THESIS FOR THE DEGREE OF LICENTIATE OF PHILOSOPHY

Resampling in network modeling of high-dimensional genomic data

Jonatan Kallus

CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
University of Gothenburg and Chalmers University of Technology
Göteborg, Sweden 2017

Resampling in network modeling of high-dimensional genomic data
Jonatan Kallus
Göteborg 2017

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
University of Gothenburg and Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Typeset with LaTeX
Printed by Ineko AB
Göteborg, Sweden 2017

Resampling in network modeling of high-dimensional genomic data

Jonatan Kallus

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
University of Gothenburg and Chalmers University of Technology

Abstract

Network modeling is an effective approach for the interpretation of high-dimensional data sets for which a sparse dependence structure can be assumed. Genomic data is a challenging and important example. In genomics, network modeling aids the discovery of biological mechanistic relationships and therapeutic targets. The usefulness of methods for network modeling is improved when they produce networks that are accompanied by a reliability estimate. Furthermore, for methods to produce reliable networks they need to have a low sensitivity to occasional outlier observations. In this thesis, the problem of robust network modeling with error control in terms of the false discovery rate (FDR) of edges is studied. As a background, existing types of genomic data are described and the challenges of high-dimensional statistics and multiple hypothesis testing are explained.

Methods for estimation of sparse dependency structures in single samples of genomic data are reviewed. Such methods have a regularization parameter that controls sparsity of estimates. Methods that are based on a single sample are highly sensitive to outlier observations and to the value of the regularization parameter. We introduce the method ROPE, resampling of penalized estimates, that makes robust network estimates by using many data subsamples and several levels of regularization. ROPE controls edge FDR at a specified level by modeling edge selection counts as coming from an overdispersed beta-binomial mixture distribution. Previously existing resampling based methods for network modeling are reviewed. ROPE was evaluated on simulated data and gene expression data from cancer patients. The evaluation shows that ROPE outperforms state-of-the-art methods in terms of accuracy of FDR control and robustness. Robust FDR control makes it possible to make a principled decision of how many network links to use in subsequent analysis steps.

Keywords: high-dimensional data, sparsity, model selection, bootstrap, genomics, graphical modeling

List of publications

This thesis is based on the work represented by the following papers:

Paper I. **Kallus, J.**, Sánchez, J., Jauhiainen, A., Nelander, S., Jörnsten, R. (2017). ROPE: high-dimensional network modeling with robust control of edge FDR. *Preprint arXiv: 1702.07685, Submitted.*

Author contributions

Paper I. Participated in model development, specified and implemented the model and supporting software, designed and generated simulated data sets, evaluated the method and compared it to other methods, drafted and edited the manuscript.

Acknowledgements

I want to thank everyone who have made this work not only possible, but also worthwhile.

Rebecka Jörnsten for inspiring discussions, your energetic positivity and our enlightening problem solving sessions. Sven Nelander for giving our work meaning by lucidly connecting it to the challenges of cancer research. Erik Kristiansson for helping me reflect on my research studies and not forget about longer term goals. José Sánchez for welcoming me into both work and social sides of the PhD-student life in the friendliest of ways. And also for, together with Alexandra Jauhiainen, participating in the development of ideas in this thesis.

Tobias, Fanny, Viktor and Olle for all our fun lunch conversations about things even more important than research. Mariana for inviting me into the lunch gang. Anna J, Anna R, Claes, Emma, Fredrik, Henrik, Henrike, Ivar, Sandra and all other colleagues for creating a friendly atmosphere.

Fredrik, Olof, Mikael and Devdatt in the machine learning group for inspiring me to apply for a PhD-student position and for sharing your research interests with me.

Rasmus Einarsson for your inquisitive questions. They force me to think more clearly. Robert and Joel for our time as undergrads and for taking on PhD studies. When you did it, then clearly I had to try as well.

Mom and dad for support and guidance since as long ago as I can remember. Daniel for our friendship.

My fiancée Josephine for believing in me and encouraging me to take on challenges. I will always remember that sunny winter day on January 15th 2017.

Contents

1 Background

This thesis treats the understanding of high-dimensional genomic data. From a statistical point of view, a key goal is to be able to make inference regarding the relevance of, and relation between, covariates (i.e. transcripts or other biological compounds). The high-dimensionality poses statistical and computational challenges in itself. Furthermore, the nature of genomic data poses challenges. Due to technical difficulty to collect such measurements, data is noisy and may suffer from unwanted variation caused by differences in laboratory procedures. Complex interactions between covariates, such as feedback loops and non-linear dependencies, calls for rich statistical models, exacerbating the challenge of high-dimensionality.

The hope is that novel statistical methods will further contribute to the understanding of the systems biology of living cells. Associations discovered in genomic data are used to form hypotheses for biomarkers for improved disease diagnosis or development of disease treatments. In order for statistical results to be useful as a guide for biological research it is important that they are accompanied by estimates of variance and accuracy.

The thesis is structured as follows. This chapter gives, first, a brief background of genomic data; what it is, why it is interesting to collect and analyse, and a description of different types of genomic data. Thereafter, the difficulties of high-dimensional data are introduced, as well as the problem of multiple hypothesis testing. Lastly, methods for finding associations in genomic data are reviewed. The second chapter defines and limits the aims of this thesis. Chapter 3 reviews resampling based methods for network modeling and gives methodological results not included in paper I. Chapter 4 gives a summary of paper I and the software package that was published along with it. Paper I and its supplementary material is included in the thesis.

1.1 Genomic data

The diversity of living organisms, and life's ability to subsist and adapt through inheritance, are astonishing. It is popularly well known that DNA transfers information about the constitution of an organism between parent and offspring. But why are cells within a multi-cell organism so different, when they contain the same DNA? How is the information in DNA put to use? How do cells respond to changes in environment and what has gone wrong when a cancerous cell starts to multiply uncontrollably? All of these questions relate to the biochemical processes taking place within the cell, from DNA transcription to protein synthesis and function (Smith and Szathmary, 2000). Genomic data consists of measurements of the abundance of substances taking part in these processes. Measurements are made on samples of tissue, on cell colonies cultured in laboratories, or, recently on single cells.

The central dogma of molecular biology (Crick, 1970) is the theory that genetic information is primarily transferred in the cell 1) from DNA to DNA through replication, 2) from DNA to RNA through transcription and 3) from RNA to protein through translation. Proteins are complex and diverse molecules responsible for functions within cells. Figuratively, DNA is the blueprint for making proteins. Due to the role of RNA as a messenger, the abundance of a specific RNA molecule corresponds to how actively a specific piece of DNA is being transcribed, and a specific protein is being constructed. A piece of DNA that gets transcribed as a single RNA molecule is called a gene, thus gene expression is measured by RNA abundance (Smith and Szathmary, 2000).

Several types of genomic data is being collected, in addition to gene expression. Variation in DNA between organisms of the same species, or between tissue within the same organism, is measured in terms of 1) single nucleotide polymorphisms (SNP, variation at a single base-pair in the DNA), 2) short insertions or deletions (indels), and 3) copy number aberrations (CNA, longer DNA regions missing or being repeated). Epigenetic marks, responsible for the vast differences between different cell types despite containing identical DNA, are measured by methylation and chromatin immunoprecipitation (ChIP). These measurements capture the type and genomic location of chemical modifications in connection to the DNA. Proteomics, the direct study of protein abundances, is challenging and cannot be conducted with current technology at a genome-wide scale. It is, however, a fast-growing field (Richardson et al., 2016).

For roughly two decades it has been possible to collect gene expression data on a massive scale. First, primarily through microarrays (Schena et al., 1995) and later through RNA-Seq (Wang et al., 2009). Human gene expression data contains measurements of the concentrations of the about 20,000 RNA molecules in a sample of biological tissue. RNA is known to exhibit complex interactions with other RNA molecules and with the DNA. TCGA (The Cancer Genome Atlas Research Network et al., 2013) makes gene expression data from thousands of cancer patients publicly available.

1.2 High-dimensional statistics

A high-dimensional data set is a data set where the number of covariates (variables measured for each observation) is far greater than the number of observations. An example is RNA-Seq gene expression data for the cancer type *glioblastoma multiforme* in TCGA. It contains measurements for 20,530 genes (covariates) in 172 tumour tissue samples from human patients (observations).

The statistical analysis of such data sets has become increasingly important due to the increased ability to collect, store and transfer vast numbers of measurements. Genomics and other areas in computational biology are important examples. For the modeling of a high-dimensional data set, even the simple linear model is too complex. Thus, the complexity of the linear model needs to be reduced further, e.g. by discarding covariates or otherwise constrain the linear model (Hastie et al., 2009).

Linear regression assumes the model $Y = X\beta + \epsilon$, where the response Y and the error ϵ are n-dimensional vectors, the parameters β is a d-dimensional vector and $X \in \mathbb{R}^{n\times d}$ is a matrix of n observations and d covariates. The elements in ϵ are independent, identically distributed, independent of X and have expectation equal to zero. We can think of Y as the gene expression of one gene and X as the gene expression of all other genes. Then β captures association between the gene represented in Y and all other genes. With the most popular estimation method *least squares*, β is estimated by minimizing the sum of squared residuals $(Y - X\beta)^T(Y - X\beta)$. When $d \leq n$, X and Y uniquely determines an estimate of β (assuming that X is full rank). In the high-dimensional case, however, the problem of estimating β is underdetermined. There exist infinitely many β such that $Y = X\beta$ and a single solution does not say anything about the relation between X and Y (Hastie et al., 2009).

To reduce model complexity a constraint can be imposed on β. Common constraints include the l_2-constraint in *ridge regression* $\sum_{i=1}^{d} \beta_i^2 < R$ (Hoerl and Kennard, 1970) and the l_1-constraint in *lasso* $\sum_{i=1}^{d} |\beta_i| < R$ (Tibshirani, 1996). Lasso has the advantage that admissible β that minimize the sum of squared residuals are, in general, such that many elements in β are equal to zero. This property of excluding less relevant covariates from the model is useful for the estimation of relevant covariates in genomic data sets. The lasso optimization is often formulated in the equivalent Lagrangian form

$$\min_{\beta}(Y - X\beta)^T(Y - X\beta)/2 + \lambda \sum_{i=1}^{d} |\beta_i|$$

with the l_1-constraint changed into an l_1-regularization term. The regularization parameter λ corresponds to the constraining parameter R (Hastie et al., 2009).

Compared to unconstrained least squares, lasso has drawbacks. First, the lasso estimate of β depends on a parameter λ. Secondly, the lasso estimation accuracy for β is less well understood (Bühlmann and van de Geer, 2011).

1.3 Multiple hypothesis testing

In mathematical statistics, decision problems are approached using hypothesis testing. When deciding if data supports an association between the expression of two genes, the *alternative hypothesis* that the association is supported is posed against the *null hypothesis* that it is not. If the probability of the observed data, or a more extreme observation, under the null hypothesis is below some threshold the alternative hypothesis is accepted. This probability is called the p-value and the threshold is commonly 0.05 (Rice, 2006). When multiple tests are performed, such as testing the association between a gene and all other genes or even the association between all pairs of genes, the classical framework is unsatisfactory. Since the probability of falsely accepting a specific hypothesis is 0.05 (if the threshold is 0.05 and the alternative hypothesis is false), we have to expect that 5% of all unassociated genes will be falsely deemed as associated. Correctly accepted alternative hypotheses risk being lost among a large number of falsely accepted alternative hypotheses. Instead of focusing on the error probability in a single test it is relevant to control the total number of errors. The family-wise error rate (FWER) is the probability that at least one alternative hypothesis is falsely accepted. The false discovery rate (FDR) is the expected proportion of accepted alternative hypotheses that are falsely accepted (Hastie et al., 2009).

Hypothesis testing relies on an assumption of the distribution of the test statistic under the null hypothesis. In large-scale multiple testing problems where the proportion of alternative cases is less than 10%, hypothesis tests can be improved by using an empirical null distribution. Empirical null distributions are overdispersed relative to a theoretical null distribution, for the following reasons:

the existence of unobserved covariates, correlations that are not accounted for in the theoretical null distribution and the existence of many real but uninterestingly small effects. The use of empirical null distribution makes an important difference in multiple testing, and rich null distributions (in comparison to commonly used theoretical null distributions) are needed to capture overdispersion (Efron, 2004).

When controlling the false discovery rate, a measure of statistical significance called the *q-value* (Storey and Tibshirani, 2003) is useful. While performing multiple hypothesis significance tests, q-values are assigned to each alternative hypothesis so that if all alternative hypotheses with $q < 0.05$ were called significant, an FDR of approximately 0.05 would be achieved. Thus, q-values have the same relation to FDR as p-values have to false positive rate.

1.4 Finding associations in genomic data

Associations in genomic data can be represented as a network, where each gene is represented by a node and nodes are connected by a link if the genes they represent are associated. Such network representations aim to raise the focus from the local associations between pairs of genes to systemic or global properties of the whole group of genes and their interactions. Network models of human gene expressions have proven useful for classification of cancer patients as well as for finding potential target genes for cancer therapies (Pe'er and Hacohen, 2011). Features at the network level that are of biological importance include genes that serve as network hubs and the network distance between them (Jörnsten et al., 2011), as well as the betweenness-centrality of nodes (i.e. network bottlenecks) (Kling et al., 2015). Such features can be predictive of survival time in cancer patients or be cancer type specific (Jörnsten et al., 2011; Kling et al., 2015).

The terms from applied fields (network, node, link) and corresponding mathematical terms (graph, vertex, edge) are used interchangeably in this thesis. A graph is defined by a set of vertices V and a set of edges E, where each edge in E is a pair of vertices in V.

This thesis concerns the estimation, from a genomic data set, of the edge set of a graph where V consists of all covariates in the data set. The estimation problem connects to previous sections 1.2 and 1.3 in that procedures based on lasso are tractable for performing such estimation, while a solution in the framework of multiple hypothesis testing with high statistical power and asymptotically correct error control is desirable.

Estimation of the edge set is a high-dimensional model selection problem. Each potential edge corresponds to a parameter in a regression model. To set a parameter to zero corresponds to not selecting the variable or edge. The lasso and related l_1-norm penalized methods are computationally and performance-wise efficient when sparsity can be assumed. Penalized methods rely on a choice of amount of penalization, an inherently hard problem. The optimal amount of penalization depends on the number of observations and variables as well as several unknown quantities such as noise, true sparsity and variable interdependence structure. It also depends on the intended use for the network model. The choice of amount of penalization corresponds to choice of model complexity in general model selection.

Traditional methods for selecting the amount of regularization, cross-validation and information criteria, are prone to overfit and sensitive to outliers (Jörnsten et al., 2011). When the goal of graphical modeling is interpretation (e.g. biomarker identification, mechanistic understanding) an accurate control of the rate of falsely discovered edges (FDR) is more important than maximizing stability or likelihood.

Many methods for estimation of biological interaction networks have been proposed in literature. Evaluation is usually performed through matching reconstructed networks with known pathways, using some degree of node closeness (e.g. path length) (Kling et al., 2015). The following sections review methods for estimation of interaction networks.

1.4.1 Graphical lasso

Assume that observations follow a multivariate Gaussian distribution, i.e. $X_i \sim N(\mu, \Sigma)$ $\forall i$, where X_i is the ith row of X, μ is the mean vector and Σ is the covariance matrix. Then, if a pair of covariates are conditionally independent given all other covariates, the corresponding element in the precision matrix Σ^{-1} is zero. This allows for the modeling of gene expression data as a graph, where two genes are connected by an edge if their partial correlation is significantly non-zero. The meaningfulness of exact zeros suggests the construction of an estimator of Σ^{-1} that tends to estimate elements to be exactly zero using a lasso penalty. The log-likelihood for $\Theta = \Sigma^{-1}$, partially maximized with respect to μ, is given by $\log\det\Theta - \mathrm{tr}(S\Theta)$, where S is the empirical covariance of X and tr is the trace operator. The graphical lasso estimates a sparse graph by solving

$$\hat{\Theta} = \arg\max_{\Theta \succeq 0} \log\det\Theta - \mathrm{tr}(S\Theta) - \lambda||\Theta||_1$$

where the constraint $\Theta \succeq 0$ means that Θ is constrained to be positive semidefinite and $||\Theta||_1$ is the sum of the absolute values of the elements in Θ. The maximization problem is convex and computationally tractable, although considerably slower to use than the methods that are reviewed next (Friedman et al., 2008; Banerjee et al., 2008).

1.4.2 Neighborhood selection

Neighborhood selection was proposed before graphical lasso but is considerably faster and can be understood as an approximation of graphical lasso. It models each covariate a with all other covariates using lasso

$$\hat{\beta}^a = \arg \min_{\{\beta:\beta_a=0\}} \frac{1}{n}(X_a - X\beta)^T(X_a - X\beta) + \lambda \sum_{i=1}^{d} |\beta_i|$$

where X_i is the ith column of X. The set $\{(i,j) : \hat{\beta}^i_j \neq 0 \vee \hat{\beta}^j_i \neq 0\}$ is the estimated edge set (Meinshausen and Bühlmann, 2006). Compared to graphical lasso, the optimization problem of neighborhood selection is computationally simpler. It is a drawback that it does not impose symmetry in gene associations, i.e. $\hat{\beta}^i_j = \hat{\beta}^j_i$. Symmetry is instead enforced after $\hat{\beta}$ has been computed.

1.4.3 WGCNA

Weighted correlation network analysis (WGCNA) takes a simpler and more direct approach. The correlation coefficient measures linear dependence between covariates. WGCNA estimates the edge set with all pairs of covariates that have an absolute correlation above a threshold τ. The parameter τ takes a role similar to the regularization parameter in lasso. A larger value of τ gives sparser network estimates (Langfelder and Horvath, 2008). This method is even faster than neighborhood selection and it is symmetric due to the symmetry of the correlation coefficient. Compared to graphical lasso and neighborhood selection, WGCNA estimates are local in the sense that the decision of connecting two covariates is based only on observations of these two covariates. The resulting correlation based network is less meaningful than a partial correlation network in the sense that the former cannot distinguish between pairs that are correlated due to dependencies to observed confounding nodes and pairs that are correlated due a direct dependency.

1.4.4 ARACNE

Correlation and linear regression both estimate linear dependencies. Thus the association of gene pairs may be missed or underestimated if their expressions are non-linearly dependent. Algorithm for the reconstruction of accurate cellular networks (ARACNE) uses measures from information theory. The mutual information of two covariates captures dependency, both linear and non-linear. Similarly to WGCNA, the network is estimated by thresholding the estimated mutual information between pairs of nodes. A post-processing step is performed that removes the link in connected triangles that has the least mutual information. This step aims to approximate a network capturing conditional dependencies (Margolin et al., 2006).

2 Aims

This thesis aims to further develop statistical methodology for the understanding of high-dimensional genomic data sets by means of graphical modeling. Several methods for estimating the edge set of such graphical models exist. The aim of this thesis is to go beyond methods that estimate a single graph (point estimates) and also beyond methods that rank edges by how supported they are by data. Instead, resampling and statistical modeling of estimates from many resamples will be used to make estimates that are stable to the existence of outlier observations, and where the false discovery rate of edges is controlled.

Accurate FDR control, in contrast to goals such as likelihood maximization or predictive performance, facilitates biomarker identification, hypothesis generation, mechanistic understanding and comparative modeling, i.e. problem domains where each inferred association needs to be individually trustable (Kling et al., 2015).

The thesis treats the analysis of single data sets. The integrative analysis of several types of genomic data (RNA, CNA, methylation etc.) is not considered. The aim is also limited to exclude special treatment of data sets where observations are divided into different groups representing e.g. patients that have different disease types.

3 Resampling based network modeling

The methods reviewed in section 1.4 are estimators of edge sets of graphs. Given a data set $X \in \mathbb{R}^{n\times d}$ and regularization parameter λ they make an estimate $\hat{S}^{\lambda}(X) \in \{0,1\}^{p}$ of an edge set. With an indexing over all pairs of covariates in X, $\hat{S}_i^{\lambda}(X) = 1$ means that the ith pair of covariates is in the estimated edge set. The number of potential edges is $p = d(d-1)/2$. When using network estimates for making biological hypotheses it is beneficial to have an understanding of the distribution of such estimates. The field of statistical inference concerns the distribution of estimates such as $\hat{S}^{\lambda}(X)$. To what extent can a network estimate be trusted? Are some or all edges strongly influenced by a few of the observations in X or are they representative of an entire population? To what extent can specific properties of the estimated network, or specific locations in it, be trusted? For non-trivial estimators $\hat{S}^{\lambda}$ these questions are difficult to answer, not only due to the unknown distribution of X.

The distribution of $\hat{S}^{\lambda}$ can be estimated using the non-parametric bootstrap or other resampling methods. The non-parametric bootstrap uses the sample X to form new samples with approximately the same distribution as X. A bootstrap sample $R(X)$ consists of n rows drawn randomly among the rows of X, with replacement. The distribution of the estimator $\hat{S}^{\lambda}$ is approximated by applying it to many resamples $R(X)$. In addition to getting an understanding of a specific estimator, this procedure can be used to compare different estimators (i.e. different levels of regularisation for a specific method or different methods). Furthermore, all of the bootstrap estimates $\hat{S}^{\lambda}(R_i(X))$, where R_i is the ith resample, constitutes a new data set that can be used for estimating the network. This route has the potential to improve error control and to improve robustness by decreasing sensitivity to single observations in X. This chapter reviews two such existing resampling based network estimators that are state-of-the-art in terms of control of the false discovery rate (FDR). Paper I contributes a new method for resample based network estimation that has more exact FDR control than existing methods and is considerably more robust than one of the state-of-the-art methods.

Bootstrapping B times and estimating a graph for each bootstrap sample yields B graphs, with equal sets of nodes but different sets of edges. Thus, each potential edge i will have appeared W_i^λ times, $W_i^\lambda \in \{0, \dots, B\}$:

$$W_i^\lambda = \sum_{j=1}^{B} \hat{S}_i^\lambda(R_j(X)) \in \{0, \dots, B\}$$

As suggested by the superscript on W, a specific estimator $\hat{S}^\lambda$ transforms a matrix X to a vector while a method $\hat{S}$ corresponds to a vector of functions of λ describing each edge's response to regularization. Figure 3.3A captures the former. It shows a histogram of how many edges that was selected k times for a specific λ. Figure 3.3B captures the latter. It shows how edges respond to varying regularization.

Simple ways to estimate a network using selection counts W^λ would be to include all edges with $W_i^\lambda > 0$ (edges selected in at least one resample) or edges with $W_i^\lambda = B$ (edges consistently selected in all resamples) or something in between (e.g. edges selected in a majority of resamples). Stability selection and bootstrap inference for network construction (BINCO), reviewed in sections 3.1 and 3.2, are more sophisticated. Resampling of penalized estimates (ROPE) introduced in paper I is first to model the sequence W_i^λ with a probability distribution. Furthermore, these three methods address the problem of selecting amount of regularization λ, by using several W^λ corresponding to a range of λ values.

3.1 Stability selection

Stability selection uses the maximum selection count for each edge over the entire range of λ values. Meinshausen and Bühlmann (2010) derive an upper FWER bound for a threshold k_t where all edges for which $\max_\lambda W_i^\lambda > k_t$ are selected (figure 3.1). It is shown in paper I that the achieved FWER is, in many cases, far below the FWER bound. This results in too conservative choices of k_t and, in turn, too sparse network estimates.

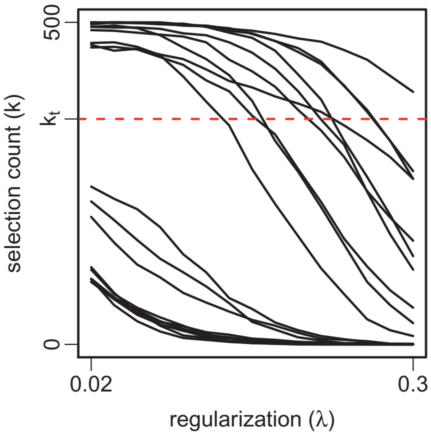

Figure 3.1: Edge selection counts k after 500 bootstraps over varying penalty parameter λ. A random subset of all edges are shown. The stability selection threshold is shown with a dashed red line. Stability selection selects all edges whose count is above a threshold k_t for at least one λ.

3.2 BINCO

BINCO (Li et al., 2013) estimates the null hypothesis distribution of edge selection counts for each value of λ (figure 3.2). The histogram estimates the distribution of edge counts, but it contains both null and alternative edges. In order to estimate a distribution that only includes potential edges that should not be included in a good network estimate, a range of selection counts is chosen that is dominated by such edges. The choice of a such range is based on the histogram having an approximate U-shape.

It is a good sign when edge selection counts are U-shaped. In an ideal case where the network estimator estimates an identical network for each bootstrap sample, each edge will get a selection count of either 0 or B. In a slightly less ideal case, edges will be selected either a small number of times or almost B times, resulting in a U-shaped histogram. It is often the case that the mode for the distribution of false edges is larger than zero. Therefore, an assumption of U-shape in the entire range is too strong. Instead histograms are assumed to be U-shaped in a range $\{c, \dots, B\}$, $0 \leq c < B$.

Li et al. (2013) states the assumption of approximate U-shape precisely as the *proper condition*. The proper condition is satisfied when the empirical probability density function for edge selection counts is U-shaped in the limit

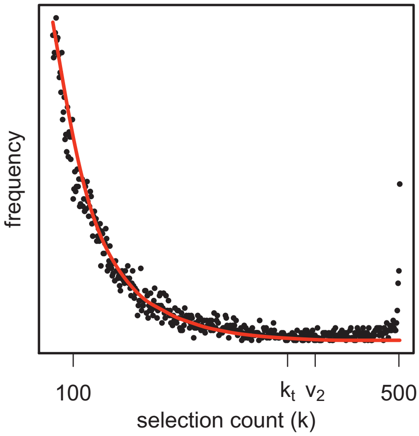

Figure 3.2: Edge selection count histogram after 500 bootstraps corresponding to one λ. The red line shows the null hypothesis distribution as estimated by BINCO. k_t shows the threshold by BINCO corresponding to an estimated FDR of 0.05. v_2 is the location of the minimum of the asymptotic distribution function estimated by BINCO.

$B \to \infty$, i.e. that when restricting the function to this interval, the function has local maxima at its end points, global minimum in the interior of its domain and no other extrema. They show that the proper condition is satisfied by selection procedures for which the selection probability tends to one uniformly for alternative edges and has a limit superior strictly less than one for null edges, as $n \to \infty$. They also show that the condition is satisfied by selection procedures that are based on resampling of consistent selection procedures, such as the lasso when the irrepresentable condition (Zhao and Yu, 2006) is satisfied.

Approximate U-shape is a condition for both BINCO and ROPE. This condition excludes problems where the generating network is either extremely sparse in relation to the variance in network estimates caused by resampling, or where several different edge sets with small mutual overlap captures the data similarly well.

BINCO makes an estimate of the mode of the null population and of the location of the minimum in the U-shaped range. Edge counts in the interval from null mode to minimum location are used to estimate the parameters of a decreasing beta-binomial density function. BINCO uses a modified beta-binomial distribution to capture overdispersion, see Li et al. (2013) or paper I.

Having an estimate of the null hypothesis distribution, and using a decision rule such that edges with selection count above a threshold k_t are accepted, the FDR corresponding to such thresholds can be estimated. The number of false selections is estimated by the mass of the null distribution to the right of k_t multiplied by p, the number of potential edges. The total number of selected edges is given by the sum of the number of edges with each selection count $k > k_t$ (the sum under the histogram to the right of the threshold). Their ratio estimates the FDR.

With a threshold k_t for every λ, each corresponding to the same estimated FDR, a decision is needed for which λ to use. BINCO uses counts from the regularization λ for which most edges are selected. That is, λ is selected to maximize estimated power.

There are drawbacks in using only the decreasing range of the histogram to estimate parameters of the null distribution. Depending on the shape of histograms, thresholds corresponding to relevant false discovery rates are often located outside the range used to fit the model. When that is the case, extrapolation of the fitted model gives an unnecessarily large variance in choice of threshold. Furthermore, presence of the alternative population in the decreasing range, especially its rightmost part, can cause an erroneous estimate of the null distribution.

3.3 Joint modeling across regularization levels

It is reasonable to assume that the distribution of edge selection counts changes smoothly when the amount of regularization is changed. Furthermore, an increase in regularization leads to a sparser network. Thus the mean of the distribution decreases when regularization increases. Lastly, the proportion of potential edges that should be included in a correct network is fixed. Instead of modeling selection counts at each level of regularization individually, these assumptions and facts can be used to fit a model globally. Such a global model can decrease variance in the estimation of model parameters that is caused by the finiteness of the number of bootstraps and observations.

These relationships between selection counts and regularization are illustrated in figure 3.3. The figure also illustrates the relationship between histograms and how the curves of individual edges changes as functions of regularization.

Numerical likelihood maximization for such a global model is challenging. Challenges include the large number of model parameters and a sound and efficient formulation of constraints that enforce smoothness in distribution change. The large number of potential edges and the possibility to perform many bootstraps ensures that there is much selection count data available to fit local models at each regularization level. This suggests that gains from enforcing smoothness across regularization levels are small. In ROPE we enforce the fact that the proportion of edges that should be included in the network are fixed regardless of λ. It is shown in paper I that this constraint decreases bias and increases robustness.

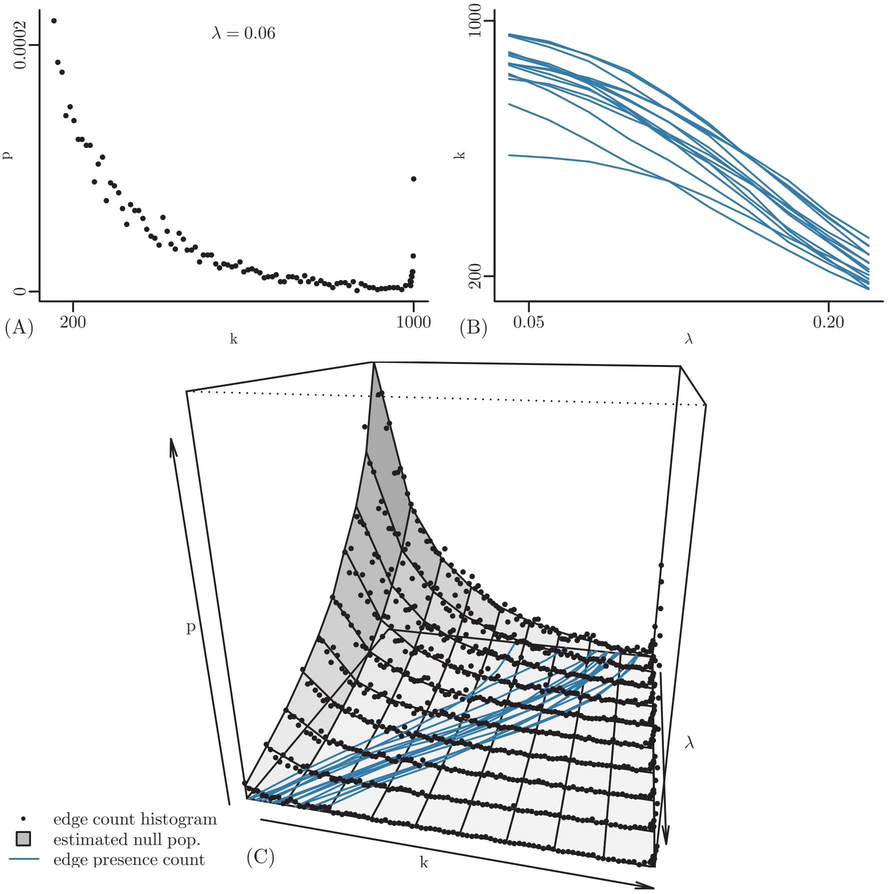

Figure 3.3: Combined two dimensional histogram of edge selection counts and penalization parameter values, along with estimated null population density (C). Edge selection counts for some individual edges are shown as functions of the penalization parameter (B, C). If all individual edges were shown, the density of curves would have corresponded to the height of the histogram. This figure shows the relationship between selection count histograms (A) and curves (B). It also shows how histograms changes smoothly with λ.

4 Summary of paper I

In paper I we introduce the method ROPE for robust network modeling with false discovery rate of edges controlled at a desired level. Like stability selection and BINCO, our method uses bootstrap samples of data to produce multiple network estimates for several values of the regularization parameter. These estimates are aggregated to selection frequencies for all edges and simultaneously analyzed across all levels of sparsity. Unlike previous methods, this global modeling approach is based on a joint beta-binomial mixture of edge selection frequencies. The edge false discovery rate estimates are based on the optimal regularization parameter value that best separates the mixture components ("true" and "false" edges) as well as information about the true level of sparsity obtained from a range of regularization levels. We show that ROPE outperforms state-of-the-art methods in terms of FDR control and robust performance across data sets. The evaluation is performed on simulated data sets and on glioblastoma tumor gene expression data from TCGA.

We propose a statistical model for selection counts, and enable a simultaneous interpretation of selection counts for different levels of regularization. The sequence $\{W_i^\lambda : i = 1, \ldots, p\}$ is modeled as coming from a mixture of beta-binomial distributions, with components capturing either the population of null edges or the population of alternative edges. Fitting this distribution makes it straight forward to choose a threshold $k_t \in [0, B]$ corresponding to a given false discovery rate such that edges i with $W_i^\lambda > k_t$ are declared significant. A range of regularization is chosen to minimize the overlap of mixture components. In this range, the ratio of alternative edges is constrained to be constant (for any regularization λ).

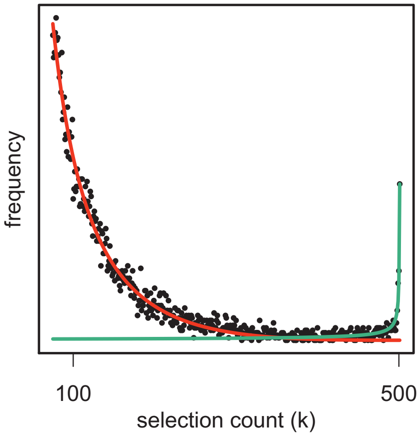

Figure 4.1: Edge selection counts histogram after 500 bootstraps corresponding to one regularization level λ. A mixture distribution with two components is estimated by ROPE. The red line shows the component that estimates null hypothesis distribution. The green line shows the component that estimates the alternative hypothesis distribution. While only the null distribution is needed to estimate the FDR of a selection threshold, having a model that captures both populations decreases bias and avoids several model estimation difficulties.

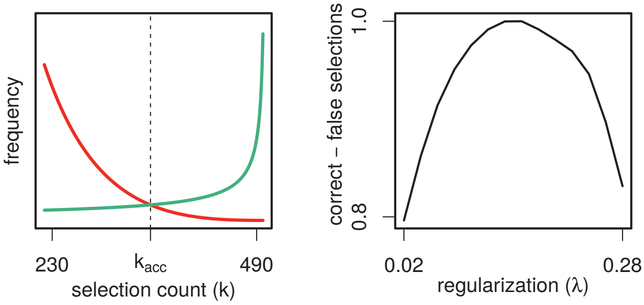

Figure 4.2: Illustration of procedure to choose which level of regularization to use for edge selection. The left panel shows the model fitted to an histogram for one level of regularization. It also shows k_{acc}, the selection count threshold that maximizes accuracy. Assuming the fitted model as truth, the right panel shows the difference between numbers of correctly and incorrectly selected edges. The difference has been normalized to have maximum 1. The procedure estimates how separated (non-overlapping) the two distributions are.

The mixture model

$$\begin{aligned} z|\pi &\sim \text{Bernoulli}(\pi) \\ y_j|\mu_j, \sigma_j &\sim \text{Beta}(\mu_j, \sigma_j),\ \ j = 0, 1 \\ W_i^\lambda|y, z &\sim \text{Bin}(B, y_z) \end{aligned}$$

is fitted to edge counts for each level of regularization λ. The model has five parameters: π the proportion of true edges, μ_1, σ_1 the mean and standard deviation of the probability of true edges to be selected and μ_2, σ_2 corresponding mean and standard deviation for false edges. The model is illustrated in figure 4.1. This model is extended to allow for overdispersion, for details see paper I.

Using the fitted models for each λ, we estimate how separated the two components are. The estimate $g(\lambda)$ is based on the difference between the number of correctly and falsely selected edges, under the fitted model.

$$g(\lambda) = \sum_{k=k_{\text{acc}}}^{B} (f_a(k) - f_n(k))$$

where f_a and f_n are the estimated distributions for alternative and null edges respectively and k_{acc} is the threshold that maximizes accuracy, given these distributions (figure 4.2).

There are two main differences between how ROPE and BINCO model selection counts. First, ROPE uses a model with a higher number of parameters. The big number of potential edges ensures that there is enough data to warrant a richer model. Second, ROPE uses a model that captures the relevant range of high selection counts (i.e. edges that are selected for most or all bootstrap samples). Thus, ROPE's threshold will not be based on an extrapolation from edges with low selection counts, and BINCO's intermediate estimation of a range of decreasing selection counts is avoided.

ROPE, BINCO and stability selection are evaluated with extensive simulation studies under a range of different variable interdependence structures. Results show consistently far more correct FDR control for simulated problems, compared to BINCO and stability selection.

The methods are also compared on public gene expression data from TCGA (The Cancer Genome Atlas Research Network et al., 2013). Selected network sizes and difference between estimates for different subsets of the gene expression data suggests that BINCO fails to control FDR while stability selection is too conservative. In the network selected by ROPE at an estimated FDR of 0.15, we found all hub genes to have documented cancer related functions.

Lastly, in the supplementary material to paper I, we apply ROPE for classification of gene expression profiles according to their primary cancer type, illustrating that ROPE can also be applied to some variable selection problems other than graphical models. There, a multinomial logistic regression model with group lasso penalty is utilized.

4.1 Software package

An implementation of our method is made available as an R package. The package gives support in choosing a regularization range, using visualizations and a heuristic for automatically deciding if histograms are U-shaped. The statistical model is fitted at each regularization step using numerical optimization of the log-likelihood function. In a second round of fitting the model, information from the optimal regularization range is used to make an estimate of mixture component sizes, based on counts from several regularization levels. The package contains several visualizations to examine goodness of model fit. The package is available at The comprehensive R archive network `https://cran.r-project.org/package=rope`.

Bibliography

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Publishing Company, Incorporated, 1st edition.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York. Autres impressions : 2011 (corr.), 2013 (7e corr.).

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Jörnsten, R., Abenius, T., Kling, T., Schmidt, L., Johansson, E., Nordling, T. E. M., Nordlander, B., Sander, C., Gennemark, P., Funa, K., Nilsson, B., Lindahl, L., and Nelander, S. (2011). Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Molecular Systems Biology*, 7:486.

Kling, T., Johansson, P., Sánchez, J., Marinescu, V. D., Jörnsten, R., and Nelander, S. (2015). Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content. *Nucleic Acids Research.*

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559.

Li, S., Hsu, L., Peng, J., and Wang, P. (2013). Bootstrap inference for network construction with an application to a breast cancer microarray study. *Ann. Appl. Stat.*, 7(1):391–417.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., and Califano, A. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(1):S7.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Pe'er, D. and Hacohen, N. (2011). Principles and strategies for developing network models in cancer. *Cell*, 144(6):864–873.

Rice, J. A. (2006). *Mathematical Statistics and Data Analysis.* Belmont, CA: Duxbury Press., third edition.

Richardson, S., Tseng, G. C., and Sun, W. (2016). Statistical methods in integrative genomics. *Annual Review of Statistics and Its Application*, 3:181–209.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470.

Smith, J. M. and Szathmary, E. (2000). *The origins of life.* Oxford university press.

Storey, J. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445.

The Cancer Genome Atlas Research Network, Weinstein, J., Collisson, E., Mills, G., Shaw, K. M., Ozenberger, B., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. (2013). The cancer genome atlas pan-cancer analysis project. *Nat Genet*, 45(10):1113–1120.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563.

Paper I

ROPE: high-dimensional network modeling with robust control of edge FDR

Jonatan Kallus[1], José Sánchez[2], Alexandra Jauhiainen[3], Sven Nelander[4], and Rebecka Jörnsten[1]

[1]Department of Mathematical Sciences, University of Gothenburg and Chalmers University of Technology
[2]Discovery Sciences, AstraZeneca, Gothenburg
[3]Early Clinical Biometrics, AstraZeneca, Gothenburg
[4]Department of Immunology, Genetics and Pathology, Uppsala University, Sweden

Abstract

Network modeling has become increasingly popular for analyzing genomic data, to aid in the interpretation and discovery of possible mechanistic components and therapeutic targets. However, genomic-scale networks are high-dimensional models and are usually estimated from a relatively small number of samples. Therefore, their usefulness is hampered by estimation instability. In addition, the complexity of the models is controlled by one or more penalization (tuning) parameters where small changes to these can lead to vastly different networks, thus making interpretation of models difficult. This necessitates the development of techniques to produce robust network models accompanied by estimation quality assessments. We introduce Resampling of Penalized Estimates (ROPE): a novel statistical method for robust network modeling. The method utilizes resampling-based network estimation and integrates results from several levels of penalization through a constrained, over-dispersed beta-binomial mixture model. ROPE provides robust False Discovery Rate (FDR) control of network estimates and each edge is assigned a measure of validity, the q-value, corresponding to the FDR-level for which the edge would be included in the network model. We apply ROPE to several simulated data sets as well as genomic data from The Cancer Genome Atlas. We show that ROPE outperforms state-of-the-art methods in terms of FDR control and robust performance across data sets. We illustrate how to use ROPE to make a principled model selection for which genomic associations to study further. ROPE is available as an R package on CRAN.

1 Introduction

Large-scale network modeling has the potential to increase our understanding of complex genomic data structures. However, the interpretability of such high-dimensional models are limited by their estimation instability and sensitivity to model tuning parameters. Network modeling is often a preliminary step toward identifying biomarkers for disease stratification or therapeutic targets (e.g. Pe'er and Hacohen, 2011). It is therefore essential that network modeling is accompanied by reliable measures of validity, e.g. false discovery rate of detected edges. Here, we focus on the network modeling of gene expression data, but the methodology is generally applicable to other genomic data sets (Kling *et al.*, 2015). Transcriptional network models aim to identify genes (transcripts) that are directly connected. How connectivity is defined depends on the method utilized. For instance, in *graphical lasso* (Friedman *et al.*, 2008) a network model is obtained through a penalized Gaussian likelihood estimate of the precision matrix (the inverse covariance matrix). Non-zero entries of this matrix identify directly connected genes as those for which the estimated partial correlation exceeds a penalization threshold. Methods like WGCNA (Langfelder and Horvath, 2008) or ARACNE (Margolin *et al.*, 2006) similarly identify connections as those for which a metric of gene-gene association (correlation for WGCNA, mutual information for ARACNE) exceeds a certain penalization threshold. Thus, common to all these methods, the complexity of the estimated network is controlled by a penalization parameter, λ, regulating the sparsity of the estimates.

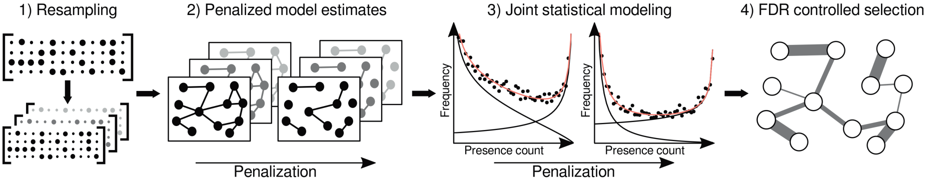
1) Resampling
2) Penalized model estimates
3) Joint statistical modeling
4) FDR controlled selection
Penalization
Frequency
Presence count
Frequency
Presence count
Penalization

Figure 1: Summary of ROPE (resampling of penalized estimates) for network modeling with control of the rate of falsely discovered edges (FDR). 1) The input data is resampled. 2) For each resample, network models are estimated with varying penalization. 3) The number of resamples in which edges are present is modeled as a mixture of "spurious" and "relevant" edges with the mixture proportion jointly estimated across penalization levels. 4) From the mixture model, each edge is assigned a q-value, the minimal FDR target for which the edge is included.

For graphical lasso, much work has focused on estimating the proper penalization for asymptotically consistent selection or optimal bias variance trade off (Meinshausen and Bühlmann, 2010; Liu *et al.*, 2010). Specifically, stability selection (Meinshausen and Bühlmann, 2010) performs model selection based on many subsamples of the data and with different levels of penalization. The method addresses selection of high-dimensional models in general and can readily be applied for selection of network models. An upper bound for the expected number of falsely selected variables (edges), family wise error rate (FWER), is derived. In practice, the estimated bound depends on the range of used penalization levels. Alternatively, one can approach the problem of proper penalization in terms of controlling false discovery rate (FDR) using subsampling or bootstrapping. Bootstrap inference for network construction (BINCO) (Li *et al.*, 2013) models the bootstrap selection frequency for spurious edges, to estimate FDR.

Other methods for selection includes StARS (stability approach to regularization selection) (Liu *et al.*, 2010) which estimates the expected probability of edges to be selected in one subsample and not in another, as a function of the penalization level. This estimate, denoted the instability of variable selection, cannot trivially be extended to control FDR. Bolasso (Bach, 2008) was the first method to combine bootstrapping and the lasso for variable selection and retains variables consistently selected for all bootstrap samples. Results focus on selection accuracy rather than false discovery control.

Here, we introduce Resampling of Penalized Estimates (ROPE) to provide robust FDR control for edge selection accompanied by a measure of validity for each edge: *q-values* (Storey and Tibshirani, 2003). q-values are assigned to each edge so that if all edges with $q < \alpha$ were retained, an FDR of α would be achieved. Thus, q-values have the same relation to FDR as p-values have to false positive rate. This results in a highly interpretable representation where the inferred network is visualized with edge widths corresponding to edge q-value. We show that ROPE outperforms state-of-the-art FDR-controlling methods through comprehensive simulation studies and application to RNA-seq expression data from the Cancer Genome Atlas (The Cancer Genome Atlas Research Network *et al.*, 2013). An easy-to-use R package is provided through CRAN.

This article is structured as follows. This section has introduced the problem at hand. Section 2 provides a detailed description of our method and a comparison with the state-of-the-art. Section 3 evaluates the method with comprehensive simulation studies and includes method comparisons on genomic data from glioblastoma tumors in TCGA. Our method finds several hub genes known to have glioblastoma associated functions, and estimates the validity of each of their connections. Section 4 concludes with the authors' thoughts on the significance of this work and directions for future research.

2 Methods

Variable selection is central to the understanding of high-dimensional data. In network modeling of genomic data, variable selection takes the form of selecting which gene-gene direct interactions (edges) to include. Traditional methods for model selection, e.g. cross validation, are unsatisfactory for high-dimensional problems, due to their tendency to overfit (Jörnsten *et al.*, 2011). Furthermore, measurement errors are expected in genomics data and high-dimensionality makes erroneous observations both influential and hard to filter. Therefore, single model estimates are not informative and resample based methods are needed.

In this article we use neighborhood selection (Meinshausen and Bühlmann, 2006) for network modeling. However, we emphasize that ROPE is applicable to any network modeling where sparsity is controlled by a tuning parameter. Neighborhood selection provides a good approximation of graphical lasso and is computationally faster. It models interactions of a gene j to other genes via the lasso.

$$\beta^j = \arg \min_{\{\beta:\beta_j=0\}} \frac{1}{n}||X_j - X\beta||_2^2 + \lambda||\beta||_1$$

where X is a matrix of n rows (observations) times d columns (genes). The parameter λ is the amount of sparsity inducing penalization. The set $\{(i,j) : \beta_j^i \neq 0 \vee \beta_i^j \neq 0\}$ is the edge set of the inferred network. Note that in network modeling of d dimensional data, the network model consists of $p = d(d-1)/2$ potential edges.

Due to estimation instability, single network estimates have limited interpretability. Therefore, it is advisable to repeat network estimation on resampled data and utilize an estimation aggregate for inference. Here, we use resampling of *randomized lasso* estimates which randomizes the amount of penalization for each individual parameter in different resamples in order to break correlations between variables. Randomized lasso in combination with resampling weakens the so-called irrepresentability conditions that data need to adhere to for consistent selection (Zhao and Yu, 2006). The amount of randomization in Randomized lasso is controlled by a weakness parameter. Weakness 1 corresponds to no randomization, while a lower weakness trades signal strength in data for a lower risk of selecting irrelevant variables (Meinshausen and Bühlmann, 2010).

Introducing some notation, let R_i be a realization of any uniform resampling procedure, most commonly subsampling with sample size $m < n$ or bootstrap, so that $R_i(X)$ is the resampled data set. Let $\hat{S}^\lambda$ be any penalized method for variable selection ($\hat{S}^\lambda(X)$ is the set of variables selected by $\hat{S}^\lambda$ given X). Let $\hat{S}_i^\lambda$ be randomization i of penalization in $\hat{S}^\lambda$. The main algorithmic input of ROPE, stability selection and BINCO is variable selection counts

$$W_j^\lambda = \sum_{i=1}^{B} \mathbb{1}[j \in \hat{S}_i^\lambda(R_i(X))] \in \{0, \ldots, B\} \tag{1}$$

for variable (edge) j over B resamples.

We now present a detailed review of the state-of-the-art FDR-controlling methods BINCO and Stability Selection. BINCO, proposed in Li *et al.* (2013), selects edges with frequency counts W_j^λ exceeding a threshold t. Parameters λ and t are chosen to maximize power while controlling FDR. For each λ, W^λ corresponds to a histogram $h^\lambda(w) = \sum_j \mathbb{1}(W_j^\lambda = w)$ (Figure 1.3). Ideally, this histogram should have two clear modes: at count 0 for spurious (null) edges and count B for the relevant (non-null) edges. For reasonable levels of regularization, $h^\lambda(w)$ is thus "U-shaped". In BINCO, the null model is estimated by fitting a powered beta-binomial distribution to h^λ in the range where h^λ is decreasing in w (defined in Equation 2, Section 2.1). By extrapolation of this null into the range of large frequency counts (dominated by non-null edges), t can be chosen for each λ to control FDR. In practice, the authors found this results in an overly liberal selection and therefore also propose a conservative modification. In conservative BINCO, the density function of the powered beta-binomial distribution is modified to be constant, instead of decreasing, to the right of the estimated minimum of the h^λ-model. This results in a larger t for a given target FDR, thus selecting fewer edges.

Stability selection (Meinshausen and Bühlmann, 2010) selects variables with $\max_{\lambda\in\Lambda} W_j^\lambda > t$ for some threshold t. That is, as long as an edge j has a frequency count exceeding threshold t for any penalization $\lambda \in \Lambda$, it is included in the model. An upper bound on the expected number of falsely selected variables, F, when $t > B/2$ is derived for $\hat{S}_i^\lambda$ randomized lasso and R_i subsampling with sample size $\lfloor n/2 \rfloor$:

$$E(F) \leq \frac{q_\Lambda^2}{(2\frac{t}{B} - 1)p},$$

where p is the number of variables and the expected number of selected variables q_Λ is estimated by $|\Lambda|^{-1} \sum_{\lambda\in\Lambda} \sum_j W_j^\lambda$. In Li *et al.* (2013), an FDR bound is derived from this by dividing both sides by the number of selected variables $\sum_j \mathbb{1}(\max_{\lambda\in\Lambda} W_j^\lambda \geq t)$. This estimate depends, not only on the threshold t, but also on the investigated range of penalization. In Li *et al.* (2013), the combination of t and Λ that selects the maximum number of edges while controlling FDR at the desired level is used.

It is a necessary condition for the applicability of both BINCO (and our method, ROPE) that the histogram h^λ is approximately U-shaped for some λ. Li *et al.* (2013) connect this condition to the irrepresentable condition, showing that satisfaction of the latter leads to U-shaped histograms. In practice, however, the BINCO procedure is sensitive to the histogram shape. First, it is sensitive to correctly estimating the end points of the decreasing range of h^λ, from which the null distribution is estimated. Second, the estimated null distribution is extrapolated into the increasing range of h^λ, where any relevant FDR controlling threshold will be. This extrapolation leads to an unnecessarily large variance for the selected threshold. Third, non-uniform presence of the alternative population (relevant edges) in the decreasing range of the histogram will cause a bias in the estimate of the null distribution. Forth, the authors warn that the method makes a too liberal selection when the minimum of the histogram is to the right of $0.8B$, which easily happens in problems that are sufficiently sparse. Stability selection, while not having the issue of sensitivity to histogram shape, has the limitation that it focuses on a worst-case guarantee, rather than an estimate of the number of false positives.

2.1 ROPE: joint model for resampled, penalized estimates

Recognizing the above limitations of state-of-the-art procedures, we here introduce ROPE, a novel joint modeling of edge presence counts across multiple penalization levels. Figure 1 summarizes the method. Specifically,

1. **Resampling of input data.** B resamples are created by resampling n observations with replacement.

2. **Generation of edge presence counts.** Edge presence counts are collected for several levels of penalization, $\lambda_j \in \Lambda$ (Equation 1). Here, we illustrate ROPE for neighborhood selection in combination with randomized lasso but, as mentioned above, other sparse network models can be used.

3. **Modeling of edge presence counts for each λ, and joint modeling across multiple λs.** We model W_i^λ, for each λ, as coming from a mixture of overdispersed beta-binomial distributions (Equation 3). For improved robustness and accuracy, we leverage the fact that the mixture proportion of null to non-null edges is constant across λ.

4. **q-value assessment and selection of final model.** Integrating information from λs where the modeled null and alternative populations are most separated (Equation 4), q-values are estimated for each edge. FDR is estimated by the probability mass of the null component to the right of threshold divided by mass of the total empirical density to the right of threshold (Equation 5).

In more detail, edge presence counts are modeled as coming from a mixture of overdispersed beta-binomial distributions. Edge selection probabilities depend not only on them being null or alternative but also on, at least, the strength of the dependence between the nodes they connect. This warrants the use of a beta-binomial distribution for each mixture component, where parameters μ represent mean edge selection probability within each component (null/alternative), and σ the variation of dependence strengths within components:

$$f_{\mathrm{BB}}(w) = \binom{B}{w} \frac{\beta(w + \frac{\mu}{\sigma}, B - w + \frac{1-\mu}{\sigma})}{\beta(\frac{\mu}{\sigma}, \frac{1-\mu}{\sigma})},$$

where β is the beta function.

For large and sparse graphs, each edge frequency count can be assumed to be independent of most other edges. (Locally, however, edge frequency counts can of course be highly correlated.) Still, the edge count histograms indicate the presence of overdispersion, likely caused by unobserved covariates, hidden correlations (not accounted for in the theoretical null distribution) and the existence of many real but uninterestingly small effects (Efron, 2004). We account for overdispersion with inflation components and modifications of the beta-binomial components. Inflation is added for both low and maximum selection counts. Since graphs are assumed to be sparse, most edges will have low selection counts. These edges are easily classified as belonging to the null so a good model fit is not important in that range. Therefore, the beta-binomial distribution that captures null edges is inflated in the range $\{0, \ldots, c^\lambda\}$ where c^λ is chosen so that 75% of edges has selection count c^λ or less. The method is not sensitive to the exact proportion of edges captured by this inflation. The distribution for alternative edges is only inflated at the maximum count B. Further overdispersion is added by raising the beta-binomial density function corresponding to the null population by an exponent γ and renormalizing, in the same vein as BINCO, yielding the density function

$$f_{\text{null}}(w) = \frac{f_{\text{BB}}(w)^\gamma}{\sum_{k=0}^{B} f_{\text{BB}}(k)^\gamma}. \tag{2}$$

The beta-binomial density function corresponding to the alternative population is modified to have zero mass in $\{0,\ldots,c^\lambda\}$ but still be continuous

$$f_{\text{alt}}(w) = \frac{(f_{\text{BB}}(w) - f_{\text{BB}}(c^\lambda))_+}{\sum_{k=0}^{B}(f_{\text{BB}}(k) - f_{\text{BB}}(c^\lambda))_+}.$$

The modification fits better with observed distributions from simulations and leads to a more conservative edge selection. Thus, W_i^λ is modeled as coming from a distribution defined by the density function

$$f(w) = (1-\pi)f_1(w) + \pi f_2(w), \tag{3}$$

$$f_1(w) = \tau_1 \frac{\mathbb{1}(w \in \{0,\ldots,c^\lambda\})}{c^\lambda + 1} + (1-\tau_1) f_{\text{null}}(w),$$

$$f_2(w) = \tau_2 \mathbb{1}(w = B) + (1-\tau_2) f_{\text{alt}}(w).$$

We impose two constraints in order to make parameters identifiable. First, the null component, $f_1(w)$, is constrained to be decreasing in its right-most part (corresponding to $\mu_1 + \sigma_1 < 1$). Secondly, the non-null, f_2, is constrained to be convex and increasing (corresponding to $\mu_2 = \sigma_2 > 0.5$). Data in $\{c^\lambda + 1, \ldots, B-1\}$ is described by five parameters $\theta = (\pi', \mu_1, \sigma_1, \gamma, \mu_2 = \sigma_2)$, where π' captures the component sizes within the range. These are estimated with numerical maximization of the log-likelihood function

$$l(\theta) = \sum_{w=c^\lambda+1}^{B-1} h^\lambda(w) \log\left((1-\pi') f_{\text{null}}(w;\theta) + \pi' f_{\text{alt}}(w;\theta)\right),$$

under the two constraints just mentioned, as well as the constraints implied by density parametrizations. Remaining parameters π, τ_1, τ_2 are then given by the estimated parameters and the data h^λ.

We have described the method for a given level of penalization λ. The choice of range of penalization Λ to fit the model for, and the unification of fits for different penalizations, remain. We propose to use selection counts from different levels of penalization λ simultaneously, in order to decrease variance in estimates of model parameters. The unknown true π, the proportion of alternative edges, is of course constant in λ. Nevertheless, we can expect $\hat{\pi}$ to have an upward bias for small λ: with too little penalization null edges will be falsely captured by the alternative mixture component. Conversely, a large penalization will push the distribution of selection counts for alternative edges leftwards into the distribution for null edges. We assume the alternative distribution to have its mode at B. Thus the upper end of Λ is the maximal penalization for which h^λ is significantly increasing in the proximity of B, i.e. h^λ is approximately U-shaped. We have included a heuristic algorithm to help identify this point in the software package. We are interested in which λ that best separates the null and alternative mixture components and for which we can thus weigh together the evidence of edge presence together across λ for better accuracy and FDR control. We define the *separation* of mixture components, for a λ, as the difference of the amount of correctly and incorrectly selected edges based on the model fit:

$$g(\lambda) = p \sum_{w=0}^{B} (\pi f_2(w) - (1-\pi) f_1(w))_+. \tag{4}$$

Let λ_a be the upper end of an approximate 0.95 bootstrap confidence interval for the location of the maximum of $g(\lambda)$. Let $\pi^* = \hat{\pi}(\lambda_a)$, i.e. a conservative estimate of the proportion of alternative edges. Next, we update the model fit for each λ with the additional constraint $\pi \leq \pi^*$, in order to incorporate the joint estimate of the proportion of alternative edges. Lastly, let λ_b be the lower end of an approximate 0.95 confidence interval for the location of the maximum of $g(\lambda)$ for the new model fits.

Using a low estimate of λ yields a conservative edge selection since constraint on π is in stronger effect there. The model fitted to selection counts for penalization λ_b, constrained to $\pi \leq \pi^*$ is used for final edge classification. A simulation presented in the next section illustrates how the simultaneous use of counts from different levels of penalization results in lower bias and lower variance (Figure 4).

The classification threshold t^λ for the given FDR target is found from the fitted model. For $t^\lambda \in \{0, \ldots, B\}$ the estimated FDR is given by

$$\widehat{\mathrm{FDR}}(t^\lambda) = \frac{p \sum_{w=t^\lambda}^{B} (1-\pi) f_1(w)}{\sum_{w=t^\lambda}^{B} h^\lambda(w)}. \tag{5}$$

where p is the number of potential edges. The final step of ROPE assigns a q-value to each edge. Given fitted parameters at the selected penalization, the q-value q_i of an edge i is $\widehat{\mathrm{FDR}}(W_i^\lambda)$. We use the upper limit of a confidence interval for q_i in order to ensure conservative estimates. Under our model, the number of type I errors approximately follows a binomial distribution with $\sum_{w=t^\lambda}^{B} h^\lambda(w)$ experiments and $\widehat{\mathrm{FDR}}(W_i^\lambda)$ success probability. Using the normal approximation of the binomial distribution, the upper 0.95 confidence bound for q_i is given by

$$\widehat{\mathrm{FDR}}(W_i^\lambda) + z_{0.975} \sqrt{\frac{\widehat{\mathrm{FDR}}(W_i^\lambda)(1 - \widehat{\mathrm{FDR}}(W_i^\lambda))}{\sum_{w=t^\lambda}^{B} h^\lambda(w)}}.$$

To conclude this section, we emphasize the methodological differences between ROPE and BINCO. First, ROPE uses a mixture model that captures both null and alternative edges, while BINCO models only the null distribution. In practice, the threshold corresponding to any relevant FDR target will be in a part of the domain where the population of alternative edges dominates. This leads to the estimation of BINCO to be based on an extrapolation, resulting, as the next section will show, in a lower stability of estimates. Furthermore, to estimate a model that only captures the null population, BINCO is forced to select a subset of data where the null population is most prevalent. This intermediate range selection contributes to the lower stability of estimates. In contrast, by modeling both null and alternative edge selection counts, ROPE can use the most relevant subset of data to fit its model parameters. Thus, extrapolation is avoided and the parameter estimates are insensitive to the exact end points of the subset range. Second, while ROPE simultaneously uses counts from different levels of penalization where the overlap of null and alternative populations is small, BINCO selects the level of regularization that selects the most edges while estimating an FDR below target. This results in lower stability of BINCO's estimates, since the selection may change due to small perturbations of the data, and in a bias of BINCO to underestimate FDR, since models with underestimated FDR tends to select more edges at a fixed FDR target. Third, overdispersion is a main modeling difficulty addressed by BINCO and ROPE. Our richer model, with greater ability to capture overdispersion, results in ROPE having a more accurate FDR control than BINCO.

3 Results

We present a comprehensive simulation study to assess the performance of ROPE and compare it with two state-of-the-art methods: BINCO and stability selection. We also present an application of the methods to gene expression data from glioblastoma cancer patients, and compare results. An application of ROPE to variable selection for a non-graphical model is provided in the supplement.

3.1 Comparison of accuracy and robustness of FDR control on simulated data

Our simulation experiment consists of data from 500-node networks of three topologies: scale-free, hubby and chain graphs. We sample standard normal data from covariance matrices corresponding to the network topologies. The signal strength is either strong (mean and standard deviation of covariances between connected nodes is 0.32 respectively 0.13) or weak (mean and standard deviation is 0.25 respectively 0.09). The scale-free networks have 495, 49 (sparse) or 990 (dense) edges. The hubby network has 20 hub nodes, each connected to between 92 and 4 other nodes. The chain network connects its 500

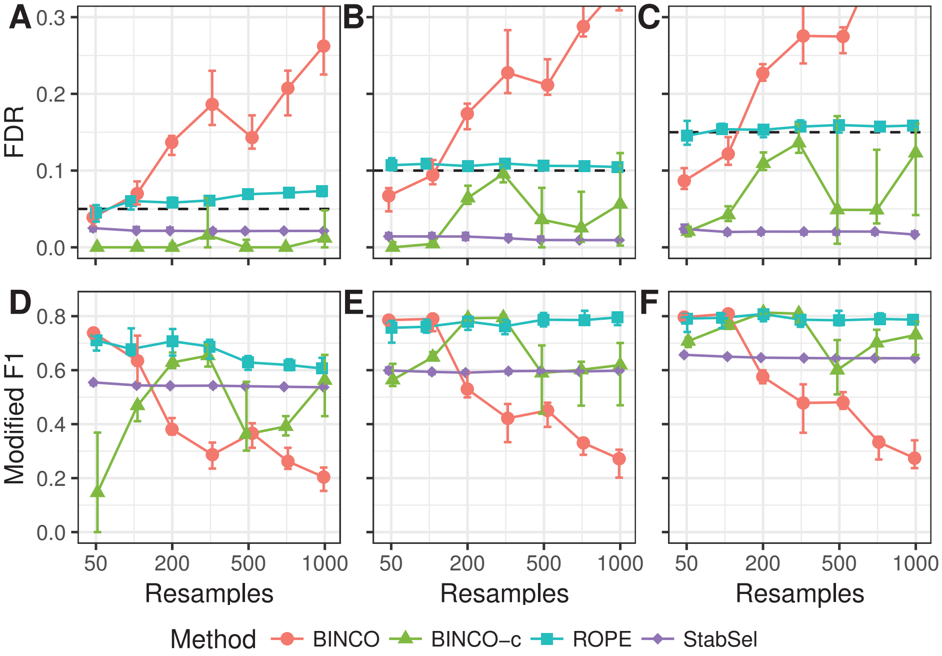

Figure 2: Validation of the proposed method on simulated data. Four methods are compared: BINCO, conservative BINCO (BINCO-c), ROPE and stability selection (StabSel). Each method has been applied for three FDR targets. Columns A-D, B-E and C-F show results for target FDR 0.05, 0.1 and 0.15, respectively. Panels A, B and C compare FDR with target FDR. ROPE achieves an FDR closest to the target. BINCO tends to make an increasingly liberal selection as the number of resamples increases. Stability selection is consistently too conservative. Panels D, E and F show the corresponding modified F1 score. ROPE scores highest overall. Points show median result (20 simulations) and whiskers represent 1.5 times IQR.

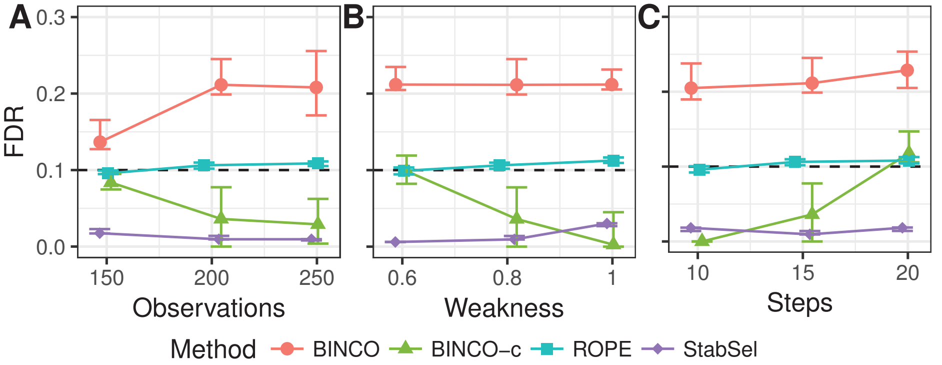

Figure 3: Examination of parameter sensitivity for the same simulated data as in Figure 2. The number of observations, weakness and number of steps in the penalization set Λ is varied, in panels A, B and C, respectively. The figure shows that ROPE performs well and gives consistent results in this parameter subspace, while stability selection is consistently conservative and BINCO and conservative BINCO give less consistent results. In general, BINCO is too liberal and conservative BINCO is too conservative. This figure shows results for a target FDR of 0.1. Results for other target FDR and settings can be found in the supplement and are in agreement with our findings here.

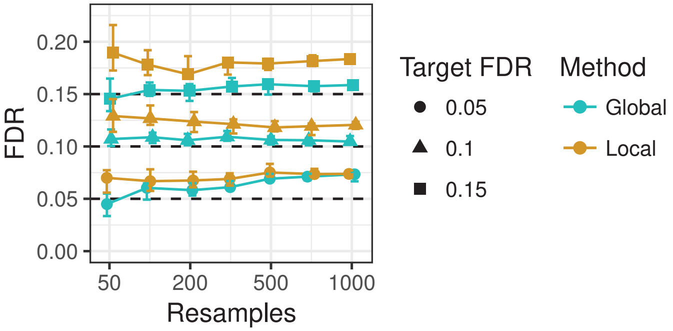

Figure 4: Comparison of ROPE with and without joint modeling of counts from different penalization levels. When counts from different levels are used (Global) the estimated FDR is closer to the target FDR and the variance between simulations is lower.

nodes into one chain of length 500. In all, this constitutes seven simulated model selection problems: three topologies, five variations of the scale-free topology. Two of these are identical to those in Li *et al.* (2013).

We generate edge presence count matrices W_j^λ for each problem by taking B bootstrap samples, and select edges for each sample using randomized neighborhood selection with penalization ranging from 0.02 to 0.3. The settings for W_j^λ, i.e. B, number of steps in Λ, weakness and n, are varied in order to assess the methods' sensitivity. We compare the methods' selections for three target FDR levels: 0.05, 0.1 and 0.15. Each combination of settings is rerun 20 times in order to assess sensitivity to randomness in subsampling. We compare target FDR with achieved FDR and score each selection with a modified F1 score

$$\mathrm{F1_m} = 2\frac{(1-\mathrm{FDR})\mathrm{TPR}}{m(\mathrm{FDR})+\mathrm{TPR}}, \quad m(\mathrm{FDR}) = \begin{cases} 1-\mathrm{FDR}, \text{if } \mathrm{FDR} \leq \mathrm{FDR}^* \\ \frac{\mathrm{FDR}}{\mathrm{FDR}^*} - \mathrm{FDR}^*, \text{otherwise} \end{cases}$$

where FDR^* is the target FDR. The denominator is modified to ensure that scores are decreasing with FDR when FDR is above target.

Results for the scale-free network with 500 nodes, 495 edges and strong signal is presented in Figures 2 and 3. In Figure 2, B is varied, while $n = 200$, weakness is 0.8 and Λ consists of 15 steps. In Figure 3, $B = 500$, while n, weakness and number of steps in Λ is varied. Results for remaining topologies and parameter combinations are presented in the supplement.

Results show that ROPE performs best in terms of modified F1, FDR and stability for the scale-free, dense scale-free, small scale-free and weak signal scale-free networks. In the chain network and the sparse scale-free network ROPE and stability selection perform similarly. Stability selection makes the most stable selections, but is generally too conservative, which is to be expected since the method is based on a bound. For the weak signal scale-free network and with a target FDR of 0.05, stability selection is too conservative to select any edge at all. BINCO and conservative BINCO both make far less stable selections than ROPE and stability selection. Furthermore, both BINCO methods are sensitive to the number of bootstraps. Logically, selection should improve when the number of bootstraps is increased. Instead, BINCO makes an increasingly more liberal selection. Similarly disconcerting, BINCO performance worsens with increased signal strength (number of observations) (Fig. 3A). Without access to the true model, it would be difficult to know how many bootstraps that should be performed to get a correct FDR control. This strong dependency between number of bootstraps, signal strength and achieved FDR makes BINCO hard to use in practice. The hubby network is the one setting where stability selection performs better than ROPE. There, ROPE makes no selection since the selection count histograms are not U-shaped. In order to examine how ROPE would perform for the hubby network if the signal were stronger, we generated additional observations, increasing the examined range from 150-500 observations to 150-1250 observations. For more than 500 observations, ROPE again yielded the highest modified F1 and the FDR closest to target.

ROPE uses selection counts from several penalization levels and, as can be seen in Figure 4, this avoids a too liberal selection and increases stability. In addition, the Figure indicates that ROPE outperforms BINCO even without the joint modeling, which emphasizes the need to model both the null and non-null edge populations as done in ROPE.

In terms of computation time, BINCO and ROPE are slower than stability selection. At each level of penalization, ROPE fits a five parameter model, while BINCO estimates the end points of an approximately decreasing range and then fits three parameters. Both take only a few seconds per penalization level on a standard desktop computer. Increasing size of networks or the number of observations does not increase computation time, since these methods use summary statistics — the number of variables having a selection count w, for each $w \in \{0, \ldots, B\}$. The computation time of stability selection, BINCO and ROPE is small compared to the time needed for resampled variable selection.

3.2 FDR controlled edge selection for a graphical model of gene expressions in the PI3K/Akt pathway of glioblastoma cancer patients

In this section, we apply ROPE to gene expression data and study the selected network. We also compare ROPE, BINCO and stability selection in terms of size of FDR controlled selections and stability. We downloaded RNA-Seq gene expressions for 172 glioblastoma multiforme cancer patients from the USCS Cancer Genomics Browser (Goldman *et al.*, 2014). The data comes from TCGA and had been normalized across all TCGA cohorts and log transformed. It contains measurements for 20,530 genes. We downloaded a list of genes in the PI3K/Akt signaling pathway from KEGG (Kanehisa and Goto, 2000). 337 genes in the gene expression data set were found in the PI3K/Akt gene list. We discarded half of the genes with lowest median absolute deviation (MAD) of expression. Remaining genes were scaled to have MAD 1. We bootstrapped the data 500 times and estimated graphical models with 12 different levels of penalization for each bootstrap sample. The weakness in randomized lasso was set to 0.8. Figure 5 shows a visualization of the final network estimated with ROPE. In the visualization we have kept all edges with an estimated q-value below 0.15, i.e. we expect that 15% of the depicted edges are false discoveries. The edge widths correspond to estimated edge q-value. Zero degree nodes are not shown. Highly connected network nodes were the epidermal growth factor receptor (EGFR, 8 links), the platelet-derived growth factor receptor alpha (PDGRA, 6 links), components of the IL2 receptor (IL2RA and IL2RG, with 7 and 3 links), vitronectin (VTN, 7 links) and tenascin R (6 links). Of these, EGFR and PDGFRA are well established glioblastoma oncogenes. TNR is a tenascin with neural restricted expression, and is likely a negative marker of glioma invasiveness (Brösicke and Faissner, 2015). By contrast VTN, which is connected to several FGF and FGFR isoforms in our network, is a pro-migratory/invasion factor (Ohnishi *et al.*, 1998). IL2, finally, has been suggested to promote growth of glioma cells (Capelli *et al.*, 1999). Our network may thus serve to prioritize hub genes for further study, as well as their functionally associated genes. Edge q-values, along with properties of methodology for subsequent analysis, may facilitate the choice of how many associations to study further.

While the correct network model of the pathway is, of course, unknown, a comparison of methods on this real data shows relevant differences. We subsampled the 500 selected models 20 times without replacement. Each subsample consists of 400 selected models. Counting edge selections within each subsample gives 20 subsampled W. Figure 6 shows a comparison of size of FDR controlled selections and of stability of selections between subsamples. BINCO selects more than 200 edges already at a target FDR of 0.0125. Stability selection selects the empty model for target FDR 0.25 and below, in agreement with the conservative behaviour observed in the simulations. BINCO and conservative BINCO show more variation between selected models for different subsamples, than ROPE and stability selection. The liberal selection by BINCO agrees with simulation results, suggesting a failure to control FDR. BINCO's lack of agreement between selections at low target FDR also suggests a failure to control FDR. The higher variability in BINCO and conservative BINCO also agrees with simulation results. We have used Fleiss' κ, an index of inter-rater agreement among many raters (Fleiss, 1971), to measure agreement between selections across subsamples.

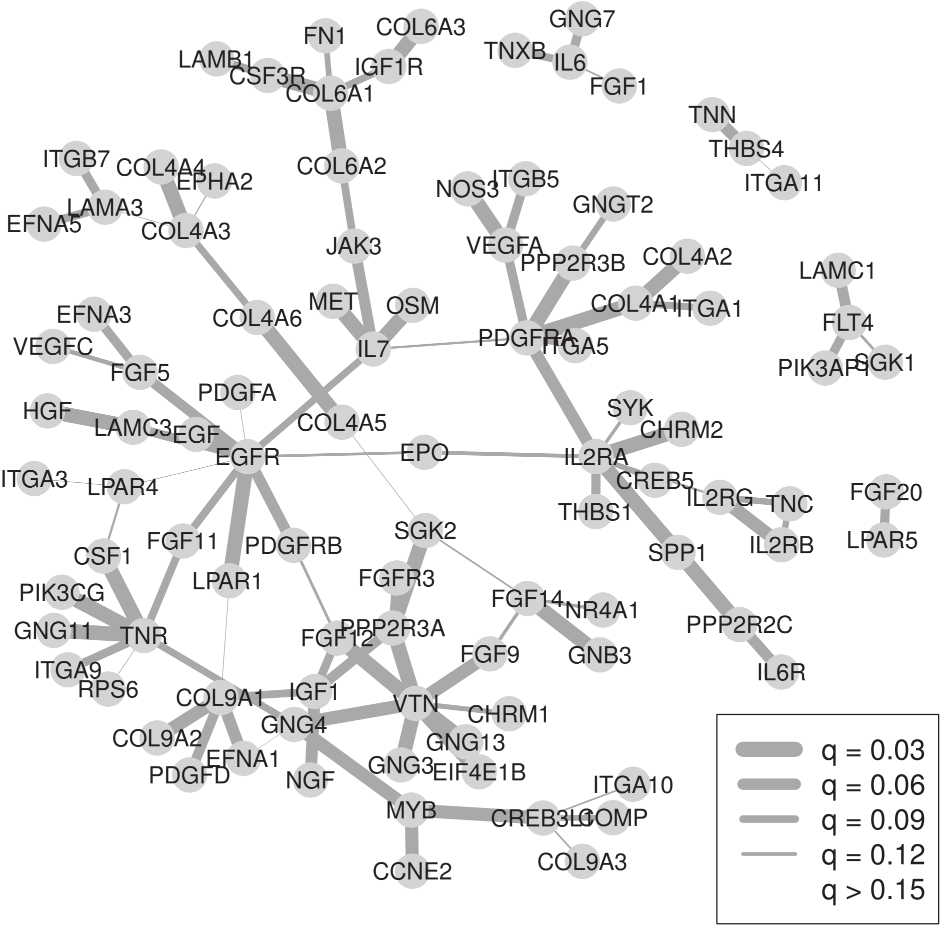

Figure 5: ROPE selection of gene connections in the PI3K/Akt pathway based on gene expressions from glioblastoma cancer patients in TCGA. Widths of edges correspond to q-values. Highly connected nodes are known to have functions associated with invasiveness and/or tumor growth in glioblastoma.

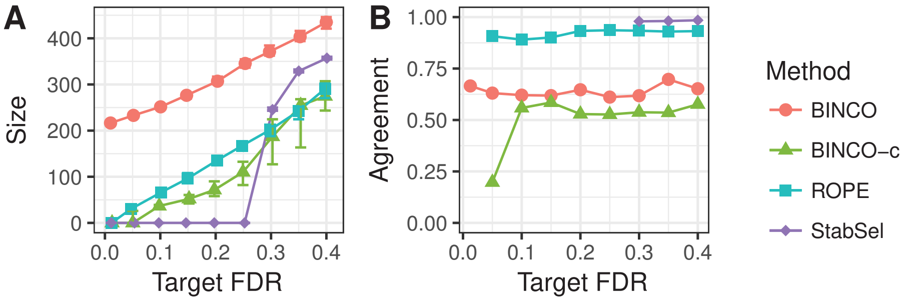

Figure 6: Method comparison on gene expressions in the PI3K/Akt pathway of glioblastoma cancer patients. Panel A shows the number of selected edges by each method for a range of target FDR. While the achieved FDR is unknown, we note that BINCO is liberal enough to select more than 200 edges even at a low target FDR of 0.0125. As expected, stability selection is conservative producing empty networks for target FDR 0.25 and below. Conservative BINCO exhibits substantial variability in network size. Panel B shows agreement within each method across 20 subsamples of W as measured by Fleiss' κ. BINCO and conservative BINCO are less stable than ROPE and stability selection. The lack of agreement for BINCO at low targets combined with a large selection size, makes it unlikely that FDR is controlled. Fleiss' κ is not defined for empty selections produced by stability selection below FDR 0.25.

4 Discussion

The problem of FDR control in high-dimensional variable selection problems is of great relevance for interpreting data from molecular biology and other fields with an abundance of complex high-dimensional data. Many methods for variable selection in high-dimensional problems exist, but they suffer from the need to tune intermediate parameters of little scientific relevance. We have introduced a method for false discovery control in network models, and presented results showing that this method outperforms existing alternatives. With the method and software package presented here, which achieve accurate and robust FDR control, we have made possible a principled selection of relevant interactions.

We did consider an alternative statistical model for selection counts where the populations of alternative and null edges were further stratified into sub populations, based on their strength or the structure of their neighborhood in the graph. We did not find such a richer model to be worth the additional cost and estimation variability. Moreover, such a model poses the additional challenge of classifying each sub population as belonging to either the null or the alternative population. We also considered strengthening the connection between statistical models across all levels of penalization. Power and stability could potentially be increased by enforcing smoothness of all model parameters across levels of penalization. But the large number of edges that are represented in each histogram suggests that improvements would be small. Furthermore, the numerical fitting of such a global model is challenging.

ROPE, BINCO and stability selection use only summary statistics, proportions of variables with each selection count. Thus, their computational complexity is not affected by an increase in the number of network nodes. Computational time is completely dominated by the preceding step of resample based estimation. However, resampling based estimation is necessary to stabilize model selection and this process is parallelizable.

Recently, methods for assigning p-values to variables in high-dimensional linear models have been proposed. See Dezeure *et al.* (2015) for a review and comparison. P-values can be used to approximate q-values (Storey and Tibshirani, 2003), and thus to control FDR. Nevertheless, due to the high instability of estimated p-values (the so called “p-value lottery”) resampling is needed when applying the reviewed methods in practice (Dezeure *et al.*, 2015). The application of this approach to graphical models is studied in Janková and van de Geer (2015). Dezeure *et al.* (2016) proposes p-value estimation for linear models based a combination of the de-sparsified lasso and bootstrap. Here, the bootstrap is not used to aggregate many, unstable estimates but to improve on p-value estimates that relied on asymptotic arguments. The dependency on a penalization parameter remains (current implementation uses a fixed penalization chosen via cross-validation). ROPE can be applied to any resampling based network selection method, including resampling of p-value based selection, and could thus improve de-sparsified lasso estimates by utilizing multiple levels of penalization.

Here, ROPE was used for FDR controlled edge selection in a single penalization parameter setting. An interesting direction for future work would be to generalize ROPE to more complex modeling settings, e.g. comparative network modeling, with multiple tuning parameters. One could approach this problem in either a sequential fashion (across tuning parameters) or generalize the distribution mixture modeling to a higher-dimensional parameter space.

Lastly, the use of richer summaries of W than histograms h^λ may improve model selection. One way is to view edge presence counts W_i^λ as functional data $W_i(\lambda)$. We have observed that these functions behave quite differently for different edges. The location and magnitude of $\min_\lambda \frac{d}{d\lambda} W_i(\lambda)$ are two examples of quantities that may facilitate edge selection. Another way is to consider correlation between edges. Edges can compete to explain the node correlation structure in a network neighborhood. Therefore, selection correlation between pairs of edges over resamples may also facilitate edge selection. Although computationally infeasible to estimate in full, the possibility to limit focus to edge pairs that are, in some sense, closely located in the network makes this an interesting direction of future research.

Acknowledgements

RJ and JK are supported in part by the Swedish Research Council. RJ is also supported in part by the Wallenberg foundation. SN is supported by the Swedish Research Council, Swedish Cancer Society and Swedish Childhood Cancer Foundation. AJ and JS are fulltime employees of Astra-Zeneca.

Conflict of interest: none declared.

References

Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM.

Brösicke, N. and Faissner, A. (2015). Role of tenascins in the ECM of gliomas. *Cell adhesion & migration*, **9**(1-2), 131–140.

Capelli, E., Civallero, M., Barni, S., Ceroni, M., and Nano, R. (1999). Interleukin-2 induces the growth of human glioblastoma cells in culture. *Anticancer research*, **19**(4B), 3147.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statist. Sci.*, **30**(4), 533–558.

Dezeure, R., Bühlmann, P., and Zhang, C.-H. (2016). High-dimensional simultaneous inference with the bootstrap. *Preprint arXiv:1606.03940*.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, **99**(465), 96–104.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, **76**(5), 378.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**(3), 432–441.

Goldman, M., Craft, B., Swatloski, T., Cline, M., Morozova, O., Diekhans, M., Haussler, D., and Zhu, J. (2014). The UCSC cancer genomics browser: update 2015. *Nucleic Acids Research*.

Janková, J. and van de Geer, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electron. J. Statist.*, **9**(1), 1205–1229.

Jörnsten, R., Abenius, T., Kling, T., Schmidt, L., Johansson, E., Nordling, T. E. M., Nordlander, B., Sander, C., Gennemark, P., Funa, K., Nilsson, B., Lindahl, L., and Nelander, S. (2011). Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Molecular Systems Biology*, **7**, 486.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27–30.

Kling, T., Johansson, P., Sánchez, J., Marinescu, V. D., Jörnsten, R., and Nelander, S. (2015). Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content. *Nucleic Acids Research*.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**(1), 559.

Li, S., Hsu, L., Peng, J., and Wang, P. (2013). Bootstrap inference for network construction with an application to a breast cancer microarray study. *Ann. Appl. Stat.*, **7**(1), 391–417.

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. In *Advances in Neural Information Processing Systems 23*, pages 1432–1440. Neural Information Processing Systems.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., and Califano, A. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**(1), S7.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, **34**(3), 1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(4), 417–473.

Ohnishi, T., Hiraga, S., Izumoto, S., Matsumura, H., Kanemura, Y., Arita, N., and Hayakawa, T. (1998). Role of fibronectin-stimulated tumor cell migration in glioma invasion in vivo: clinical significance of fibronectin and fibronectin receptor expressed in human glioma tissues. *Clinical & experimental metastasis*, **16**(8), 729–741.

Pe'er, D. and Hacohen, N. (2011). Principles and strategies for developing network models in cancer. *Cell*, **144**(6), 864–873.

Storey, J. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**(16), 9440–9445.

The Cancer Genome Atlas Research Network, Weinstein, J., Collisson, E., Mills, G., Shaw, K. M., Ozenberger, B., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. (2013). The cancer genome atlas pan-cancer analysis project. *Nat Genet*, **45**(10), 1113–1120.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.

Supplementary material: ROPE: high-dimensional network modeling with robust control of edge FDR

Jonatan Kallus[1], José Sánchez[2], Alexandra Jauhiainen[3], Sven Nelander[4], and Rebecka Jörnsten[1]

[1]Department of Mathematical Sciences, University of Gothenburg and Chalmers University of Technology
[2]Discovery Sciences, AstraZeneca, Gothenburg
[3]Early Clinical Biometrics, AstraZeneca, Gothenburg
[4]Department of Immunology, Genetics and Pathology, Uppsala University, Sweden

1 FDR controlled variable selection for a multinomial logistic regression classifier of gene expression profiles

In our final experiment, we apply ROPE to model selection for a non-graphical model. In particular, we demonstrate the use of ROPE for a multinomial logistic regression classifier for classifying the primary cancer type of a gene expression profile. We downloaded RNA-Seq gene expression profiles consisting of measurements of 20,530 genes for 9,755 cancer patients from the USCS Cancer Genomics Browser. The data comes from TCGA. We removed profiles corresponding to cancer types for which less than 100 observations were present in the data set, in order to reduce the chance of drawing bootstrap samples without all classes represented. The resulting data set consists of 9,256 observations and 20,530 variables. Each observation is classified as having one of 24 primary cancer types. We drew 100 bootstrap samples and fitted generalized linear models with lasso penalization and multinomial response to each bootstrap sample. We used grouped lasso penalization so that each variable is either selected for all classes or excluded entirely. For each bootstrap sample, one model was fitted for each of 22 levels of penalization, ranging from 0.015 to 0.039. Lower penalization resulted in non-convergence when fitting the model and higher penalization resulted in histograms not being U-shaped. The resulting matrix W of 22 times 20,530 variable inclusion counts was used with ROPE to make an FDR controlled selection of genes whose expression level is predictive of primary cancer type. 86, 118 and 133 genes were selected at the 0.05, 0.1 and 0.15 FDR level, respectively. The selected genes are presented in Table 1. This experiment shows that ROPE can be applied to some variable selection problems other than edge selection in graphical models.

2 Additional simulation results

Figures 1 to 56 show results from all simulations. For each simulation setting, four parameters are varied one by one (number of bootstraps B, number of penalization levels, number of observations n and weakness in randomized lasso). For each varied parameter, FDR and modified F1 are shown for each method and three target FDR: 0.05, 0.1 and 0.15. A detailed description of simulation settings and interpretation of results is given in the main article.

	gene	q-value		gene	q-value		gene	q-value
1	ATP5EP2	0.025	46	SFTA3	0.025	91	KRT74	0.051
2	AZGP1	0.025	47	SFTPA1	0.025	92	LYPLAL1	0.051
3	BCL2L15	0.025	48	SFTPB	0.025	93	MSX1	0.051
4	C10orf27	0.025	49	SLC6A3	0.025	94	MUC5B	0.051
5	C14orf105	0.025	50	SOX17	0.025	95	PTGER3	0.051
6	C8orf85	0.025	51	SPRYD5	0.025	96	RNF212	0.051
7	CALML3	0.025	52	ST6GALNAC1	0.025	97	SLC5A6	0.051
8	CDH16	0.025	53	TBX5	0.025	98	SLCO1A2	0.051
9	CDHR1	0.025	54	TCF21	0.025	99	C6orf223	0.058
10	CDX1	0.025	55	TFRC	0.025	100	ERBB3	0.058
11	CFHR2	0.025	56	TG	0.025	101	FOXF1	0.058
12	DPPA3	0.025	57	TMEFF2	0.025	102	IRX1	0.058
13	DSG3	0.025	58	TPO	0.025	103	NACAP1	0.058
14	EBF2	0.025	59	TRPS1	0.025	104	PHOX2A	0.058
15	EMX2	0.025	60	TSIX	0.025	105	C2orf80	0.065
16	FLJ45983	0.025	61	TYR	0.025	106	MMD2	0.065
17	FOXE1	0.025	62	UPK1B	0.025	107	SLC22A2	0.065
18	FTHL3	0.025	63	UPK2	0.025	108	APCS	0.071
19	FUNDC2P2	0.025	64	ZNF134	0.025	109	GJB1	0.071
20	FXYD2	0.025	65	ZNF280B	0.025	110	LOC285740	0.071
21	HAND2	0.025	66	FABP7	0.035	111	BCAR1	0.078
22	HOXA9	0.025	67	HOXC8	0.035	112	ACTC1	0.084
23	INS	0.025	68	KRT20	0.035	113	CTAGE1	0.091
24	IRX2	0.025	69	MAP7	0.035	114	ESR1	0.091
25	IRX5	0.025	70	MS4A3	0.035	115	GFAP	0.091
26	ITGA3	0.025	71	MUC16	0.035	116	HKDC1	0.091
27	KIAA1543	0.025	72	NOX1	0.035	117	PLA2G2F	0.091
28	KLK2	0.025	73	NTRK2	0.035	118	SOX10	0.091
29	LGSN	0.025	74	PAX3	0.035	119	PPARG	0.103
30	LOC407835	0.025	75	PRO1768	0.035	120	C21orf131	0.109
31	LOC643387	0.025	76	SERPINB3	0.035	121	DLX6	0.109
32	MAB21L2	0.025	77	SYCP2	0.035	122	GAL3ST3	0.109
33	NACA2	0.025	78	C14orf115	0.043	123	HNF1B	0.109
34	NDUFA4L2	0.025	79	C14orf19	0.043	124	KRT5	0.109
35	PA2G4P4	0.025	80	C1orf172	0.043	125	SPINK1	0.109
36	PAX8	0.025	81	FGL1	0.043	126	ARHGEF33	0.115
37	PHOX2B	0.025	82	GATA3	0.043	127	C1orf14	0.115
38	POU3F3	0.025	83	HOXA11	0.043	128	APOA2	0.121
39	PRAC	0.025	84	KRT7	0.043	129	LRRN4	0.121
40	RFX4	0.025	85	PRHOXNB	0.043	130	SOX2	0.121
41	RPL17	0.025	86	SCGB2A1	0.043	131	WNT3A	0.127
42	RPL39L	0.025	87	FLJ32063	0.051	132	GJB7	0.133
43	RPS4Y1	0.025	88	FOXA2	0.051	133	NASP	0.144
44	SCGB2A2	0.025	89	HECW2	0.051	134	ATCAY	0.150
45	SERPINB13	0.025	90	KLK3	0.051	135	DDR1	0.150

Table 1: The 135 transcripts with lowest q-value as selected with ROPE for a multinomial logistic classifier of expression profiles by cancer type.

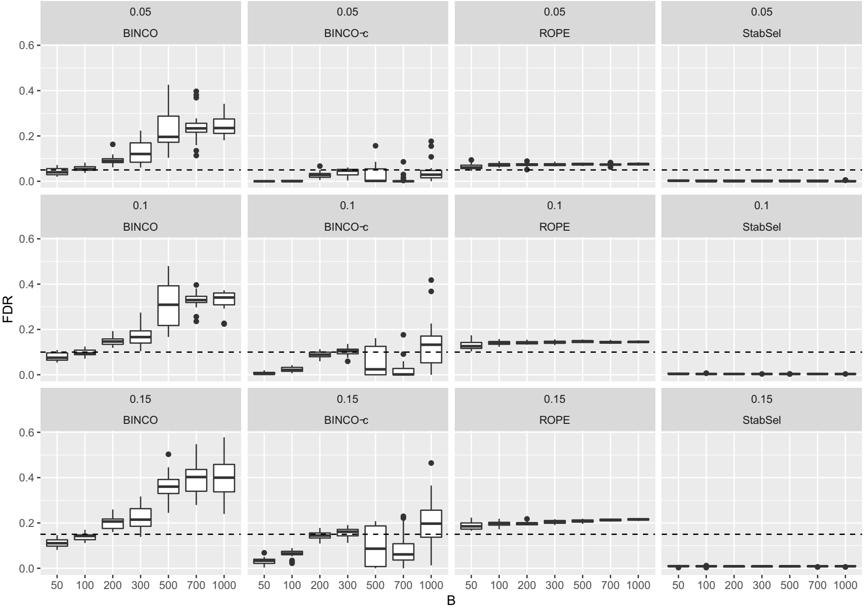

Figure 1: Network topology: chain, steps: 15, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

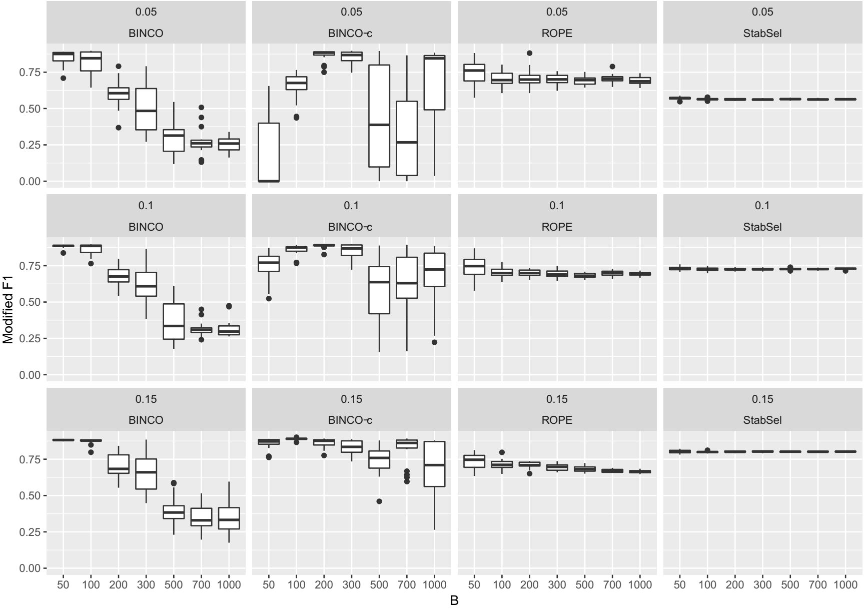

Figure 2: Network topology: chain, steps: 15, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

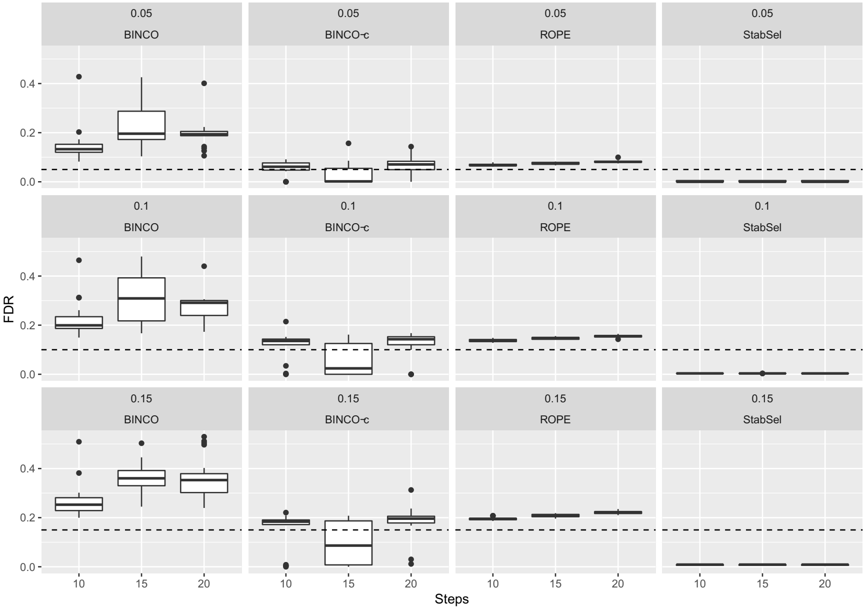

Figure 3: Network topology: chain, $B = 500$, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

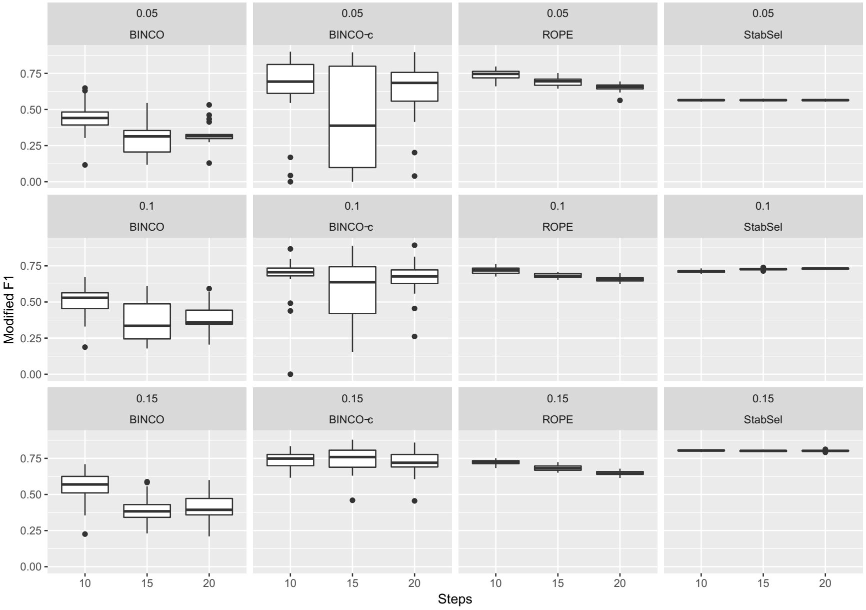

Figure 4: Network topology: chain, $B = 500$, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

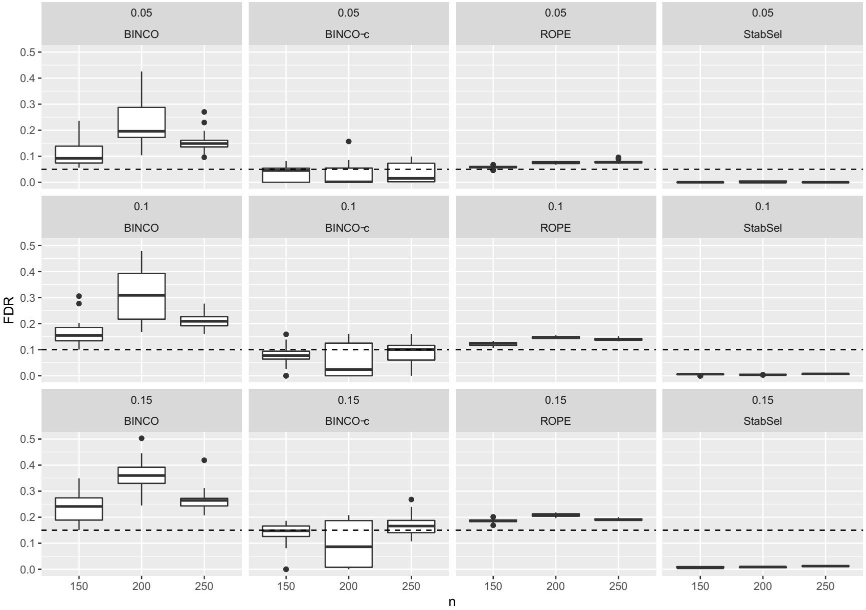

Figure 5: Network topology: chain, $B = 500$, steps: 15, weakness: 0.8, facet titles: target FDR and method.

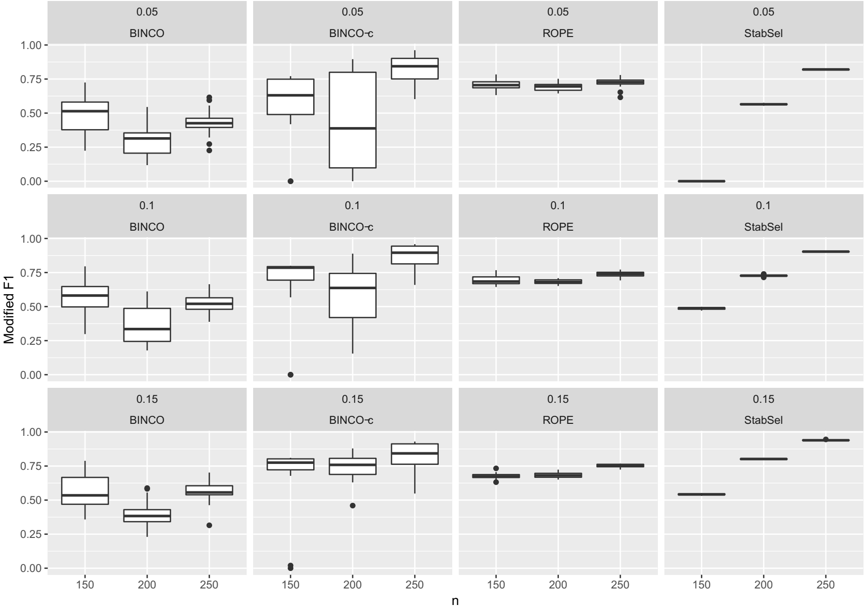

Figure 6: Network topology: chain, $B = 500$, steps: 15, weakness: 0.8, facet titles: target FDR and method.

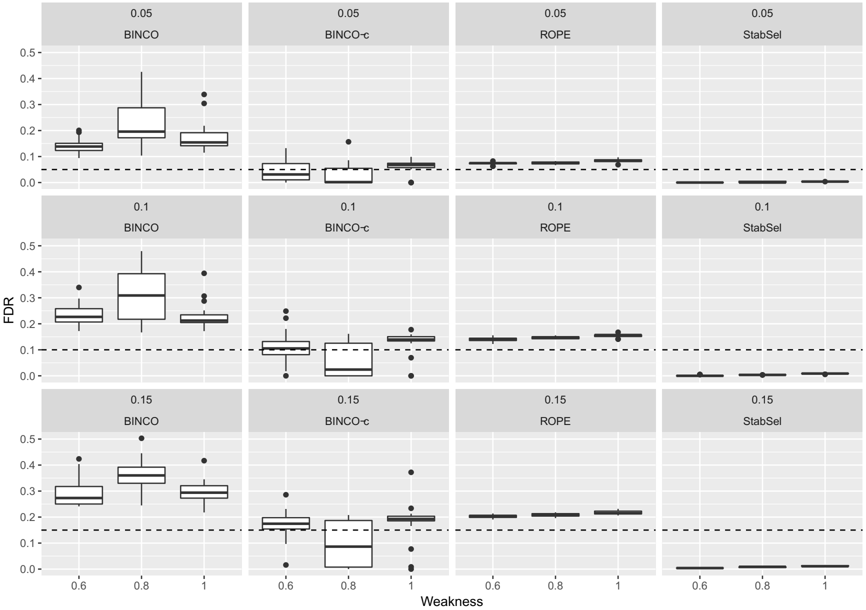

Figure 7: Network topology: chain, $B = 500$, steps: 15, $n = 200$, facet titles: target FDR and method.

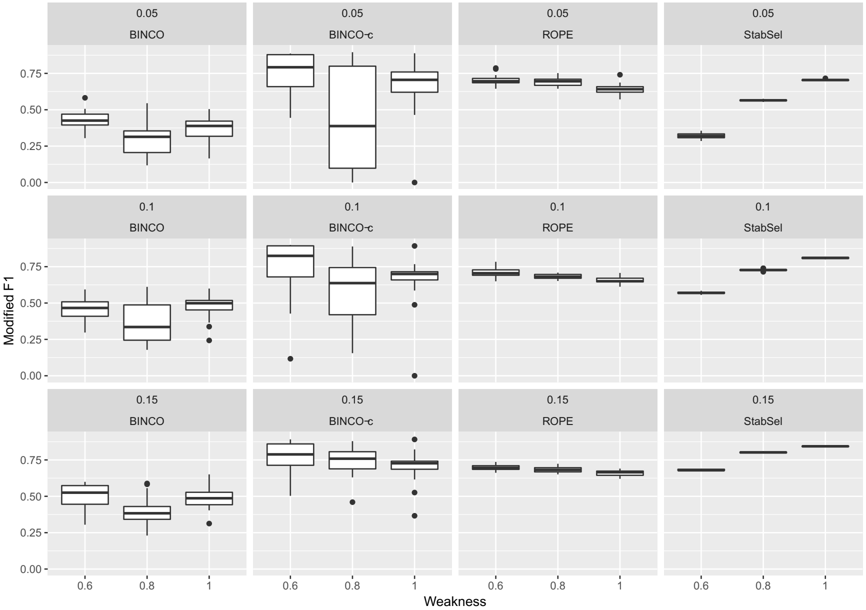

Figure 8: Network topology: chain, $B = 500$, steps: 15, $n = 200$, facet titles: target FDR and method.

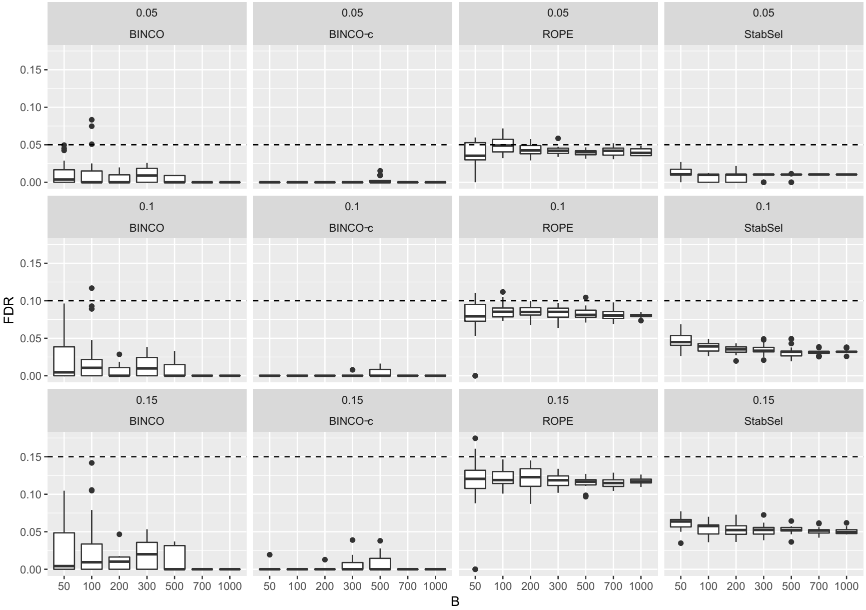

Figure 9: Network topology: dense, steps: 15, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

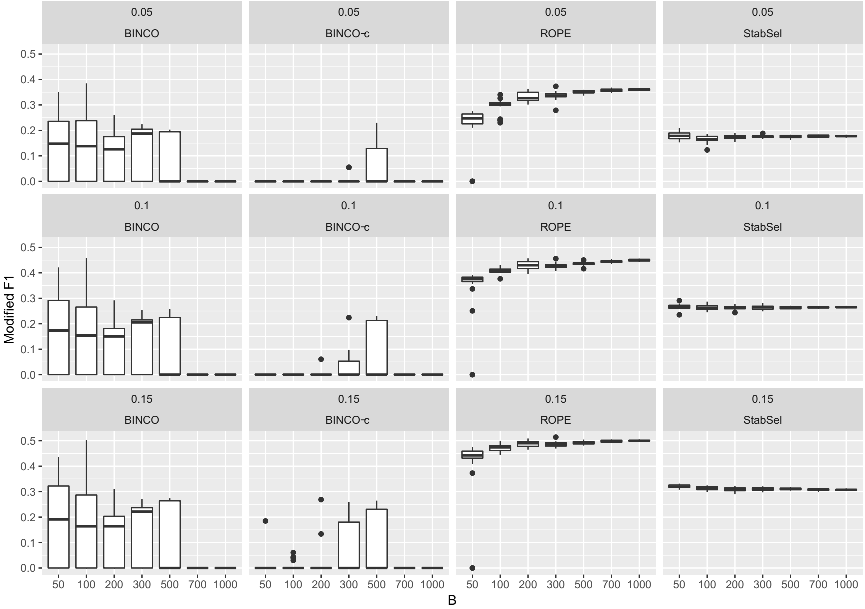

Figure 10: Network topology: dense, steps: 15, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

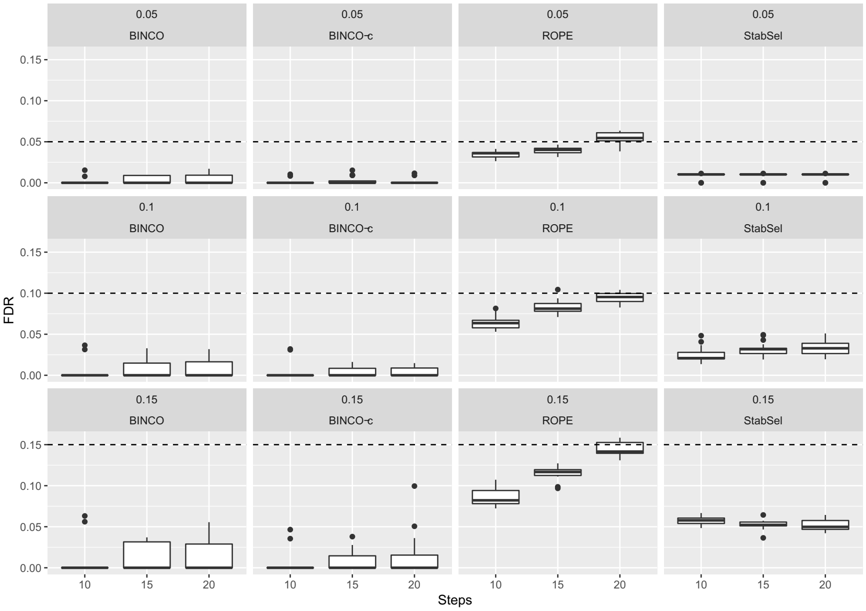

Figure 11: Network topology: dense, $B = 500$, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

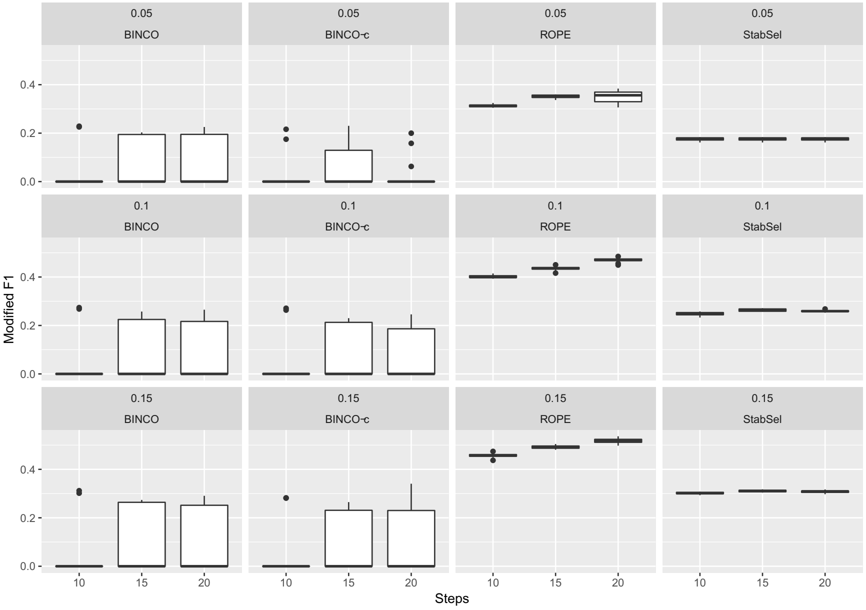

Figure 12: Network topology: dense, $B = 500$, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

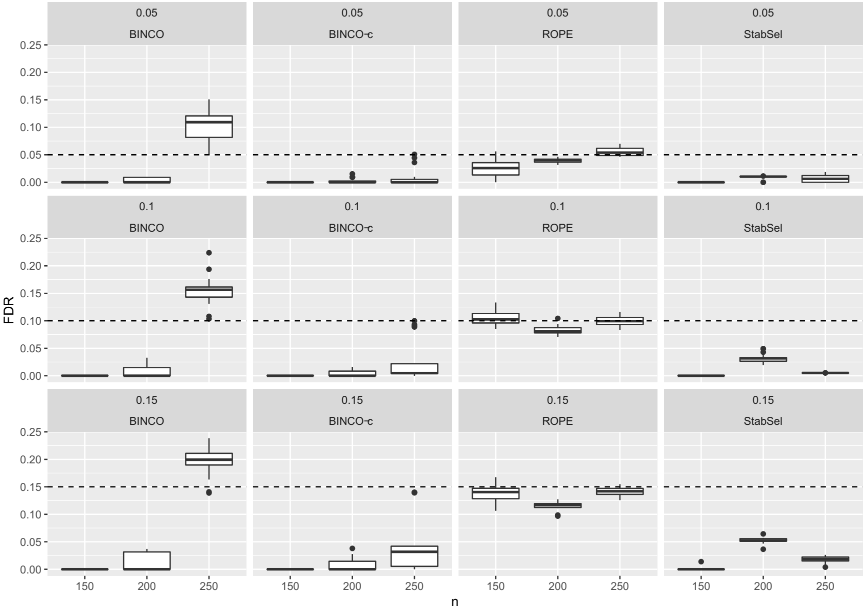

Figure 13: Network topology: dense, $B = 500$, steps: 15, weakness: 0.8, facet titles: target FDR and method.

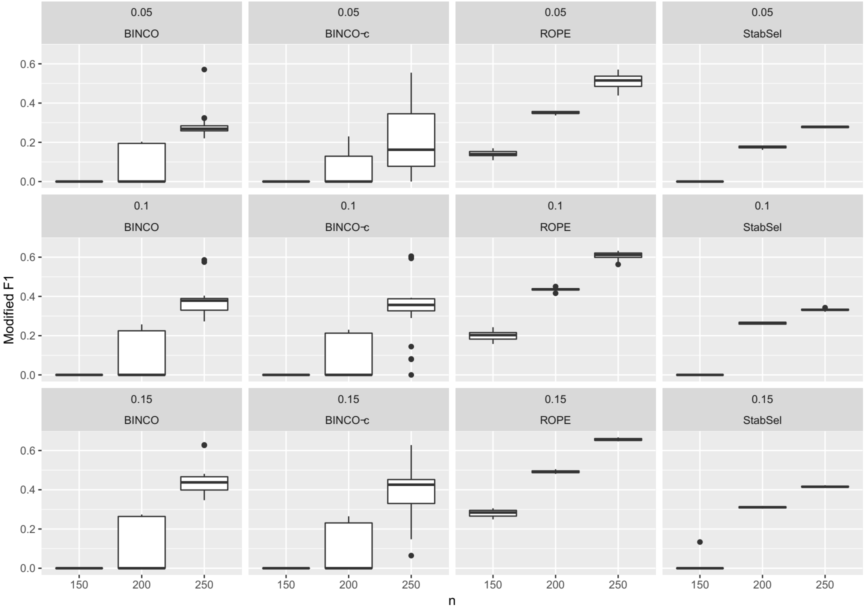

Figure 14: Network topology: dense, $B = 500$, steps: 15, weakness: 0.8, facet titles: target FDR and method.

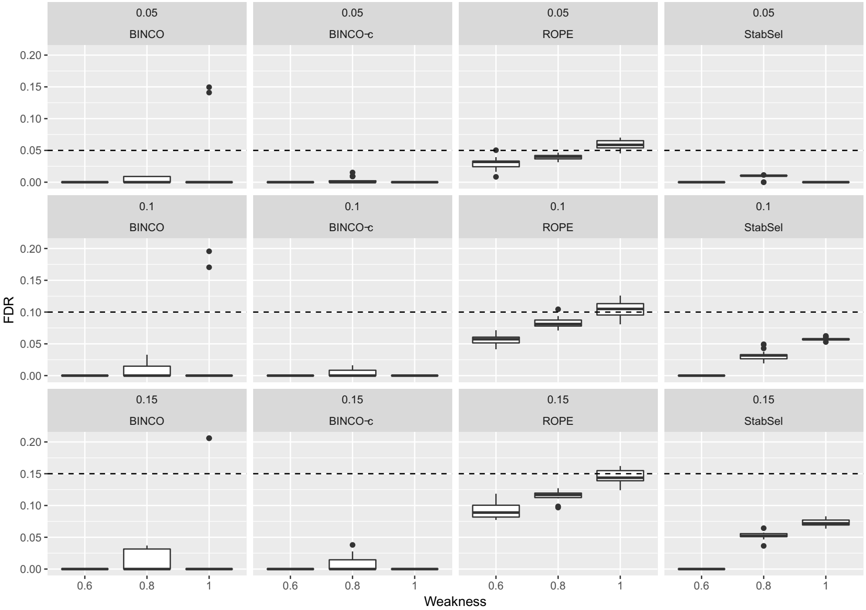

Figure 15: Network topology: dense, $B = 500$, steps: 15, $n = 200$, facet titles: target FDR and method.

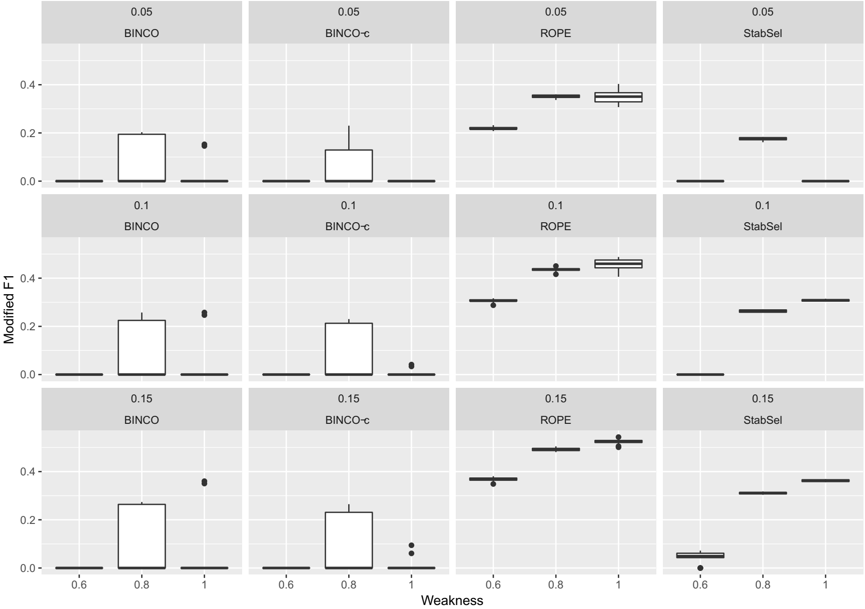

Figure 16: Network topology: dense, $B = 500$, steps: 15, $n = 200$, facet titles: target FDR and method.

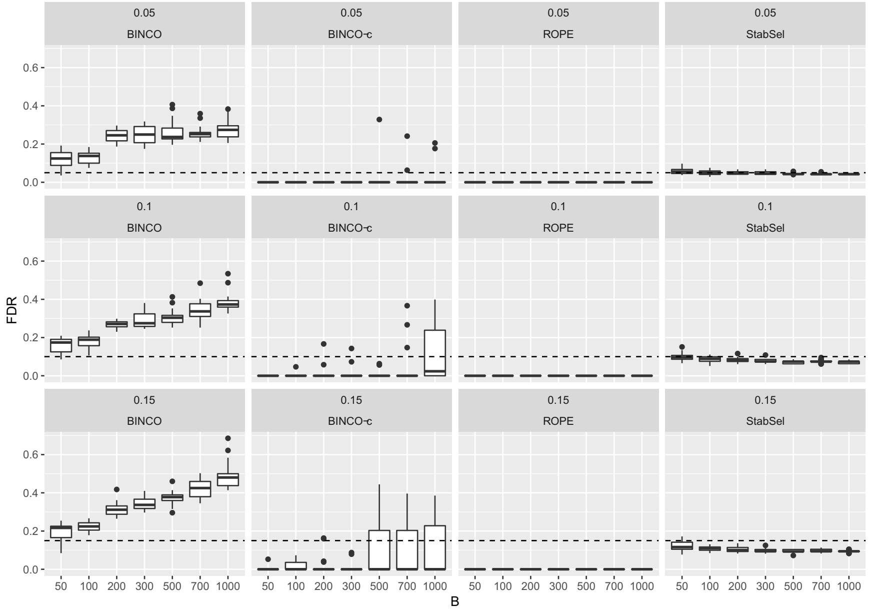

Figure 17: Network topology: hubby, steps: 15, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

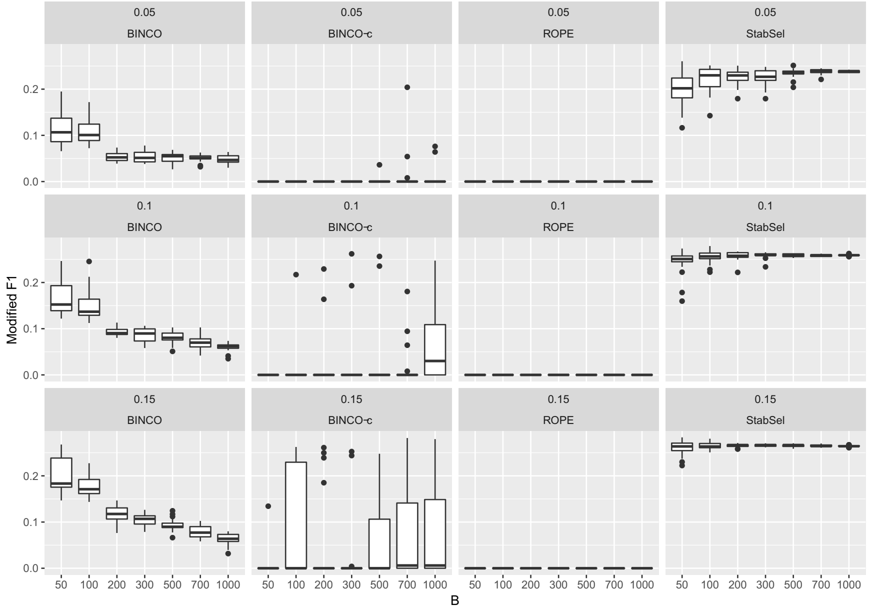

Figure 18: Network topology: hubby, steps: 15, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

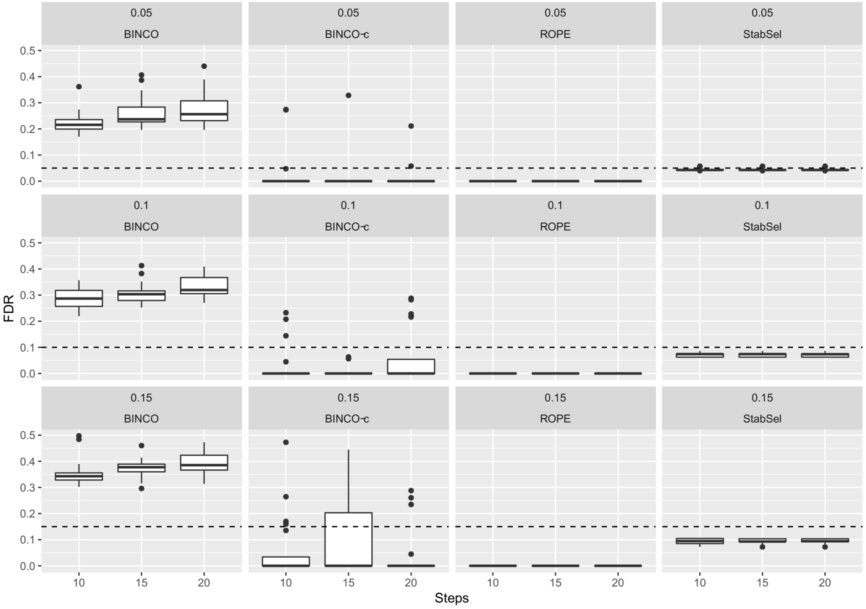

Figure 19: Network topology: hubby, $B = 500$, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

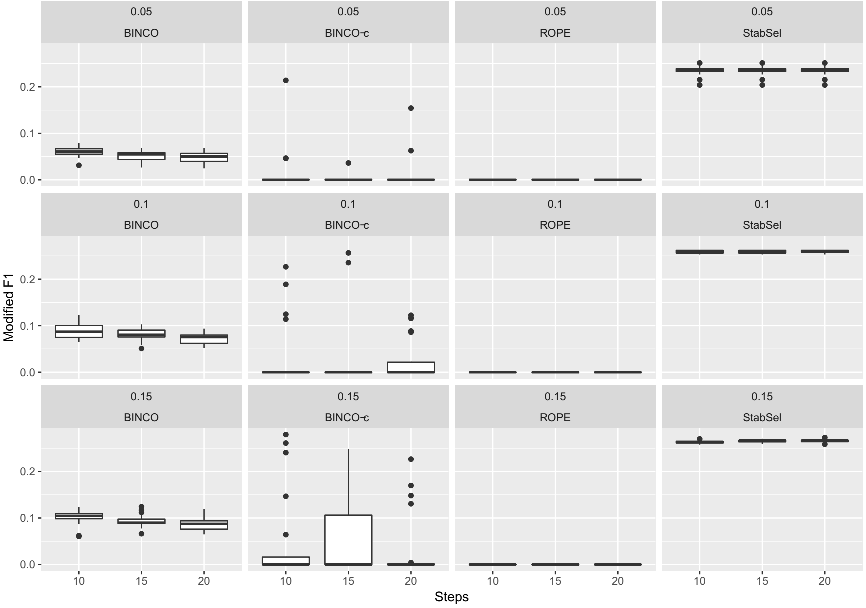

Figure 20: Network topology: hubby, $B = 500$, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

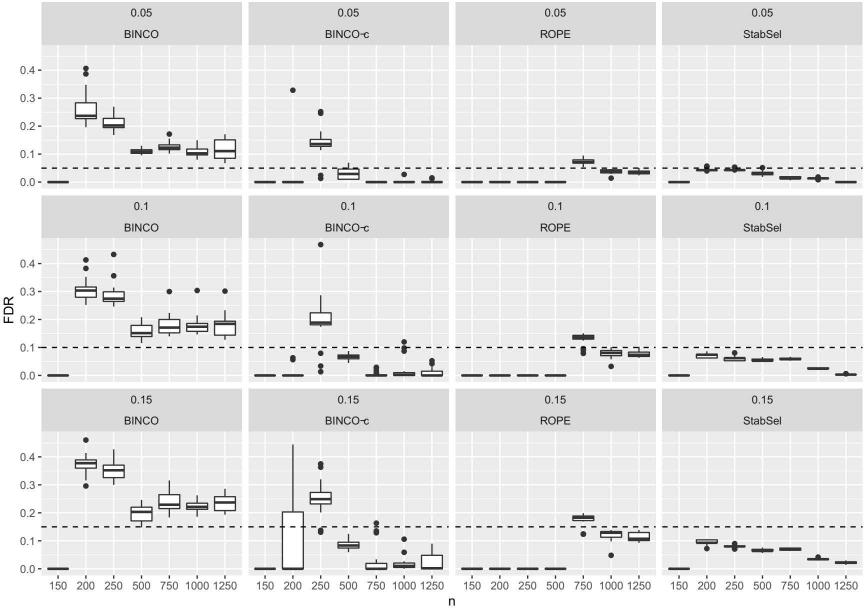

Figure 21: Network topology: hubby, $B = 500$, steps: 15, weakness: 0.8, facet titles: target FDR and method.

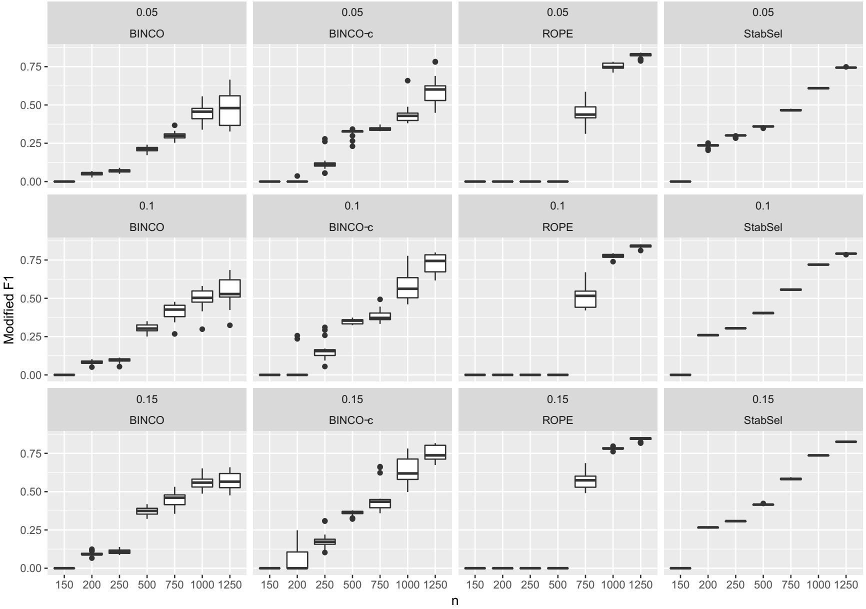

Figure 22: Network topology: hubby, $B = 500$, steps: 15, weakness: 0.8, facet titles: target FDR and method.

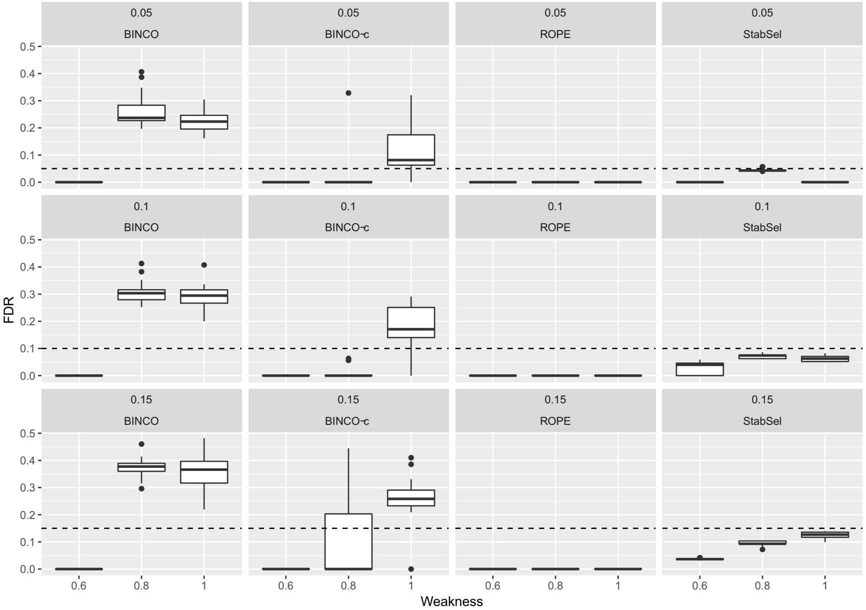

Figure 23: Network topology: hubby, $B = 500$, steps: 15, $n = 200$, facet titles: target FDR and method.

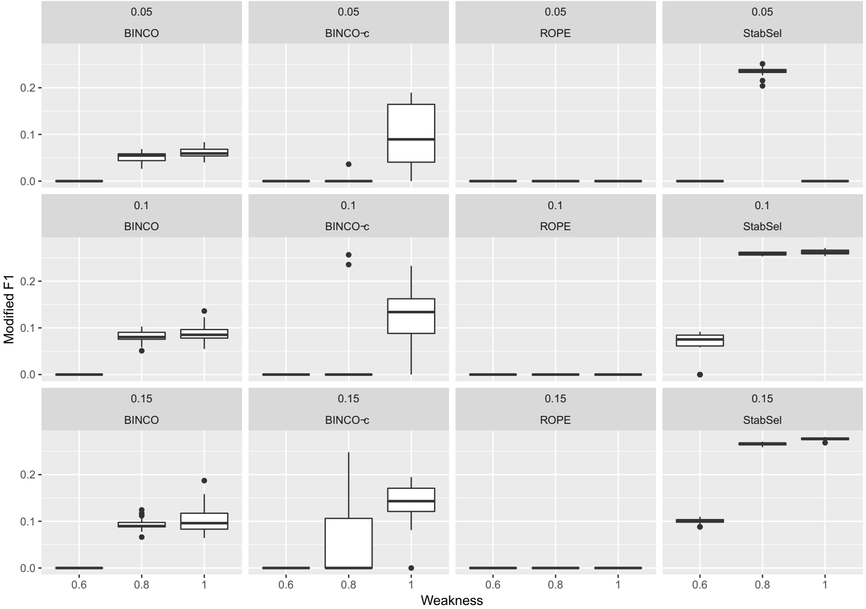

Figure 24: Network topology: hubby, $B = 500$, steps: 15, $n = 200$, facet titles: target FDR and method.

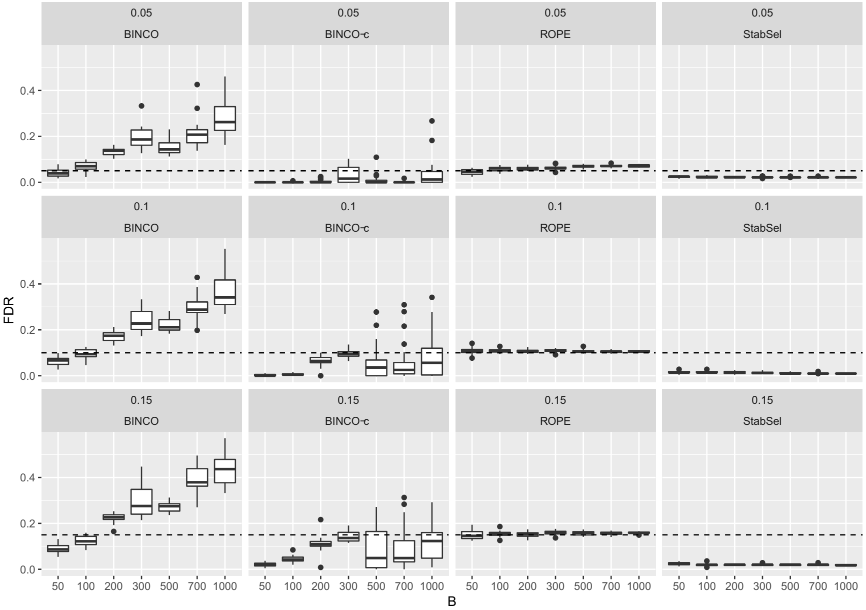

Figure 25: Network topology: scale-free, steps: 15, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

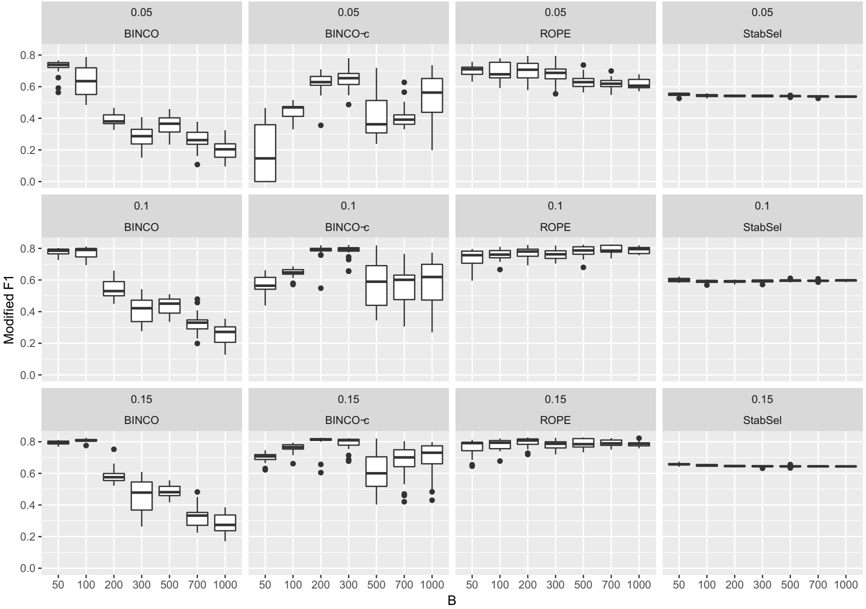

Figure 26: Network topology: scale-free, steps: 15, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

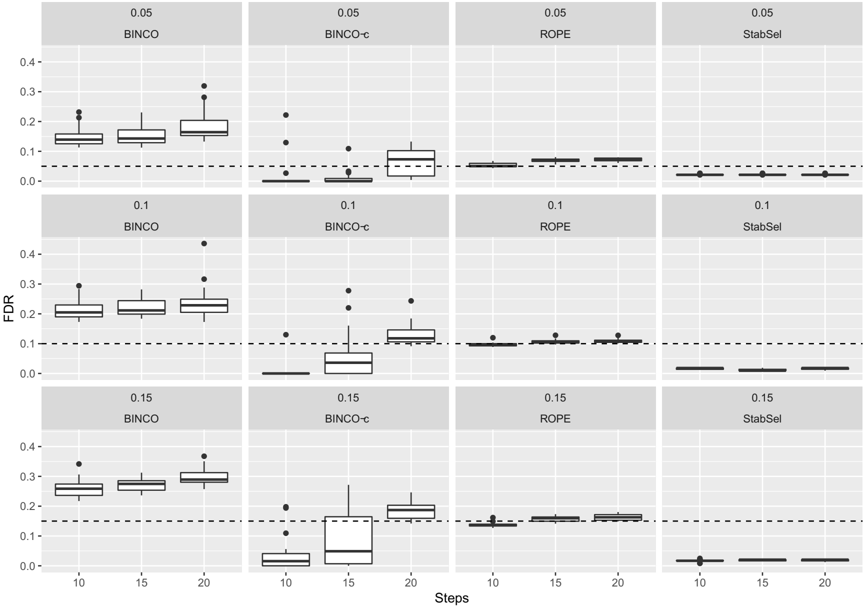

Figure 27: Network topology: scale-free, $B = 500$, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

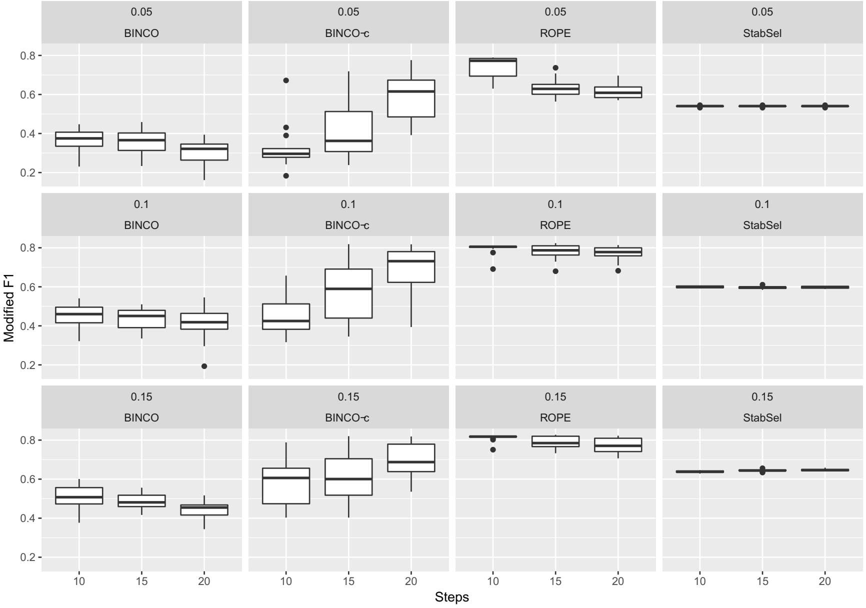

Figure 28: Network topology: scale-free, $B = 500$, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

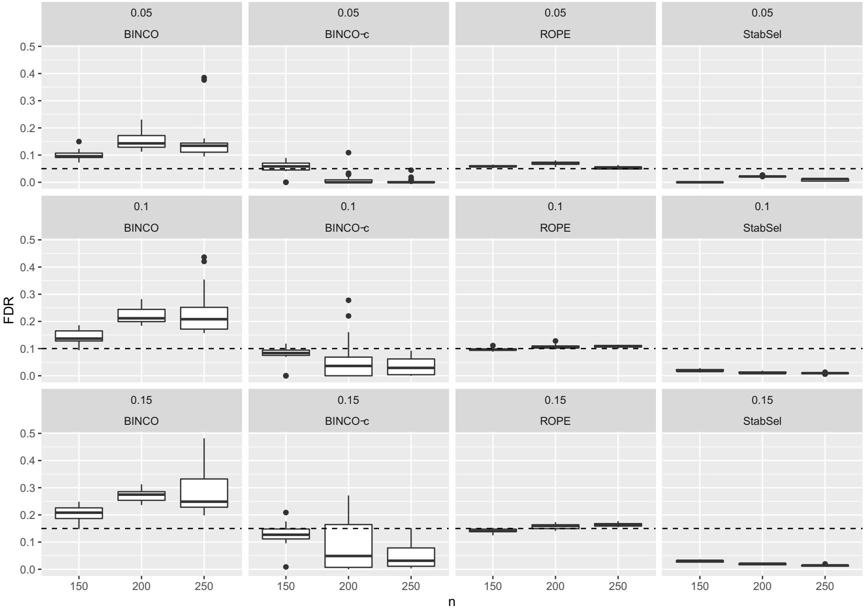

Figure 29: Network topology: scale-free, $B = 500$, steps: 15, weakness: 0.8, facet titles: target FDR and method.

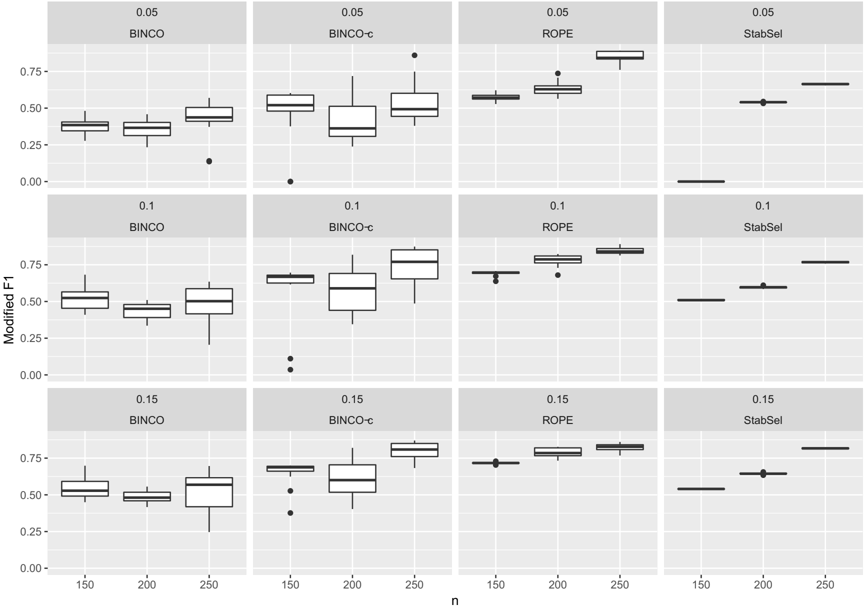

Figure 30: Network topology: scale-free, $B = 500$, steps: 15, weakness: 0.8, facet titles: target FDR and method.

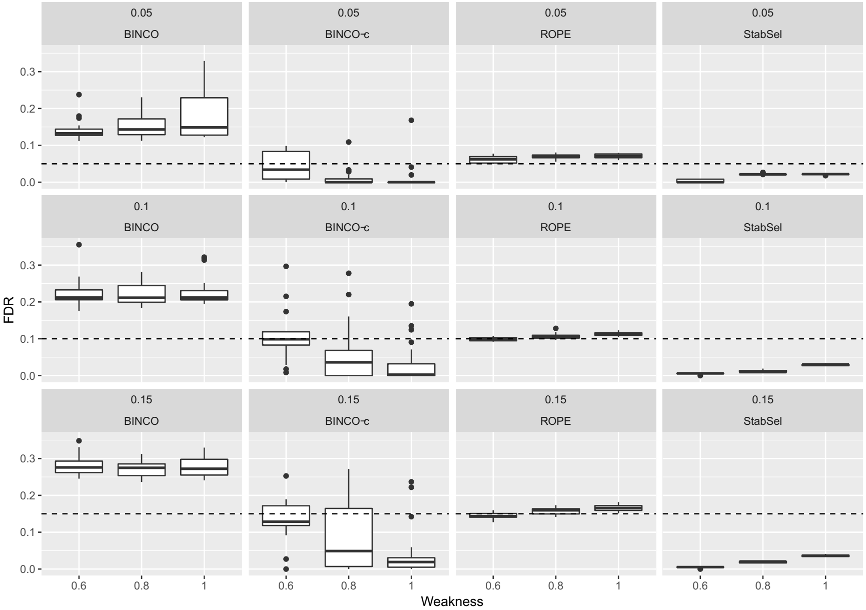

Figure 31: Network topology: scale-free, $B = 500$, steps: 15, $n = 200$, facet titles: target FDR and method.

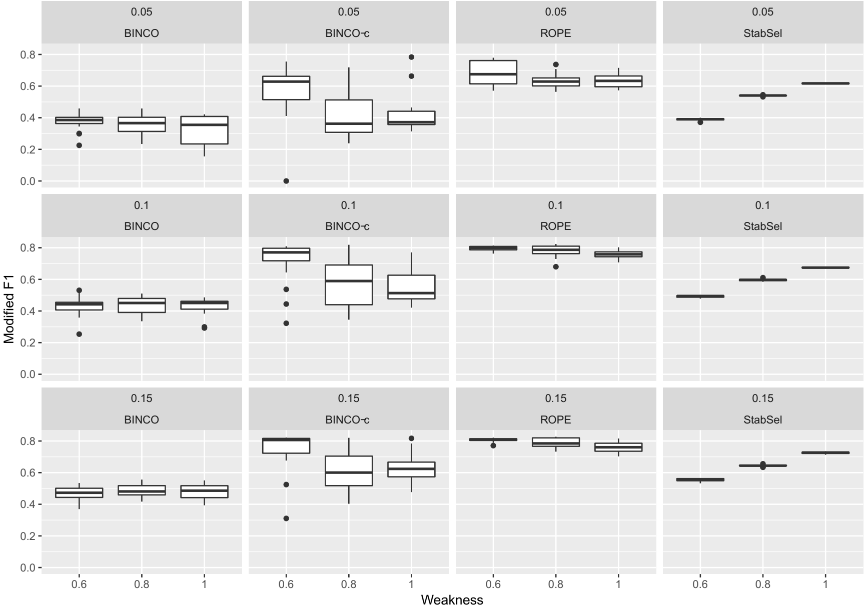

Figure 32: Network topology: scale-free, $B = 500$, steps: 15, $n = 200$, facet titles: target FDR and method.

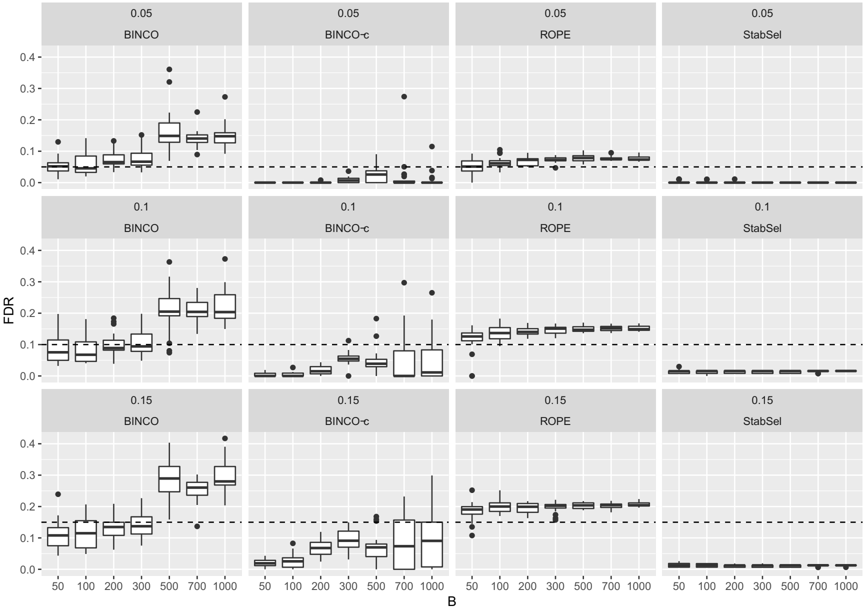

Figure 33: Network topology: small, steps: 15, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

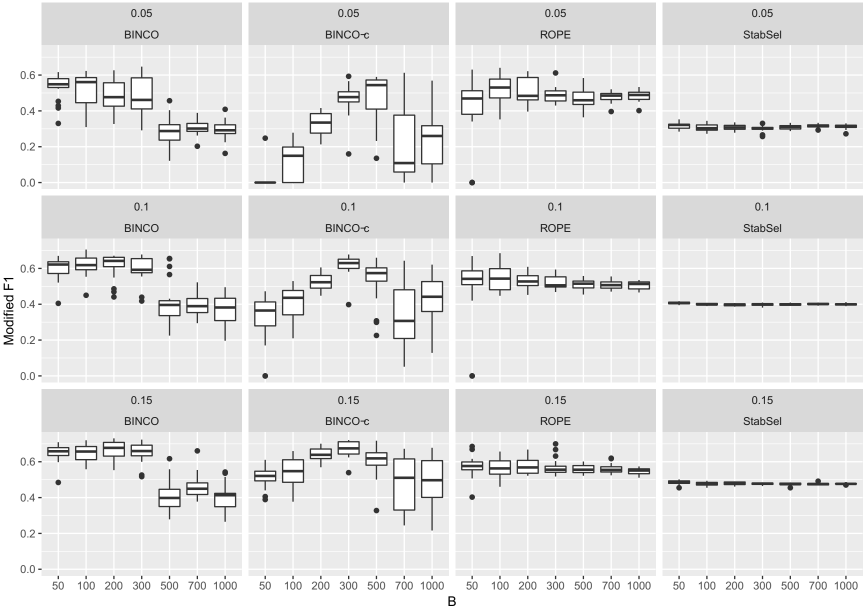

Figure 34: Network topology: small, steps: 15, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

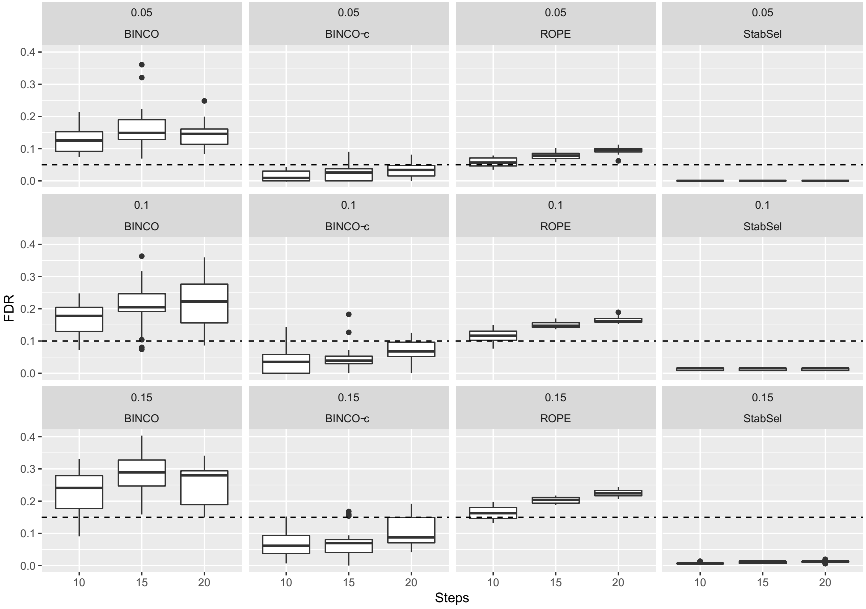

Figure 35: Network topology: small, $B = 500$, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

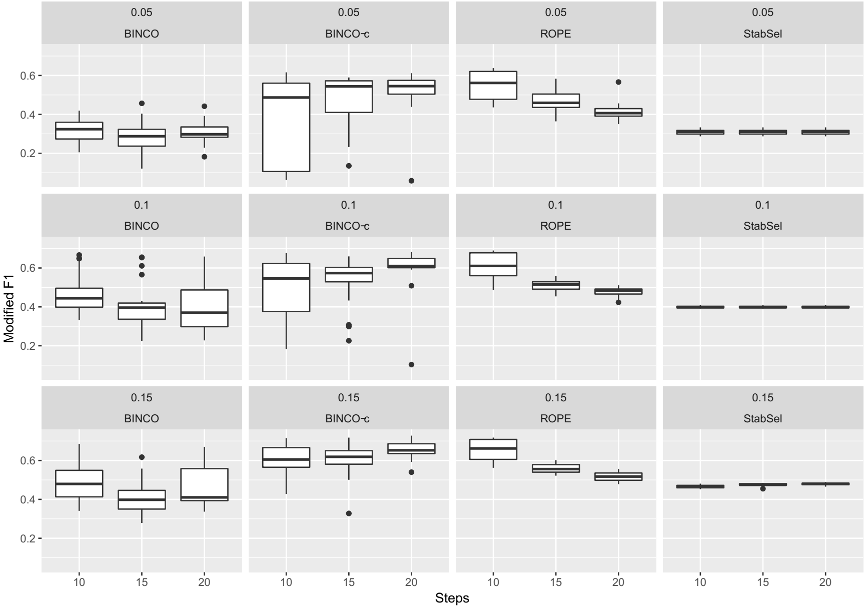

Figure 36: Network topology: small, $B = 500$, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

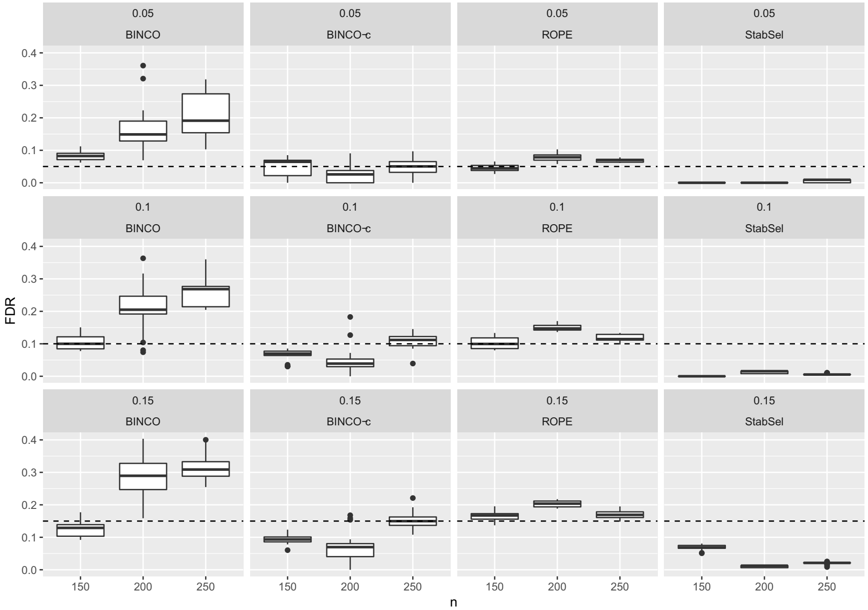

Figure 37: Network topology: small, $B = 500$, steps: 15, weakness: 0.8, facet titles: target FDR and method.

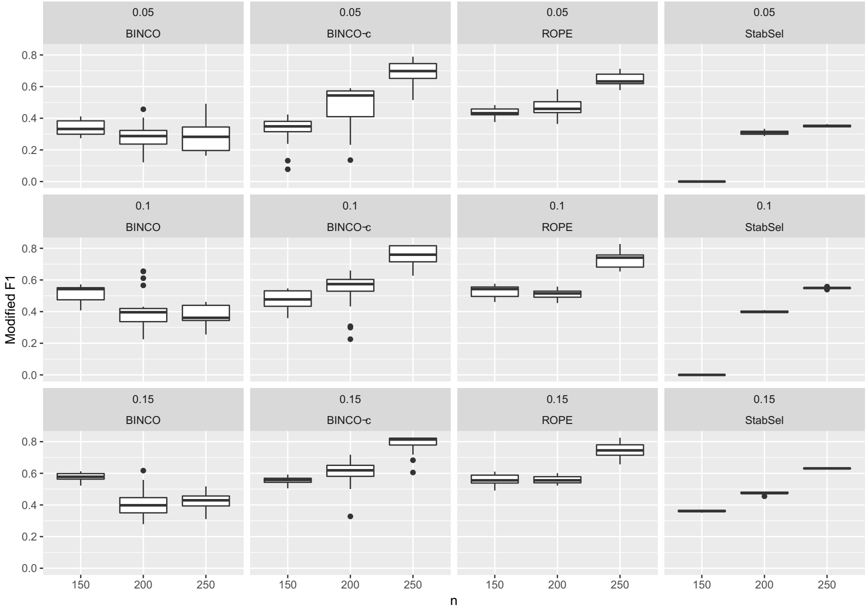

Figure 38: Network topology: small, $B = 500$, steps: 15, weakness: 0.8, facet titles: target FDR and method.

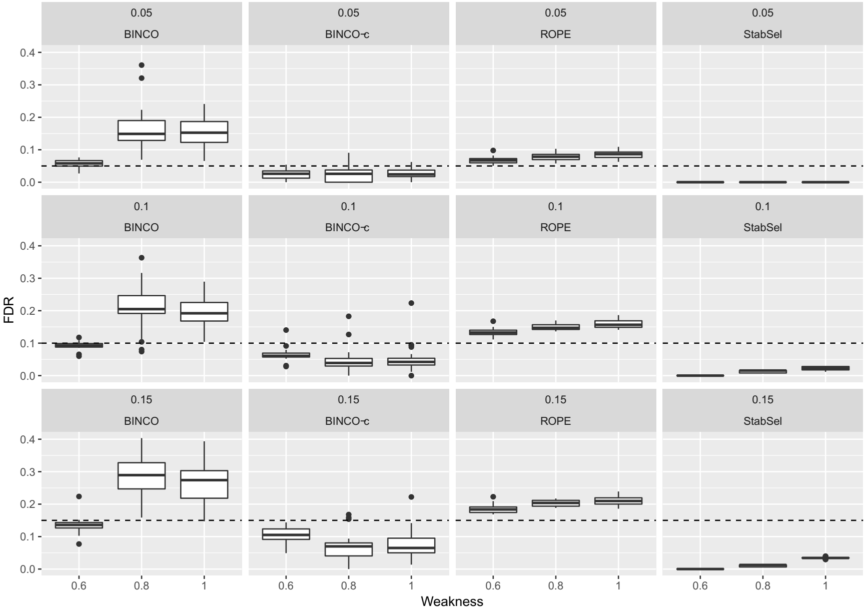

Figure 39: Network topology: small, $B = 500$, steps: 15, $n = 200$, facet titles: target FDR and method.

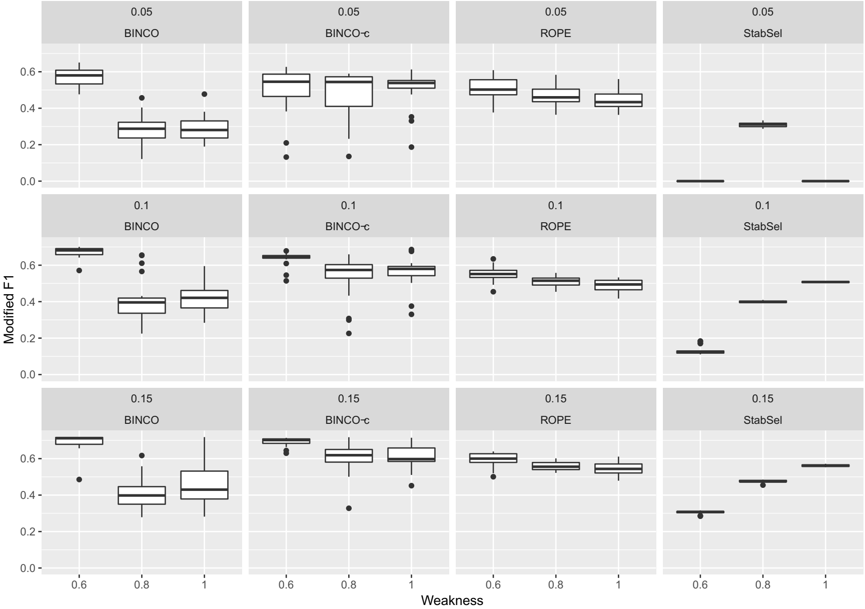

Figure 40: Network topology: small, $B = 500$, steps: 15, $n = 200$, facet titles: target FDR and method.

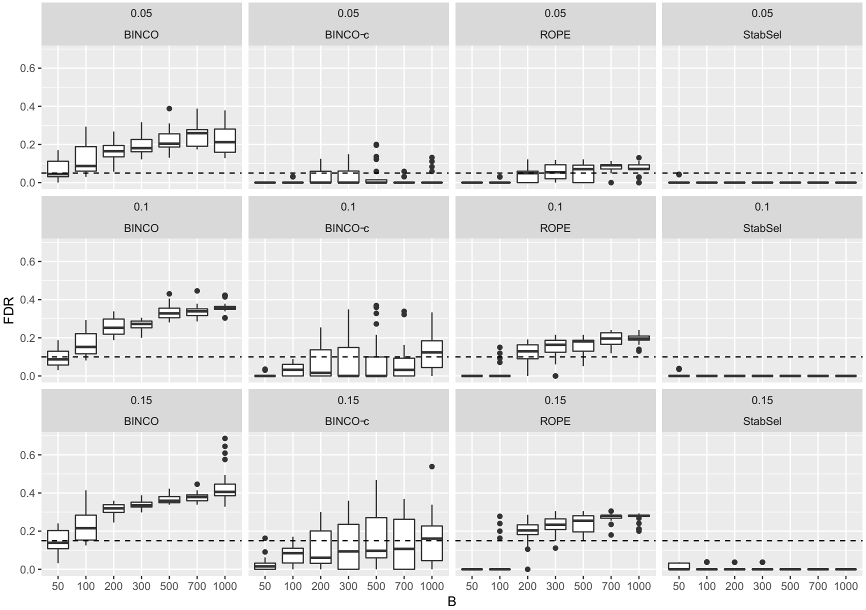

Figure 41: Network topology: sparse, steps: 15, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

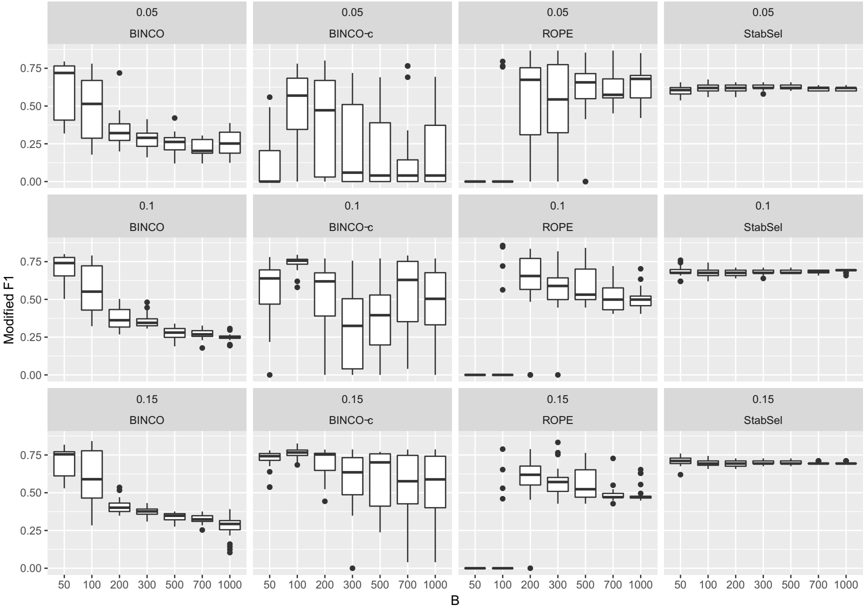

Figure 42: Network topology: sparse, steps: 15, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

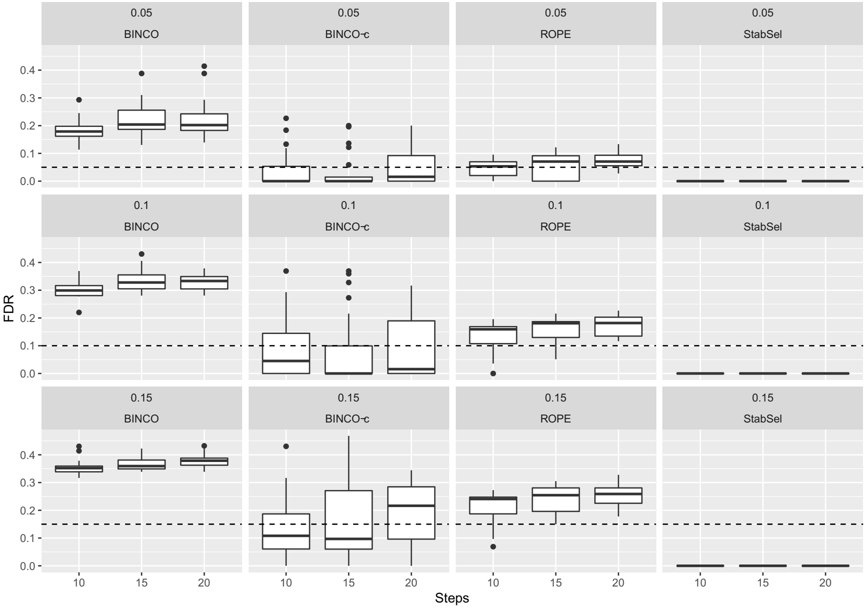

Figure 43: Network topology: sparse, $B = 500$, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

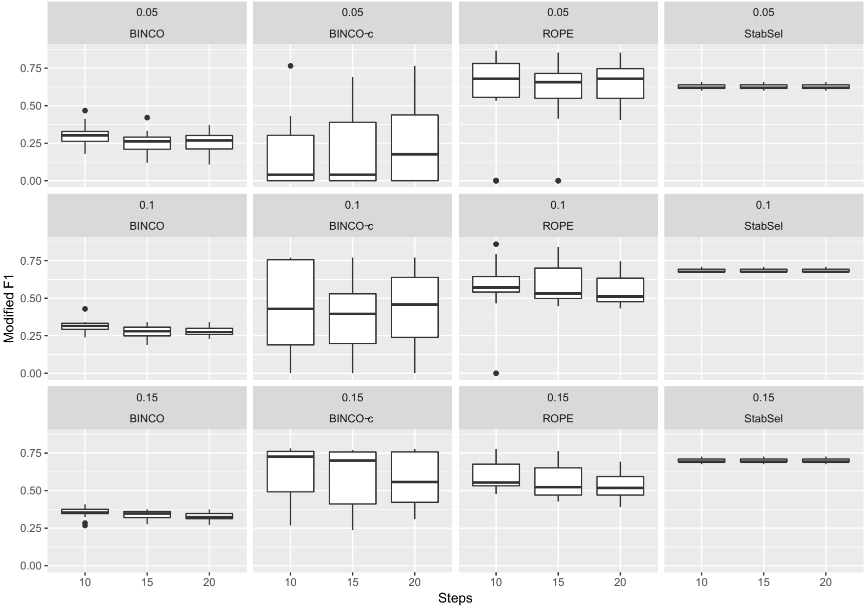

Figure 44: Network topology: sparse, $B = 500$, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

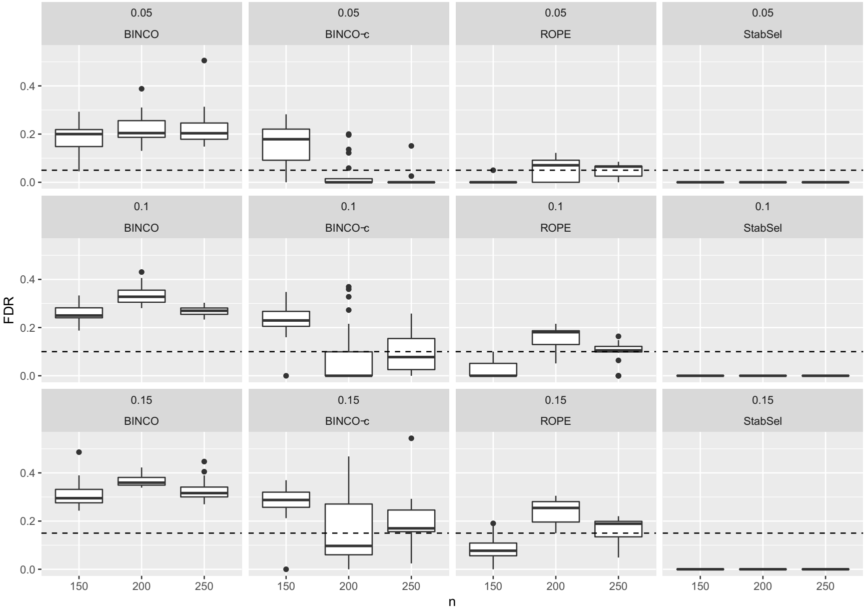

Figure 45: Network topology: sparse, $B = 500$, steps: 15, weakness: 0.8, facet titles: target FDR and method.

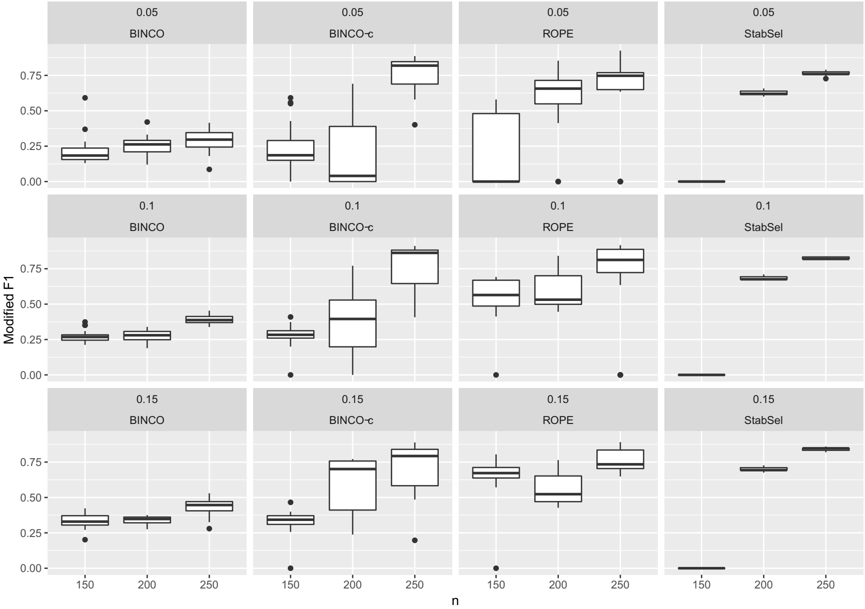

Figure 46: Network topology: sparse, $B = 500$, steps: 15, weakness: 0.8, facet titles: target FDR and method.

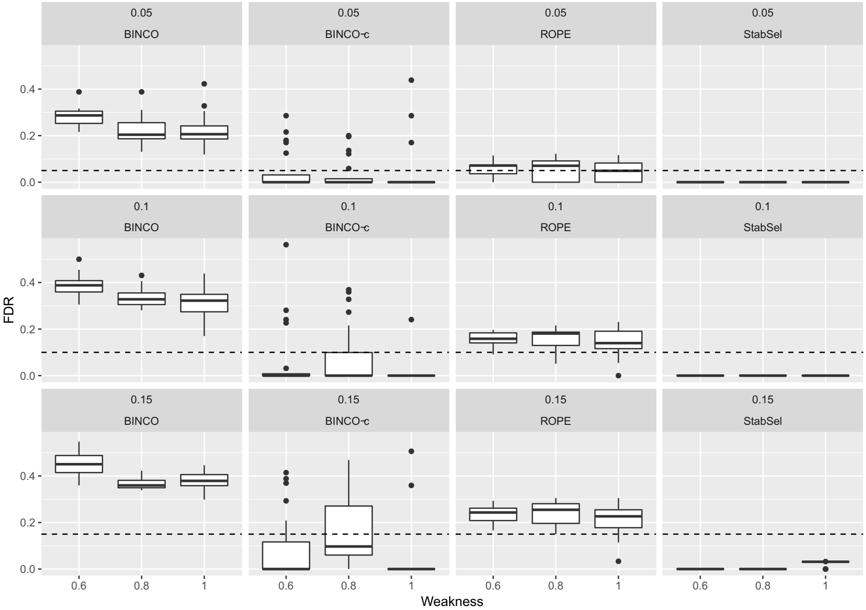

Figure 47: Network topology: sparse, $B = 500$, steps: 15, $n = 200$, facet titles: target FDR and method.

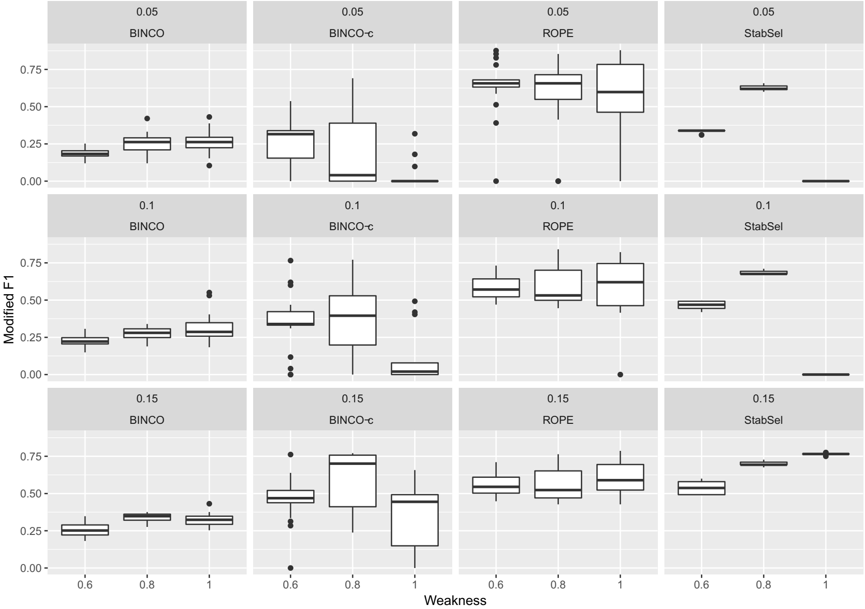

Figure 48: Network topology: sparse, $B = 500$, steps: 15, $n = 200$, facet titles: target FDR and method.

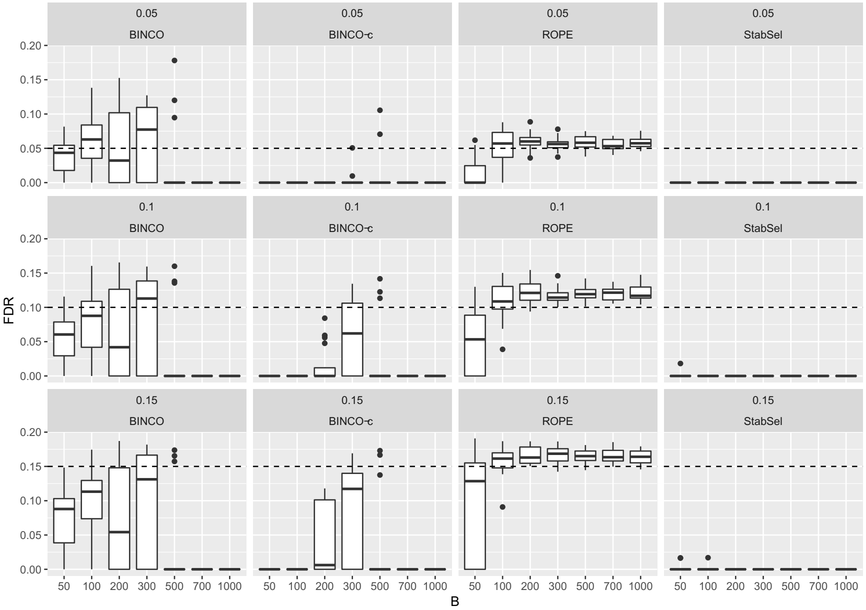

Figure 49: Network topology: weak, steps: 15, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

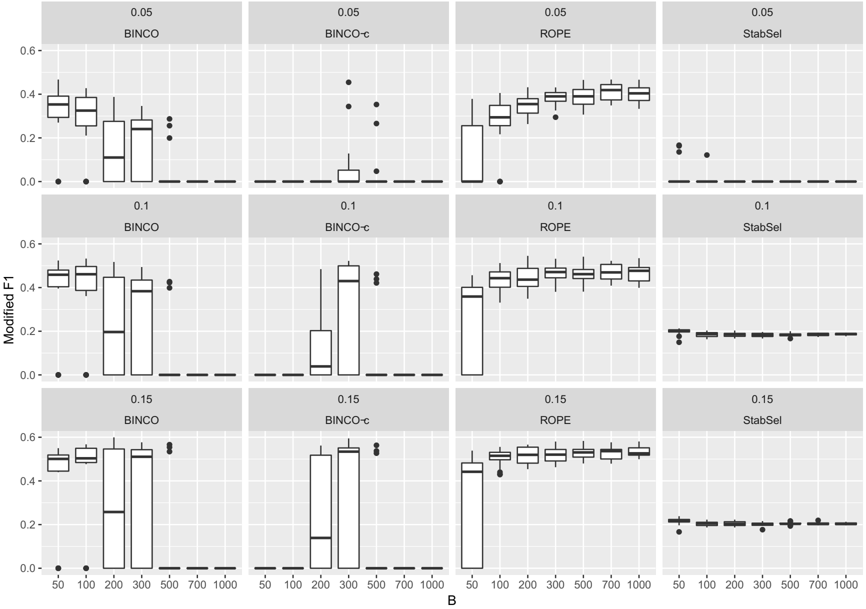

Figure 50: Network topology: weak, steps: 15, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

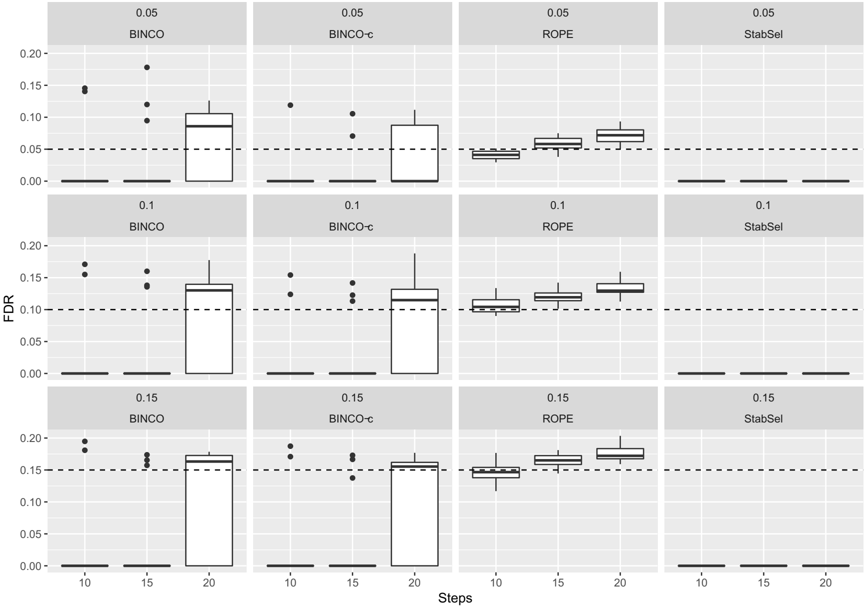

Figure 51: Network topology: weak, $B = 500$, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

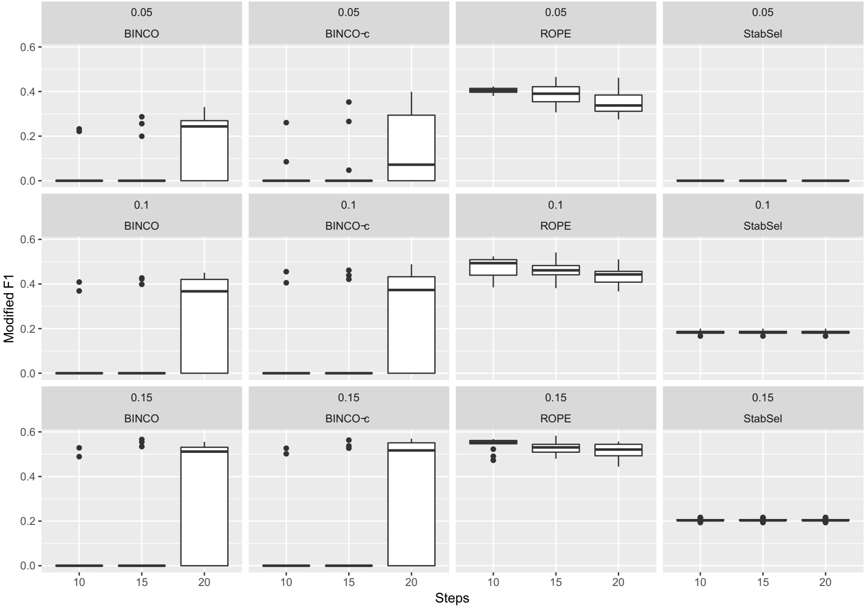

Figure 52: Network topology: weak, $B = 500$, $n = 200$, weakness: 0.8, facet titles: target FDR and method.

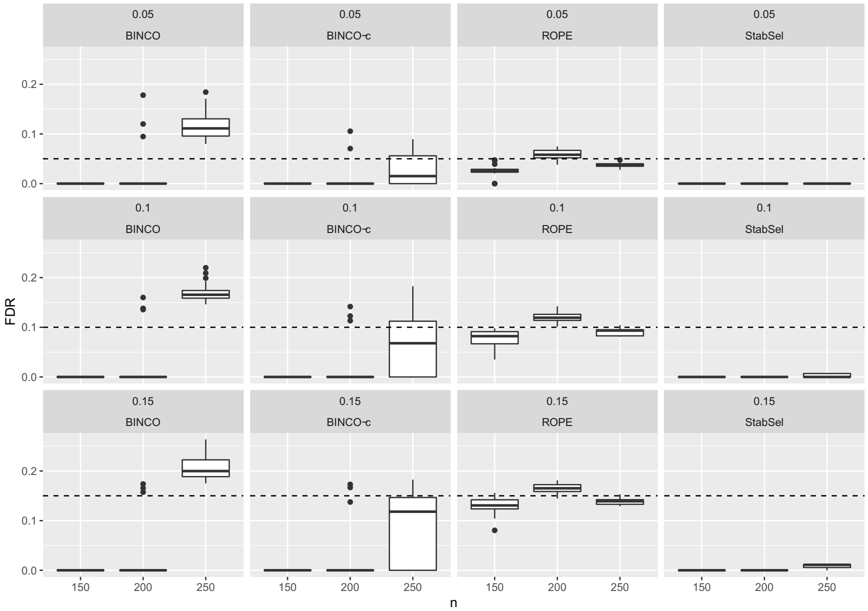

Figure 53: Network topology: weak, $B = 500$, steps: 15, weakness: 0.8, facet titles: target FDR and method.

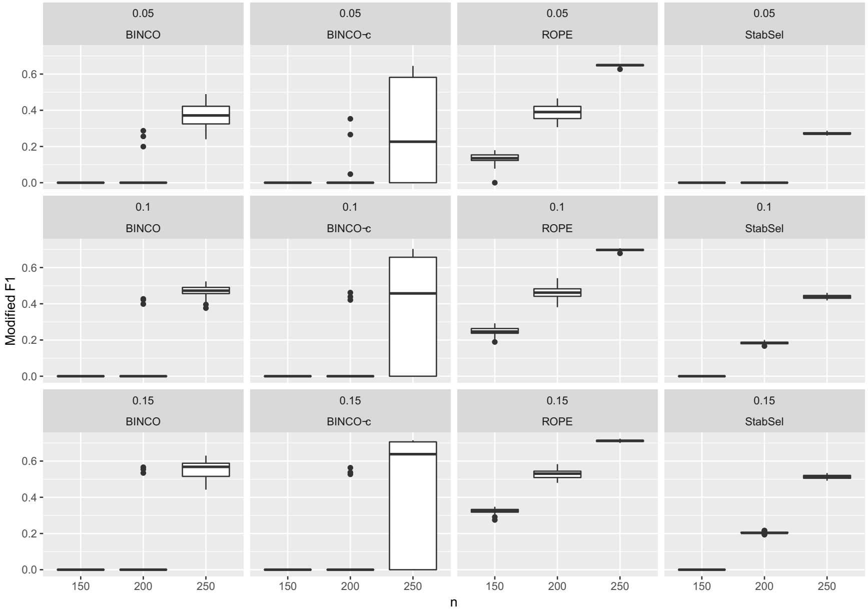

Figure 54: Network topology: weak, $B = 500$, steps: 15, weakness: 0.8, facet titles: target FDR and method.

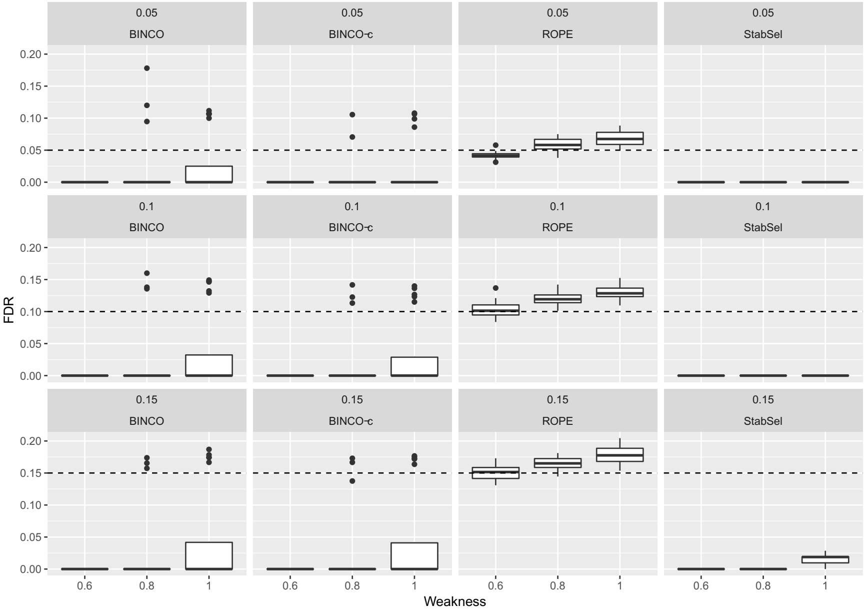

Figure 55: Network topology: weak, $B = 500$, steps: 15, $n = 200$, facet titles: target FDR and method.

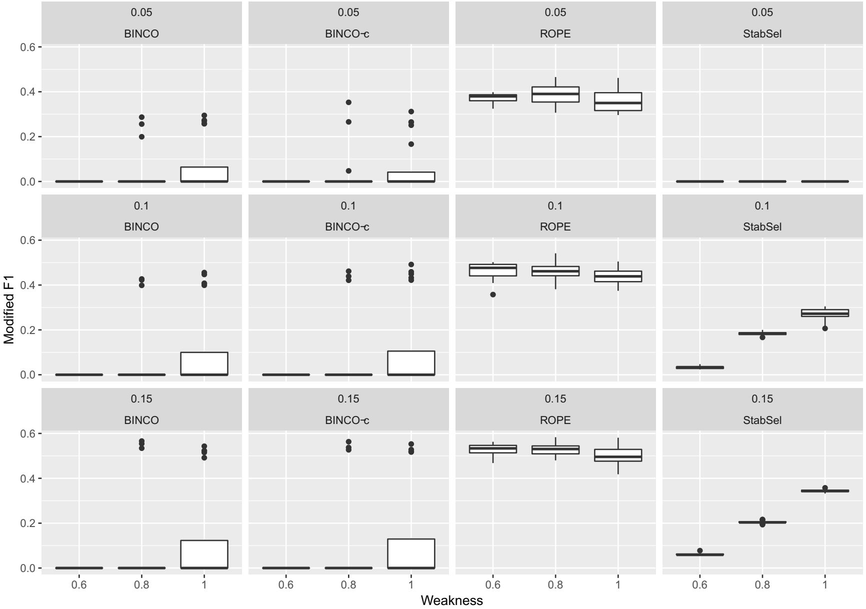

Figure 56: Network topology: weak, $B = 500$, steps: 15, $n = 200$, facet titles: target FDR and method.