Graphical Model Comparison for Gene Expression Networks in Asthma vs Healthy Samples

Yixin Zheng

P8124 Final Project

1 Goal

The main scientific question is: Are genetic regulatory networks really different in different disease categories? What is different? More precisely, our goal is not to infer causal genetic regulation, but rather to compare the estimated undirected association structures, as represented by GGMs, across disease categories and graph-learning methods.

2 Motivation

Studying the differences in genetic regulatory networks between disease and healthy conditions helps us understand how asthma changes the body's overall (system-level) biological regulation.

Although changes in individual gene expression have been widely studied, examining how genes interact in conditional dependence networks can reveal patterns of dysregulation that may be missed when analyzing genes one at a time, as in marginal analyses.

Given that in high-dimensional gene expression data network inference depends on the graphical modeling method used, comparing Gaussian graphical models (GGMs) between disease groups and estimation methods helps evaluate the robustness and consistency of inferred network differences. This is crucial for a proper biological interpretation and to help guide future methodological research.

3 Data

3.1 Data source

We performed transcriptome profiling analysis of ArrayExpress data E-MTAB-1425. This dataset contains normalized gene expression from human lymphoblastoid cell lines of children recruited through a proband with asthma.

3.2 Preprocessing and exclusions

The normalized expression matrix contains 54,675 gene features across 395 samples. Sample-level annotations for these 395 samples can be found in the SDRF file.

1. Sample selection We then cleaned the SDRF annotations and restricted to samples labeled as asthma or normal (dropping other/unknown). The 'other' category was omitted for its small size and unclear disease definition.

2. Dimensionality reduction Next, a subset of $p = 800$ genes with the highest variance in the pooled data are kept. This dimension reduction improves the stability of high-dimensional GGM estimation when $p \gg n$.

We use this fixed set of p genes to ensure that all subsequent comparisons (i.e. methods, tuning, and permutation inference) are made focusing on the same feature

Standardization We next normalize each gene with the pool mean and standard deviation obtained from all asthma and normal samples. We scale genes to the same level such that any associations estimated are independent of scale.

Final datasets Then, we split the scaled matrix into asthma and normal groups for model fitting.

Table 1 shows the final sample sizes.

Table 1: Sample sizes after exclusions

Group	n	Notes
Asthma	258	disease=`asthma`
Healthy	134	disease=`normal`

4 Methods

4.1 Overview and comparison design

An undirected graph is built on $p = 800$ variables using 2 approaches:

(i) graphical lasso (sparse precision matrix estimation)

(ii) FastGGM (scaled-lasso based inference of partial correlations).

To allow for a clearer comparison we control the sparsity of the networks by matching graph density.

For each method, we tune settings so that the final networks for each disease group contains approximately 800 undirected edges. This strategy helps prevents signal of disease effects to be confused as mere differences in edge counts.

We selected the target edge count $K = 800$ to ensure that all resulting networks turned out sparse enough to avoid nearly complete graphs, but also dense enough to retain the hub structure, and allow for meaningful overlap comparisons.

4.2 Approach 1: Graphical lasso (Gaussian MRF; parameter estimation)

Within each group, let $X \in \mathbb{R}^{n \times p}$ be the standardized expression matrix and let S be the sample covariance.

The graphical lasso estimates a sparse precision matrix $\widehat{\Omega}$ by solving

$$\widehat{\Omega}(\lambda) = \underset{\Omega \succ 0}{\arg\min} \{-\log\det(\Omega) + \mathrm{tr}(S\Omega) + \lambda\|\Omega\|_{1,\mathrm{off}}\}$$

The tuning parameter λ controls the strength of the ℓ_1 penalty on off-diagonal entries of Ω: larger values of λ correspond to fewer edges, which trade-off between model complexity and sparsity.

For each disease group we fit a regularization path using `huge(method="glasso")` and choose the index of the undirected graph that has an edge count closest to 800. The same λ value may lead groups to have distinct levels of sparsity due to sampling variability and covariance structure. To address this, we define the resulting edge-matched graph by keeping the top K edges, as ordered based on the magnitude of partial correlations implied by $\widehat{\Omega}$, where $K = 800$. This step makes the network structure comparable across disease groups with the relative strength order of estimated conditional dependencies by Glasso kept.

4.3 Approach 2: FastGGM (partial correlation inference; graph construction)

FastGGM produces (i) estimates of pairwise partial correlations and (ii) corresponding p-values for conditional dependence tests based on scaled-lasso regression (Wang et al. 2016). In a Gaussian graphical model, partial correlations are given in terms of their precision matrix Ω via

$$\rho_{ij\cdot\mathrm{rest}} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\,\omega_{jj}}}, \qquad i \neq j,$$

so an edge corresponds to statistical evidence that $\rho_{ij\cdot\mathrm{rest}} \neq 0$

FastGGM fits nodewise scaled-lasso regressions using a theoretically motivated default penalty parameter

$$\lambda = \sqrt{\frac{2\log(p/\sqrt{n})}{n}},$$

which is tuned to sample size n and dimension p and is designed to control false discoveries in high-dimensional settings. As n is different in disease groups, the obtained λ (and hence the raw graph densities) varies between groups.

In theory, edges can be chosen by using Benjamini–Hochberg false discovery rate thresholding to the FastGGM p-values. Yet, for our data BH-FDR with $\alpha = 0.10$ returns large differences in graph densities between diseases groups (9264 edges in asthma versus 3912 in normal). To ensure that subsequent differences between network topologies rather than due to trivial differences in number of edges, we also generate density-matched graphs which preserves the top-K edges with smallest FastGGM p-values for each group of size $K = 800$.

Even though in theory the BH-FDR control is appealing, in practice we encountered a severe imbalance between FDR-selected edge counts across groups. So comparing the topologies directly will give us an incorrect answer. The Top-K thus embodies a deliberate trade-off: giving up direct interpretation of error-rate in favor of controlled structural comparison.

4.4 Network summaries

For each group and method, we compute:

i) edge count

ii) degree distribution

iii) overlap between asthma and healthy graphs within a method,

iv) a readable top-degree induced subgraph (for visualization only).

We report edge-set intersection size and symmetric difference size.

4.5 Edge-level statistical test: permutation test for differential partial correlations (Glasso)

To investigate the scientific question at the level of edges, we formally estimate whether or not the conditional associations are distinct between asthma and healthy samples.

Hypotheses For each edge (i, j) in the test set E_{test}, we test

$$H_{0,ij} : \rho^{(\text{asth})}_{ij\cdot-ij} = \rho^{(\text{norm})}_{ij\cdot-ij} \quad \text{versus} \quad H_{A,ij} : \rho^{(\text{asth})}_{ij\cdot-ij} \neq \rho^{(\text{norm})}_{ij\cdot-ij}$$

Together, these hypotheses define a globally defined null condition that is restricted to the hub:

$$H_0^{\text{global}} : \ \rho^{(\text{asth})}_{ij\cdot-ij} = \rho^{(\text{norm})}_{ij\cdot-ij} \quad \forall (i, j) \in E_{\text{test}}$$

Test set Using the edge-matched glasso graphs, we define E_{union} as the union of edges selected in either group.

To reduce multiplicity and align inference with hub-focused network summaries, we restrict testing to edges incident to hub nodes.

Let H denote the union of the top-30 nodes by degree in each group, and define

$$E_{\text{test}} = \{(i, j) \in E_{\text{union}} : \ i \in H \text{ or } j \in H\}$$

If this filtering does not produce any edges, we set $E_{\text{test}} = E_{\text{union}}$

Test statistic and permutation procedure For each $(i, j) \in E_{\text{test}}$, we calculate the observed difference in partial correlation:

$$\Delta^{\text{obs}}_{ij} = \rho^{(\text{asth})}_{ij\cdot-ij} - \rho^{(\text{norm})}_{ij\cdot-ij},$$

where partial correlations are derived from the estimated precision matrices.

We then perform a label permutation test under the null hypothesis that asthma/normal labels are interchangeable (as justified by pooled scaling): shuffle disease labels, refit glasso during each permutation split using fixed λ (to avoid model complexity drifting across permutations), and calculate $\Delta^{(b)}_{ij}$.

The two-sided empirical p-value is

$$p_{ij} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\left(|\Delta^{(b)}_{ij}| \geq |\Delta^{\text{obs}}_{ij}|\right)}{B + 1}$$

Sensitivity to λ and multiple testing In order to adjust for λ sensitivity, we calculate the empirical p-values with the asthma and normal selected values and merge them conservatively combine them by

$$p_{ij} = \max\Big(p_{ij}^{(\lambda_{\text{as}})},\, p_{ij}^{(\lambda_{\text{no}})}\Big)$$

We control false discovery rate across tested edges using Benjamini–Hochberg (BH) adjustment.

Computational considerations In practice we used $B = 200$ permutations and obtained an empirical p-value resolution of $1/(B+1) \approx 0.004975124$. This provides enough power to moderate size effects and large effects, yet minimizes the sensitivity to very small edge-wise differences.

5 Results

5.1 Exploratory structure

Figure 1 shows the correlation heatmap of the top 100 high-variance genes. The bright diagonal is expected, as each gene has a correlation of 1 with itself. Off-diagonal cells tend to have mid-level colors, indicating near-zero correlations. Local lighter blocks with moderate positive correlation among small sets of genes are few, and scattered darker 'streaks' indicate negative correlation among a few gene pairs.

The overall dominance of near-zero off-diagonal correlations implies that strong marginal dependence is limited to modules of relatively small size. This result corresponds to what one would expect from a sparse underlying Gaussian graphical model, with many gene pairs exhibiting marginal weak association after conditioning on the other genes.

5.2 Tuning and density matching

Figure 2 shows the glasso edge-count path. The curves rise up rapidly at the beginning, such that tiny perturbations in path index can lead to big changes in the edge counts.

After that point, the curve levels out, and the edge counts changes more gradually. This indicates that sparsity is sensitive to tuning in some regimes, a small change in the index can quickly densify the graph.

The horizontal reference line at 800 represents the target of matching density. For each group, the path index is selected so that the graph obtained is as close as possible to this value.

Here, density matching is used so that later differences in overlap and degree profiles are not solely related to change in sparsity.

5.3 Network-level differences across disease groups

Figure 3 compares degree distributions after density matching ($K = 800$).

In both analytical approaches, the resulting distribution is heavily skewed to the right: Most nodes have very small degrees (around 0 - 5), while a small number of nodes form a long tail with higher degree, which implies hub-like behaviour.

When FastGGM (TopK) is used, the asthma and normal group curves are very similar in the bulk. However, the normal group shift a little more to the right, and has a slight more mass at low-to-moderate degrees (around 2 to 7), consistent with a mild redistribution of connectivity (as opposed to pooling into one hub).

When Glasso (|pcor| Top) is used, the tail behavior differs more visibly: now, asthma curve exhibits slightly more mass in the degree range around 10 - 20, while normal shows a small bump around very high degrees (around 30), suggesting that differences may appear through how connectivity is distributed into hubs, few stronger hubs or many mid-sized hubs, even when the overall edge count stays constant.

Table 2 presents the edge overlap summaries for the density-matched networks. When the number of edges is fixed, a larger bar means there are more shared edges and greater stability between the disease groups. Both groups have 800 edges in each approach, so the intersection size $|E_\cap|$ directly measures how stable the estimated edge set is across disease groups.

Glasso showed much higher overlap (570/800 $\approx 71\%$) than FastGGM (238/800 $\approx 30\%$).

FastGGM resulted in much lower cross-group edge overlap under density matching. This suggests there are larger differences in the estimated conditional dependence structure between asthma and normal samples.

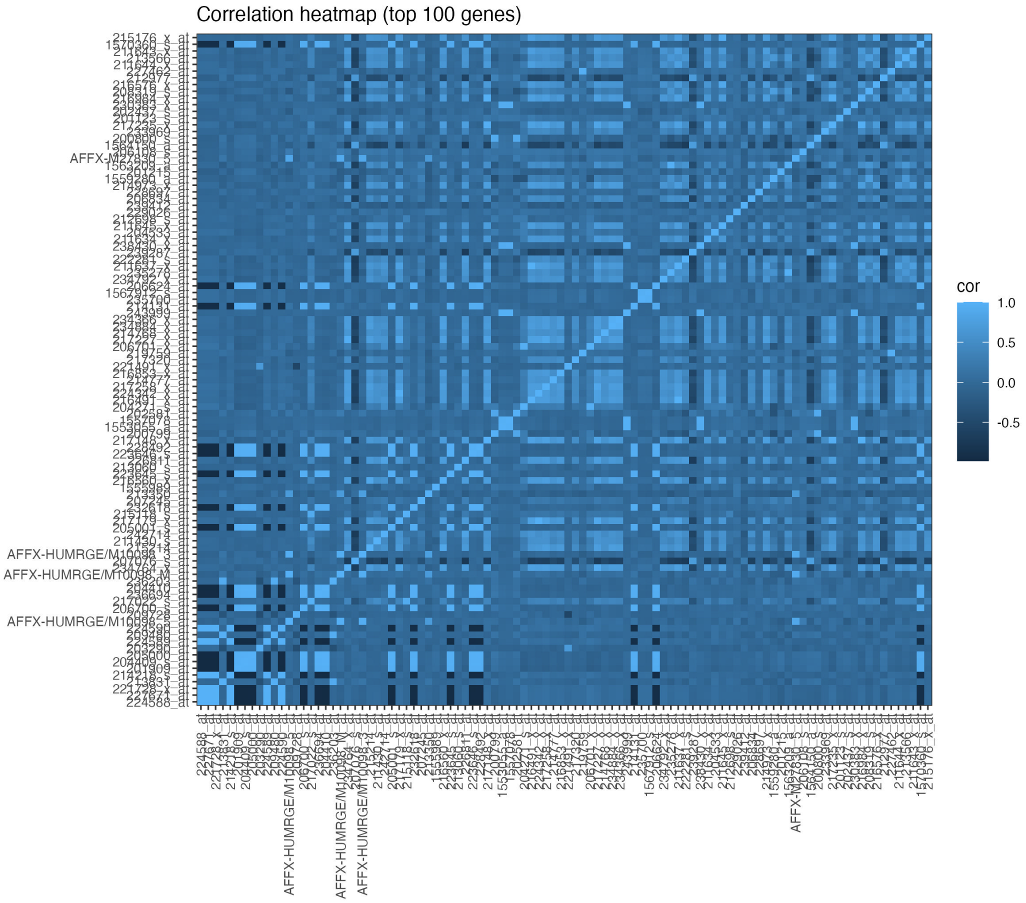


Figure 1: Correlation heatmap (top 100 high-variance genes) for EDA

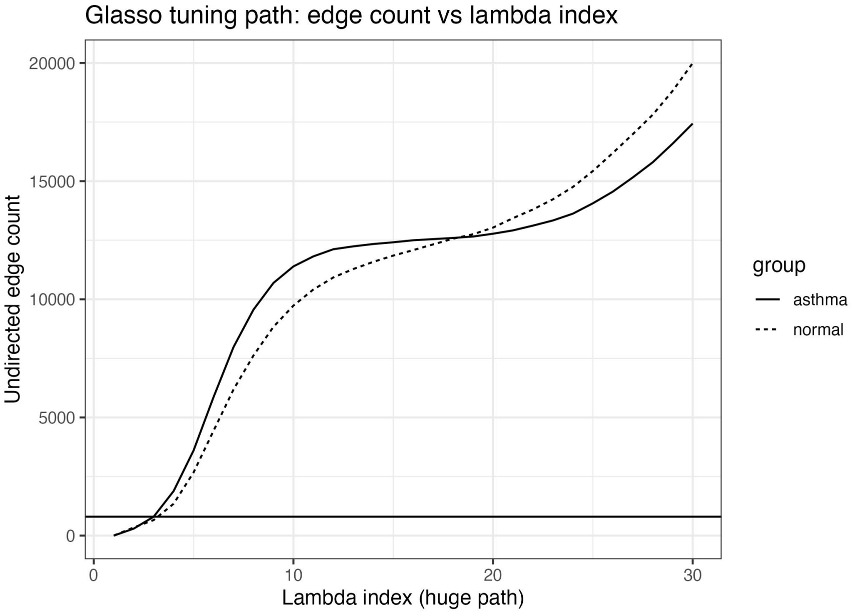


Figure 2: Glasso tuning path: undirected edge count vs. λ

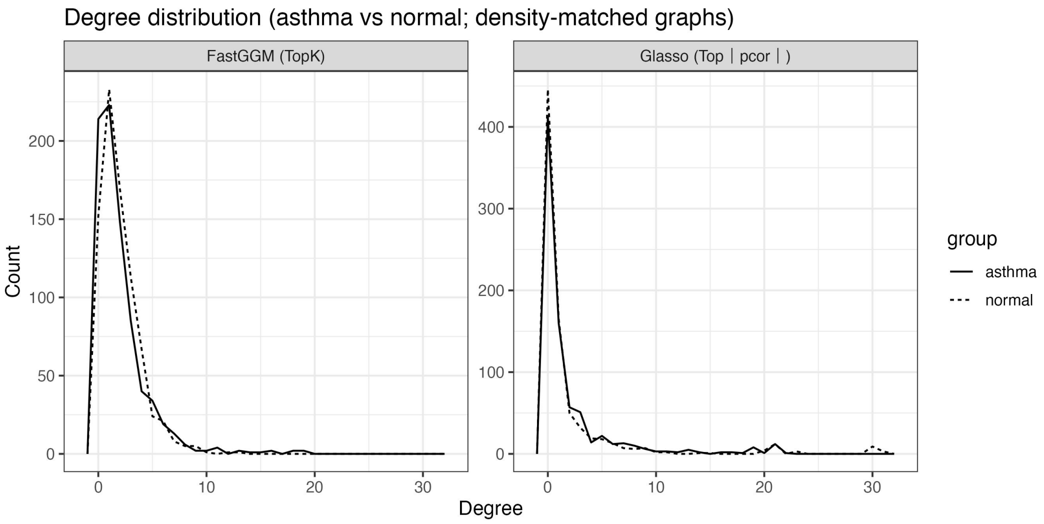


Figure 3: Degree distributions by disease group and method

These differences may be due to biological heterogeneity or because the methods vary in how sensitive they are to weak partial correlations.

Similarly, the symmetric difference $|E_{\triangle}|$ is larger for FastGGM, which means a greater fraction of edges are specific to each group.

Table 2: Edge overlap between asthma and normal graphs (density-matched at $K = 800$)

Method	$\lvert E_{\text{asth}}\rvert$	$\lvert E_{\text{norm}}\rvert$	$\lvert E_{\cap}\rvert$	$\lvert E_{\triangle}\rvert$
FastGGM (TopK)	800	800	238	1124
Glasso (Top|pcor|)	800	800	570	460

5.4 FastGGM FDR edge counts (motivating density matching)

Table 3 shows the raw FastGGM BH-FDR graphs without density matching, which helps keep the analysis transparent. At $\alpha = 0.10$, the asthma graph is much denser than the normal graph, and the intersection (929) is still small compared to the number of edges in the asthma graph. Because of this big difference, comparing the graphs directly could be misleading unless density is controlled. This is why the Top-K density-matching strategy described above is used.

Table 3: FastGGM BH-FDR graphs (no density matching) at $\alpha = 0.10$

	$\lvert E_{\text{asth}}\rvert$	$\lvert E_{\text{norm}}\rvert$	$\lvert E_{\cap}\rvert$	$\lvert E_{\triangle}\rvert$
FastGGM (BH-FDR)	9264	3912	929	11318

5.5 Readable network visualizations

For each method and condition we also present a top-degree induced subgraph (Figure 4) to make the results more understandable/ easier to read. There are two main patterns. In the glasso subgraphs (top), asthma and normal exhibit a node structure in the central part that is more compact, with some small satellite nodes. The normal panel has a more obvious hub-and-spoke pattern, with a tighter central hub and fewer large secondary hubs. The asthma panel, in contrast, has a higher concentration of nodes moderately large around the core.

The connections are more evenly situated in the FastGGM subgraphs (bottom row).

Both the asthma and normal panels have lots of small and medium sized nodes within components, which agrees with FastGGM's tendency to draw more group-specific edges at equivalent density.

Note: these plots are for pattern recognition not formal inference. As node labels are not displayed, the description focuses on hubs and components instead of specific gene names.

5.6 Edge-level statistical test: permutation results (Glasso)

We evaluated differential partial correlations on a subset of edges called the hub-filtered test set, which is defined as the subset of edges that are incident to the union of the top-degree nodes in two edge-matched glasso graphs.

This hub-filtered test set consisted of 518 edges.

Using $B = 200$ permutations and BH FDR correction, the smallest empirical p-value observed was $\min p_{\text{emp}} = 0.004975124$ (the resolution limit $1/(B+1)$ with $B = 200$), while the smallest BH-adjusted q-value was $\min q_{\text{FDR}} = 0.3964791$. No edges met a conventional FDR threshold (e.g., $q \leq 0.10$).

We therefore fail to reject the hub-constrained global null of no difference edges following adjustment for multiple testing. This probably shows weak edge-wise effects are spread across many gene pairs, or that there is limited power because the search space is high-dimensional. Results were robust to the choice of tuning parameter λ such that empirical p-values were were combined conservatively across group-specific λ. This method favors control of false positives over power.

6 Conclusions

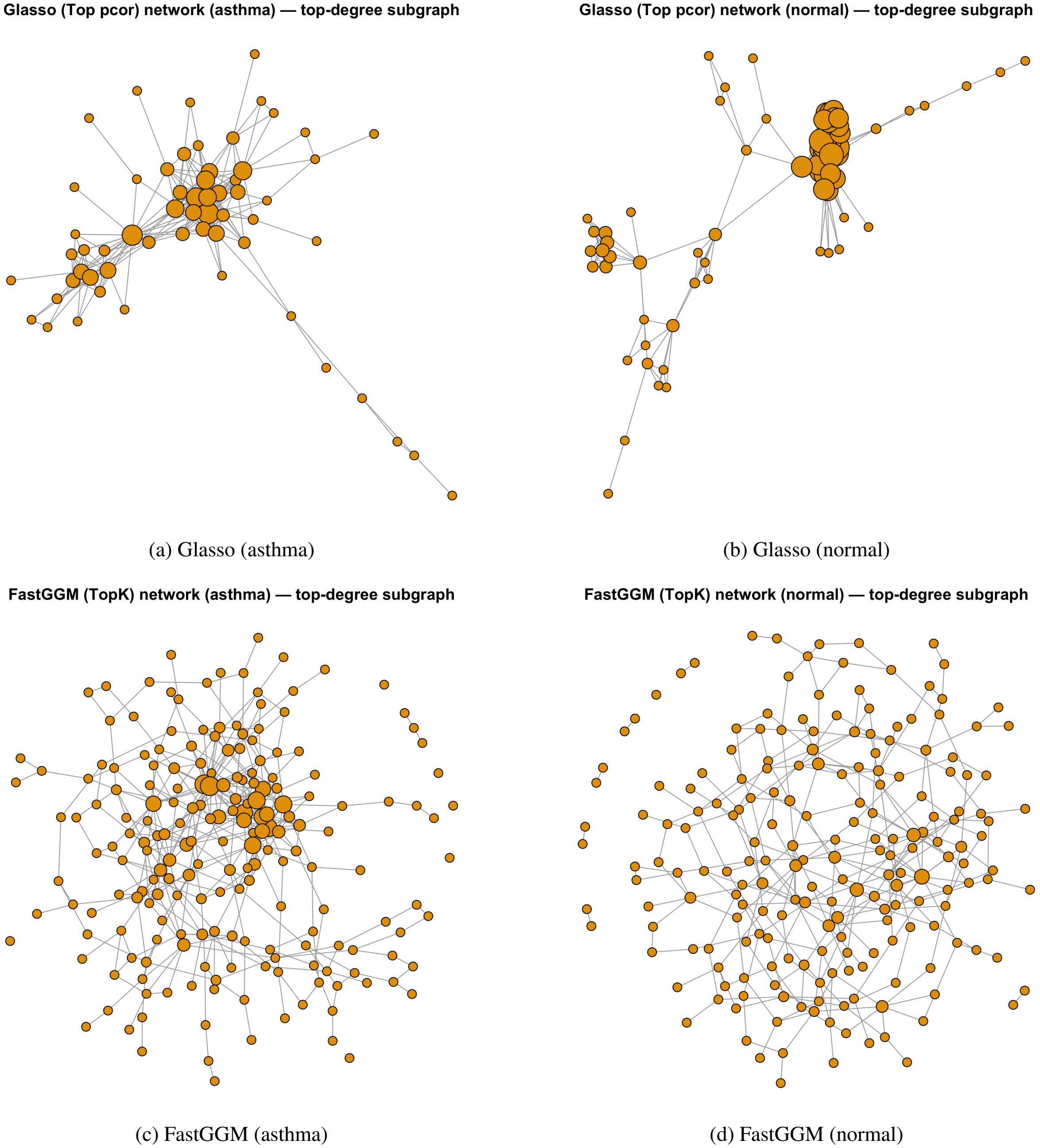
We applied two methods for learning Gaussian graphical models (Glasso and FastGGM) to the same high-variance gene subset, network-wise rather than edge-wise differences between asthma and healthy populations can be identified. After the graph density is equated, then we also observe a few consistent bias in the distribution of degrees and the hub arrangement per disease groups.

Glasso exhibited considerably larger cross-group edge overlap (570/800 $\approx 71\%$), which indicating a more stable conditional dependence structure. Compared to this, FastGGM with much lower overlaps (238/800 $\approx 30\%$), which indicates higher degree of network rewiring between two groups (asthma and healthy) with the same sparsity. Similarly, these discrepancies is also evident in the overall degree profile and top-degree subgraph visualization from which it is apparent that the connectivity among hubs tends to be distributed, rather than merely modified few major edges, depending on disease states.

At the edge level, no differences were detected even after a permutation test on edges around hubs. After FDR correction, none of the partial correlations were significant. This result implies that disease-specific effects are likely to be little and spread out over many small effects.

Overall, our results suggest that asthma is related to multivariate gene dependence structure. They also show that network-level summaries are more important than edge-by-edge analysis at high dimensions for graphical models.

genetic regulatory networks differ between disease categories, and the difference lies mainly in network organization (edge overlap, degree distribution, and hub structure) rather than in a few individually significant gene–gene interactions.



Glasso (Top pcor) network (asthma) — top-degree subgraph

Glasso (Top pcor) network (normal) — top-degree subgraph

(a) Glasso (asthma)

(b) Glasso (normal)

FastGGM (TopK) network (asthma) — top-degree subgraph

FastGGM (TopK) network (normal) — top-degree subgraph

(c) FastGGM (asthma)

(d) FastGGM (normal)

Figure 4: Top-degree subgraph visualizations by method and disease group.

References

Wang, T., Ren, Z., Ding, Y., Fang, Z., Sun, Z., MacDonald, M. L., Sweet, R. A., & Chen, W. (2016). FastGGM: An efficient algorithm for the inference of Gaussian graphical model in biological networks. *PLOS Computational Biology*, 12(2), e1004755.