On Generalized Minimum Intervention Covers

Sanjeev Tewani (st3186) Nihaar Shah (ns3413)

May 15, 2020

1 Introduction

We make headway into the problem of generalizing minimum intervention covers to the more realistic setting where there are constraints on the experimental distributions available. For illustration, consider a drug study involving a candidate which is highly interactive with other therapies. In any clinical study, patients may already be taking a range of medications, and a researcher may find it best to model all the drugs taken by a patient as an intervention, rather than as merely observational variables. If a total of 30 drugs are taken across all test subjects, the total experimental space will clearly exceed the set of trials any company would be willing to undertake. To add to the problem, in such a setting optimal therapies may involve combinations of drugs, such that a researcher's target may be to identify very many queries and not just one.

In recent years the literature has made significant progress on two sides of this problem: first, the existence of a sound and complete algorithm for general identifiability of individual queries has been demonstrated [Lee and Bareinboim, 2019]; and second, an efficient algorithm for identifying the minimum intervention cover has been found [Kandasamy, 2019]. Each of these proposes a unique graphical condition which facilitates the problem at hand. We make two primary contributions in this paper which show that the two tasks are tightly-linked. First, leveraging completeness results already proved in the literature we show that one can efficiently identify a minimum intervention cover using a restricted set of experimental queries. Second, we prove that graphical structures which demonstrate that a particular query is not g-identifiable (the *thicket*) imply the existence of a structure (the *bush*) which can be used to efficiently enumerate required distributions for g-identifying said query. As such, we provide one solution to inverse problem of g-identifiability.

The format of this paper is as follows: in the first section, we lay out all of the prerequisite definitions identically as they appear in [Shpitser and Pearl, 2008], [Kandasamy, 2019], and [Lee and Bareinboim, 2019]. Familiar readers should feel free to skip this section. In the second section, we define the problem and offer examples which illustrate our setting. In the third section we prove our two results, and in the fourth section we compute the resulting algorithm and discuss its runtime. The fifth section concludes.

2 Preliminaries

As is traditional in the literature, we are concerned with semi-Markov Structural Causal Models represented as DAGs $\mathcal{G}$ over a set of observable variables V. Furthermore, nodes in V are endowed with the usual kinship operations pa, an, ch, de and their closures Pa, An, Ch, De. The following graphical structures will be of use here:

C-forest. A semi-Markovian graph $\mathcal{G}$ with root set R is said to be an R-rooted c-forest if $\mathcal{G}$ is a c-component with a minimal number of edges.

The following, originally appearing in [Shpitser and Pearl, 2008], is of importance in classical identification:

Hedge A hedge is a pair of R-rooted c-forests $\langle F, F' \rangle$ such that $F' \subseteq F$.

[Lee and Bareinboim, 2019] builds upon the graphical structures above by proposing two new structures specifically adapted to the generalized problem.

Hedgelet decomposition The hedgelet decomposition of a hedge $\langle F, F' \rangle$ into hedgelets $\{F(W)\}_{W \in C(F'')}$ where each hedgelet $F(W)$ is a subgraph of F made of (i) $F[V(F') \cup W]$ and (ii) $F[De(W)_F]$ without bidirected edges.

Thicket Let R be a non-empty set of variables and $\mathbf{Z}$ be a collection of sets of variables in $\mathcal{G}$. A thicket $\mathcal{T} \subseteq \mathcal{G}$ is an R-rooted c-component consisting of a minimal c-component over R and hedges $F_{\mathcal{T}} = \{\langle F_Z, \mathcal{T}[R] \rangle | F_Z \subseteq \mathcal{G} \setminus Z, Z \cap R = \emptyset\}_{Z \in \mathbf{Z}}$. Let X, Y be disjoint sets of variables in $\mathcal{G}$. A thicket $\mathcal{T}$ is said to be formed for $P_x(y)$ in $\mathcal{G}$ with respect to $\mathbf{Z}$ if $R \subseteq An(Y)G_{\underline{X}}$ and every hedgelet of each hedge $\langle F_Z, \mathcal{T}[R] \rangle$ intersects with X.

Thickets are complete for the task of general identification. Formally, from [Lee and Bareinboim, 2019]

g-Identifiability for X, Y disjoint sets of variables, $\mathbf{Z} = \{Z\}_{i=1}^{m}$ a collection of sets of variables, and $\mathcal{G}$ a causal diagram, $P_x(y)$ is said to be g-identifiable from $\mathbf{Z}$ in $\mathcal{G}$ if $P_x(y)$ is uniquely computable from distributions $\{P(V|do(z))\}_{Z \in \mathbf{Z}, z \in \mathcal{X}_z}$ in any causal model which induces $\mathcal{G}$. Here, $\mathcal{X}_z$ is the domain of the variable Z and the domain of all variables is $\mathcal{X}$.

By complete, we mean a query $P_x(y)$ is *not* identifiable in $\mathcal{G}$ if and only if one can find a thicket which is formed for $P_x(y)$ with respect to the set $\mathbf{Z}$.

A heretofore unrelated structure emerges in the construction of a minimum intervention cover, which formally is

Minimum intervention cover a set of interventional distributions (an information *set*) for a causal graph $\mathcal{G}$, such that it (i) identifies the set $\cup_{S \subseteq V}\{P_s(V \setminus S|do(s))\}$ (it is an intervention *cover*), and (ii) there exists no intervention cover of G of smaller size.

Roughly speaking, the structure factors a maximal c-component for a subset of $\mathcal{G}$ by isolating the confounded parents of the c-component from the unconfounded parents:

Bush For a given causal graph $\mathcal{G}$, let $\mathbf{A}$ and $\mathbf{B}$ be disjoint subsets of the observable variables $\mathbf{V}$. Then **A, B** form a bush in $\mathcal{G}$, if

$$
\begin{aligned}
&(i)|B| \neq 0 \\
&(ii)B \in C(V \setminus Pa(B)) \\
&(iii)\forall A_i \in A, A_i \in Pa(B) \text{ and } C(\{A_i\} \cup B) = \{\{Ai\} \cup B\} \\
&(iv)\forall P_i \in Pa(B) \setminus A, C(\{P_i\} \cup B) = \{\{Pi\}, B\}.
\end{aligned}
$$

Graphically, this means that the set B is not empty, B is a maximal C-component in the graph $\mathcal{G}[V \setminus Pa(B)]$, each element in A is a parent of the set B and shares a bidirected edge with B, and each element of $Pa(B) \setminus A$ does not share a bidirected edge with B. By construction, no other node in the graph can share a directed or bidirected edge with B.

[Kandasamy, 2019] prove that from every bush, one can construct an *Informative Intervention Set* denoted. The authors show that the set of all queries in a graph or component is identifiable by an interventional set $\mathbf{Z}$ if and only $\mathbf{Z}$ contains at least one intervention from each $IIS((A, B))$.

IIS Given a causal graph $\mathcal{G}$ and a bush and assignment pair $((A, B), do(pa(B) = b))$ of $\mathcal{G}$, the informative intervention set $IIS(do(A = a), do(pa(B)\ a = b), B)$, is a joint information set that contains the set of all interventions I such that 1) I intervenes on all the vertices of A with

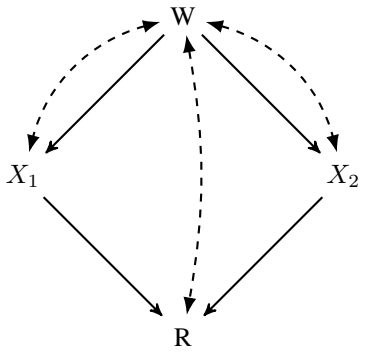

Figure 1: A graph containing two hedgelets.

the assignment a; 2) I does not intervene on any vertex in B; and 3) I and $do(pa(B) = b)$ are consistent on $pa(B)$. When no ambiguity arises (the specific assignments are left general), we refer to the set of interventions as $IIS((A, B))$

From the definitions above, it is not immediately apparent that there is any meaningful relationship between the bush and the graphical structures defined elsewhere in the identification literature. In particular, hedges and thickets refer to shared rootset and are formed from minimal c-components, while bushes nest according to the parent set A and are maximal c-components. However, in a sense focusing on the rootset and causal paths is critical for understanding what can be *identified*; focusing on the parent set aids in understanding what the parent set is able to *identify*.

An example of this distinction is illustrated below: As elaborated in [Lee and Bareinboim, 2019], this graph 1 is not a hedge, but it contains two hedges which are themselves hedgelets. It is also a thicket with respect to $\{\{X_1\}, \{X_2\}\}$ for the query $P_{X1,X2}(R)$. By completeness, there exists no way to write said query in terms of only interventions on X_1 or X_2.

We can read this same fact off of the bush factorization of the c-component. Note that the graph contains 5 bushes: $(\{\}, \{W, X_1, X_2, R\})$, $(\{W\}, \{X_1, X_2, R\})$, $(\{W\}, \{X_1\})$, $(\{W\}, \{X_2\})$, and $(\{W\}, \{R\})$. The existence of these bushes correspond to the determination that only the distributions $P(X_1, X_2, R|do(W))$ and $P(X_1, X_2, R, W)$ are required to identify all the possible interventional distributions, and in fact that no other distributions are helpful for identifying all underlying queries. With this, it is plain to see by Rule 2 of do-calculus [Shpitser and Pearl, 2008] that $P_{X_1,X_2,W}(R) = P_W(R|X_1, X_2))$ and that the P_W distribution is sufficient for the task at hand.

A very important feature of bushes is the following, initially appearing as Lemma 5 in [Kandasamy, 2019]:

Lemma 5 Let A', B' form a bush for $\mathcal{G}$. Then for all $A \subset A'$, there is a bush A, B for $\mathcal{G}$ such that $B \supseteq B' \cup (A' \setminus A)$

Furthermore, one can prove that none of the experimental distributions required to answer all queries in one bush can be used to answer all queries in a nested bush, which illustrates the complexity of the problem at hand when one is attempting to produce a minimum intervention with limited experiments available.

3 Discussion

As such, we seek to bridge these converging paths in the literature. This line of research is started in the appendix of [Kandasamy, 2019], which demonstrates that if a bush (A, B) exists then there is no hedge formed for $P_A(B)$. To the best of our knowledge, no relationship has

been shown between bushes and hedges' more complicated generalization, the thicket. We show first that there exists a constructive method to identify a bush from a thicket.

Additionally, we seek to find a minimum intervention cover in a setting where only a restricted set of experimental data is attainable. That is, for a causal graph $\mathcal{G}$ and a set of experimental distributions $\{P(V|do(z))\}_{Z\in\mathbf{Z},z\in\mathcal{X}_z}$, we seek to fully identify the minimum subset required to identify $\cup_{S\subseteq V}\{P_s(V \setminus S|do(s))\}$, or to discover when such a task is not doable.

With respect to our first result, we note that the contributions of this are three-fold:

First, and most critically, this demonstrates that, given a thicket, we can construct an interventional distribution separate from the distributions the thicket is formed for, and we can enumerate the sets which are viable for general identification of the associated bushes. This completely characterizes the set of experiments required to answer all related questions. In a sense, we have solved one important inverse problem of gID.

Secondly, there are structures which are generalizations of the thicket, such as the S-thicket [Lee and Bareinboim], for which there are no analogous bush-like structures. Further research in this direction is warranted to establish whether or not such inverse solutions can be obtained in even more general settings. As many of these more general settings are also more realistic causal data science settings, it may be of significant interest to identify analogues to the bush in these settings.

Lastly, a constructive solution also offers the potential for a more efficient algorithm for determining the presence of thickets in a c-component, by significantly reducing the search space required for identification. For one, since every thicket implies a bush, one can enumerate all bushes (efficiently, by the algorithm in [Kandasamy, 2019]) and observe that if a thicket exists, both X and Y are contained within the bottom of the bush. Although the task of gID is $O(mn^4)$ for a single query, adaptations to the algorithm below (which do not terminate when a required distribution is missing) may be able to more efficiently identify all thickets. However, this remains to be seen.

Note that for the purposes of the second result of this paper, we do not actually require the above, since [Kandasamy, 2019] demonstrate that bushes are complete for the task of finding a minimum intervention cover. Marrying this with the completeness of g-ID, we know that therefore there are no thickets if we have a distribution from each informative intervention set. However, the completeness results do not guarantee that there is still an efficient algorithmic construction of a minimum intervention cover, which we show is still possible.

For the second task, we note first that by the existence of an algorithm for the gID of individual queries, a brute force algorithm for the question at hand is available. The algorithm takes $O(mn^4)$ and, since there are some $|\mathbf{Z}|2^n$ joint queries of interest (every variable can be the target of a query or not, and distributions must be chosen from $\mathbf{Z}$), the brute force algorithm operates in $O(mn^4|\mathbf{Z}|2^n)$ time. See [Lee and Bareinboim, 2019] for a proof of the runtime for a single query of interest.

Faster brute force algorithms $O(mn^5|\mathbf{Z}|)$ would exist if one only desires to identify if a cover is possible, since one only need to check that singletons are identifiable. However, as proved in [Kandasamy, 2019], an efficient $construction$ of a minimum intervention cover with no restrictions is exponential only in the size of the degree of vertices in the graph and the size of c-components, offering a considerable savings when compared to the brute-force algorithm.

4 Analysis

From the 4-node above example, it would seem reasonable to assert that, given a thicket, we may be able to construct a bush using the sets X, Y, R, and Z. We see that in some tricky examples, the set Z may not even be part of the same c-component, and therefore it is not immediately

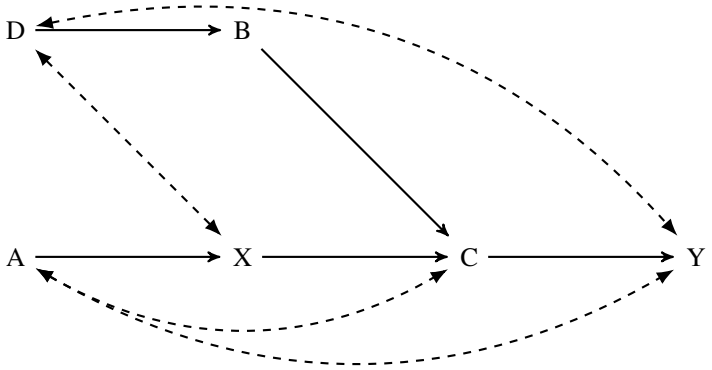

Figure 2: Thicket formed for $P_X(Y)$ with respect to B.

obvious that one can posit such a relationship. Concretely, consider the thicket formed for $P_X(Y)$ with respect to B in fig 2, which originally appears in the appendix of [Lee and Bareinboim, 2019]:

In this example, the nodes $\{A, X, C\}$ form the top of a hedge and D, Y are the root set comprising the bottom of the hedge. The node B is absent from the c-component entirely, and hence any bush structure that could be formed for this graph. The top of the hedge A, X, C, cannot be formed into the top of a bush, for example: C does not share a bidirected edge with the set $\{D, Y\}$ and X is not a parent of any of $\{D, Y\}$. Indeed, the sole non-degenerate bush for the graph is given as $(\{A\}, \{D, X, Y\})$ (its nested child with top $\{\}$ is *degenerate* because the top is the empty set). Apart from the observational distribution, interventions on $\{A\}$, $\{A, B\}$, $\{A, C\}$, and $\{A, B, C\}$ are all informative for the problem of general identification of all queries associated with these variables. None of these sets were readily apparent from the sets identified in the thicket; however, we see that the set X is not in the top of the bush, and the sets $Z \in \mathbf{Z}$ needn't even be part of the same c-component.

With this insight, we can now prove that given any thicket in a graph $\mathcal{G}$, we can construct a bush which validates the existence of the thicket:

Theorem For any thicket in $\mathcal{G}$, there exists a bush which affirms the existence of the thicket. *Proof* Concretely, assume that in the graph $\mathcal{G}$ there exists an R-rooted thicket formed for the distribution $P_X(Y)$ with respect to the available distributions Z. Let B be the maximum c-component of the graph $\mathcal{G}[V \setminus Pa(X \cup Y)]$ and let $A = (Pa(B) \cap G \setminus B)$. Then we have that no other node not in A shares a bidirected edge with B. Split A into the nodes A' which have a bidirected edge with B and A'' which do not. Certainly B is non-empty. By construction A only contains parents of B, and we admit to A' only those with a bidrected edge to B. B is then the maximal c-component in the graph $\mathcal{G}[V \setminus A']$, because the only excluded variables (those in A'') do not share a bidirected edge with B by construction. Therefore, (A', B) is a bush for the graph $\mathcal{G}$. Additionally, we can state the following:

Corollary For any thicket in $\mathcal{G}$, we can construct the set of interventions required for general identification of the associated C-component. *Proof* by Claim 2 and Theorem 4 of [Kandasamy, 2019], we know that all distributions associated with the bush are identifiable by a set of distributions $\mathbf{Z}$ if and only if $\mathbf{Z}$ contains an interventional distribution from the Informative Intervention Set. The informative intervention set can be read off from the constructed bush and the graph $\mathcal{G}$ as per [Kandasamy, 2019], or one of it's nested bushes computed according to Lemma 5 of [Kandasamy, 2019].

Lastly, this result also gives us an efficient and straightforward check for whether or not a General Minimum Intervention Cover is possible, since the algorithm will merely halt if this isn't possible.

This dramatically simplifies the computation required by the brute force approach.

5 Algorithm

We propose an algorithm for computing an actual MIC following the methodology laid out in [Kandasamy, 2019] (i.e. constructing a node out of each bush in a constructed alternative graph $\hat{\mathcal{G}}$, and finding a minimum vertex coloring). Our algorithm proceeds almost identically, with modifications only made to the definition of a conflict edge. For concreteness, we will re-state the algorithm without proof as it is defined in [Kandasamy, 2019], and only add proofs where necessary. We show that the worst case scenario doesn't change when the set of interventional distributions is restricted, because if a required distribution is not available the algorithm merely fails.

From the graph $\mathcal{G}$, construct an undirected graph $\hat{\mathcal{G}}$ as follows: for each bush and assignment pair $((A_i, B_i), do(pa(B_i) = b))$, associate a vertex in $\hat{\mathcal{G}}$. An edge exists between two vertices W_i and W_j if and only if there exists no intervention involving interventions in $\mathbf{Z}$ which identifies both $P(B_i|do(pa(B_i)) = b)$ and $P(B_j|do(pa(B_j)) = b')$. These edges are called *conflict edges*, and are defined below:

Conflict Edge There exists a conflict edge between two distinct vertices $W_i = ((A_i, B_i), do(pa(b_i)))$ and $W_j = ((A_i, B_i), do(pa(b_i)))$ if one of the following conditions holds:

1. $A_i \cap B_j$ is non-empty
2. $A_j \cap B_i$ is non-empty
3. $do(a_i)$ and $do(pa(b_j))$ are inconsistent
4. $do(a_j)$ and $do(pa(b_i))$ are inconsistent
5. $\nexists Z \in \mathbf{Z}$ such that
   $A_i, A_j \subseteq Z$
   $Z \cap B_i = Z \cap B_j = \emptyset$,
   $Z \setminus Pa(B_i) \setminus (B_i) \neq \emptyset, Z \setminus Pa(B_j) \setminus (B_j) \neq \emptyset$

Consistent here means that two assignments $\mathbf{x}, \mathbf{y}$ of $\mathbf{X}, \mathbf{Y}$ agree on all of the vertices in $\mathbf{X} \cap \mathbf{Y}$. Thus, inconsistent means that the assignments do not agree on *all* such vertices.
In [Kandasamy, 2019], a conflict edge exists whenever the sets in question are overlapping and the variable assignments are consistent. In our setting, a conflict edge will still emerge in all of these settings; however, there will also be a conflict edge if there does not exist a $Z \in \mathbf{Z}$ which can identify both bushes. Note that for two overlapping bushes (A', B') and (A, B) such that $A' \subset A$, either $Pa(B) \supset Pa(B')$ or $B' \cap A \setminus A \neq \emptyset$. If the first condition is true, then any Z which identifies (A, B) will also identify (A', B') by Theorem 4 of [Kandasamy, 2019], and if the second condition is true, it will be caught as a conflict edge by construction.

Therefore, the primary difference between our setting and that considered in [Kandasamy, 2019] is that for non-overlapping bushes, we should not assume that an experimental distribution Z which identifies both is available. Such a Z is available if Z contains both sets A, A' and Z is contained within both sets $pa(B), pa(B')$, hence the last condition of our definition of a conflict edge.

Let $C = \{W_{c_i}\}_{i=1}^{k}$ be any minimum vertex coloring of the graph $\hat{\mathcal{G}}$. Then the minimum vertex coloring is directly related to the minimum intervention cover, as spelled out in [Kandasamy, 2019].

[Kandasamy, 2019] point out that for a graph with bounded c-component size p and bounded vertex degree d, the degree of any fixed node in the graph $\hat{\mathcal{G}}$ cannot exceed $2^{2p(1+d^2)}|\mathcal{X}|^{2pd^3}$. Note that the fifth condition for defining a conflict edge above is only relevant when the other four are not met; furthermore, the derivation of this bound is a worst case scenario which includes

the case that no bushes are able to share interventional sets. This occurs when, for example, the bidirected arrows in the c-component in question form a complete graph. Thus the worst-case scenario as stated already subsumes our additional constraint, and we see that the bound on the cardinality of the minimum intervention cover is the same.

The only remaining difference is in the construction of the actual distributions in the MIC; [Kandasamy, 2019] choose the smallest set which satisfies the required conditions, but we must track for each node in $\hat{\mathcal{G}}$ the available Zs. This is itself not a simple procedure, and the best way to do this is the following: for every pair of nodes in $\hat{\mathcal{G}}$ which satisfy the first four conditions but not the last, maintain a set of distributions in $\mathbf{Z}$ which are in the Informative Information Set and are consistent with the node. After constructing the minimum intervention cover, for every color c find the hitting set of sets in the color c. The hitting set is minimal by definition, hence, our algorithm returns a minimal result.

While this is an NP-complete problem in general, in fact we can appeal yet again to the structure of the bush to compute this efficiently. The efficiency of this step is outside of the scope of this paper, but we suspect it offers considerable gains.

Note again that the condition 5 can only be activated after conditions 1-4 pass. The first four conditions being satisfied imply that the tops of two bushes A and A' overlap, and they share the same set $Pa(B) = Pa(B')$. Note that by Lemma 5 of [Kandasamy, 2019], if $|A| \neq |A'| - 1$, then there exists another bush with $A' \subset A'' \subset A$ which can also be identified by the same available interventional distributions. As such, and as alluded to by the algorithm to compute all bushes, the set of all bushes for a c-component form a tree structure, indexed by the sets A and their assignments. If we associate with each edge in the tree the set of interventions which identify both bushes, we need only compute the hitting set over edges in the tree. This is a total of $\sum_{i=0}^{p} |\mathcal{X}|^i \binom{k}{i} - 1 = (|\mathcal{X}|+1)^p - 1$ total edges when the size of all c-components is bounded by p. By Lemma 7 of [Kandasamy, 2019], operating with a restricted set will at worst double the lower bound of the MIC for a each c-component, but does not impact the upper bound on the MIC.

Algorithm 1 MIC(G)-original

IB $\leftarrow$ FINDALLBUSHES(G)
$\mathbf{W} \leftarrow$ VERTEXCONSTRUCTION(IB,G)
Let $\mathbf{W}$ be the vertex set of G^{CO}
Construct edges of G^{CO}
Find a minimum vertex coloring of G^{CO}
Let $\mathbf{W}_c$ be the vertices colored using color c
For each color c, let $I_c = Pr[\mathbf{V} \setminus \mathbf{A}_c | do(\mathbf{a_c}]$
where $\mathbf{A_c} = (\bigcup_{(\mathbf{A},\mathbf{B},pa(b)\in W_c)} A)$
and $\mathbf{a_c} = (\bigcup_{((\mathbf{A},\mathbf{B}),\mathbf{pa(b)})\in \mathbf{W_c}} \mathbf{a})$
return $\bigcup_{\mathbf{c}} \mathbf{I_c}$

Algorithm 2 VertexConstruction(IB,G)-modified

Result: Pairs P
Initialize pairs $P = \{\}$
for $(\mathbf{A}, \mathbf{B}) \in IB$ **do**
  for $\mathbf{pa}(\mathbf{B}) \in \Sigma^{|\mathbf{Pa(B)}|}$ **do**
    If $\nexists Z \in \mathbf{Z}$ such that $Z \in IIS((A, B))$, FAIL
    Add a new vertex $((\mathbf{A}, \mathbf{B}), \mathbf{pa}(\mathbf{B}))$ to P
  end
end
return P

Algorithm 3 MIC-modified

Result: $\cup_c Z_c$
IB $\leftarrow$ FINDALLBUSHES(G)
W $\leftarrow$ VERTEXCONSTRUCTION(IB,G)
Let W be the vertex set of G_{CO}
InformationSets ={}
Construct the edges of G_{CO} using the Definition of a Conflict Edge
Find a minimum vertex coloring of G_{CO}.
Let W_c be the vertices colored using color c.
Build a tree over W_c where for W, W' in W_c, W' is a child of W if A' contains every element of A except 1
for *each color c* **do**
    $Z_c = \{\}$
    Traverse W_c as such:
    For every leaf node $W' \in W_c$, if $IIS(W') \cap IIS(W)$ is non-empty, then add $IIS(W') \cap IIS(W)$ to Z_c and delete all nodes below W_c. If not, add $IIS(W')$ to Z_c and delete the leaf node, repeating until all nodes are removed. Lastly, compute the hitting set of Z_c
end
Return $\cup_c Z_c$

6 Conclusion

We demonstrate a link between two graphical structures specifically formed for the task of g-Identifiability in causal graphs. One structure, the thicket, is formed for individual queries, while another is formed for identifying all queries in a cleverly-partitioned c-component. We show that because a thicket implies a bush, one can solve the "inverse" problem of g-Identifiability as it appears in [Lee and Bareinboim, 2019] directly by construction. This suggests algorithmic solutions to problems often faced by researchers, involving optimally choosing what experiments must be conducted next. We leave as an open avenue of research whether or not these results can be expanded to more challenging – and realistic – settings, such as g-transportability [Lee and Bareinboim].

We furthermore explore one specific problem in the literature: that of finding the minimum number of required distributions for all queries in a causal graph when the set of queries is restricted. We show that this is solvable using an almost identical infrastructure to [Kandasamy, 2019]. Restricting the set of queries available can increase the lower bound on the size of the minimum intervention cover in general, but it cannot impact the upper bound, which is in any case exponential only in a polynomial of the degree of nodes in the graph and the size of the largest c-component. Since both of these are likely to be limited in any graph solvable by human researchers, this is a significant improvement over the brute force solution of applying the g-ID algorithm for all queries. As a last result, it has been shown in the literature that the smallest MIC is not a small set, but that in a bad scenario restricting available queries does not incur a significant cost in the size of the MIC. Researchers in rich settings involving non-manipulable variables or who are solving optimization problems around which experiments to run can make significant use of our results.

References

Lee and Bareinboim. General identifiability with arbitrary surrogate experiments. 2019.

et al Kandasamy. Minimum intervention cover of a causal graph. 2019.

Shpitser and Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 2008.

Correa Lee and Bareinboim. Generalized transportability: Synthesis of experiments from heterogeneous domains.