Analysis of correlation-based biomolecular networks from different omics data by fitting stochastic block models

Katharina Baum[1,2], Jagath C. Rajapakse[3,*], Francisco Azuaje[1,*]

[1] Bioinformatics and Modelling, Luxembourg Institute of Health, Luxembourg, 1A-B rue Thomas Edison, 1445 Strassen, Luxembourg

[2] Mathematical Modelling of Cellular Processes, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Robert-Rössle-Str. 10, 13125 Berlin, Germany

[3] School of Computer Science and Engineering, Nanyang Technological University, N4-2a06, 50 Nanyang Avenue, Singapore 639798

[*] Corresponding authors email: asjagath@ntu.edu.sg, francisco.azuaje@lih.lu
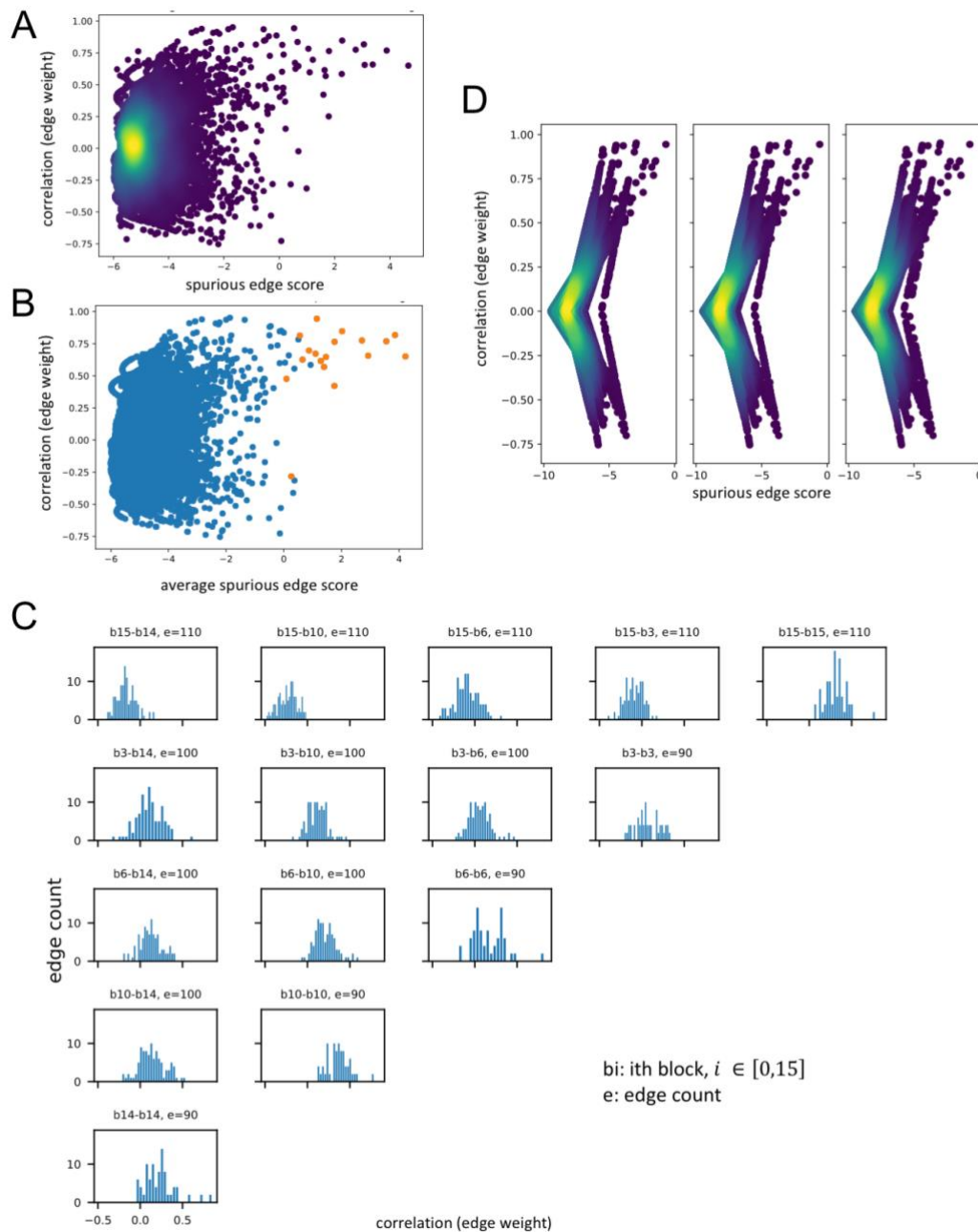
Supplementary Figure S1. The weighted SBM seems not appropriate for edge prediction from edge confidence scores for fully connected networks. We fitted the fully connected correlation-based metabolite network (162 nodes, 13041 edges, data from Budczies et al., 2013) to a weighted degree-corrected SBM with prior assumption of a real-normal distribution of edge weights between blocks for 50 random initializations using the Python module graph-tool (Peixoto, 2014). Overall, 16 blocks are predicted for the best fitting SBM. **(A)** Edge weight (correlation) vs. spurious edge confidence scores for the best fitting SBM. Against expectations on edge confidence scores, mainly edges with high weight are predicted as spurious. **(B)** Edge confidence scores averaged over the three best fitting SBMs show similar results as the single best fit. Edges with score > 0 in all three SBMs are highlighted in orange. **(C)** Edge weight distributions between the largest 5 blocks of the SBM from (A). Edges with high weight are outliers in the edge weight distributions and consequently are predicted as spurious. **(D)** Edge confidence scores for the best fitting weighted SBM with a real-exponential distribution as prior assumption for the edge weights. In contrast to the real-normal assumption, only 6 blocks are predicted and different edge confidence scores are obtained, which are visibly dominated by their block association. Again, against expectations, edges with high weights tend to be predicted most spurious. This shows the strong influence of the prior on global SBM characteristics and edge prediction for fully connected weighted networks.
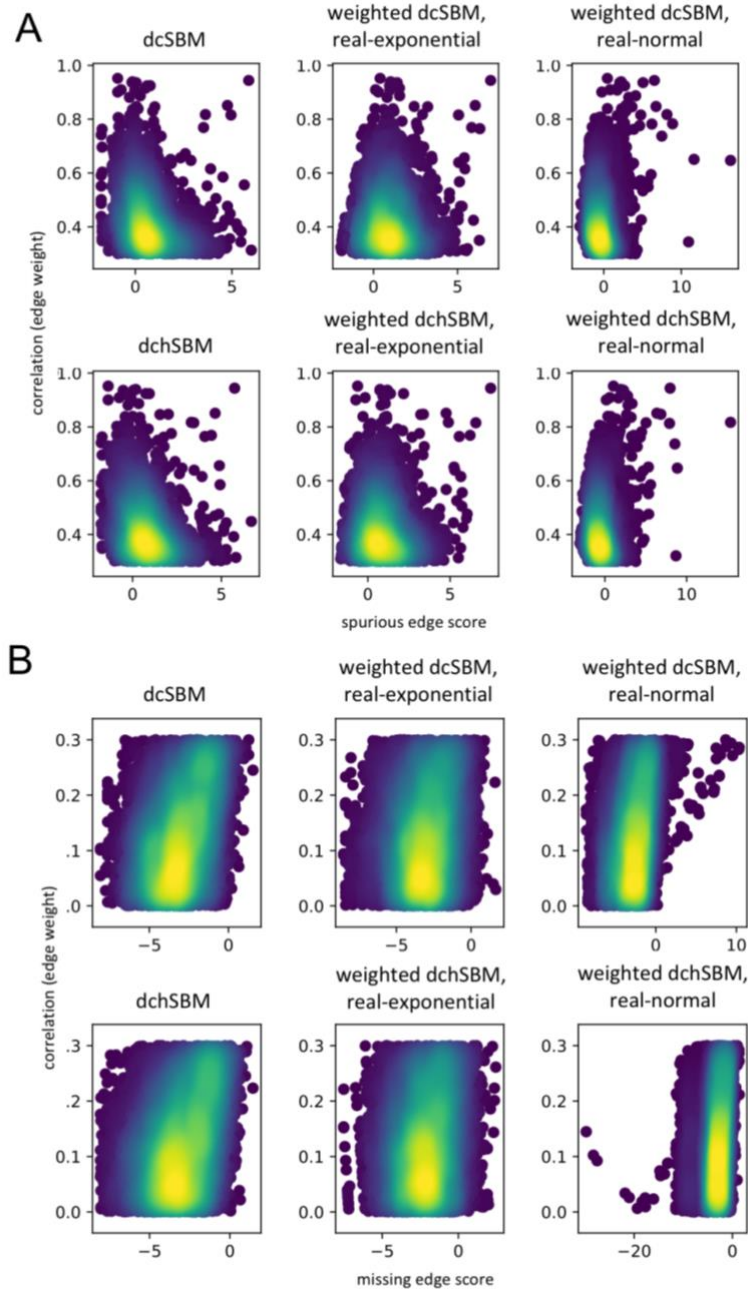
Supplementary Figure S2. Edge predictions for a reduced weighted network with planar or hierarchical weighted SBMs. We fitted the correlation-based metabolite network reduced by neglecting edges with |correlation| < 0.3 to the unweighted, degree-corrected planar (dcSBM) or hierarchical SBM (dchSBM) or to weighted SBMs (dcwSBM or dchSBM) using a real-exponential distribution (middle) or a real-normal distribution (right) as prior assumptions for the edge weight distribution in the Python module graph-tool. **(A)** Absolute edge weights (correlations) versus the according SBM-derived spurious edge confidence scores. **(B)** Absolute edge weights (correlations) versus the according SBM-derived missing edge confidence scores. Edge predictions are overall in accordance with expectations on edge relevance given by correlations, i.e., edges with large absolute correlation are preferentially predicted as missing and not as spurious.