaz, and J.M. Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615, 2004.

S.L. Lauritzen, A.P. Dawid, B.N. Larsen, and H.-G. Leimer. Independence properties of directed markov fields. *Networks*, 20(5):491–505, 1990.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan Kaufmann, 1988.

J. Pearl. A solution to a class of selection-bias problems. Technical report, UCLA Cognitive Systems Laboratory, 2012.

P. Spirtes, C. Meek, and T. Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence*, pages 499–506, 1995.

J. Tian, J. Pearl, and A. Paz. Finding minimal d-separators. Technical report, UCLA Cognitive Systems Laboratory, 1998.

A.S. Whittemore. Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 328–340, 1978.