Appendix A: Supplementary Material

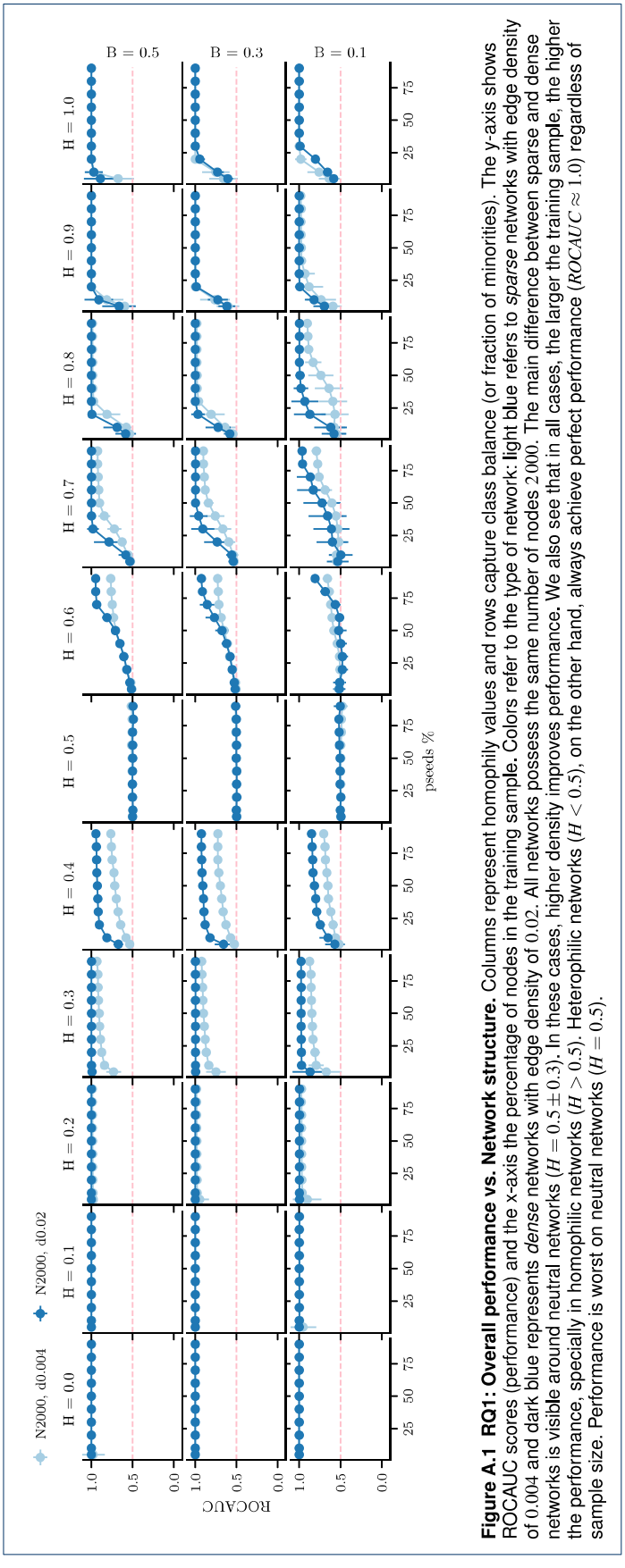


Figure A.1 RQ1: Overall performance vs. Network structure. Columns represent homophily values and rows capture class balance (or fraction of minorities). The y-axis shows ROCAUC scores (performance) and the x-axis the percentage of nodes in the training sample. Colors refer to the type of network: light blue refers to *sparse* networks with edge density of 0.004 and dark blue represents *dense* networks with edge density of 0.02. All networks possess the same number of nodes 2000. The main difference between sparse and dense networks is visible around neutral networks ($H = 0.5 \pm 0.3$). In these cases, higher density improves performance. We also see that in all cases, the larger the training sample, the higher the performance, specially in homophilic networks ($H > 0.5$). Heterophilic networks ($H < 0.5$), on the other hand, always achieve perfect performance ($ROCAUC \approx 1.0$) regardless of sample size. Performance is worst on neutral networks ($H = 0.5$).

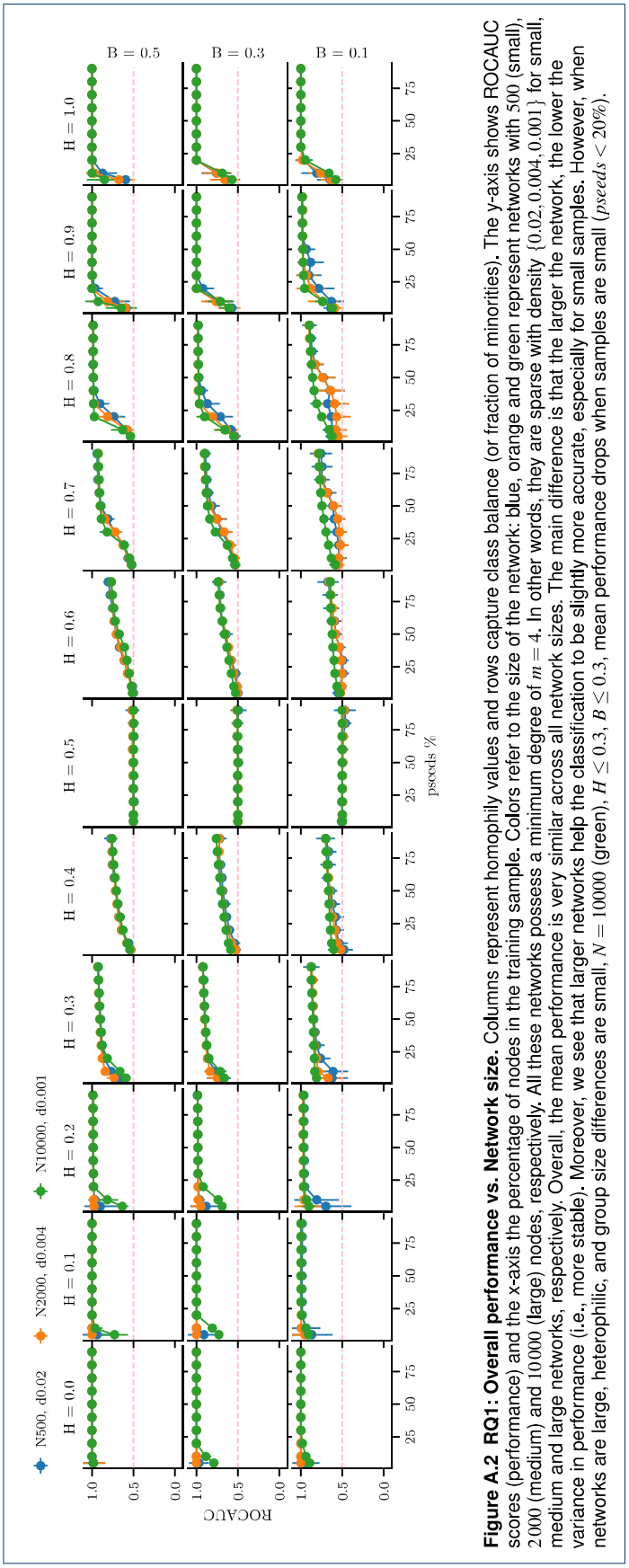


Figure A.2 RQ1: Overall performance vs. Network size. Columns represent homophily values and rows capture class balance (or fraction of minorities). The y-axis shows ROCAUC scores (performance) and the x-axis the percentage of nodes in the training sample. Colors refer to the size of the network: blue, orange and green represent networks with 500 (small), 2000 (medium) and 10000 (large) nodes, respectively. All these networks possess a minimum degree of $m = 4$. In other words, they are sparse with density $\{0.02, 0.004, 0.001\}$ for small, medium and large networks, respectively. Overall, the mean performance is very similar across all network sizes. The main difference is that the larger the network, the lower the variance in performance (i.e., more stable). Moreover, we see that larger networks help the classification to be slightly more accurate, especially for small samples. However, when networks are large, heterophilic, and group size differences are small, $N = 10000$ (green), $H \leq 0.3$, $B \leq 0.3$, mean performance drops when samples are small ($pseeds < 20\%$).

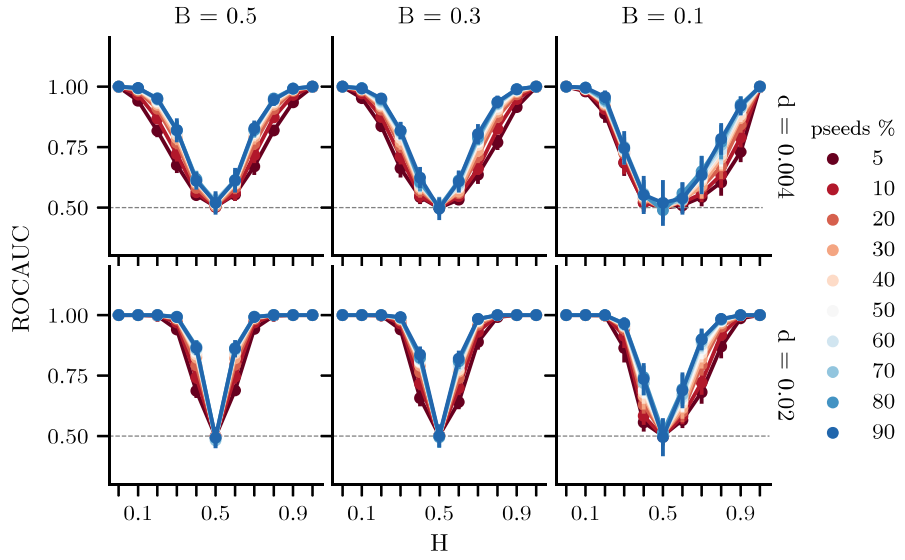


Figure A.3 RQ1: Network structure, sample size and overall performance using random node sampling and LINK classifier. These results reflect the interplay between network structure, sample size and classification performance. In Figure 2 results were obtained using the network-only Bayes classifier (nBC), while in this figure results were obtained using the LINK classifier. Overall, the main patterns persist: Classification on neutral networks is as good as a random classifier, and classification performance is best in the extreme heterophilic and homophilic regimes. However, results using LINK are more consistent across samples sizes (i.e., performance differences between small and large samples are small), and even within small training samples (i.e., variance is low). Also, ROCAUC scores using nBC are slightly higher than the ones obtained with LINK.

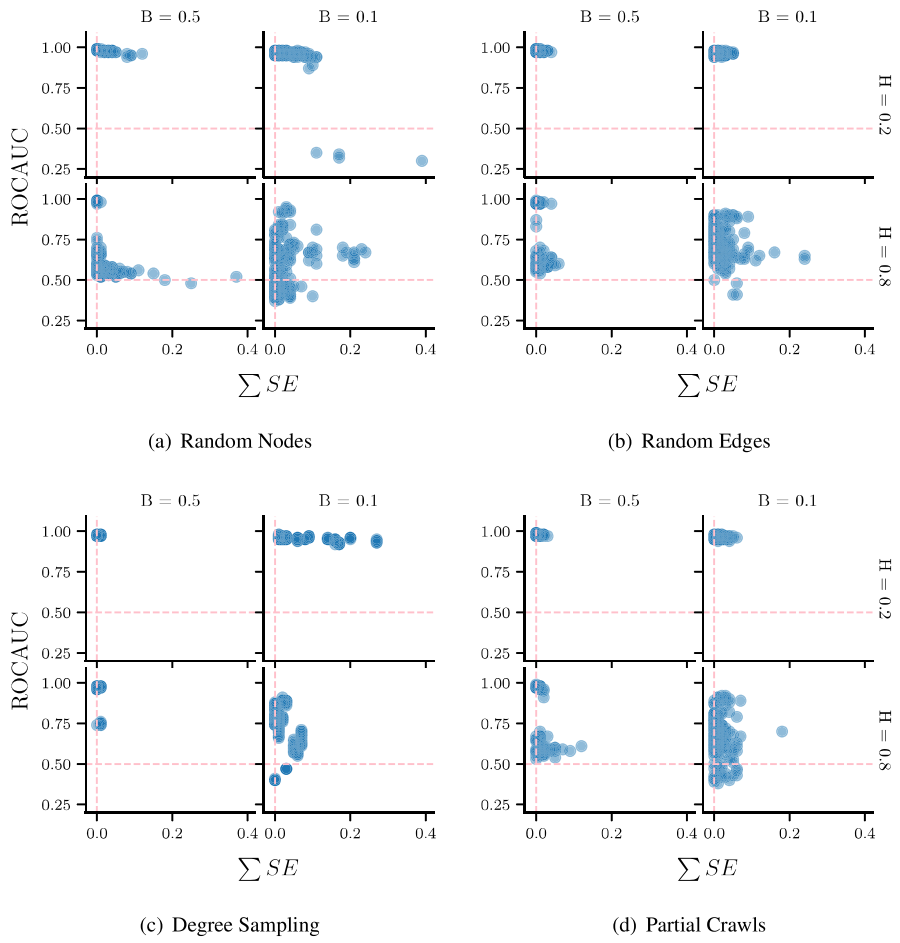
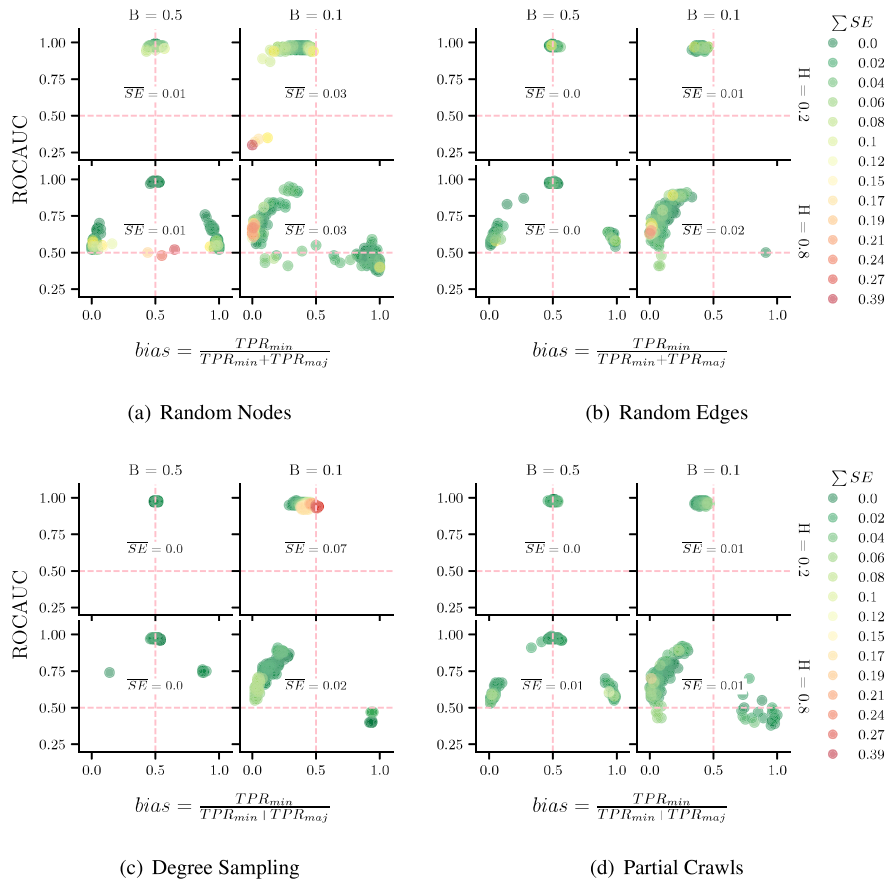


Figure A.4 RQ2: Overall performance vs. overall parameter estimation on small samples. These are the sum of squared estimation errors of class priors and conditional probabilities (x-axis) with their respective ROCAUC scores (y-axis). These are observations on small samples ($pseeds \leq 30\%$) using different sampling techniques: (a) random nodes, (b) random edges, (c) degree sampling, and (d) partial crawls. We see that across sampling techniques, heterophilic networks are consistent w.r.t performance, except a few cases in random nodes. In contrast, performance in homophilic networks is not consistent (i.e., high variance), and low estimation errors do not necessarily map to high performance. Overall, estimation errors are higher when sampling by random nodes and random edges.



(a) Random Nodes (b) Random Edges

(c) Degree Sampling (d) Partial Crawls

Figure A.5 RQ1, RQ2, RQ3: Overall performance, bias, and parameter estimation on small samples. This plot captures the relationship between network structure (class balance B as columns, homophily H as rows), sample size (only small samples $pseeds \leq 30\%$), parameter estimation (sum of estimation errors SE as colors), bias (x-axis) and classification performance (y-axis) using random node sampling. $\overline{SE}$ values refer to the mean sum of estimation errors in the corresponding type of network. We see the following patterns: (i) Heterophilic networks achieve the highest ROCAUC scores and their parameter estimation is almost perfect (i.e., sum of errors is close to zero; green), except a few cases in degree sampling where ROCAUC is high and errors are high (as shown in Section 5.2 and Section 7, this is due to erroneous class prior estimation). (ii) Unbiased results on heterophilic networks correlates with high performance and low estimation errors on unbalanced networks. Results on heterephilic and unbalanced networks can be slightly biased towards majority nodes without affecting much performance and estimation of model parameters. (iii) Unbiased classification in homophilic networks correlates with high performance and perfect parameter estimation. This only occurs when networks are balanced. Unbiased results often correlate with low performance. When homophilic networks are unbalanced, this bias can affect minorities but also majority nodes.

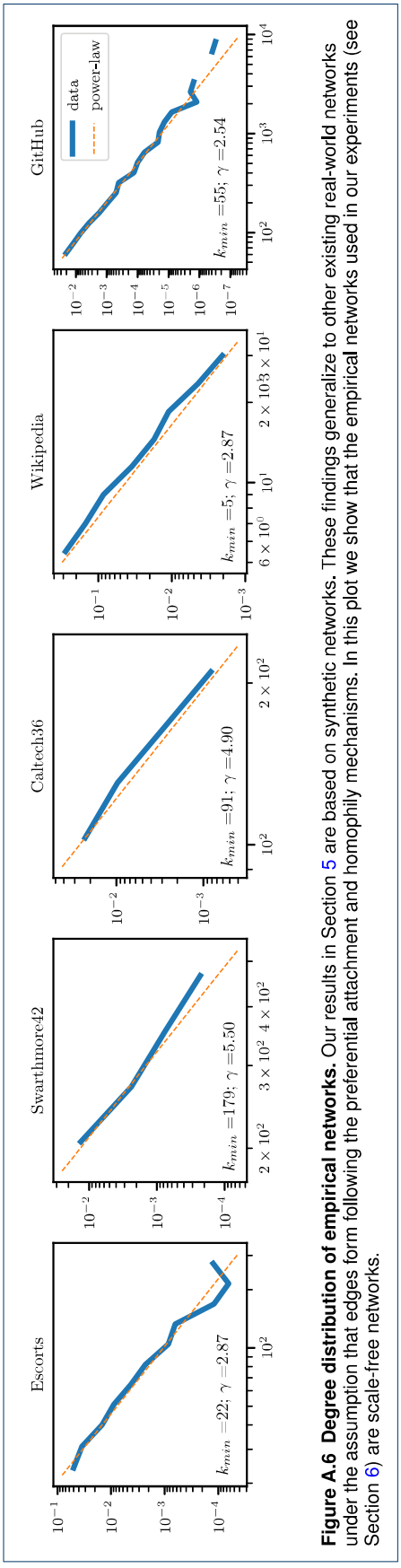


Figure A.6 Degree distribution of empirical networks. Our results in Section 5 are based on synthetic networks. These findings generalize to other existing real-world networks under the assumption that edges form following the preferential attachment and homophily mechanisms. In this plot we show that the empirical networks used in our experiments (see Section 6) are scale-free networks.