A Note on Structure Compatibility for Large Scale Structure Learning

Sung-Ho Kim
Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology, Daejeon, S. Korea
e-mail: sungkim@kaist.ac.kr
Namgil Lee
Department of Information and Statistics, Kangwon National University, Chuncheon, S. Korea
e-mail: namgil.lee@kangwon.ac.kr

***Abstract*—Suppose that we are given two statistical model structures given in graphs. We are interested in testing whether they are from one source model or data. If the models share a source, we say that the models are compatible. In the paper, we present methods of testing compatibility of two model structures provided that the two structures share at least two nodes. The model structure represents causal or associative relationships between random variables (or nodes in graphs). Two testing methods will be proposed. One is by comparing structures of the intersection part of the two models, and the other is by using what we call union graphs. A union graph is obtained by merging the given structures with some additions and deletions of edges under a specified condition. We then check if the given structures are possible from the union graph. The methods are illustrated through examples. We aim to develop a method of structure learning by using as many pieces of structure information as possible. In this line of work, the pieces of information given in graphs need be checked for compatibility among themselves. This is the reason why this small piece of work is so crucial to the success of our future work.**

***Keywords-combined model structure; Markovian subgraph; structural discrepancy; union graph*combined model structure; Markovian subgraph; structural discrepancy; union graph.**

I. Introduction

Independence graphs have been extensively used in multivariate data analysis to understand the Markov properties underlying joint distributions. In an independence graph, random variables are represented as nodes, where edges are absent between two nodes if the corresponding random variables are conditionally independent. This study focuses on undirected independence graphs, characterized by undirected edges and no predefined node ordering. For comprehensive overviews of independence graphs, see [1].

In scenarios involving two distinct data sources, where only part of a whole multivariate system can be observed from each source, it is necessary to combine models inferred separately from each source to construct a unified model for the entire system. We investigate the problem of combining two probabilistic graphical models, represented by their respective graph structures, into a single, larger graphical model. While substantial research has focused on combining Bayesian networks [2]–[4], comparatively less attention has been given to the combination of undirected graphical models. Notable exceptions include works addressing the combination of conditional log-linear model structures [5], studies on Markovian subgraphs under undirected decomposable graphical models [6], and strategies for combining decomposable and non-decomposable undirected graphical model structures [7]. Further, Massa and Lauritzen [8] explored the properties of combined distribution families, applying their findings to Gaussian graphical models.

The computational complexity associated with searching for graphical model structures increases significantly with the number of variables. Thus,, it is beneficial to check the feasibility of a joint model prior to undertaking computationally intensive procedures for combining marginal model structures. Dawid and Studeny [9] introduced the concept of compatibility for merging objects while preserving as much conditional independence as possible. Kim [7] defined graphical compatibility, for sets of graphical model structures using Markovian subgraphs.

This work proposes a novel methodology for testing the structure compatibility of two undirected graphical model structures, where the compatibility is defined in Definition 2.

The paper is organized in 4 sections. In Section 2, we introduce some notation and terminologies along with a few lemmas as preliminary to the main results. The relationship between probability models and graphs is briefly described. Section 3 is of main results proposed for testing structural incompatibility. In Section 4, we conclude the paper with remarks for summary and plans for future works.

II. Preliminaries

The set of nodes in a graph G is denoted by $V(G)$, while the set of edges is denoted by $E(G)$. For a graph $G = (V, E)$ and two nodes $i, j \in V$, a path between i and j is a sequence of edges in E that connects i and j. For example, a path can be represented as $\{(i, v_1), (v_1, v_2), \dots, (v_{m-1}, j)\} \subset E$. For subsets of nodes $A_1, A_2 \subset V$, a path $\{(i, v_1), (v_1, v_2), \dots, (v_{m-1}, j)\}$ is an $A_2 \setminus A_1$-path between i and j if $v_1, v_2, \dots, v_{m-1} \in A_2 \setminus A_1$.

For a graph $G = (V, E)$ and a subset of nodes $A \subset V$, the induced subgraph of G on A is defined as $G_A = (A, E \cap (A \times A))$. Another type of subgraph, called the Markovian subgraph, is defined as follows.

Definition 1 (Markovian subgraph [6])**.** *For a graph $G = (V, E)$ and a subset of nodes $A \subset V$, the Markovian subgraph of G upon A is defined as $G\_A = (A, E\_A)$ such that*

$(i, j) \in E\_A$ if and only if $(i, j) \in E \cap (A \times A)$ or there exists a $V \setminus A$-path between i and j.

If a graph G' is a Markovian subgraph of a graph G, we write as

$$G' \subset_M G \quad \text{or} \quad G \supset_M G'.$$

Before moving further, we need more notations. For a given graph $G' = (V', E')$ and a set of nodes $W \supset V'$, we define two sets of graphs as follows:

$$\begin{aligned} I^W(G') = \{G \,|\, W \supset V(G) \supset V(G') & \text{ such that, for } \{i,j\} \subseteq \\ & V(G'), (i,j) \notin E(G') \text{ implies that } (i,j) \notin E(G) \\ & \text{and that there is no } V(G) \setminus V(G')\text{-path between} \\ & i \text{ and } j \text{ in } G\}. \\ D^W(G') = \{G \,|\, W \supset V(G) \supset V(G') & \text{ such that, for } \{i,j\} \subseteq \\ & V(G'), (i,j) \in E(G') \text{ implies that } (i,j) \in E(G) \\ & \text{ or that there is at least one } V(G) \setminus V(G')\text{-path} \\ & \text{between } i \text{ and } j \text{ in } G\}. \end{aligned}$$

The superset W is required only to set an upper bound on the node set so that the sets $I^W(G')$ and $D^W(G')$ are well-defined. The following lemma is immediate from Definition 1.

Lemma 1. *For graphs G and G' with $W \supset V(G) \supset V(G')$, $G \supset_M G'$ if and only if $G \in I^W(G') \cap D^W(G')$.*

We can now show that the Markovian subgraph relation is transitive as in

Lemma 2. *If $G \supset_M G' \supset_M G''$, then $G \supset_M G''$.*

Proof of Lemma 2. Assume that $G \supset_M G' \supset_M G''$. Let W be a set such that $W \supset V(G)$. Then we have from the assumption and Lemma 1 that $G \in I^W(G') \cap D^W(G')$ and that $G' \in I^W(G'') \cap D^W(G'')$. It suffices to show that $G \in I^W(G'') \cap D^W(G'')$.

Since $G' \in I^W(G'')$, for nodes i and j in $V(G'')$, $(i,j) \notin E(G'')$ implies that $(i,j) \notin E(G')$ and that there is no $V(G') \setminus V(G'')$-path between i and j in G'. Since $G \in I^W(G')$, we can further derive that $(i,j) \notin E(G'')$ implies that $(i,j) \notin E(G)$ and that

$$\begin{array}{l} \text{there does not exist any } V(G') \setminus V(G'')\text{-path} \\ \text{between } i \text{ and } j \text{ in } G' \text{ nor any } V(G) \setminus V(G')\text{-path} \\ \text{between } i \text{ and } j \text{ in } G. \end{array} \quad (1)$$

Statement (1) implies that there does not exist any $V(G) \setminus V(G'')$-path between i and j in G: if the latter path exists, at least one of the two paths in statement (1) must exist by the definition of Markovian subgraph. Thus, we have $G \in I^W(G'')$.

From the condition of the lemma, we also have $G \in D^W(G')$ and $G' \in D^W(G'')$. It follows that $(i,j) \in E(G'')$ implies one of the following: either $(i,j) \in E(G)$, there exists a $V(G') \setminus V(G'')$-path between i and j in G', or there exists a $V(G) \setminus V(G')$-path between i and j in G. Since $V(G) \supset V(G') \supset V(G'')$, we can regard each of $V(G') \setminus V(G'')$-path between i and j in G' and $V(G) \setminus V(G')$-path between i and j in G as a $V(G) \setminus V(G'')$-path between i and j in G. , we have $G \in D^W(G'')$.

Therefore, it follows that $G \in I^W(G'') \cap D^W(G'')$, which completes the proof. □

Since Markovian subgraphs are determined uniquely, we have

Lemma 3. *For a graph $G = (V, E)$ and two subsets of nodes A and B such that $A \subset B \subset V$, the following relationship holds:*

$$G_{\_A} = (G_{\_B})_{\_A}.$$

We will briefly look into the relationship between probability models and graphs. For a given probability distribution P with its independence graph G, the relationship between a marginal probability distribution and a Markovian subgraph can be described as follows. A probability distribution P is said to be Markov with respect to G if P satisfies all the conditional independences (i.e., Markov properties) represented by G. The set of all probability distributions that are Markov with respect to G is denoted by $\mathcal{M}(G)$. If the joint distribution P satisfies the Markov properties associated with a graph G, then for any subset of nodes $A \subset V$, the marginal distribution P_A must satisfy the Markov properties expressed by the Markovian subgraph $G_{\_A}$ [7]. This relationship can be formally stated as follows.

Corollary 1 ([7])**.** *Let $G = (V, E)$ be an undirected graph, and suppose the distribution P of a random vector X_V satisfies $P \in \mathcal{M}(G)$. For a subset $A \subset V$, let $X_V = (X_A, X_{V \setminus A})$, and let P_A denote the marginal distribution of the random vector X_A. Then, the following holds:*

$$P_A \in \mathcal{M}(G_{\_A}).$$

For two probability distributions $P_1 \in \mathcal{M}(G_1)$ and $P_2 \in \mathcal{M}(G_2)$, the P_1 and P_2 are said to be (strongly) compatible if there exists a joint probability distribution $P \in \mathcal{M}(G)$ for some undirected graph G, such that P_1 and P_2 are the marginal distributions of P [9]. Using Corollary 1, the concept of compatibility for probability distributions can be extended to the compatibility of graph structures as follows.

Definition 2 (Structure Compatibility)**.** *Two undirected graphs G_1 and G_2 are said to be compatible if there exists a graph G such that G_1 and G_2 are Markovian subgraphs of G.*

Further terminologies related to combined models are introduced below.

Definition 3 (Combined Model Structure (CMS) [7])**.** *A graph $G = (V, E)$ is referred to as a combined model structure (CMS) of two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ if $V = V_1 \cup V_2$ and G_1 and G_2 are Markovian subgraphs of G. A CMS is called maximal if adding any additional edge results in a graph that is no longer a CMS. The set of all maximal CMSs of G_1 and G_2 is denoted by $G_1 \oplus G_2$.*

It is obvious that the concepts of structure compatibility and existence of a CMS are equivalent.

III. MAIN RESULTS

In this section, we devise rules for testing incompatibility between graphs by applying graph theory and Markov properties. To avoid trivial cases, we assume that two graphs share at least two nodes. Consider two graphs, $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, for which $C = V_1 \cap V_2$ and $|C| \geq 2$.

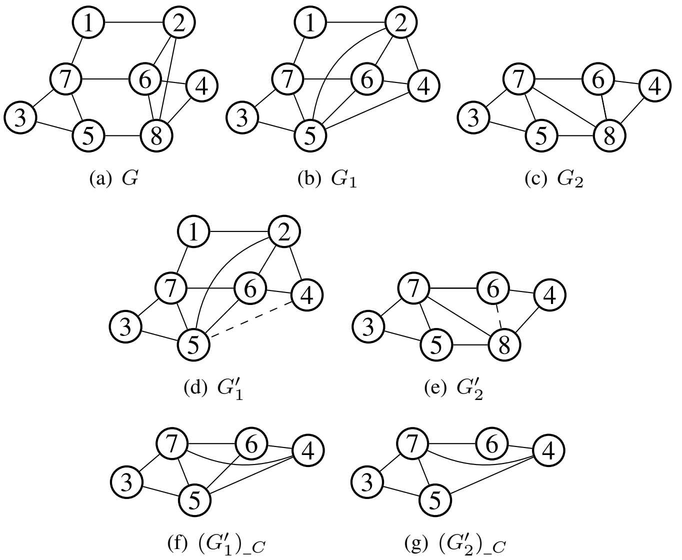
A. Checking Discrepancy in intersection part

The first of our methods of compatibility test is by comparing $(G_1)_C$ and $(G_2)_C$. It is intuitive that, if the two Markovian subgraphs G_1 and G_2 on C are not the same, G_1 and G_2 are not from the same model structure. This observation is stated formally as follows.

Theorem 1. *For graphs G_1 and G_2 with $C = V_1 \cap V_2$, if $(G_1)_{\_C} \neq (G_2)_{\_C}$, then $G_1 \bigoplus G_2 = \emptyset$.*

Proof of Theorem 1. Suppose there exists a graph $G = (V, E)$ such that G_1 and G_2 are Markovian subgraphs of G. Then, by Lemma 3, we have $(G_1)_{\_C} = G_{\_C} = (G_2)_{\_C}$. This contradicts the condition of the theorem, implying that G_1 and G_2 cannot have a Markovian supergraph such as G. From this follows the desired result that $G_1 \bigoplus G_2 = \emptyset$. □

An example of the two graphs that satisfy the condition of this theorem is given in Figure 1. We can easily check that the Markovian subgraphs G_1 and G_2 of G in the figure satisfy the equality $(G_1)_{\_C} = (G_2)_{\_C}$ for $C = \{3, 4, 5, 6, 7\}$. However, the other two graphs G'_1 and G'_2, which were obtained by removing the edges $(4, 5)$ from G_1 and $(6, 8)$ from G_2, are not compatible according to Theorem 1.

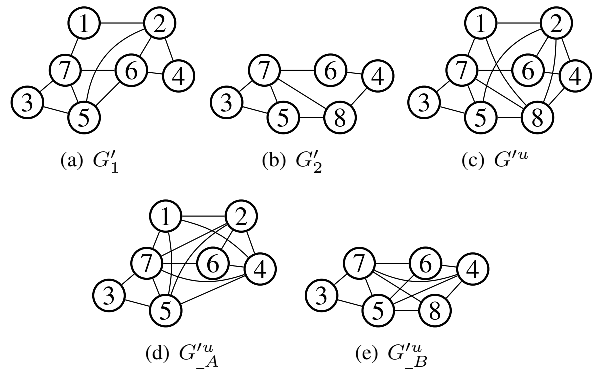


(a) G (b) G_1 (c) G_2

(d) G'_1 (e) G'_2

(f) $(G'_1)_{\_C}$ (g) $(G'_2)_{\_C}$

Figure 1. Two graphs, G'_1 and G'_2, satisfying the condition of Theorem 1. For a graph G, its Markovian subgraphs G_1 and G_2 are obtained upon the node sets $A = \{1, 2, \cdots, 7\}$ and $B = \{3, 4, \cdots, 8\}$ respectively. Edge $(4, 5)$ is removed from G_1 into G'_1 and edge $(6, 8)$ is removed from G_2 into G'_2. The removed edges are dashed in G'_1 and G'_2. For the set $C = A \cap B$, $(G'_1)_{\_C} \neq (G'_2)_{\_C}$.



(a) G'_1 (b) G'_2 (c) G'^u

(d) $G'^u_{\_A}$ (e) $G'^u_{\_B}$

Figure 2. An example where a Markovian subgraph of union(G_1, G_2) does contain none of G_1 and G_2 as a subgraph. G'_1 and G'_2 are carried over from Figure 1 with dashed edges erased. $G'^u_{\_A}$ and $G'^u_{\_B}$ are Markovian subgraphs of G'^u upon A and B respectively. Note that G'**??**$_1 \not\subseteq G'^u_{\_A}$.

B. Using Union Graphs

A union graph, $G^u = \text{union}(G_1, G_2)$, of two graphs G_1 and G_2 is defined as $G^u = (V^u, E^u)$ where $V^u = V_1 \cup V_2$ and $(i, j) \in E^u$ if and only if either $(i, j) \in E_1 \cap E_2$ or $i \in V_1 \setminus V_2$ and $j \in V_2 \setminus V_1$. That is, a union graph is formed by adding edges between nodes in $V_1 \setminus V_2$ and those in $V_2 \setminus V_1$ [7].

The following lemma provides a theoretical basis for testing incompatibility using union graphs.

Lemma 4. *For a graph $G = (V, E)$, let A and B be subsets of V such that $A \cup B = V$. Then*

$$G \subseteq union(G_{\_A}, G_{\_B}).$$

Proof of Lemma 4. $G_A \subseteq G_{\_A}$ and $G_B \subseteq G_{\_B}$. Thus, it follows that $G \subseteq \text{union}(G_A, G_B) \subseteq \text{union}(G_{\_A}, G_{\_B})$. □

Theorem 2. *Let $G^u = union(G_1, G_2)$ and $V_i = V(G_i)$ for $i = 1, 2$. If there exists i such that $G_i \not\subseteq (G^u)_{\_V_i}$, then $G_1 \bigoplus G_2 = \emptyset$.*

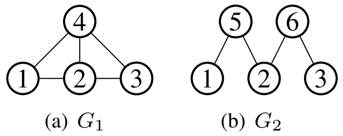
Proof of Theorem 2. Suppose that $G_1 \bigoplus G_2 \neq \emptyset$ and let H be a graph in $G_1 \bigoplus G_2$. Then, by Lemma 4, $H \subseteq G^u = \text{union}(G_1, G_2)$. Since $G_i \subseteq H_{\_V_i}$ for $i = 1, 2$, we have $G_i \subseteq (G^u)_{\_V_i}$ for $i = 1, 2$, which contradicts the condition of the theorem.

Therefore, under the condition of the theorem, it must hold that $G_1 \bigoplus G_2 = \emptyset$. □

An example of this theorem is given in Figure 2 where G'_1 and G'_2 are carried over from Figure 1 with dashed edges erased. After constructing the union graph G'^u of G'_1 and G'_2, we checked if $G'_i \subseteq G'^u_{\_V_i}$ holds and found that $G'_1 \not\subseteq G'^u_{\_A}$ and $G'_2 \subseteq G'^u_{\_B}$.

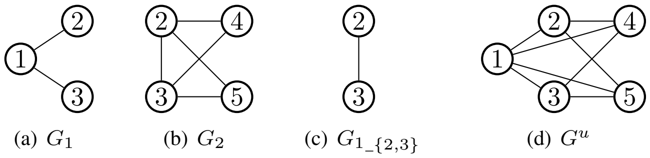
For any two graphs G_1 and G_2 with $V(G_i) = V_i$, $i = 1, 2$, and $C = V_1 \cap V_2$, the discrepancy, $(G_1)_{\_C} \neq (G_2)_{\_C}$, does not necessarily imply existence of $i \in \{1, 2\}$ such that $G_i \not\subseteq \text{union}(G_1, G_2)_{\_V_i}$. For G_1 and G_2 in Figure 3 with $C = \{1, 2, 3\}$, we have $(G_1)_{\_C} \neq (G_2)_{\_C}$ but

$$G_i \subseteq \text{union}(G_1, G_2)_{\_V_i} \text{ for } i = 1, 2. \tag{2}$$



(a) G_1 (b) G_2

Figure 3. An incompatible pair of graphs. The incompatibility of this pair is confirmed by Theorem 1 but not by Theorem 2.



(a) G_1 (b) G_2 (c) $G_{1\_\{2,3\}}$ (d) G^u

Figure 4. Two graphs G_1 and G_2, the Markovian subgraph $G_{1\_\{2,3\}}$ on the common nodes 2 and 3, and the union graph G^u.

However, $(G_1)_{\_C} = (G_2)_{\_C}$ has much to do with (2) as shown below.

Theorem 3. *For two graphs G_i, $i = 1, 2$, let $V_i = V(G_i)$ and $C = V_1 \cap V_2$. If $(G_1)_{\_C} = (G_2)_{\_C}$, then (2) holds true.*

Proof of Theorem 3. Let E^* be the set of the edges whose nodes are in C only and each of which appears in only one of G_1 or G_2. We will denote by G^+ the graph whose node and edge sets are given respectively by $V^+ = V_1 \cup V_2$ and $E^+ = E(G_1) \cup E(G_2) \setminus E^*$. Then, under the condition of the theorem, we have

$$G^+_{\_V_i} = G_i, \text{ for } i = 1, 2. \tag{3}$$

Note that $G^+_{\_V_i}$ is obtained by adding to the induced subgraph $G^+_{V_i}$ of G^+ all the edges in $E((G_i)_C) \setminus E(G^+_C)$. This is why we have the above equation.

By definition, $G^+ \subseteq \text{union}(G_1, G_2)$, since $\text{union}(G_1, G_2)$ is obtained by adding to G^+ all the edges between the nodes in $V_1 \setminus C$ and those in $V_2 \setminus C$. From this follows the desired result (2). □

We can see an example of this theorem in Figure 4. In the figure, $V_1 = \{1, 2, 3\}$, $V_2 = \{2, 3, 4, 5\}$, and $C = \{2, 3\}$. We have $(G_1)_{\_C} = (G_2)_{\_C}$ and $G_i \subseteq G^u_{\_V_i}$, $i = 1, 2$. Note however in the figure that G_1 and G_2 are not compatible. This simple example indicates that the conditions of Theorems 1 and 2 are sufficient for incompatibility of a pair of graphs but not necessary.

IV. Conclusion and future work

In this work, we presented two methods for incompatibility test. One of them is by checking structural discrepancy in the intersection part of two model structures and the other is by using union graphs. If any of the given graphs is not contained in the corresponding Markovian subgraph of the union graph, we may conclude that the graphs are not compatible.

The methods are devised based on graph and statistical theories. If the graphs are incompatible, it means that they are not from a unified graph. If we regard the graphs as model structures, the incompatibility implies that the models are not from a source of data. Experiments show that these testing methods are very useful in large scale structure learning since we can save our time in structure learning by avoiding incompatible model structures at the early stage of structure learning.

This work is yet an early stage of large scale statistical (structure) learning from big data. We will search for more methods for incompatibility test, then we will develop methods for using pieces of structure information obtained from different sets of data towards the large scale learning.

Kim and Kim[4] used pieces of structural information for structure learning and improved preceding learning methods. We aim to develop a method of structure learning by using as many pieces of structure information as possible. In this line of work, the pieces of information given in graphs need be checked for compatibility among themselves. In this way we could keep the quality of the structure information at a high level.

Acknowledgements

This work was supported by National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. RS-2024-00358572, No. RS-2024-00336424).

References

[1] S. L. Lauritzen, *Graphical Models*. Oxford, UK: Clarendon Press, 1996.

[2] D. Danks, C. Glymour, and R. Tillman, “Integrating locally learned causal structures with overlapping variables,” in *Advances in Neural Information Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., vol. 21, Curran Associates, Inc., 2008.

[3] S. Triantafillou and I. Tsamardinos, “Constraint-based causal discovery from multiple interventions over overlapping variable sets,” *Journal of Machine Learning Research*, vol. 16, no. 66, pp. 2147–2205, 2015.

[4] G. Kim and S. Kim, “Marginal information for structure learning,” *Statistics and Computing*, vol. 430, no. 30, pp. 331–349, 2020. DOI: 10.1007/s11222-019-09877-x.

[5] S. E. Fienberg and S.-H. Kim, “Combining conditional log-linear structures,” *Journal of the American Statistical Association*, vol. 94, no. 445, pp. 229–239, 1999. DOI: 10.1080/01621459.1999.10473838.

[6] S.-H. Kim, “Properties of Markovian subgraphs of a decomposable graph,” in *MICAI 2006: Advances in Artificial Intelligence*, A. Gelbukh and C. A. Reyes-Garcia, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 15–26. DOI: 10.1007/11925231_2.

[7] S.-H. Kim, “Learning model structures based on marginal model structures of undirected graphs,” KAIST, BK21 Research Report 09-04, Mar. 2009.

[8] M. S. Massa and S. L. Lauritzen, “Combining statistical models,” in *Algebraic methods in statistics and probability II*, ser. Contemporary Mathematics, M. A. G. Viana and H. P. Wynn, Eds., Providence, RI: American Mathematical Society, 2010, pp. 239–259.

[9] A. P. Dawid and M. Studený, “Conditional products: An alternative approach to conditional independence,” in *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*, D. Heckerman and J. Whittaker, Eds., ser. Proceedings of Machine Learning Research, vol. R2, PMLR, 1999, pp. 27–35.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org