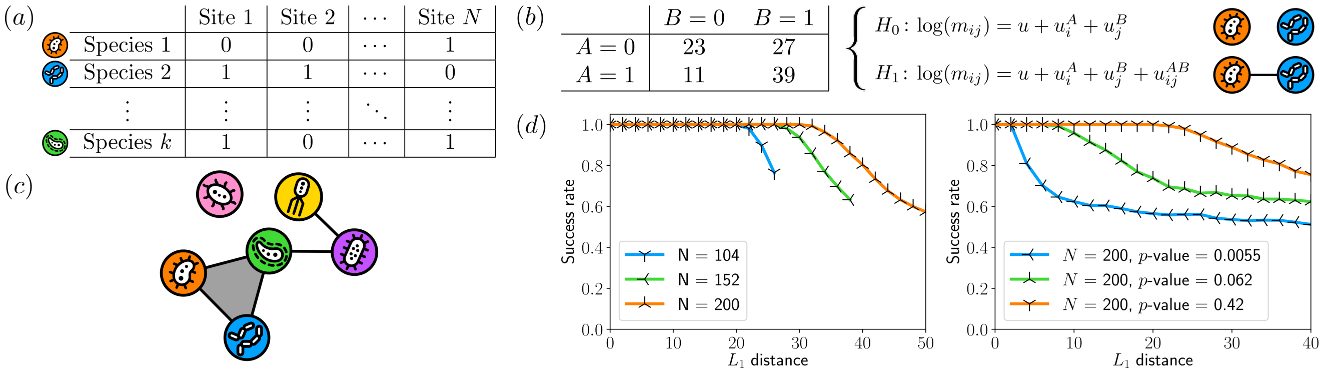
Inferring higher-order co-occurrence patterns and simplicial complexes from presence/absence data

Xavier Roy-Pomerleau[1,2], Louis J. Dubé[1,2], and Patrick Desrosiers[1,2,3]

1. Département de physique, de génie physique et d'optique, Université Laval, Québec (QC), G1V 0A6, Canada
2. Centre interdisciplinaire de modélisation mathématique de l'Université Laval, Québec (QC), G1V 0A6, Canada
3. Centre de recherche CERVO, Québec (QC), G1J 2G3, Canada

Recent work has shown that considering higher-order interactions improve the explanatory power of theoretical network models in various situations [1,2]. Yet, whether relationships between nodes are dyadic or not is rarely explicit in real data sets. This is particularly true in ecological studies [3] where the interactions between species are often assessed using presence/absence data, similar to the binary table of FIG. (a). To address this problem, we have developed a general framework to detect higher-order co-occurrence patterns from presence/absence data. Our method relies on log-linear models and hypothesis testing. As illustrated in FIG. (b), we first build contingency tables for all pairs of "species" in the presence/absence matrix and use the independence test, which provides a network of statistically significant pairwise co-occurrences. Then, log-linear models [4] are fitted on $2 \times 2 \times 2$ contingency tables formed by cliques of three interconnected nodes within the network. Cliques in which triple co-occurrence patterns can only be explained by the addition of higher-order terms in the log-linear model are promoted to the rank of 2-simplices. This procedure is then repeated on contingency tables of higher dimensions until the desired order is obtained. The resulting structure is a simplicial complex [1] in which simplices of dimension two and over represent higher-order co-occurrence patterns. An example is given in FIG. (c). To quantify the performance on noisy data, we perturb the entries of manually designed contingency tables for which the conclusion of the hypothesis test is known. For a specific L_1 distance (i.e. sum of absolute differences) from the designed table, we generate perturbed tables that contain the same number of observations as the original table. The performance analysis is illustrated in FIG. (d). Our results show that a higher number of observations render the statistical analysis more robust to noise regardless of the table's dimensions. However, the robustness to noise varies largely from one table to the other and constitutes a warning that real data have to be collected with uttermost care and rigour before concluding that higher-interactions exist.

[1] C. Giusti, R. Ghrist, and D.S. Bassett (2016). *Journal of computational neuroscience*, **41**, 1–14.
[2] R. Lambiotte, M. Rosvall, and I. Scholtes (2019). *Nature physics*, **15**, 313—320.
[3] J. Comte, C. Lovejoy, S. Crevecoeur, and W.F. Vincent (2016). *Biogeosciences*, **13**, 175–190.
[4] S.E. Fienberg and A. Rinaldo (2007). *Journal of Statistical Planning and Inference*, **137**, 3430–3445.



(a)

	Site 1	Site 2	⋯	Site N
Species 1	0	0	⋯	1
Species 2	1	1	⋯	0
⋮	⋮	⋮	⋱	⋮
Species k	1	0	⋯	1

(b)

	$B=0$	$B=1$
$A=0$	23	27
$A=1$	11	39

$$\begin{cases} H_0: \log(m_{ij}) = u + u_i^A + u_j^B \\ H_1: \log(m_{ij}) = u + u_i^A + u_j^B + u_{ij}^{AB} \end{cases}$$

(c)

(d)

FIG. (a) Representation of presence/absence matrix for k species (nodes) and N sites (observations or samples). (b) Contingency table for species A and B in the case of $N = 100$ sites together with the hypotheses H_0 and H_1 to be tested. The symbol m_{ij} refers to entries in the table and the u terms are parameters to be fitted. If H_0 is more plausible than H_1, then both species are considered independent and no link is introduced in the network. Otherwise, a link is added to the network. (c) Representation of a simplicial complex where simplices of dimension 1 and above represent significant co-occurrence patterns. (d) Success rate for designed $2 \times 2 \times 2$ tables. On the left, the model of total independence, $H_0 : \log(m_{ijk}) = u + u_i^A + u_j^B + u_k^C$, fits the designed tables perfectly. On the right, we test the model of no second-order interaction, $H_0 : \log(m_{ijk}) = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC}$, where a p-value below $\alpha = 0.01$ suggests a higher-order co-occurrence. Otherwise, pairwise interactions are sufficient. Each point is an average over 10 samples of 1000 perturbed tables for a given L_1 distance. The variance (not shown) is of the size of the curves. On the left, as N increases, the last value of L_1 for which the success rate is 1.0 are 20, 26, and 30. On the right, these values are 2, 6, and 20.