IUST Iran University of Science and Technology
Iranian Journal of Electrical and Electronic Engineering
Journal Homepage: ijeee.iust.ac.ir

Optimal Passive Experiment Design for Full Identification of Causal Structure Learning

R. Safaeian*(C.A.), M. Tabandeh**

Abstract: Directed Acyclic Graphs stand as one of the prevailing approaches for representing causal relationships within a set of variables. With observational or interventional data, certain undirected edges within a causal DAG can be oriented. Performing intervention can be done in two different settings, passive and active. Here, we prove that an optimal intervention set can be obtained based on the minimum vertex cover of a graph. We propose an algorithm that efficiently identifies such an optimal intervention set for chordal graphs within polynomial time. Performing intervention on this optimal set recovers all the undirected edges in graph G, regardless of the underlying ground truth DAG. Furthermore, we present an algorithm for evaluating the performance of passive algorithms. This evaluation provides insights into how many intervention steps of a specific algorithm are required to recover all edges in the causal graph for any possible underlying ground truth in the equivalence class. Experimental findings underscore that the number of nodes in the optimal intervention set increases with growing the number of nodes in a graph, where the edge density is fixed, and also increases with the rising edge density in a graph with a fixed number of nodes.

Keywords: Causal Structure Learning, Passive Setting, Full Identification

1 Introduction

Revealing the causal relationships within a dataset is a fundamental objective in empirical sciences. Without understanding how variables in a complex system influence each other, it becomes challenging to predict the system's behavior when subjected to specific interventions.

In scientific literature, causal relationships among variables are commonly represented using directed acyclic graphs (DAGs). In DAG representation, a directed edge from node v_1 to node v_2 indicates that v_1 is a direct cause of v_2. Through observational data, and assuming faithfulness and causal sufficiency, it is possible to identify the true causal graph within a Markov equivalence class (MEC). The MEC is a set of all DAGs that encode the same conditional independences among variables. It has been demonstrated that all DAGs exhibit identical skeletons and sets of v-structures[1], within a given MEC. Given the set of v-structures and skeleton, four rules are introduced in [10] for further edge orientation, known as Meek rules. By applying these rules, more edges in the graph can be oriented, while ensuring that no directed cycles or new v-structures are introduced.

Applying the rules iteratively recovers more edges. However, some edges may not be recovered even after applying Meek rules. The resulting graph is referred to as the essential graph, summarizing all DAGs within the MEC. Directed edges in the essential graph maintain consistent orientation across all DAGs within MEC. Undirected edges in the essential graph result in opposite directions for that edge in at least two DAGs within the MEC.

Iranian Journal of Electrical & Electronic Engineering, 2024.
Paper first received 27 January 2024 and accepted 11 April 2024.
*The author is with the Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran.
E-mail: ramin.safaeian@ee.sharif.edu.
**The author is with the Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran.
E-mail: tabandeh@sharif.edu.
Corresponding Author: Ramin Safaeian.

[1] A v-structure forms with nodes X, Y, and Z if X and Y directly cause Z and are not directly connected to each other.

The primary standard procedure for uncovering the complete causal structure entails conducting interventions within the system. Observational data facilitates the recovery of the essential graph, while interventional data aids in orienting undirected edges. The objective here is to orient additional edges, given the absence of v-structures and directed cycles in the graph. Again, Meek rules serve as a foundation for recovering more edges in the graph. Here, we focus on the node selection algorithm for intervention. In the causality context, a problem of experiment design is defined as devising algorithms for selecting intervention nodes and performing intervention on those nodes for recovering undirected edges in the graph. In the following, we discuss proposed solutions for the experiment design problem in previous work.

Theoretically, recovering the complete causal structure is possible with enough interventions, but in many practical applications, challenges arise due to time and cost constraints. The objective of the experiment design problem is to create a series of experiments, each entailing the execution of multiple interventions concurrently. Some work allowed for a fixed maximum number of interventions per experiment, while others perform only one intervention per experiment.

Passive learning and active learning represent different approaches within the experiment design problem. In the active learning scenario, experiments are performed sequentially, and the results of each experiment are used to guide the design of subsequent ones. Perfect randomized interventions can orient all edges which are connected to an intervened variable. Using these newly oriented edges, Meek's rules can orient additional edges within resulting chain components. In active settings, the goal is to identify the complete causal structure with fewer interventions. Passive learning operates within a limited budget for selecting interventions, utilizing the derived MEC from observational data and conducting experiments simultaneously. The objective is to devise a sequence of experiments with minimal interventions to orient as many edges as possible in the causal structure.

In the passive learning scenario, [7] proposed a straightforward method that involves listing all DAGs within a MEC and selecting a minimal set of variables for interventions, ensuring the recovery of all potential underlying DAGs from the intervention outcomes. [4] introduced a greedy algorithm for selecting target interventions by sampling DAGs from the MEC. This study estimates the average number of edges that can be oriented as a result of interventions. [8] proposed an algorithm for experiment design considering a cost associated with intervening on each variable. They demonstrated that the optimal solution can be obtained in polynomial time when the causal structure is a tree or a clique tree. In [13], the author presents Meek functions as a solution to causal orientation learning problems. The functions exhibit favorable properties, allowing for an accelerated application of Meek rules. They apply Meek functions efficiently by employing a dynamic programming approach. Furthermore, they introduce a lower bound for number of undirected edges that can be oriented through intervention by employing the proposed Meek functions.

This paper aims to develop an algorithm to find the optimal intervention set, in terms of number of nodes, which performing intervention on, recovers all undirected edges in the graph. Here is a summary of the main contributions of the paper:

- We state a proposition to find the minimal intervention set, which recovers all the undirected edges in the passive setting after intervention.
- We propose an algorithm that computes the optimal intervention set in a polynomial time on chordal graphs.
- We propose an algorithm that identifies how many interventions a passive setting algorithm needs for recovering all the undirected edges in a graph, regardless of what underlying ground truth DAG is.

Following is the organization of the paper: Section 2 provides problem definition and introduces relevant terminologies. Section 3 presents the proposed approach for finding the optimal intervention set, and an evaluation method for different algorithms. The experimental results are reported in Section 4, and the paper concludes in Section 5, discussing potential future research directions.

2 Preliminaries

2.1 Graph Terminology

A graph G is defined as a pair $G(\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V}$ represents the set of nodes (vertices) in the graph and $\boldsymbol{E}$ represents the set of edges. An undirected edge between two nodes $v_1, v_2 \subseteq \boldsymbol{V}$ is denoted as $v_1 - v_2$. This edge does not have a specific direction and can be directed in both directions. An oriented edge originating from node v_1 and terminating at node v_2 is represented as $v_1 \rightarrow v_2$. This manuscript assumes that there can be, at most, a single edge (whether directed or undirected) connecting any pair of nodes within the graph. This ensures that the graph does not contain multiple edges between the same pair of nodes. For a node v in the graph, $neigh(v)$ represents the set of nodes that are directly connected to v. In other words, $neigh(v)$ contains all the neighbors of v, which are the nodes directly connected to v by either an undirected or directed edge.

A clique in an undirected graph is a subset of nodes where every pair of nodes in the subset are adjacent. In other words, a clique is a fully connected subgraph. A partially directed acyclic graph (PDAG) is a graph without any directed cycle [12]. In addition, a directed acyclic graph (DAG) is a PDAG, like G, where all the edges in G are directed.

Any sub-graph which has the formation of $v_1 \rightarrow v_2 \leftarrow v_3$ is called as a v-structure. In addition, a chain graph is a graph without any partially directed cycle. After eliminating all directed edges in a chain graph, what remains are undirected disjoint chain graphs, known as chain components. A graph is deemed chordal if there is a chord present in any cycle with a length exceeding three.

2.2 Causal Model

The causal relationships among variables $X = X_1, \ldots, X_n$, are commonly represented by a directed acyclic graph (DAG) G. Each variable in X corresponds to a node in G, and an arrow from one node to another, e.g., $v_1 \rightarrow v_2$, indicates that the variable represented by v_1 directly influences the variable represented by v_2. This DAG, which captures the causal relationships, is referred to as a "causal DAG".

The Markov property is a fundamental concept in probabilistic graphical models. In the context of a joint distribution P across X, Markov property holds true with respect to graph G when any variable within G becomes independent of its non-descendants when conditioned on its parents. This property ensures that the conditional independence relationships encoded by G hold in the distribution P.

Given the assumptions of faithfulness and causal sufficiency, the Markov property allows for the inference of any conditional independence within the joint distribution P. In other words, the Markov property characterizes the conditional independence assertions implied by the causal structure.

The collection of all Directed Acyclic Graphs (DAGs) that represent identical conditional independence relationships is referred to as a Markov equivalence class (MEC). Each MEC is associated with an essential graph, wherein the skeleton of this essential graph matches the skeletons of all DAGs within the MEC. In the essential graph, an edge is directed if and only if it has the same direction in all DAGs of the MEC. Furthermore, the essential graph captures the common structure shared by all DAGs in the MEC.

Removing of all directed edges from the essential graph, while preserving the directed ones, results in another graph known as the chain component. In other words, each chain component corresponds to a subset of variables that have a chain-like causal relationship. There are no unidentified v-structures (collider structures) within a chain component, since all the v-structures have been identified and represented by directed edges in the essential graph.

We use the term UCCG to refer one of the chain components of an essential graph. It represents a subset of variables with a chain-like causal relationship, where the presence of undirected edges indicates conditional dependencies between the variables.

In the context of inferring the causal structure from observational data, the essential graph representing the true underlying DAG can be reconstructed through the execution of conditional independence tests on the observational distribution P [14]. These tests allow us to identify the absence of direct causal relationships between variables.

However, to further determine the orientations of undirected edges and fully recover the causal structure, intervention in the system is required. The notion of intervention used here is referred to as a "hard intervention" as defined by [2] and [11]. When a random variable X is intervened, it obtains values from a randomized distribution independent of its parents' values.

An independent set $\boldsymbol{I}$ in a graph G is a subset of vertices such that for every pair of vertices v_1 and v_2 in $\boldsymbol{I}$, there is no edge between v_1 and v_2 in G. A maximum independent set in a graph refers to an independent set that includes the greatest possible number of vertices among all independent sets within that graph. A vertex cover $\boldsymbol{C}$ of a graph G is a subset of vertices such that for every edge $v_1 - v_2 \in \boldsymbol{E}$, we have one of the followings: $v_1 \in \boldsymbol{C}$, $v_2 \in \boldsymbol{C}$, or $\{v_1, v_2\} \subset \boldsymbol{C}$. For a graph G, $VertexCover(G)$ represents the set of all vertex cover sets with respect to graph G. Minimum vertex cover is a vertex cover that minimizes the number of vertices needed to cover all the edges in the graph (See Figure 1 for different vertex cover sets of an example UCCG). In other words, a minimum vertex cover is the minimum number of nodes that we can select from a graph in which for any edge between two nodes in that graph, at least one of the nodes is in the minimum vertex cover set.

2.3 Problem Definition

In this manuscript, we are looking to calculate the minimum number of interventions which is needed for recovering the orientation of all undirected edges in an essential graph, in the passive setting. With this number of interventions, we can discover all undirected edges in an essential graph, regardless of what the orientation of underlying ground truth DAG is. In addition, we are looking for an algorithm to compare different algorithms in the passive setting, in terms of a number of interventions that they need to recover all undirected edges for any possible underlying ground truth DAG.

After obtaining samples from a causal system, there are different methods to obtain the essential graph. There are constraint-based algorithms such as IC ([11]), PC and FCI

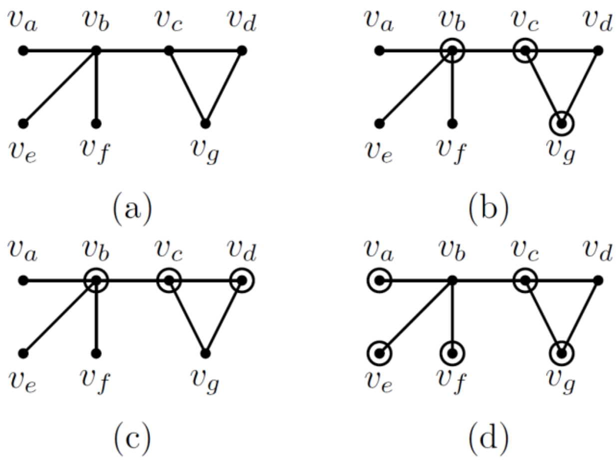


Fig. 1 Examples of different vertex cover sets: (a) a UCCG (G), (b) A minimum vertex cover set of G, (c) Another minimum vertex cover set of G, (d) A vertex cover set of G. Nodes in the circles represent the nodes in the respective vertex cover sets.

([14]), or a score-based methods ([10], [1]). Chordal chain components of the essential graph can be obtained by removing all directed edges in the essential graphs. It has been proved that the orientations of edges within one chain component do not yield any insights into the orientations of edges in other chain components ([5]). Based on this, [6] has shown that intervening on the chain component is equivalent to intervening on the essential graph itself. Therefore, we can just focus on undirected chordal chain graphs (UCCGs).

3 Proposed Approach

In this section, we propose a method to calculate the minimum number of interventions which is needed to recover all the undirected edges in a UCCG in the passive setting. We prove that we need at least this number of interventions to recover all the edges, regardless of what the ground truth DAG orientation is. We provide a theorem that enables us to find the set of nodes for full recovery of orientations and we prove that the proposed set is optimal. This theorem is stated in subsection 3.1. In addition, we introduce an algorithm for identifying the optimal set of nodes. Taking advantage of the proposed theorem in subsection 3.1, we present another algorithm to compare different algorithms in terms of the number of interventions they must perform to recover all the undirected edges, without any knowledge of underlying ground truth DAG. We describe this algorithm in subsection 3.2.

In this paper, we assume the availability of an infinite number of samples from both interventional and observational distributions. This assumption enables the exact recovery of the essential graph corresponding to the true causal DAG.

3.1 Optimal interventions set

In this section, we focus on the passive setting and we target to recover all of the undirected edges in a UCCG. Here, the goal is to recover any plausible underlying ground truth DAG with the optimal intervention set in terms of the number of interventions. Proposition 1 states that any set of nodes, which are superset or equal to a vertex cover of a graph, can recover all the undirected edges in that graph.

Proposition 1. *Given a UCCG $G = (\boldsymbol{V}, \boldsymbol{E})$, and considering a passive setting for performing intervention, any set of nodes $\boldsymbol{C} \subseteq \boldsymbol{V}$ can recover all of the undirected edges in $\boldsymbol{E}$, if and only if $\boldsymbol{C}$ is a vertex cover of UCCG $\boldsymbol{G}$.*

Proof. For "if" part, we should prove if we can recover all the edges by a set of interventions like $\boldsymbol{C}$, then $\boldsymbol{C}$ is a vertex cover. We prove this by contradiction. By contradiction, we know $\boldsymbol{C}$ can recover all the edges, but it is not a vertex cover. As $\boldsymbol{C}$ is not a vertex cover, we have at least one edge like $v_k - v_l \in \boldsymbol{E}$ in which $v_k \notin C$ and $v_l \notin \boldsymbol{C}$. Now, assume that we have node v_l as root in the underlying ground truth DAG. In this case, for recovering edge $v_k - v_l$, we should intervene node v_k or v_l. This is because none of the Meek rules are applicable in this case. However, we know that $v_k \notin \boldsymbol{C}$ and $v_l \notin \boldsymbol{C}$. Thus, we cannot recover this edge, which is in contradiction with the assumption of recovering all the edges.

Now, we prove the "only if" statement by contradiction. By contradiction, after performing an intervention on every node in $\boldsymbol{C}$, we will have at least one unoriented edge like $v_i - v_j \notin \boldsymbol{E}$. This implies that neither the v_i nor the v_j are in $\boldsymbol{C}$. Otherwise, after performing an intervention on $\boldsymbol{C}$, the edge $v_i - v_j \in \boldsymbol{E}$ should be oriented. However, this violates the definition of the vertex cover and the proof is complete. ■

Proposition 1 states that intervention on one of the vertex covers of UCCG G recovers all the undirected edges in E, in the passive setting. Having this, the next theorem states the minimum number of interventions that we need to recover all the undirected edges, regardless of the underlying causal DAG. Theorem 1 leverages Proposition 1 to provide the optimal intervention set for full recovery in the passive setting.

Theorem 1 (Optimal intervention set). *Given a UCCG $G = (\boldsymbol{V}, \boldsymbol{E})$, the set of minimum number of interventions, which is needed for recovering all of the edges in $\boldsymbol{E}$, in a passive setting, denoted by $\boldsymbol{I}^*$, can be obtained as follows:*

$$\boldsymbol{I}^* \in \underset{\boldsymbol{C} \in VertexCover(G)}{\operatorname{argmin}} |\boldsymbol{C}|$$

Proof. We need to prove that performing intervention on all the nodes in $\boldsymbol{I}^*$, can recover any ground truth DAG in the Markov equivalence class of G. Based on the Proposition 1, this statement holds, as any selected $\boldsymbol{I}^*$ is a vertex cover. In addition, we must prove that $\boldsymbol{I}^*$ is the minimum number of interventions required. This means that for any other set like $\boldsymbol{I}_F$ that can recover any possible ground truth DAG in the Markov equivalence class of G by performing interventions on all the nodes in $\boldsymbol{I}_F$, we have $|\boldsymbol{I}^*| \leq |\boldsymbol{I}_F|$. Based on Proposition 1, any set like $\boldsymbol{I}_F$ that can recover all the DAGs in the Markov equivalence class of G must be a vertex cover of G. Here, $\boldsymbol{I}^*$ is a set with a minimum number of nodes that have been selected through all possible vertex cover on G. Thus, we have $|\boldsymbol{I}^*| \leq |\boldsymbol{I}_F|$. ∎

Theorem 1 states that one of the minimum vertex covers of a UCCG like G is enough for recovering all orientations in G. This theorem provides an exact lower bound on the number of interventions needed for recovering the orientations of all undirected edges in a UCCG in the passive setting. In addition to determining the number of interventions, this theorem also specifies the set of nodes for intervention. In this part, we want to propose an algorithm for obtaining the $\boldsymbol{I}^*$. In [3], the author proves that we can find the minimum vertex cover for a chordal graph in a polynomial time. As any UCCG graph is a chordal graph, we take advantage of this result and propose the algorithm based on this to calculate the minimum intervention set by means of using the minimum vertex cover. We propose Algorithm 1 to calculate one of the minimum intervention sets of a UCCG G. We call this algorithm as the "optimal" algorithm.

Algorithm 1 proposes a method to obtain the minimum vertex cover set in the UCCG G. In this Algorithm, in Line 4, we generate a DAG with the skeleton of UCCG G. Please note that this generated DAG, D, can be any of the DAGs inside the MEC of G. Note that any algorithm can be used for generating DAGs, for example a rooted partition method in [15]. We label every single node in D in Line 5, according to the following mechanism. The greatest label is assigned to a node without any child. Then, we remove this node and all the parents of this node. In the next step, again, we find the node without any child. This node would be the second greatest number. By iterating on this node selection and removal, the source node gets labeled "1" and the sink node gets labeled "$|\boldsymbol{V}|$". In Line 6, we initialize set $\boldsymbol{M}$ by an empty set. From Lines 7 to 12, we repetitively run some functions to find the minimum vertex cover. In particular, in line 8, we select the maximum label in the label set $\boldsymbol{O}$ and put it in the variable m. In Line 9, variable $\boldsymbol{J}$ is a set of all nodes which are neighbors of the node m. In Line 10, we remove the selected node m and the neighbor nodes from the set $\boldsymbol{O}$. In Line 11, we put all the selected nodes m in each iteration of the while loop in the variable $\boldsymbol{M}$. Please note

Algorithm 1 Intervention set for full recovery in passive setting ($MinVertexCover(G)$)

1: **Input:** $G = (\boldsymbol{V}, \boldsymbol{E})$, where G is a UCCG
2: **Output:** $\boldsymbol{I}^* = MinVertexCover(G)$
3: **Function** $MinVertexCover(G)$
4: $\quad D \leftarrow DAGGeneration(G)$
5: $\quad \boldsymbol{O} \leftarrow Label(D)$
6: $\quad \boldsymbol{M} \leftarrow \{\}$
7: $\quad$ **While** $\boldsymbol{O} \neq \{\}$ do
8: $\qquad m \leftarrow \max(\boldsymbol{O})$
9: $\qquad \boldsymbol{J} \leftarrow \{j \in \boldsymbol{O} \backslash \{m\} | v_j \in neigh(v_m)\}$
10: $\qquad \boldsymbol{O} \leftarrow \boldsymbol{O} \backslash (\{m\} \cup \boldsymbol{J})$
11: $\qquad \boldsymbol{M} \leftarrow \boldsymbol{M} \cup \{m\}$
12: $\quad$ **End While**
13: $\quad$ **return** $\boldsymbol{V} \backslash \boldsymbol{M}$
14: **End Function**

that set $\boldsymbol{M}$ is a maximum independent set and the complement of this set is the minimum vertex cover set. In Line 13, algorithm returns the minimum vertex cover set, which is the complement of the obtained set $\boldsymbol{M}$ during the running algorithm. Please refer [3] for the proof of Algorithm 1.

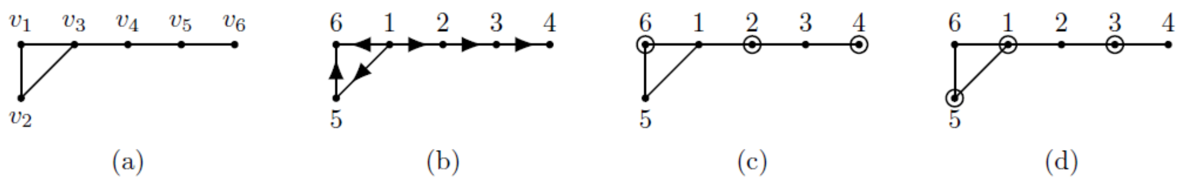
Example 1. *We provide an example in Figure 2 for finding the optimal intervention set using the proposed Algorithm 1. Given a UCCG in 2(a), we generate a DAG in 2(b). We show the maximum independent set of the graph G, denoted by $\boldsymbol{M}$, in 2(c). Figure 2(d) shows the minimum vertex cover set with circles around the nodes on UCCG G which are in the optimal intervention set. In Figure 2(e), We summarize the values that different variables get by running Algorithm 1.*

In this example, we obtained three nodes for the minimum vertex cover set, which means at least three nodes need to be intervened for full identification. In Figure 2(f), we depict all the possible cases for selecting two nodes from six nodes. Every graph is a representative of one of those two nodes. We provide an example DAG for each of those in which we are not able to recover all the undirected edges in G. Blue lines in Figure 2(f) are the edges that cannot be recovered with the two selected nodes for the intervention in that specific DAG.

In the next part, we propose an algorithm that gets an algorithm like A as input and returns the number of nodes that algorithm needs for full identification in the passive settings.

3.2 Algorithm Comparison

In this section, we want to devise an algorithm which enables us to evaluate the capability of different algorithms. First, we explain the main purpose of



Iteration 0	$m = null$	$\mathbf{J} = \{\}$	$\mathbf{O} = \{1,2,3,4,5,6\}$	$\mathbf{M} = \{\}$
Iteration 1	$m = 6$	$\mathbf{J} = \{1,5\}$	$\mathbf{O} = \{2,3,4\}$	$\mathbf{M} = \{6\}$
Iteration 2	$m = 4$	$\mathbf{J} = \{3\}$	$\mathbf{O} = \{2\}$	$\mathbf{M} = \{6,4\}$
Iteration 3	$m = 2$	$\mathbf{J} = \{\}$	$\mathbf{O} = \{\}$	$\mathbf{M} = \{2,4,6\}$

(e)

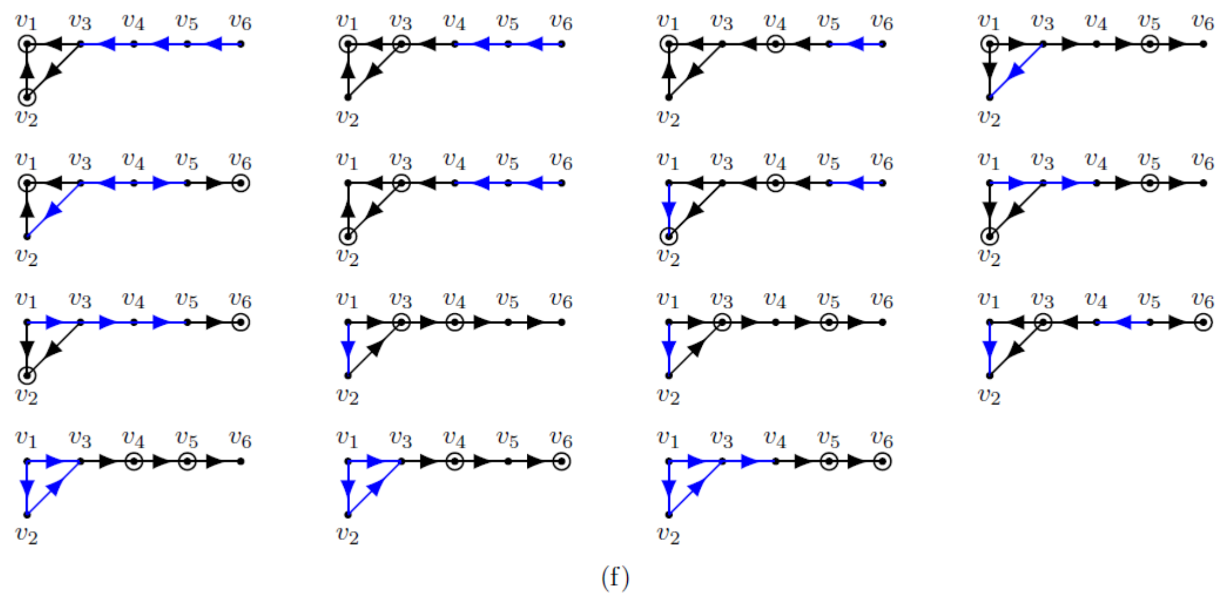


Fig. 2 *Example of computing minimum vertex cover using Algorithm 1: (a) A UCCG ($\boldsymbol{G}$). (b) A DAG based on the MEC of $\boldsymbol{G}$. (c) maximum independent set of $\boldsymbol{G}$. (d) minimum vertex cover set of $\boldsymbol{G}$. (e) output of each iteration for running the algorithm. (f) set of all possible selections of two nodes for intervention, and one example DAG that we cannot identify completely with those two interventions.*

evaluation. Consider an algorithm like A in the passive experiment setting. After running algorithm A on a given graph G, we obtain a set of nodes for intervention, called $\boldsymbol{T}$. There is no information on what the underlying ground truth DAG is. We perform single interventions based on the order in set $\boldsymbol{T}$. For each DAG in the MEC of G, we need to perform different number of interventions in $\boldsymbol{T}$ to recover orientations of all undirected edges in graph G. We want to know how many intervention steps we should take from algorithm A to guarantee recovering all orientations in G, regardless of underlying causal DAG orientation. In the following, we propose Lemma 1 which selects such a minimum set from set $\boldsymbol{T}$. We called this minimum set as sufficient intervention set, which is the subset of the proposed set of intervention nodes by algorithm A.

Lemma 1 (Sufficient intervention set). *Consider any algorithm A that performs interventions in the passive setting on a given UCCG G, recovering all undirected edges in G. Then, the intervened nodes selected by the algorithm A is a vertex cover set of UCCG G.*

Proof. *Based on Proposition 1, we can recover any possible underlying causal DAG if the intervention set is a vertex cover.* ∎

The minimal set of interventions from the output of algorithm A, which is a vertex cover, is called as a sufficient intervention set. We propose Algorithm 2 which selects the sufficient intervention set from the suggested intervention nodes by a passive algorithm.

In Algorithm 2, $|sufficient(A, G)|$ shows the number of interventions for full recovery for algorithm A on graph

Algorithm 2 Assessment of performance of an algorithm in the worst case

1: **Input:** Algorithm A
2: **Output:** $Sufficient(A, G)$
3: **Function** $Sufficient(A, G)$
4: $\boldsymbol{S} \leftarrow null$
7: **While** $\boldsymbol{S} \notin VertexCover(G)$ do
8: $\boldsymbol{I} \leftarrow$ Run one step of algorithm A
9: $\boldsymbol{S} \leftarrow \boldsymbol{S} \cup \boldsymbol{I}$
12: **End While**
13: **return** $\boldsymbol{S}$
14: **End Function**

G. The better algorithm shows a lower value on the same graph G. Here, we explain Algorithm 2. In Line 4, we initialize set $\boldsymbol{S}$, which is a sufficient intervention set to fully recover $\boldsymbol{G}$. In Lines 5 to 8, we run the steps of Algorithm A, and we add the obtained nodes from Algorithm A in each step to the sufficient intervention set $\boldsymbol{S}$. We terminate running Algorithm A, whenever $\boldsymbol{S}$ is a vertex cover of graph G.

4 Experiment

In this section, we perform several experiments that compare the optimum intervention set in the passive settings based on our algorithm with two naive approaches. The ratio of the number of edges in the graph to the maximum possible number of edges for a certain size defines what we refer to as "edge density."

4.1 Chordal graph generation

In our experiment, we generated chordal graphs using a method described in [9]. For each point in our plot, we collected data from 150 instances of chordal graphs.

4.2 Comparison

In this part, we compared the optimal intervention set with intervention sets provided by two naive algorithms. In our experiment, each algorithm selects the intervention set first and then performs an intervention on every single node inside the intervention set. In particular, we compared our algorithm with the "Random" and "MaxDegree" algorithms. The "Random" algorithm chooses an intervention set through random node selection, and the "MaxDegree" algorithm selects intervention nodes based on node degrees. Nodes with the highest degree will be the first nodes that are going to be intervened. For both of these algorithms, we select an intervention set that has the same number of nodes as the graph. However, for the optimal intervention set, we are selecting the nodes for the intervention based on Algorithm 1.

We compared our optimal proposed algorithm, with two baselines, in terms of the average number of interventions, which is shown in Figure 3. The average number of interventions for these three methods is given in Figure 3. The plots show the average number of interventions that we need to recover all possible orientations in the graph, in the passive setting.

The left column in Figure 3 shows plots based on the different edge densities where we generated the chordal graphs with the number of nodes in the set $\{45, 65, 85, 105\}$. Experiments in the left column show that increasing the edge density will increase the average number of interventions that we need. In addition, we see that increasing the number of nodes increases the average number of needed interventions when the edge density is fixed. This is somehow expected because a dense graph needs more intervention to be recovered. The right column in Figure 3 shows different curves based on the different number of nodes. In each of these figures, we investigated the effect of increasing edge density when we have a fixed number of nodes in the graph. For achieving full recovery, increasing the edge density leads to a higher average number of interventions required.

5 Conclusion

In this paper, we focused on the computing optimal intervention set, in the passive experiment design problem which appears in many causal discovery tasks. We showed that any underlying causal DAG can be recovered with a specific intervention set in the passive setting if and only if the intervention set is a vertex cover for the graph. We considered just chordal graphs in our investigation. In addition, we provided an algorithm that enables us to evaluate different algorithms in terms of their capability to select a minimum number of interventions for recovering all the undirected edges in the graph, regardless of the orientation of the underlying ground truth DAG. Investigation of the same problem in the active setting is an interesting future research direction.

References

[1] D. M. Chickering, “Optimal structure identification with greedy search.”, Journal of machine learning research, vol 3, pp. 507–554, 2002.

[2] F. Eberhardt, C. Glymour, and R. Scheines, “On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables”, Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI), pp. 178–184, 2005.

[3] F. Gavril, “Algorithms for Minimum Coloring, Maximum Clique, Minimum Covering by Cliques, and Maximum Independent Set of a Chordal Graph”, SIAM Journal on Computing 1.2, pp. 180–187, 1972.

[4] A. Ghassami, S. Salehkaleybar, N. Kiyavash, and E. Bareinboim, “Budgeted experiment design for

causal structure learning", International Conference on Machine Learning, pp. 1724–1733, 2018.
[5] A. Hauser and P. Buhlmann, "Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs". Journal of Machine Learning Research 13(1), pp. 2409–2464, Aug. 2012.
[6] A. Hauser and P. Buhlmann, "Two optimal strategies for active learning of causal models from interventional data", International Journal of Approximate Reasoning 55(4), pp. 926–939, June 2014.
[7] Y. He and Z. Geng, "Active learning of causal networks with intervention experiments and optimal designs", Journal of Machine Learning Research 9, pp. 2523–2547, Nov. 2008.
[8] M. Kocaoglu, A. Dimakis, and S. Vishwanath, "Cost-optimal learning of causal graphs", International Conference on Machine Learning, 2017.
[9] L. Markenzon, O. Vernet, and L. Henrique Araujo. "Two Methods for the Generation of Chordal Graphs", Annals of Operations Research. 157, pp. 47–60, 2008.
[10] C. Meek, "Causal inference and causal explanation with background knowledge", Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 403–410, 1995.
[11] J. Pearl, Causality. Cambridge university press, 2009.
[12] J. Peters, D. Janzing, and B. Scholkopf, "Elements of causal inference: foundations and learning algorithms", MIT Press, 2017.
[13] R. Safaeian, S. Salehkaleybar, and M. Tabandeh, "Fast causal orientation learning in directed acyclic graphs", International Journal of Approximate Reasoning 153, pp. 49–86, 2023.
[14] P. Spirtes, C. Glymour, and R. Scheines. Causation, prediction, and search. MIT press, 2000.
[15] Y. He, J. Jia, Bin Yu. Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs, The Journal of Machine Learning Research, pp 2589–2609, 2015.

Ramin Safaeian received the M.S degree in Digital Electronic Engineering from Sharif University of Technology, Iran, in 2012, and B.Sc. degree in Electrical and Electronic Engineering from Zanjan university in 2010. He is currently Ph.D. student in Digital Electronic Engineering at Sharif University of Technology. His research area is causality and machine learning.

Mahmoud Tabandeh received the Ph.D. degree in Computers and Electronics from the university of California, Berkeley, in 1974, and M.Sc. degree in Control Systems from Louisiana State University, in 1969, and B.Sc. degree in Electronics from Institute National des Sciences Appliquees, in 1967. He is currently working as an associate Professor in the Department of Electrical and Electronic Engineering, Sharif University of Technology, Tehran, Iran.

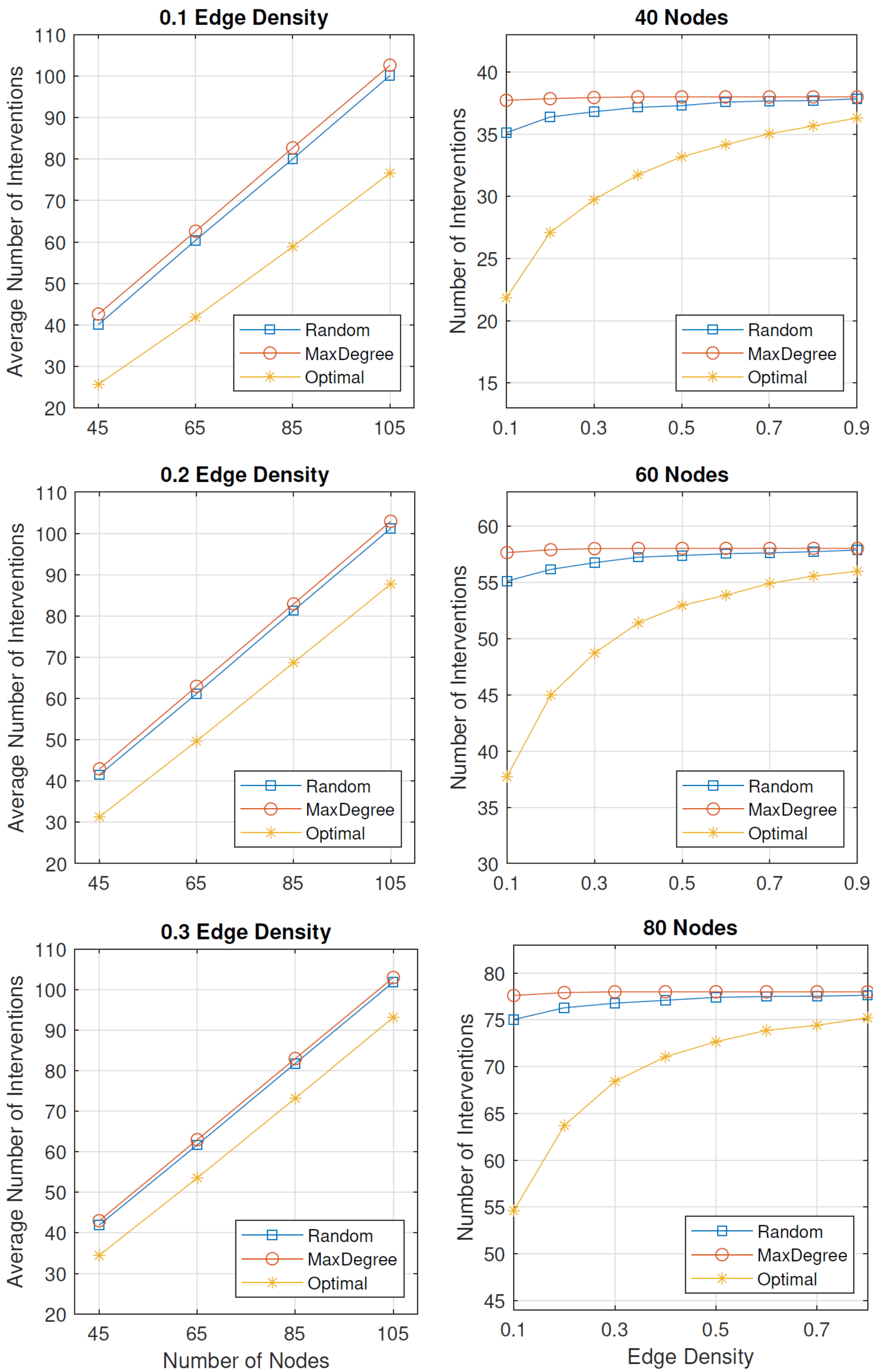


Fig. 3 Comparison of the average number of interventions which is needed for full recovery between Random, MaxDegree and Optimal algorithms. Left column shows the graphs with different edge densities, while the right column corresponds to graphs with different numbers of nodes.