INNOV 2025, Sep. 28- Oct. 2, 2025. Lisbon, Portugal

# A Note on Structure Compatibility for Large Scale Structure Learning

Sung-Ho Kim
Korea Advanced Institute of Science and Technology (KAIST)
email: sungkim@kaist.ac.kr

September 30, 2025

# Short Biography

Current Position: Professor Emeritus of Statistics, Dept of Mathematical Sciences, Korea Advanced Institute of Science and Technology(KAIST).

Education: B.Sc. in Math (Seoul National Univ, 1980, S. Korea)/ Ph.D. in Statistics (Carnegie Mellon University, U.S.A, 1989)

Research fields: Multivariate Analysis, Graphical modelling, Structure learning, Structure combination, Large scale modelling

Professional experience: Educational Testing Service (research scientist, 1989-1993, USA.), KAIST (Professor, since 1993, S. Korea).

Professional membership: International Statistical Institute, Elected member, Since Jan. 2003.

Professional activity: Over 70 research articles in peer-reviewed journals, over 30 invited talks in conferences and universities, and over 200 evaluations of research articles.

# Outline

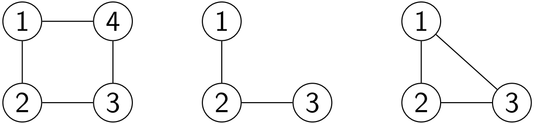
# 1. Introduction

# Back ground and Goal

- Data can originate from multiple sources, each involving different subsets of variables.
- Sparse data problems (e.g., data size for a model of 40 binary variables)
- Graphical Models from different sources of data can be combined.
- Goal: Find conditions sufficient for graph combination
- This work proposes some sufficient conditions for graph combination

# 2. Graphs and Markov Properties

- Graph $G = (V, E)$: $V$ is a set of nodes, $E \subset V \times V$ is a set of edges.
- Induced subgraph: $G_A = (A, E \cap (A \times A))$ for $A \subset V$.

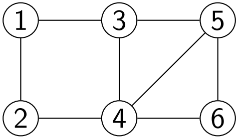



(a) $G$ (b) $G_A$ (c) $G\_A$

Figure 1: Two types of subgraph of $G$ on $A = \{1, 2, 3\}$.

# Graph basics(2/2)

- Separation: $A$ and $B$ are separated by $S$ if all the paths from $A$ to $B$ pass through $S$.

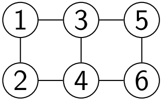


- Let $X = \{X_1, \ldots, X_n\}$ be random variables with joint distribution $P$.
- For $A = \{1, 2\}$, we will write $X_A$ for $(X_1, X_2)$.
- Let $V = \{1, 2, \cdots, n\}$. For two random vectors $X_A$ and $X_B$ with $A, B \subseteq V$ and $A \cap B = \emptyset$, we say that $X_A$ and $X_B$ are stochastically independent if
$$P(x_{A \cup B}) = f_1(x_A) f_2(x_B).$$
- Now suppose $A \cap B \neq \emptyset$ and let $A \cap B = C$. We say that $X_{A \setminus B}$ and $X_{B \setminus A}$ are conditionally independent given $X_C$ (or $X_{A \setminus B} \perp X_{B \setminus A} | X_C$) if
$$P(x_{A \setminus B}, x_{B \setminus A} | x_C) = f_3(x_{A \setminus B}) f_4(x_{B \setminus A}).$$

# Markov Properties(2/5)

- $P$ is globally Markov w.r.t. $G$ if:
  $X_A \perp X_B \mid X_S$ whenever $A$ and $B$ are separated by $S$ in $G$.
- $M(G)$: A set of distributions globally Markov to $G$.
- $G$ is a perfect map of $P$ if $P \in M(G)$ and all conditional independencies in $P$ are encoded in $G$.

# Markov Properties(3/5)

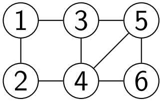



(a) $G_1$

$$P_1(x_1, \cdots, x_6) = g_1(x_1, x_2)g_2(x_1, x_3)g_3(x_2, x_4)g_6(x_3, x_4)g_7(x_3, x_5) \times g_8(x_4, x_6)g_9(x_5, x_6).$$

$$\begin{aligned} P_1(x_1, \cdots, x_6 | x_3, x_6) &= g_1(x_1, x_2)g_2(x_1, x_3)g_3(x_2, x_4)g_6(x_3, x_4)g_7(x_3, x_5) \\ &\quad \times g_8(x_4, x_6)g_9(x_5, x_6)/P(x_3, x_6). \\ &= h_1(x_1, x_2)h_2(x_2, x_4)h_3(x_5) \end{aligned}$$

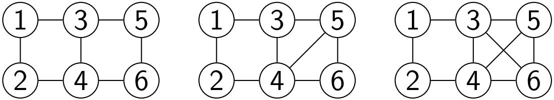
# Markov Properties(4/5)



(b) $G_2$

$$P_2(x_1, \cdots, x_6) = g_1(x_1, x_2)g_2(x_1, x_3)g_3(x_2, x_4)g_4(x_3, x_4, x_5) \times g_5(x_4, x_5, x_6).$$

$$\begin{aligned} P_2(x_1, \cdots, x_6 | x_3, x_6) &= g_1(x_1, x_2)g_2(x_1, x_3)g_3(x_2, x_4)g_4(x_3, x_4, x_5) \\ &\quad \times g_5(x_4, x_5, x_6)/P(x_3, x_6). \\ &= h_1(x_1, x_2)g_3(x_2, x_4)h_4(x_4, x_5) \end{aligned}$$



(c) $G_1$ (d) $G_2$ (e) $G_3$

Figure 2: Graphs of 6 nodes

$$
\begin{aligned}
P_1(x_1,\cdots,x_6) &= g_1(x_1,x_2)g_2(x_1,x_3)g_3(x_2,x_4)g_6(x_3,x_4)g_7(x_3,x_5) \\
&\quad \times g_8(x_4,x_6)g_9(x_5,x_6). \\
P_2(x_1,\cdots,x_6) &= g_1(x_1,x_2)g_2(x_1,x_3)g_3(x_2,x_4)g_4(x_3,x_4,x_5)g_5(x_4,x_5,x_6). \\
P_3(x_1,\cdots,x_6) &= g_1(x_1,x_2)g_2(x_1,x_3)g_3(x_2,x_4)g_{10}(x_3,x_4,x_5,x_6).
\end{aligned}
$$

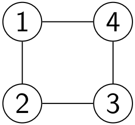
==>

1. $G_i$ is a perfect map of $P_i$ for $i = 1, 2, 3$.
2. $P_1, P_2$, and $P_3$ are all in $M(G_3)$.
3. $P_1, P_2$ are in $M(G_2)$. $P_1 \in M(G_1)$.

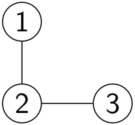
# 3. Combined model structure and Markovian combination

# Markovian Subgraphs

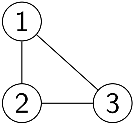
- Given $G = (V, E)$ and $A \subset V$, define the Markovian subgraph $G_{\_A} = (A, E_{\_A})$:
  $(i, j) \in E_{\_A}$ if there exists a $(V \setminus A)$-path between $i$ and $j$ in $G$.
- Independence properties of marginal distribution $P_A$ are captured by $G_{\_A}$.
- If $P$ is globally Markov w.r.t. $G$, then $P_A$ is globally Markov w.r.t. $G_{\_A}$. In other words, if $P \in M(G)$, then $P_A \in M(G_{\_A})$.



(a) $G$



(b) $G_A$



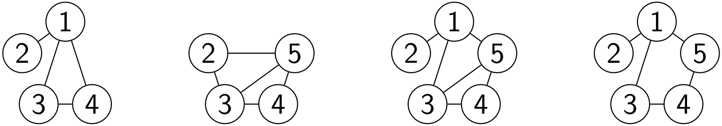
(c) $G_{\_A}$

# Combined model structure

## Definition 1

*Given two graphs* $G_1 = (V_1, E_1)$ *and* $G_2 = (V_2, E_2)$*, then we say that a graph* $G = (V, E)$ *is a combined model structure (CMS) of* $G_1$ *and* $G_2$ *if the following conditions (called CMS conditions) hold:*

1. $V = V_1 \cup V_2$,
2. $G_{\_V_1} = G_1$ *and* $G_{\_V_2} = G_2$.




(d) $G_1$ (e) $G_2$ (f) $G^1_{CMS}$ (g) $G^2_{CMS}$

Figure 3: Two CMS's of $G_1$ and $G_2$. $G^1_{CMS}$ is a maximal CMS.

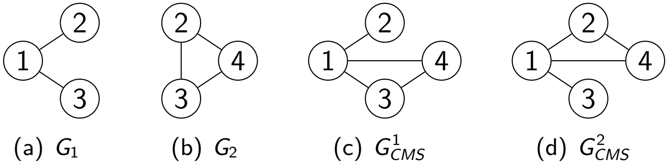



Figure 4: Two maximal CMS's of $G_1$ and $G_2$.

# Union graph

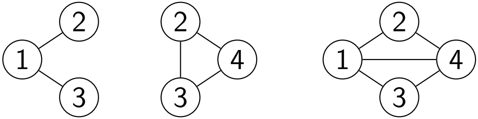
For an edge set $E$ of a graph and a node set $A$, we define $E_A$ as

$$E_A = \{(u, v) | (u, v) \in E \text{ and both } u \text{ and } v \text{ are in } A\}.$$

## Definition 2

*Given two graphs* $G_1 = (V_1, E_1)$ *and* $G_2 = (V_2, E_2)$, *let* $C = V_1 \cap V_2$. *then the union graph* $G = (V, E)$ *of the two graphs is defined such that the following holds:*

1. $V = V_1 \cup V_2$,
2. $E = E_1 \cup E_2 \cup \{(u, v) | u \in V_1 \setminus V_2,\ v \in V_2 \setminus V_1\} \setminus ((E_1)_C \Delta (E_2)_C)$.




(a) $G_1$ (b) $G_2$ (c) $Union(G_1, G_2)$

Figure 5: Union graph of $G_1$ and $G_2$.

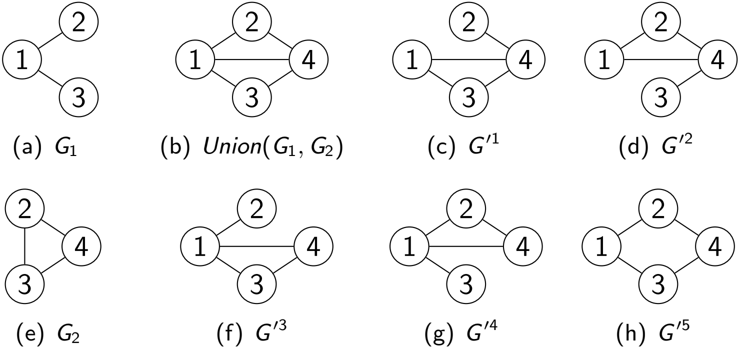
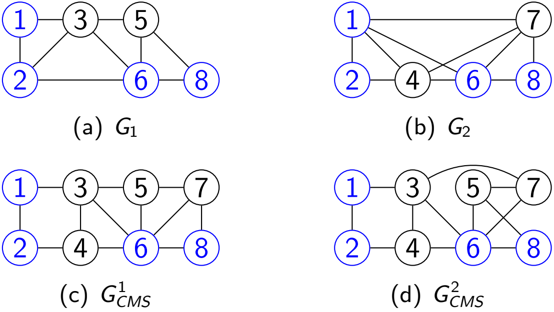



Figure 6: The union graph of $G_1$ and $G_2$ in Figure 8 is in panel (a). The subgraphs of $G'$ to be examined for separateness are in (c), (d), (f), (g), and (h).

# Example of Graph Combination



(a) $G_1$ (b) $G_2$

(c) $G^1_{CMS}$ (d) $G^2_{CMS}$

Figure 7: Graphs $G_1$ and $G_2$ are merged

# 4. Structure Compatibility

**Definition 1 (Structure Compatibility)**

Two undirected graphs $G_1$ and $G_2$ are said to be structure compatible if there exists a graph $G$ such that $G_1$ and $G_2$ are Markovian subgraphs of $G$.

**Theorem 1**

*Two undirected graphs $G_1$ and $G_2$ are structure compatible if and only if a CMS of $G_1$ and $G_2$ exists.*

# Example of five random variables

Table 1: Examples of probability tables for five Bernoulli random variables $X_1, X_2, X_3, X_4$, and $X_5$.

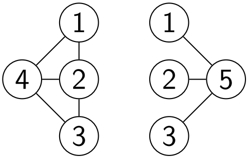
$P_1$

| | $X_3$ | 0 | | | | 1 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $X_2$ | 0 | | 1 | | 0 | | 1 | |
| | $X_1$ | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| $X_4$ | 0 | 0.010 | 0.010 | 0.050 | 0.090 | 0.020 | 0.020 | 0.030 | 0.054 |
| | 1 | 0.040 | 0.080 | 0.075 | 0.095 | 0.040 | 0.080 | 0.135 | 0.171 |
| | Total | 0.050 | 0.090 | 0.125 | 0.185 | 0.060 | 0.100 | 0.165 | 0.225 |

$P_2$

| | $X_3$ | 0 | | | | 1 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $X_2$ | 0 | | 1 | | 0 | | 1 | |
| | $X_1$ | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| $X_5$ | 0 | 0.020 | 0.020 | 0.080 | 0.080 | 0.030 | 0.030 | 0.120 | 0.120 |
| | 1 | 0.030 | 0.070 | 0.045 | 0.105 | 0.030 | 0.070 | 0.045 | 0.105 |
| | Total | 0.050 | 0.090 | 0.125 | 0.185 | 0.060 | 0.100 | 0.165 | 0.225 |

# Model structures for the Bernoulli variables in Table 1




(a) $G_1$ (b) $G_2$

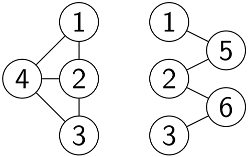
Figure 8: Model structures of $P_1$ and $P_2$ in the index order

# 5. Main result

# Rule 1: Discrepant common test

**Theorem 2 (Discrepant common test)**

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




(a) $G_1$ (b) $G_2$

# Example of Rule 1

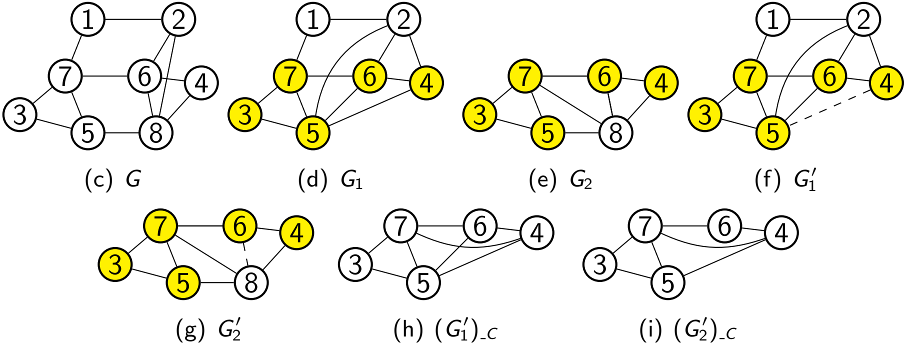



Figure 9: Two graphs, $G_1'$ and $G_2'$, satisfying the condition of Theorem 2. For a graph $G$, its Markovian subgraphs $G_1$ and $G_2$ are obtained upon the node sets $A = \{1, 2, \cdots, 7\}$ and $B = \{3, 4, \cdots, 8\}$ respectively. Edge $(4, 5)$ is removed from $G_1$ into $G_1'$ and edge $(6, 8)$ is removed from $G_2$ into $G_2'$. The removed edges are dashed in $G_1'$ and $G_2'$. For the set $C = A \cap B$, $(G_1')_{-C} \neq (G_2')_{-C}$.

# Rule 2: Union graph test

**Theorem 3 (Union Graph Test)**

*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 $G_1 \bigoplus G_2 = \emptyset$.*

# Example of Rule 2

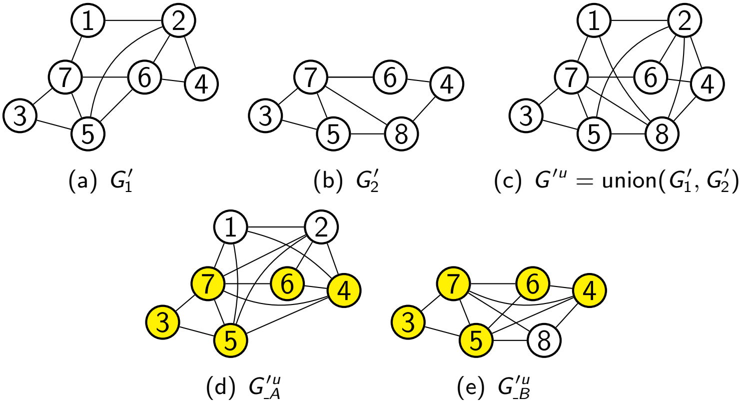



Figure 10: An example where a Markovian subgraph of union$(G'_1, G'_2)$ does contain none of $G'_1$ and $G'_2$ as a subgraph. $G'_1$ and $G'_2$ are carried over from Figure 9 with dashed edges erased. $G'^u_{\_A}$ and $G'^u_{\_B}$ are Markovian subgraphs of $G'^u$ upon $A$ and $B$ respectively. Note that $G'_1 \not\subseteq G'^u_{\_A}$.

# 6. Conclusion

1. Structure combination is a way of learning model structures based on a set of marginal model structures.
2. Markov properties play a key role in structure compatibility test.
3. In combining marginal model structures, checking structure compatibility will save us time by guiding us avoid unnecessary combining procedures.
4. More test rules for checking structure compatibility are to be sought for.

# Selected References


1. Dawid, A.P. and Lauritzen, S.L. (1993) Hyper Markov laws in the statistical analysis of decomposable graphical models, *Annals of Statistics* 21: 1272–1317.
2. Kim, S. (2006) Properties of markovian subgraphs of a decomposable graph, *MICAI 2006, Lecture Notes in Artificial Intelligence, LNAI* 4293. Advances in Artificial Intelligence. Alexander Gelbukh and Carlos Alberto Reyes-Garcia (Eds.), pp. 15-26.
3. Kim, G. and Kim, S. (2020) Marginal information for structure learning, *Statistics and Computing* 30(2): 331-349.
4. Kim, G., Lee, N., and Kim, S. (2025) A sequential approach for combining model structures of undirected graphical models, Submitted.
5. Massa, M.S. and Lauritzen, S.L. (2010) Combining statistical models, *Contemporary Mathematics* 516: 239-259.

# Thanks a lot for Your Attention!