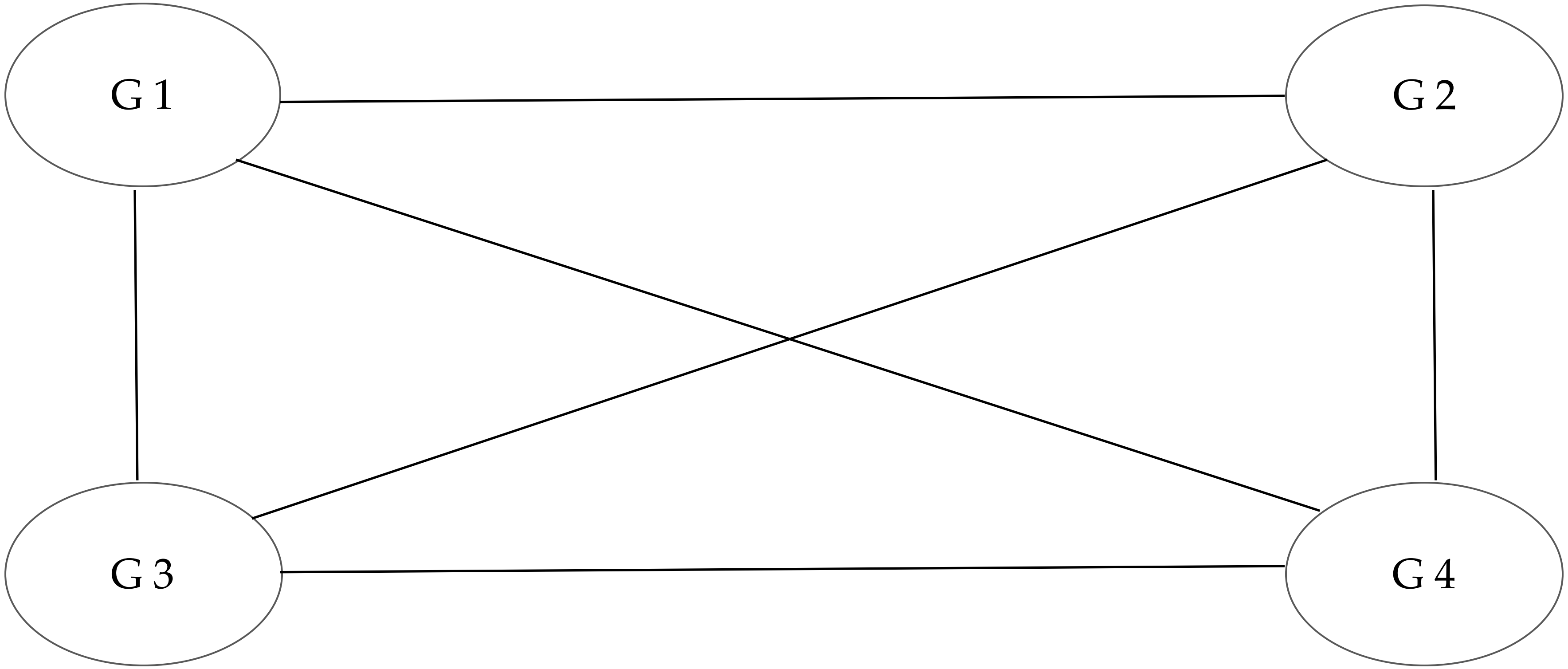
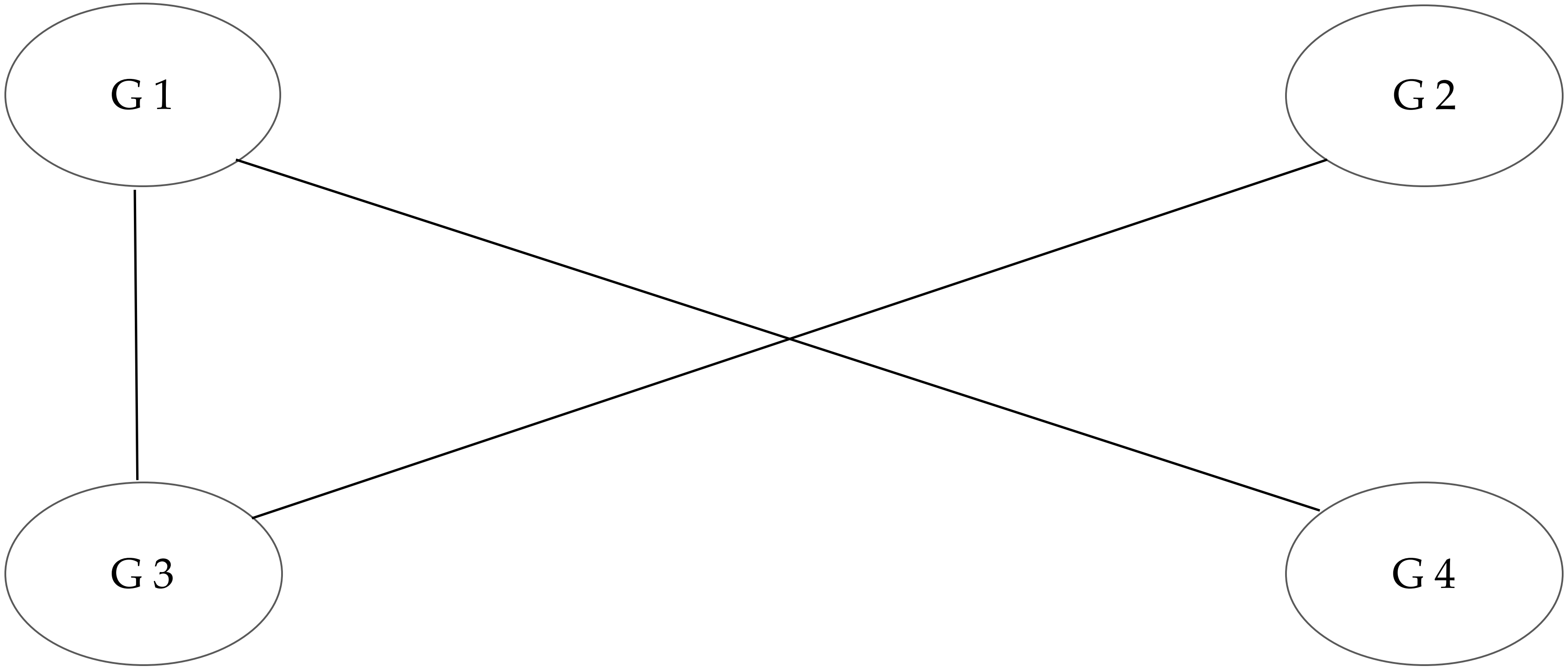
INFERRING GENE REGULATORY NETWORKS USING THE IMPROVED MARKOV BLANKET DISCOVERY ALGORITHM

W. Liu, Y. Jiang, L. Peng, X. Sun, W. Gan, Q. Zhao · Interdisciplinary Sciences: Computational Life Sciences· 2022

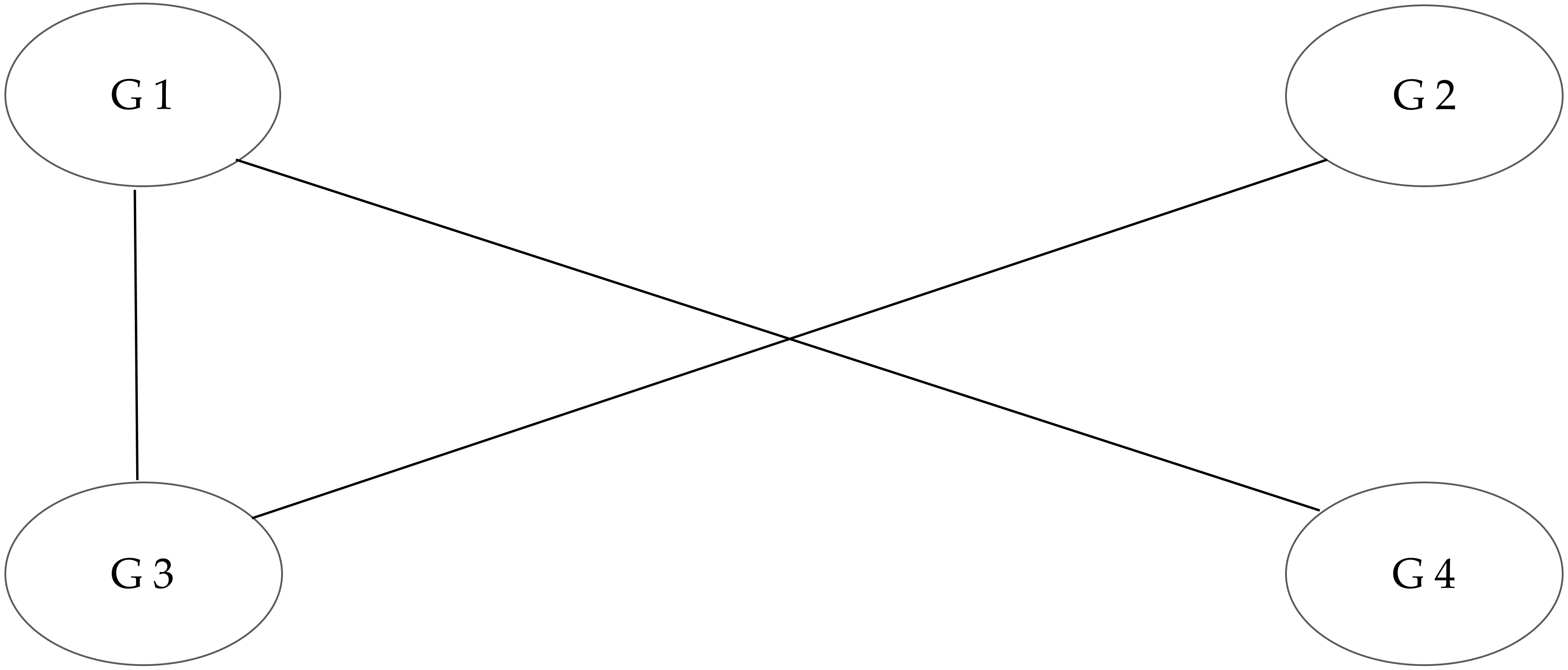
By Soel Micheletti · March 2023

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH


G 1
G 2
G 3
G 4


G 1
G 2
G 3
G 4

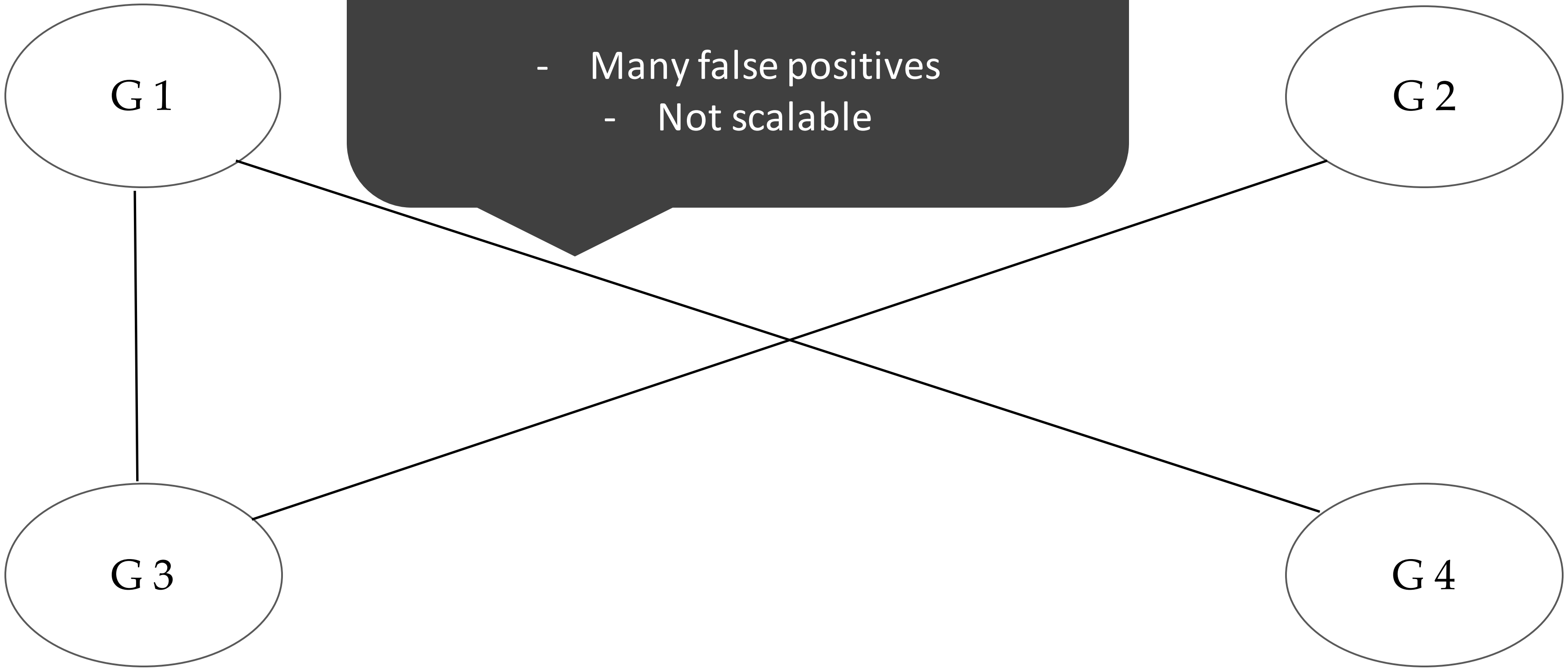
Undirected, unweighted gene co-expression



Undirected, unweighted gene co-expression

Information theoretic scores suffer mainly from two drawbacks:

- Many false positives
- Not scalable



IMBDANET

Step 1: IMBDA · **Step 2: pruning** · **Step 3: fine tuning**

IMBDANET

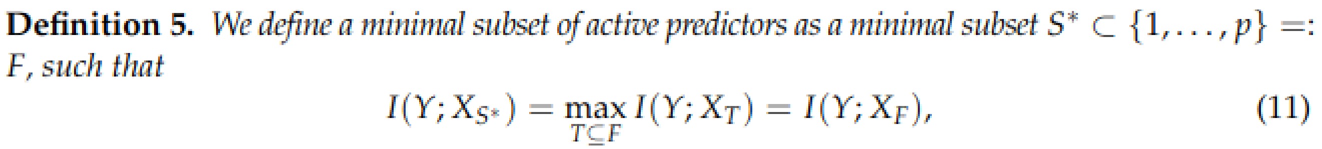
Step 1: IMBDA · **Step 2: pruning** · **Step 3: fine tuning**

For each gene, it uses an information theoretic score to compute its Markov Blanket.

IMBDANET

Step 1: IMBDA · Step 2: pruning · Step 3: fine tuning

For each gene, it uses an information theoretic score to compute its Markov Blanket.



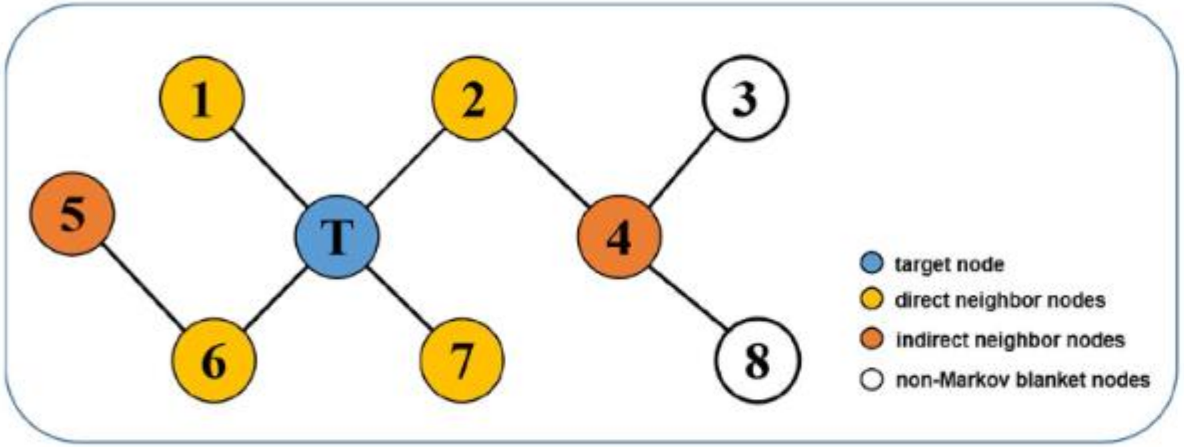
Definition 5. *We define a minimal subset of active predictors as a minimal subset $S^* \subset \{1, \dots, p\} =: F$, such that*

$$I(Y; X_{S^*}) = \max_{T \subseteq F} I(Y; X_T) = I(Y; X_F), \tag{11}$$

IMBDANET

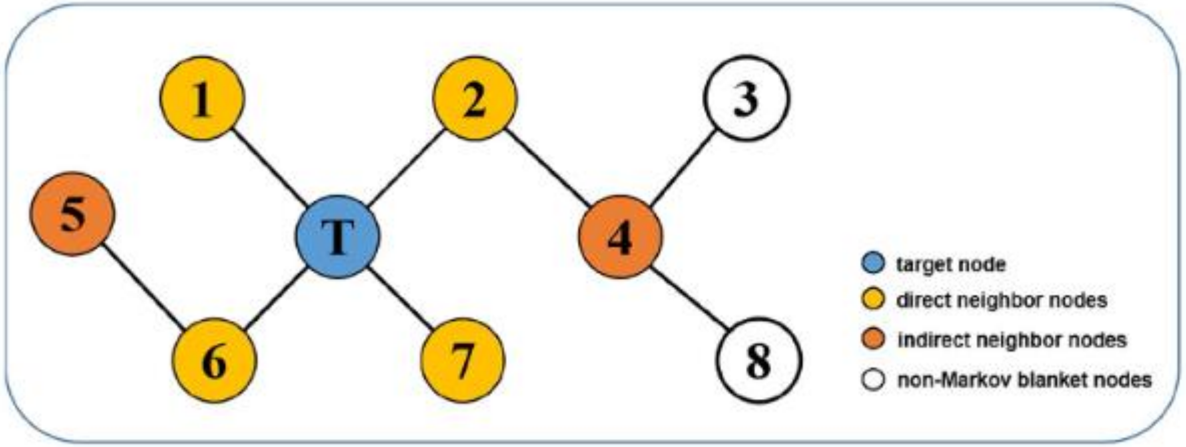
Step 1: IMBDA · Step 2: pruning · Step 3: fine tuning

For each gene, it uses an information theoretic score to compute its Markov Blanket.



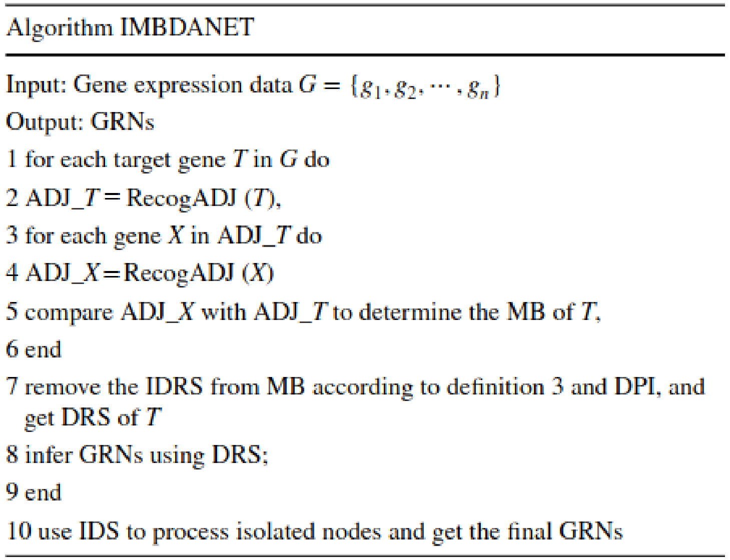
IMBDANET

Step 1: IMBDA · Step 2: pruning · Step 3: fine tuning

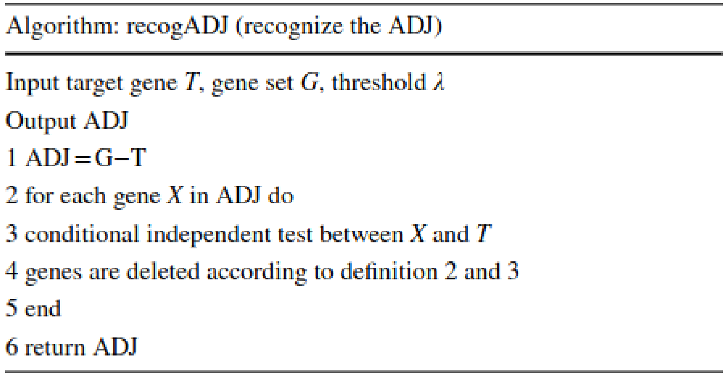


For each gene, it uses an information theoretic score to compute its Markov Blanket.

Iterative approach (PC style)



```
Algorithm IMBDANET
Input: Gene expression data $G = \{g_1, g_2, \cdots, g_n\}$
Output: GRNs
for each target gene $T$ in $G$ do
    ADJ_$T$ = RecogADJ ($T$),
    for each gene $X$ in ADJ_$T$ do
        ADJ_$X$=RecogADJ ($X$)
        compare ADJ_$X$ with ADJ_$T$ to determine the MB of $T$,
    end
    remove the IDRS from MB according to definition 3 and DPI, and get DRS of $T$
    infer GRNs using DRS;
end
use IDS to process isolated nodes and get the final GRNs
```



```
Algorithm: recogADJ (recognize the ADJ)
Input target gene $T$, gene set $G$, threshold $\lambda$
Output ADJ
ADJ = G−T
for each gene $X$ in ADJ do
    conditional independent test between $X$ and $T$
    genes are deleted according to definition 2 and 3
end
return ADJ
```

$$\lambda = \min(\mathrm{MIM}) + \mathrm{var}(\mathrm{MIM}) \times (\mathrm{mean}(\mathrm{MIM}) - \min(\mathrm{MIM}))$$

IMBDANET

Step 1: IMBDA · **Step 2: pruning** · Step 3: fine tuning

The MB contains both directly and indirectly related nodes. To remove indirectly related nodes, they use

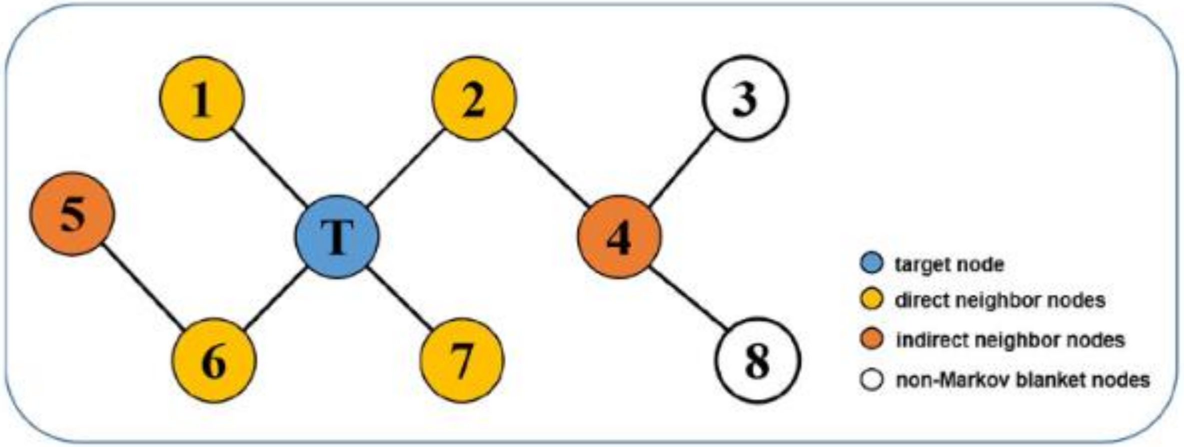
1. Data Processing Inequality
2. Definition of indirectly related nodes (dependency only when conditioning on collider)

IMBDANET

Step 1: IMBDA · **Step 2: pruning** · Step 3: fine tuning

The MB contains both directly and indirectly related nodes. To remove indirectly related nodes, they use

1. Data Processing Inequality
2. Definition of indirectly related nodes (dependency only when conditioning on collider)



IMBDANET

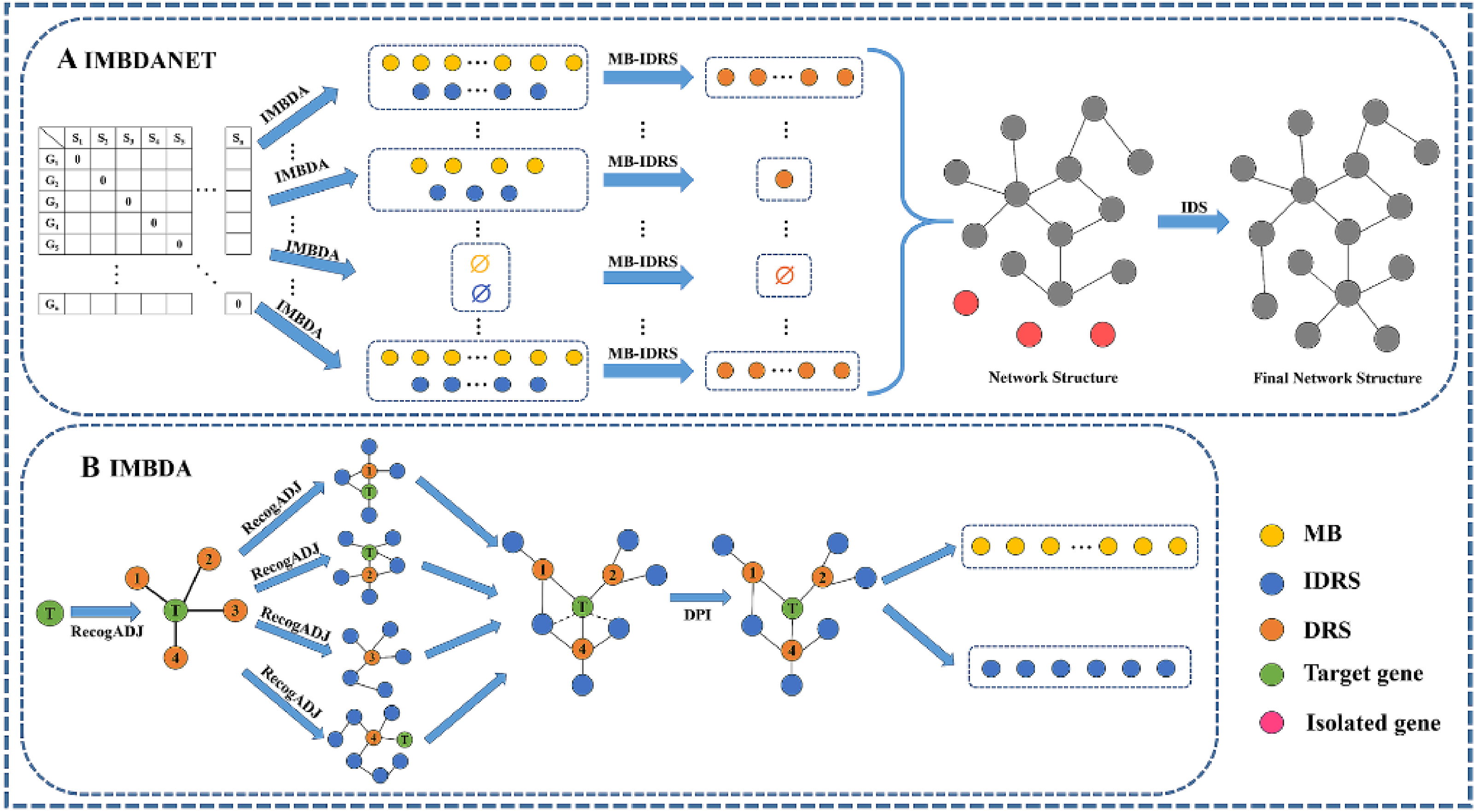
Step 1: IMBDA · Step 2: pruning · **Step 3: fine tuning**

To avoid isolated genes, genes with no neighbors after pruning are connected to the "most relevant genes" according to the IDS metric.

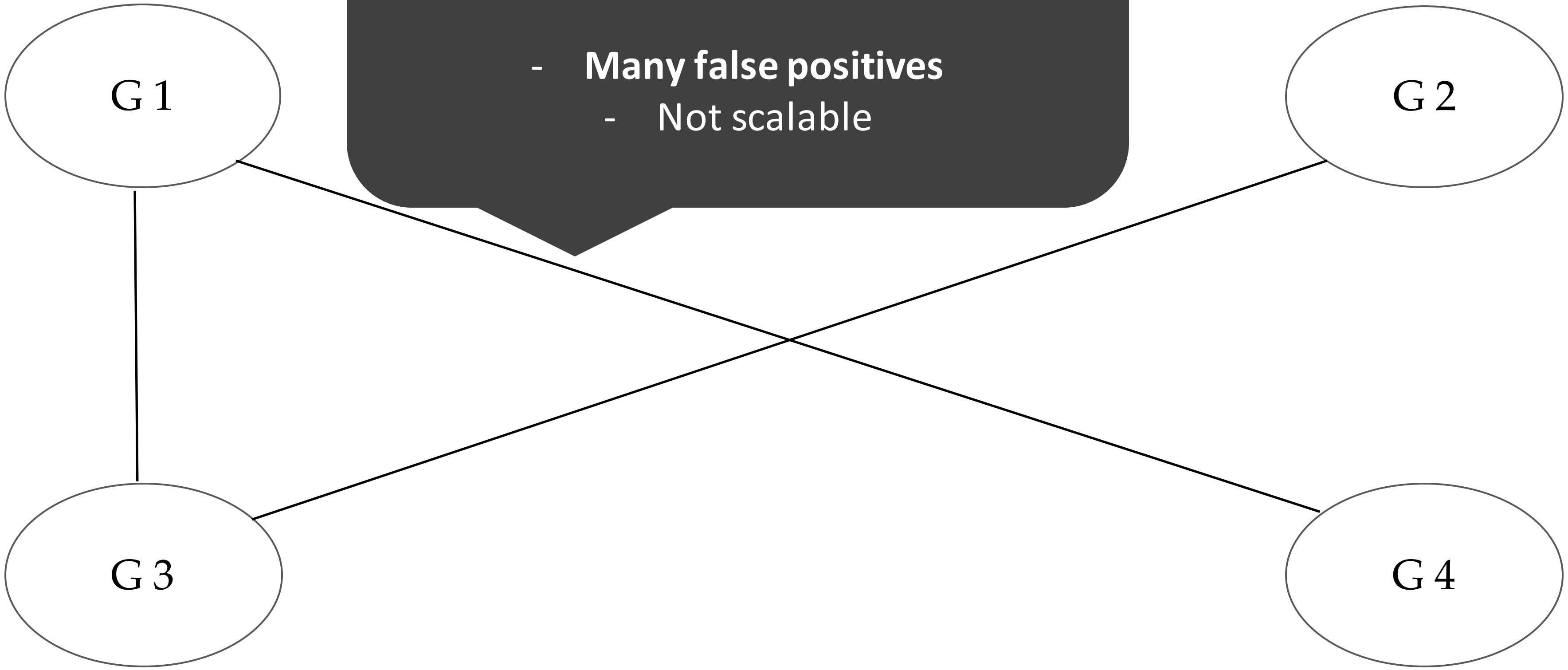
$$\mathrm{IDS}(g_i, g_c) = \sum_{g_c \in ISO; g_i, g_j \in G}^{i \neq j} \left[H(g_c | g_i) - H(g_c | g_i, g_j) \right]^2 \mathrm{MI}(g_i, g_j)$$

IMBDANET

Step 1: IMBDA · Step 2: pruning · Step 3: fine tuning

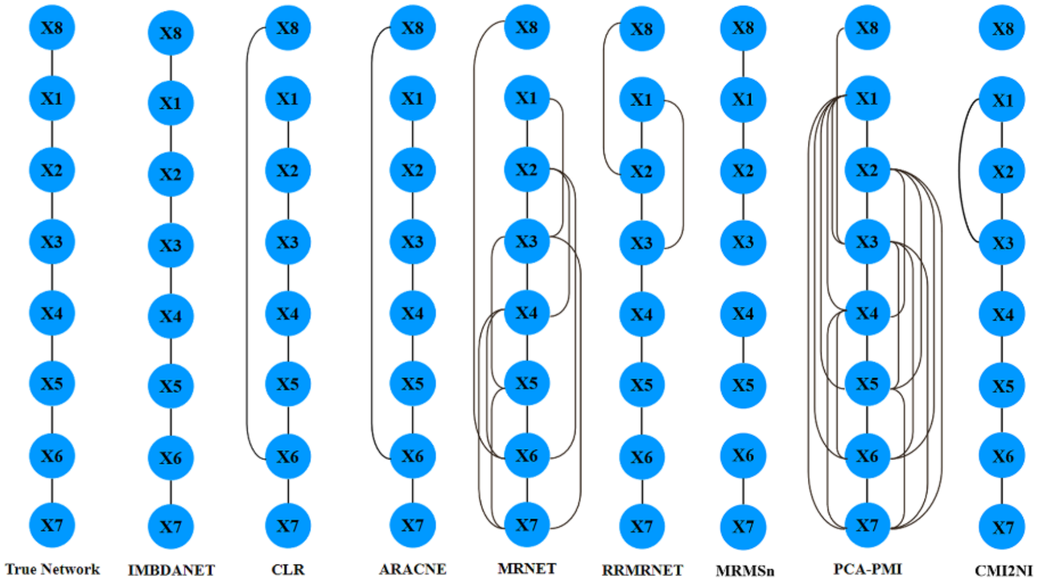


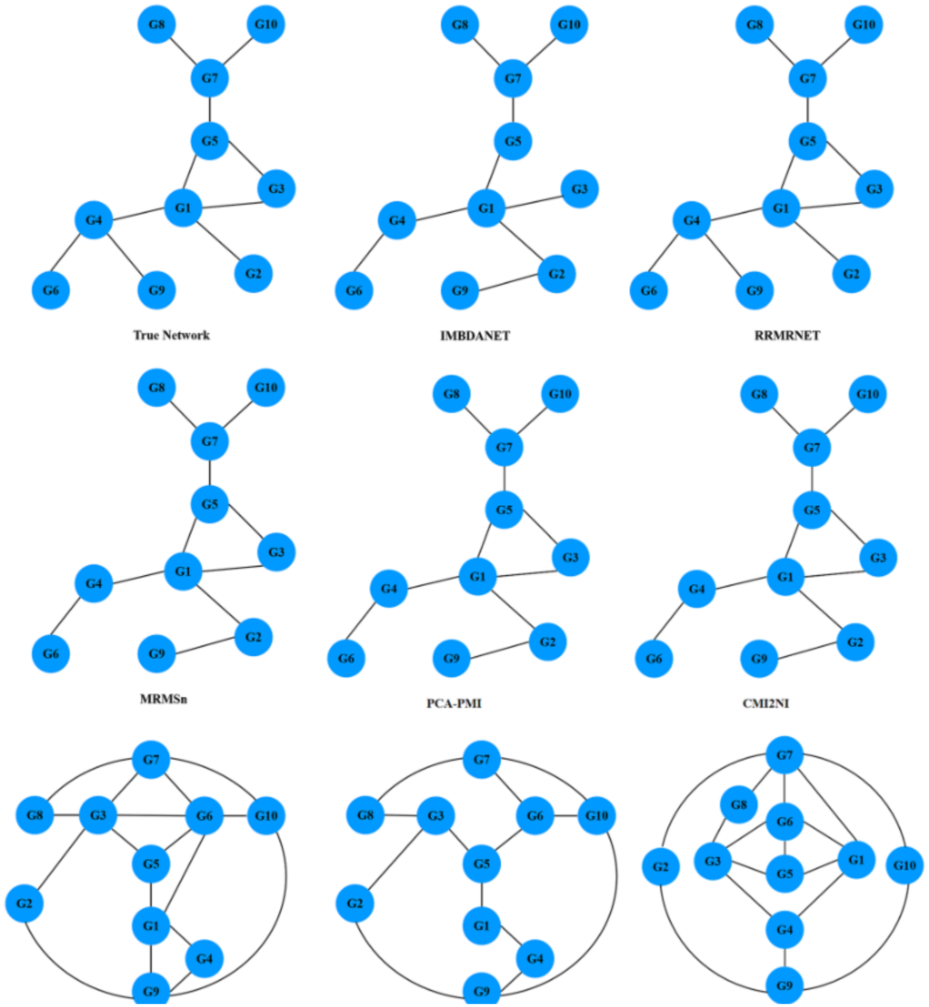
Undirected, unweighted gene co-expression



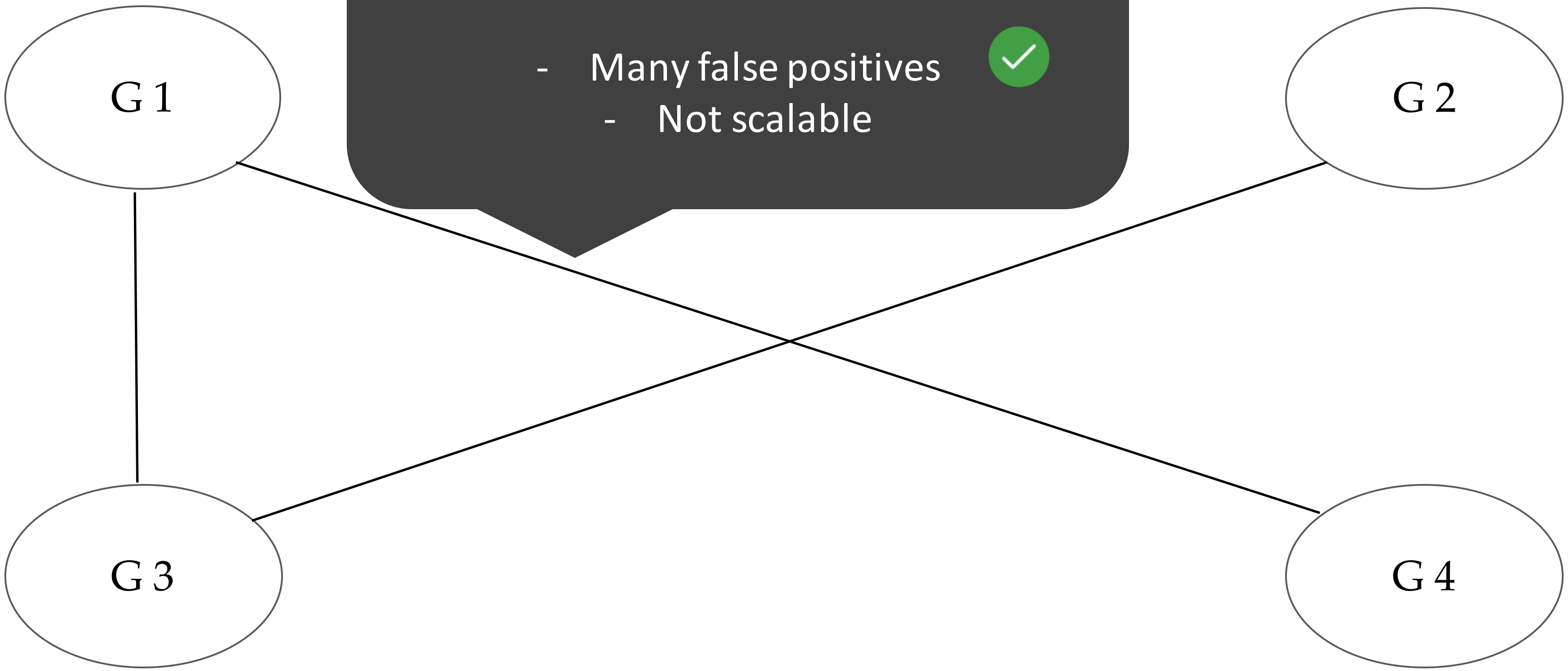
Information theoretic scores suffer mainly from two drawbacks:

- **Many false positives**
- Not scalable


X8
X1
X2
X3
X4
X5
X6
X7
True Network
IMBDANET
CLR
ARACNE
MRNET
RRMRNET
MRMSn
PCA-PMI
CMI2NI


G1
G2
G3
G4
G5
G6
G7
G8
G9
G10
True Network
IMBDANET
RRMRNET
MRMSn
PCA-PMI
CMI2NI

Undirected, unweighted gene co-expression



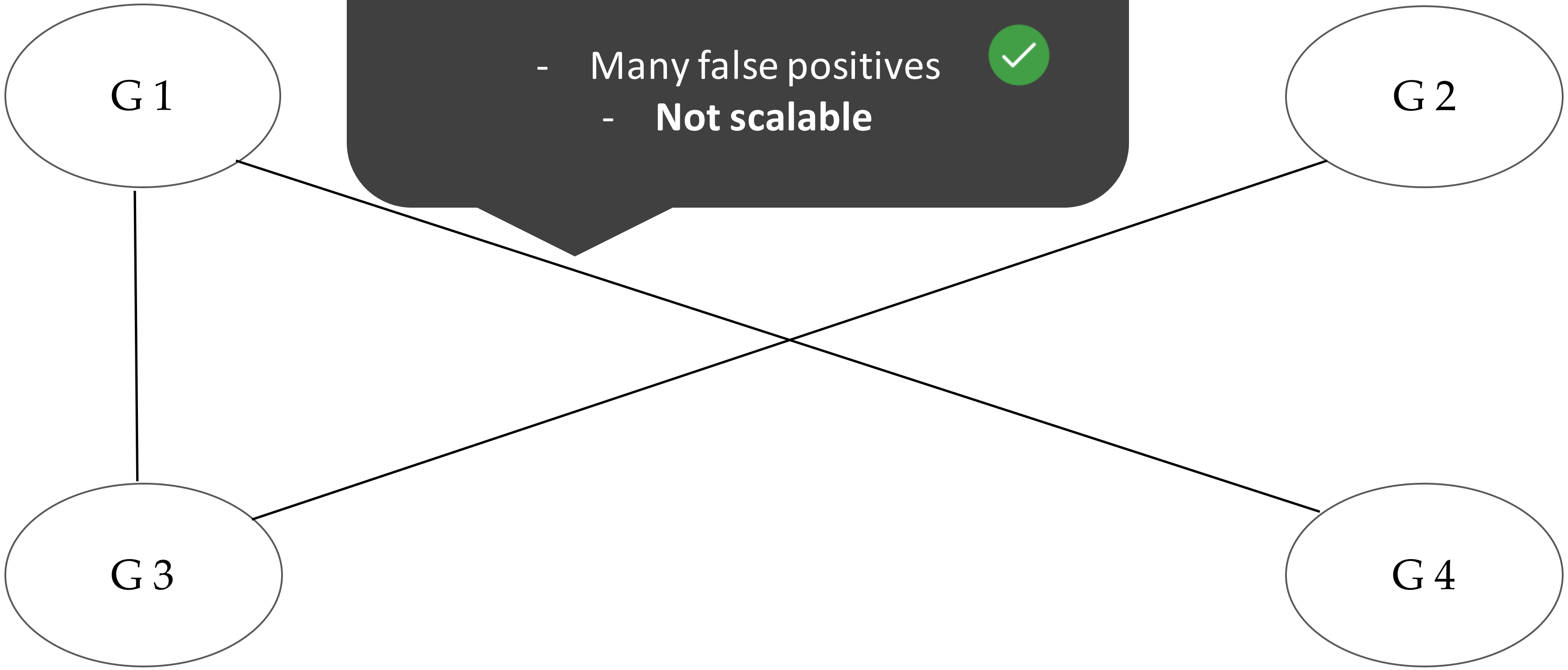
Information theoretic scores suffer mainly from two drawbacks:

- Many false positives ✓
- Not scalable

Undirected, unweighted gene co-expression

Information theoretic scores suffer mainly from two drawbacks:

- Many false positives ✓
- **Not scalable**

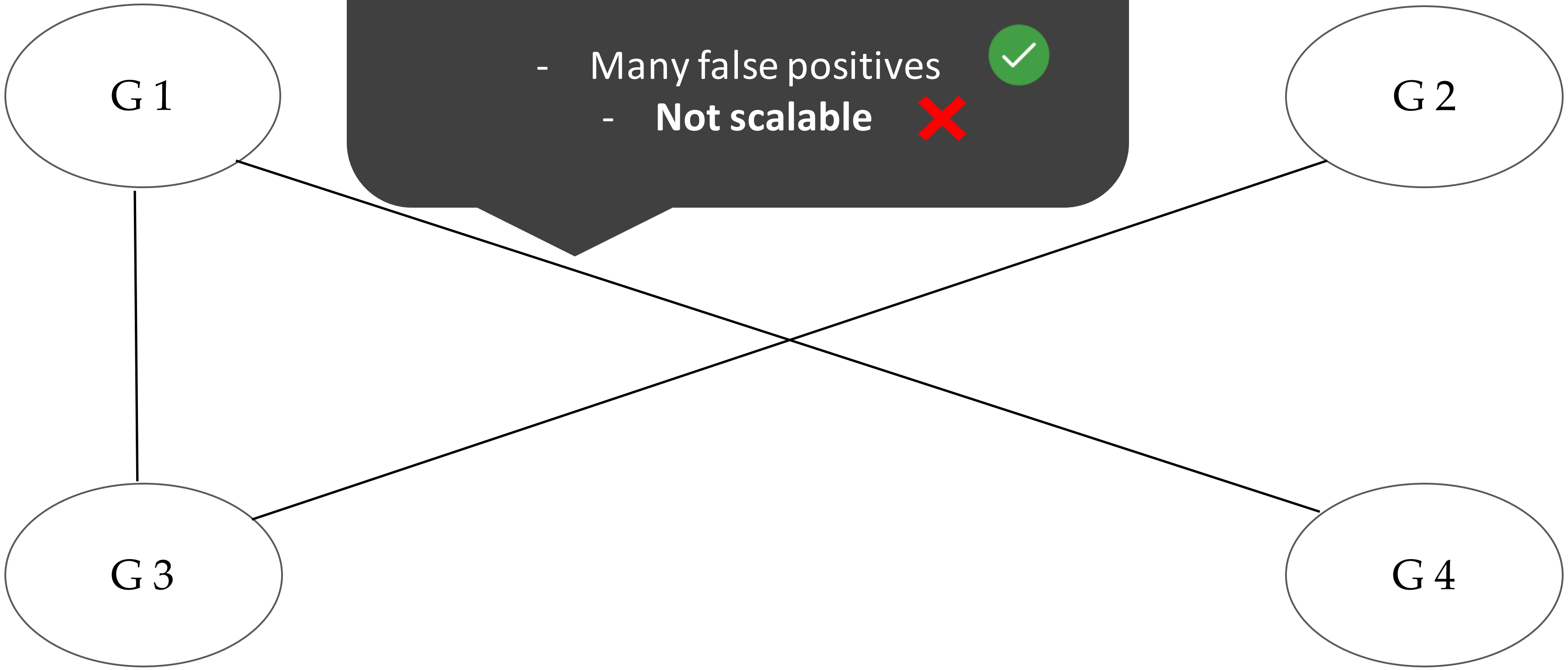


Datasets	Variables	Samples	Type	Nodes	Edges
Reaction chain with four species	4	100	Simulated	4	3
Reaction chain with eight species	8	250	Simulated	8	7
InSilicoSize10-Yeast1-null-mutants	10	10	Simulated	10	10
InSilicoSize50-Yeast1-null-mutants	50	50	Simulated	50	77
InSilicoSize100-Yeast1-null-mutants	100	100	Simulated	100	166
SOS	9	9	Real	9	24

Undirected, unweighted gene co-expression

Information theoretic scores suffer mainly from two drawbacks:

- Many false positives ✓
- **Not scalable** ✗



My takes from the paper

- Interesting application of MBs to GRN
- Some details about the algorithms are not explicitly stated in the paper
- Lack of theoretical insights