Reviewed Report on the Comparison of Secondary Extinctions and Largest Impact on Shortest Path (LIASP) in random artificial communities

Rafael Menezes e Amanda Campos

01/06/2023

Introduction and Summary

Once extinct, keystone species cause a disproportional impact on their community relative to the removal of other species from the same community. Here we use 2 approaches to estimate a keystone species. In a dynamical approach, we remove a species from a community and measure the impact of its removal in terms of secondary extinctions, which refer to the extinctions of other species (nodes) that occur following the removal of a single species. This measure is based on the temporal evolution of the community after the removal of the organism (node). In a structural approach, we use LIASP (largest impact on shortest path) to quantify the increase in the shortest path of the community following the removal of a single species (node). In this approach, the impact of removing a species can be further classified into direct and indirect effects: the former is due to the disconnection of the node from its former neighbors, and the latter is related to changes in the shortest path that passed through the removed node. The dynamical (Secondary Extinctions metric, or SE) and structural (LIASP metric) approaches are complementary and provide insights from different perspectives into the consequences of removing a species for the ecological community. In this report, we will compare these metrics, LIASP and SE, by generating artificial communities with a clustered small-world topology, using the Klemm-Eguiluz algorithm, and simulating the dynamical evolution of the network using a generalized Lotka Volterra model.

Here we found that, generally, there is a low correlation between LIASP and SE metrics in terms of identity and amount of keystone organisms (node) identified in the artificial ecological communities. However, we found that the proportion of positive interactions was relevant for the precision of LIASP in predicting the same keystone as SE metric, indicating that both metrics convey more similar information in mutualistic/comensalistic networks as opposed to antagonistic ones, even thou the mechanism undelying this pattern is obscure.

Methods

We artificially generated ecological communities to explore the relationship between the LIASP and SE metrics. We generated 150 networks with varying levels of connectance (from 0.01 to 0.1) and proportion of positive interactions in the network (from 0% to 100%). We calculated the SE value and the LIASP for each node in each network.

Topological Features of the Artificial Communities

Microbial interaction networks reconstructed from data are typically clustered, sparse and present small-world behavior (Berry and Widder, 2014). To simulate networks with these characteristics, we used the algorithm proposed by Klemm-Eguíluz (Klemm and Eguíluz, 2002) and implemented in the R package

`seqtime` (Faust et al, 2018). We generated artificial ecological networks with 100 populations and, to produce the clustering effect, we set the clique number parameter to 10. To match the connectance reported for microbial networks, we adjusted the interaction matrix connectance to either 0.1, 0.05, or 0.01. The small connectance values could result in some isolated organisms, which also appears in the co-occurrence networks in our work (Ferreira et al, in prep.).

Dynamical Features of the Artificial Communities

The generalized Lotka-Volterra (gLV) model is a widely used system of ordinary differential equations (ODE) model in theoretical ecology. It assumes that the growth rate of a population is a linear function of the abundances of other populations, with the interaction intensity and signal given by the interaction intensities a_{ij}. The basal growth rates r_i indicate the per-capita growth rate of a population when grown in isolation, which depends on the abiotic conditions of the environment. Denoting the vector of populations' abundances by $\vec{N}$, the matricial formulation of the gLV is given by:

$$\frac{d\vec{N}}{dt} = \operatorname{diag} \vec{N} \left(\vec{r} + \mathbf{A}\vec{N} \right)$$

Here, a_{ij} are the elements of the interaction matrix $\mathbf{A}$. To generate the interaction matrix, we used the directed and weighted network produced by the method described in the previous section. The presence of an interaction from one organism to another was indicated by the existence of an edge in the network. The interspecific interaction strengths were assigned by sampling values from a uniform distribution ranging from 0 to 1. After setting all interaction strengths, a subset of interactions was randomly chosen to be negative, in order to obtain a variable proportion of positive interactions on the network ranging from 0% to 100%.

If the interaction matrix is invertible, the non-zero equilibrium of the gLV is:

$$\vec{N}^\star = -\mathbf{A}^{-1}\vec{r} \iff \vec{r} = -\mathbf{A}\vec{N}^\star$$

Using this relationship, we established a connection between the equilibrium abundances of the populations, the growth rates, and the interaction matrix. To generate a species abundance distribution (SAD), we followed the lognormal distribution with mean 1.8 and standard deviation 1.9, which was found to be the AIC best model for most communities in a study with over 20,000 samples (Shoemaker et al., 2017). We assumed that the equilibrium abundances of the population followed the SAD and defined $\vec{r}$ using the refered relationship.

To ensure a stable community, we calculated the largest real eigenvalue of the interaction matrix with no intraspecific competition. If it was positive, we set the intensity of intraspecific competition to 101% of this value. This procedure stabilizes the community, making the coexistence of all species (inner equilibrium) feasible and stable.

LIASP metric

based on Roberto Andrade's revision of the manuscript

The dissimilarity (δ) between two networks (α and β) ($\delta(\alpha, \beta)$ was measured by summing over a function that accounts for the difference of the shortest path between each pair of nodes in each network $f(m_{i,j}(\alpha), m_{i,j}(\beta))$ as:

$$\delta(\alpha,\beta) = \frac{1}{N(N-1)} \sum_{i=1,j=1,i\neq i}^{N} f\Big(m_{i,j}(\alpha), m_{i,j}(\beta)\Big).$$

and we consider the functional form

$$f\Big(m_{i,j}(\alpha),\, m_{i,j}(\beta)\Big) = \left| \frac{1}{m_{i,j}(\alpha)} - \frac{1}{m_{i,j}(\beta)} \right|,$$

so that is related to the absolute difference in the efficiencies $E(\alpha)$ and $E(\beta)$ of the networks being compared (Latora, 2001). In this equation, N is the number of nodes, and $m_{i,j}$ is the element of the dissimilatity matrix defined as the length of the shortest path between nodes i and j, if they were connected. If i and j were not connected, the element $\frac{1}{m_{i,j}}$ was defined to be zero. As a consequence the diagonal elements $m_{i,i}$ vanish identically. When the network β is obtained by a process of removing edges from network α, we get $\delta(\alpha,\beta) = E(\alpha) - E(\beta)$, since in this case we have that $m_{ij}(\alpha) \leq m_{ij}(\beta) \Longrightarrow m_{ij}(\alpha)^{-1} - m_{ij}(\beta)^{-1} \geq 0$.

To simplify the interpretation of the dissimilarity metric, we introduce the concept of the Largest Influence of the Average Shortest Path (LIASP) of a node, which can be adapted to situations in which other definitions of $m_{i,j}$ are used. In this work the LIASP of a given node k is defined as:

$$LI_k = \frac{\delta(\alpha,\beta)}{E(\beta)}$$

With the expression above, LIASP measures the decrease in the efficiency of the network after removal of all connections related to a specific node. For taxa that do not participate in the exchange of information within the network, such as isolated non-interacting taxa, this metric is zero. If the metric is 1 for a particular taxa, it indicates that the removal of this taxa would have a significant impact on the network, effectively doubling the distances between all pairs of taxon in the original network.

The dissimilarity measure is sensitive to changes in the shortest paths between every pair of nodes. In particular, this index incorporates direct (decrease in efficiency due to the removal of paths between node k and its neighbors) and indirect (decrease in efficiency due to the increase in shortest paths among the remaining nodes) effects contributing to the equivalent total expansion in network path lengths caused by the removal of that node. Since the amount of direct and indirect effects can provide insights into the network's function and dynamics, such as the role of a specific node in the microbial community or how its removal will affect the rest of the network. We further divide the LIASP into its direct and indirect components, as follows

$$LI_k^{dir} = \frac{1}{E(\beta)} \frac{2}{N(N-1)} \sum_{i,j,j\neq i} \left(\frac{1}{m_{kj}(\alpha)} - \frac{1}{m_{kj}(\beta)} \right)$$
$$LI_k^{ind} = LI_k - LI_k^{dir}$$

We considered a taxon a structural keystone if its LIASP index is above the median plus twice the standard deviation of the LIASP index distribution across the network. By using this threshold, we aimed to select only those taxa that have a disproportional impact on the network's topology when removed.

Secondary Extinctions

Secondary extinctions were measured by removing a node (primary extinction) and integrating the gLV model until a certain time point. Then, all additional populations that went bellow a certain pre-defined abundance threshold are counted as extinct, that is, they undergone secondary extinction. In this study,

the threshold was set to $N_{min} = 1 \times 10^{-6}$ and the integration was performed until $t = 3 \times 10^3$ or until the solver had performed 10000 steps. To perform the numerical integration, we used the `bdf` solver provided by the R library `deSolve`, with default parameters and an user-specified jacobian matrix.

We calculated the number of secondary extinctions for each node removed in the community. Similarly to LIASP, we considered a taxon a dynamical keystone if the number of secondary extinctions it caused was above the median plus twice the standard deviation of the secondary extinctions distribution across the network.

Results

We generated networks with connectance values of 0.01, 0.05, or 0.1, and with a varying proportion of positive interactions that ranged from 0% (representing fully competitive networks) to 100% (representing mutualistic networks), with intermediate values of 25%, 50%, and 75%. We created 10 different networks with 100 organisms for each combination of connectance and proportion of positive interactions, leading to a fully factorial simulation design with a total of 150 networks and 15,000 simulated populations.

Precision of the Metrics

For each network, we recorded the nodes identified as keystones by LIASP and secondary extinctions. Consider the following network example:

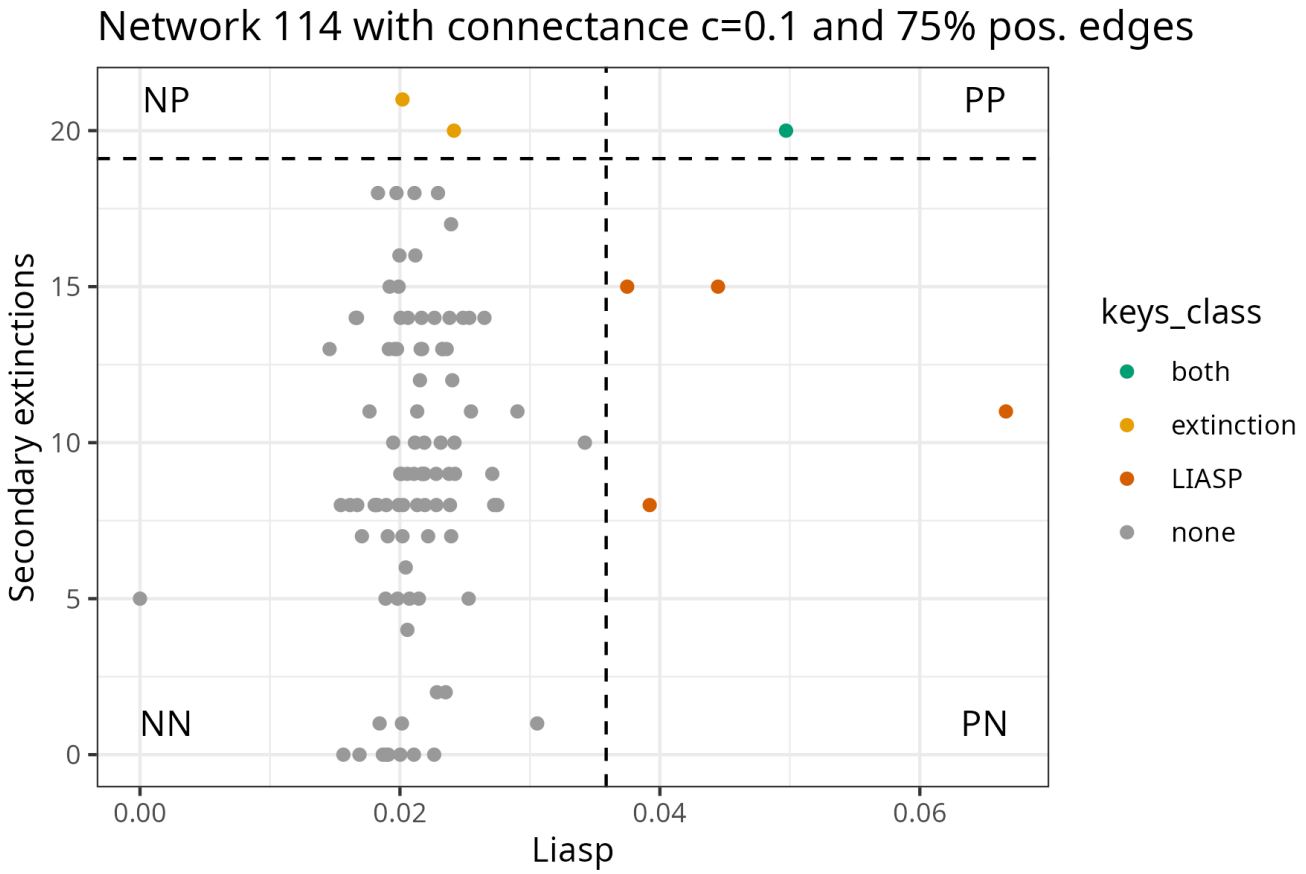


Figure 1: *The LIASP index of each node is plotted in the x-axis against the SE index in the y-axis. The dashed lines indicate the "keystone" threshold, given by the median plus two standard deviations of each metric. We colored the points following their classification as keystones by both metrics. Points to the right of the vertical dashed line indicate populations classified as keystones by the LIASP metric and the points above the dashed horizontal line indicate populations classified as keystones by the SE metric. In this network, for instance, only one species/node was considered keystone by both metrics. The set of populations classified*

as keystones only by LIASP, only by secondary extinctions, by both, and by none is indicated by PN, NP, PP, and NN regions indicated in the figure.

To measure the amount of agreement of both metrics (PP and NN), we calculate the precision of one metric trying to predict the other. The LIASP precision is the proportion of LIASP keystones that are also classified as keystones by the secondary extinction metric, and, correspondingly, the secondary extinction precision is the proportion of secondary extinctions keystones that are also classified as keystones by the LIASP metric. The precision indexes are:

$$\text{Precision}_{LIASP} = \frac{PP}{PP + PN} \quad \text{Precision}_{ext} = \frac{PP}{PP + NP}$$

In the figure above (network 144) for instance, we have:

$$\text{Precision}_{LIASP} = \frac{1}{1+4} = 20\% \quad \text{Precision}_{ext} = \frac{1}{1+2} = 33\%$$

Thus, in this example, a population being classified as a keystone by one of the metrics is not a good predictor that it will be classified as a keystone by the each other. In the results section, we'll see that this is the general case.

Precision

The precision of either metric when trying to predict the other was generally low, with most networks having precision 0 for at least one metric. In this scenario, for a given network it is likely that no population is predicted to be keystone by both metrics. Figure 2 shows the histograms of the precision of both metrics.

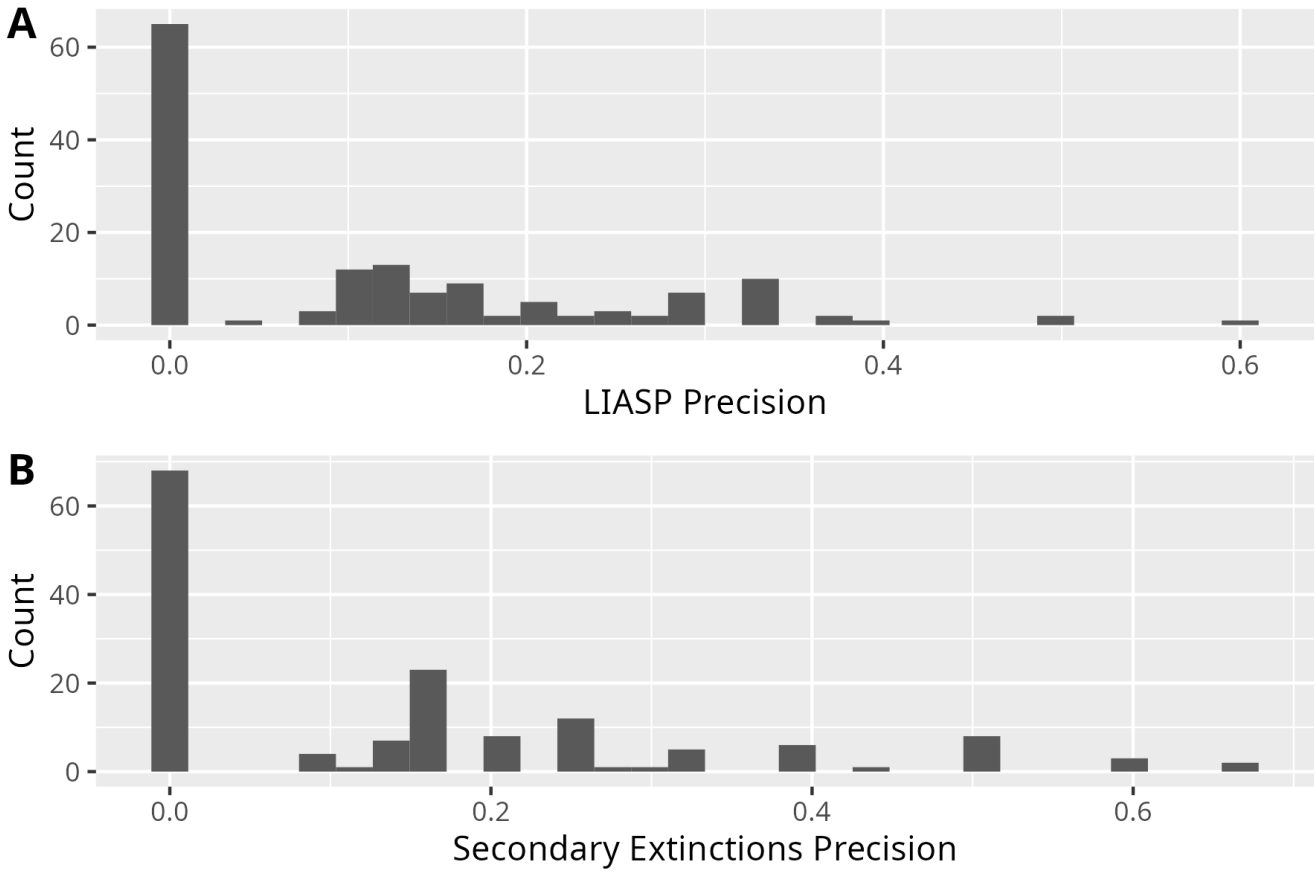


Figure 2: *(A-B) histograms with the precision of the metric in predicting the other. In this context, the precision indicates the proportion of populations classified as keystone by the one metric were also*

classified as keystone by the other metric. The LIASP precision indicates how useful LIASP is in predicting the keystones identified by secondary extinctions, and vice-versa.

A scatter plot of the precision (Fig. 3) indicate that there is a correlation between LIASP and secondary extinctions precision ($\rho = 0.81, p < 0.001$), which indicate that this agreement might be a property of the network. Since we varied both the proportion of positive interactions and the connectance of the networks, we investigated how these measures were correlated with LIASP precision. The connectance of the network did not influence the precision of LIASP when predicting the keystones indicated by secondary extinctions (Fig. 4A; $\rho = -0.03,\ p > 0.05$). However, the proportion of positive interactions was very relevant for the precision (Fig 4B; $\rho = 0.41,\ p < 0.001$). This indicates that the LIASP and secondary keystone metrics convey more similar information in mutualistic/comensalistic networks as opposed to antagonistic ones.

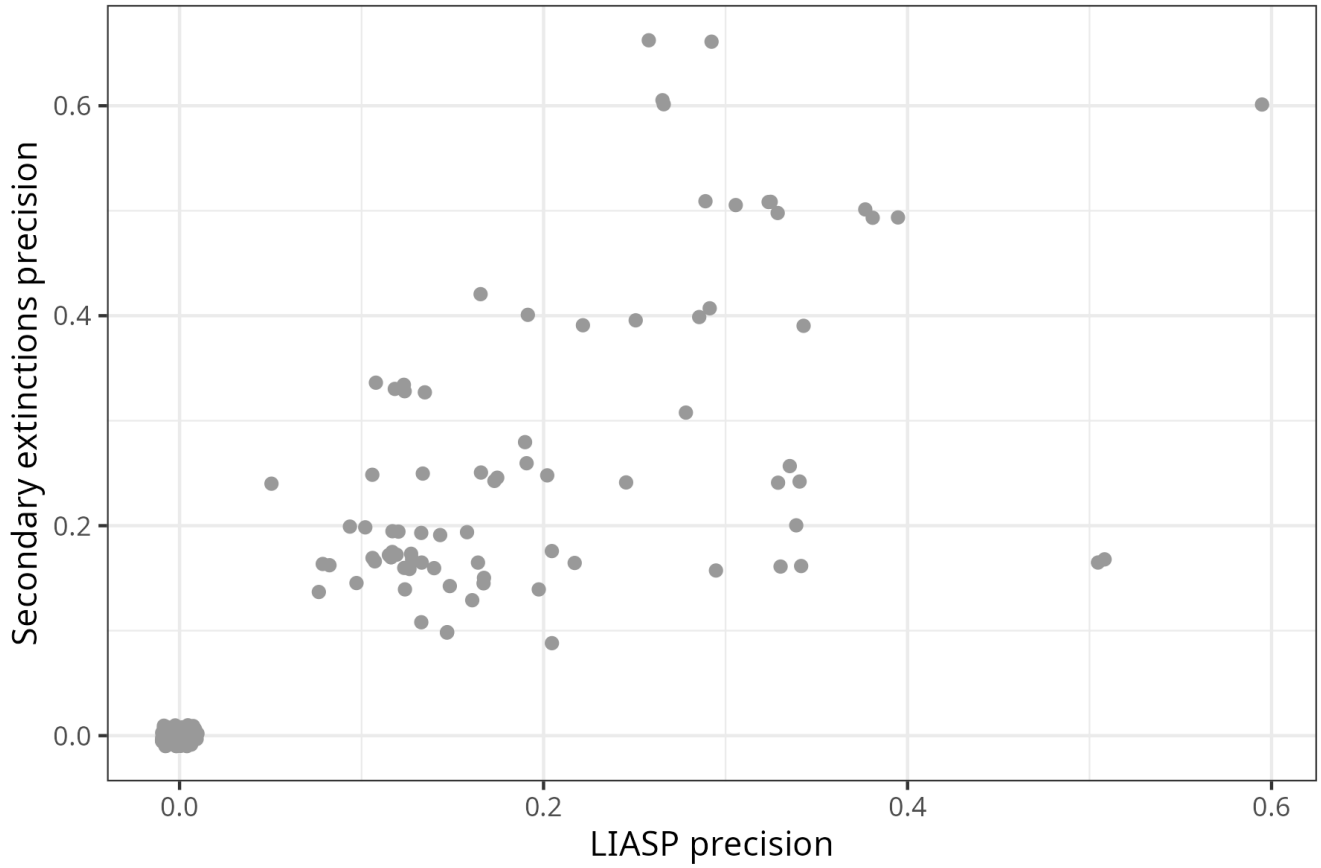


Figure 3: *The secondary extinctions precision is plotted against the LIASP precision. The points were jittered (0.01) in both directions to aid in the visualization of clustered points. Note the large amount of points aggregated at (0,0).*

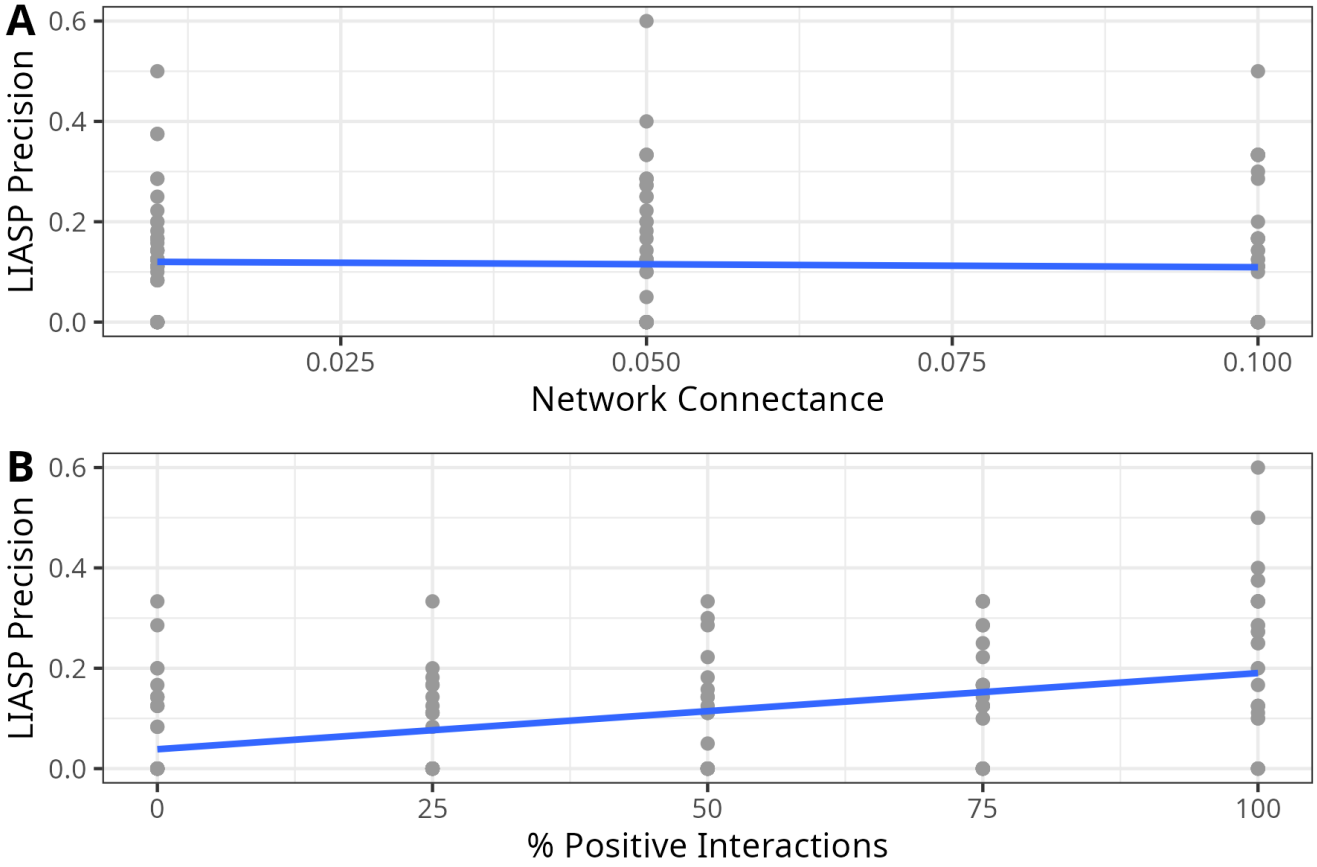


Figure 4: *LIASP precision as a function of network connectance (A) and the proportion of positive interactions in the network (B). The prediction for each network is plotted in a grey dot and the regression is plotted as a blue line.*

Correlation between LIASP and Secondary Extinction metrics

Each node in a given network has an associated LIASP and SE index, thus we estimated the Pearson correlation between these two metrics for the nodes of each network and recorded the p-value of this correlation. We discretized the p-value by using the levels of significance $\alpha_1 = 0.001, \alpha_2 = 0.01, \alpha_3 = 0.05$. Each correlation was classified as "significant at the α_{min} level" or as "non-significant" as appropriate. The significance value α_{min} for a correlation with p-value p_i was determined as the minimum α_j such that $p_i < \alpha_j$ was true. A large correlation value occurs when both LIASP and SE metrics classify as keystone and non-keystone the same species/nodes, whereas small correlation values indicate that the metrics differ in the keystone classification.

For 69% of the networks there were no significant correlation between the number of secondary extinctions and the LIASP metric. In Figure 5 we present a histogram of the estimates of the correlations for each network, colored by the level of significance of this estimate. In Figure 6, we show the regression between these two metrics for all the networks, divided by the connectance and proportion of positive interactions in each network.

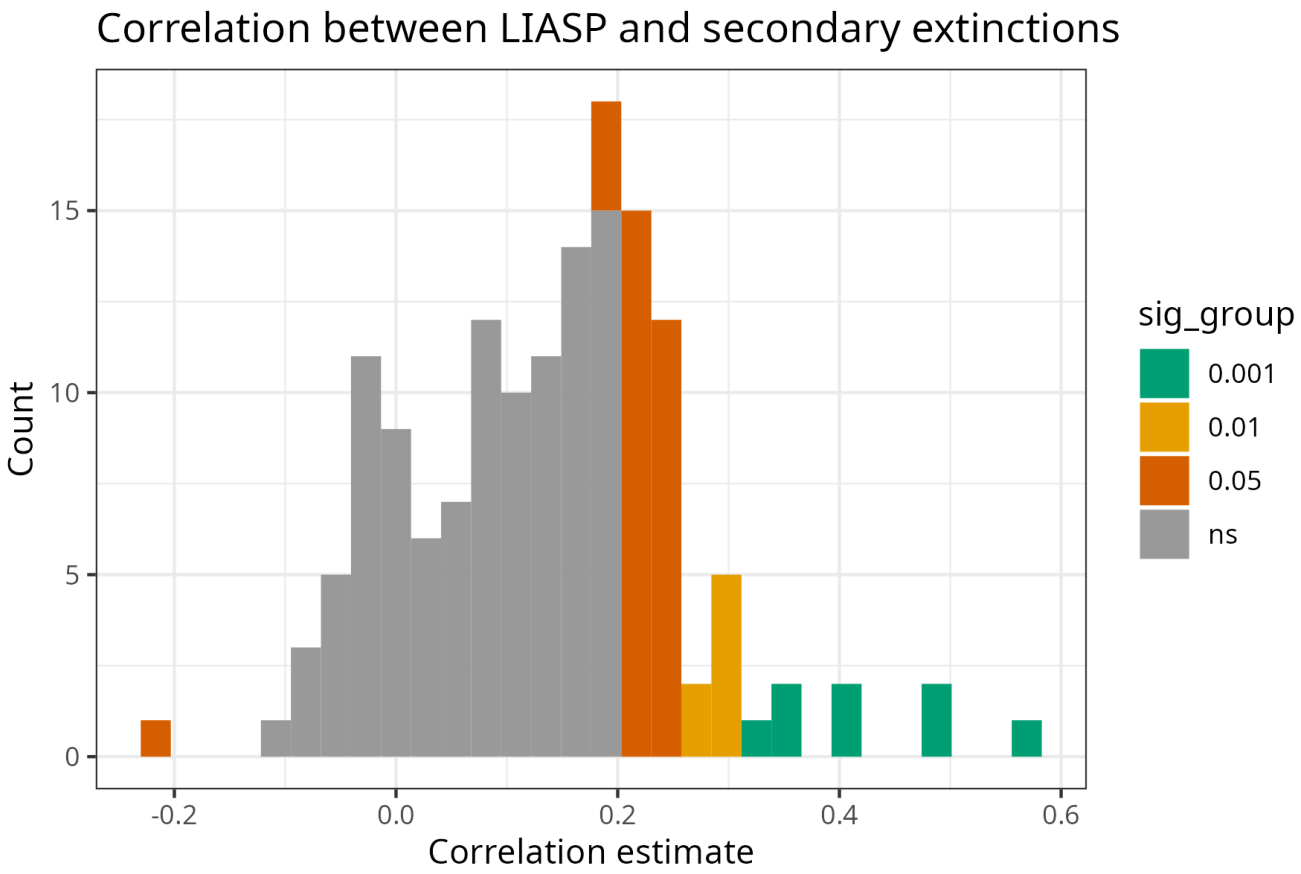


Figure 5: *A histogram of the estimates of the correlations between the number of secondary extinctions and the LIASP metric in each network. The histogram is colored by the significance level of each estimate.*

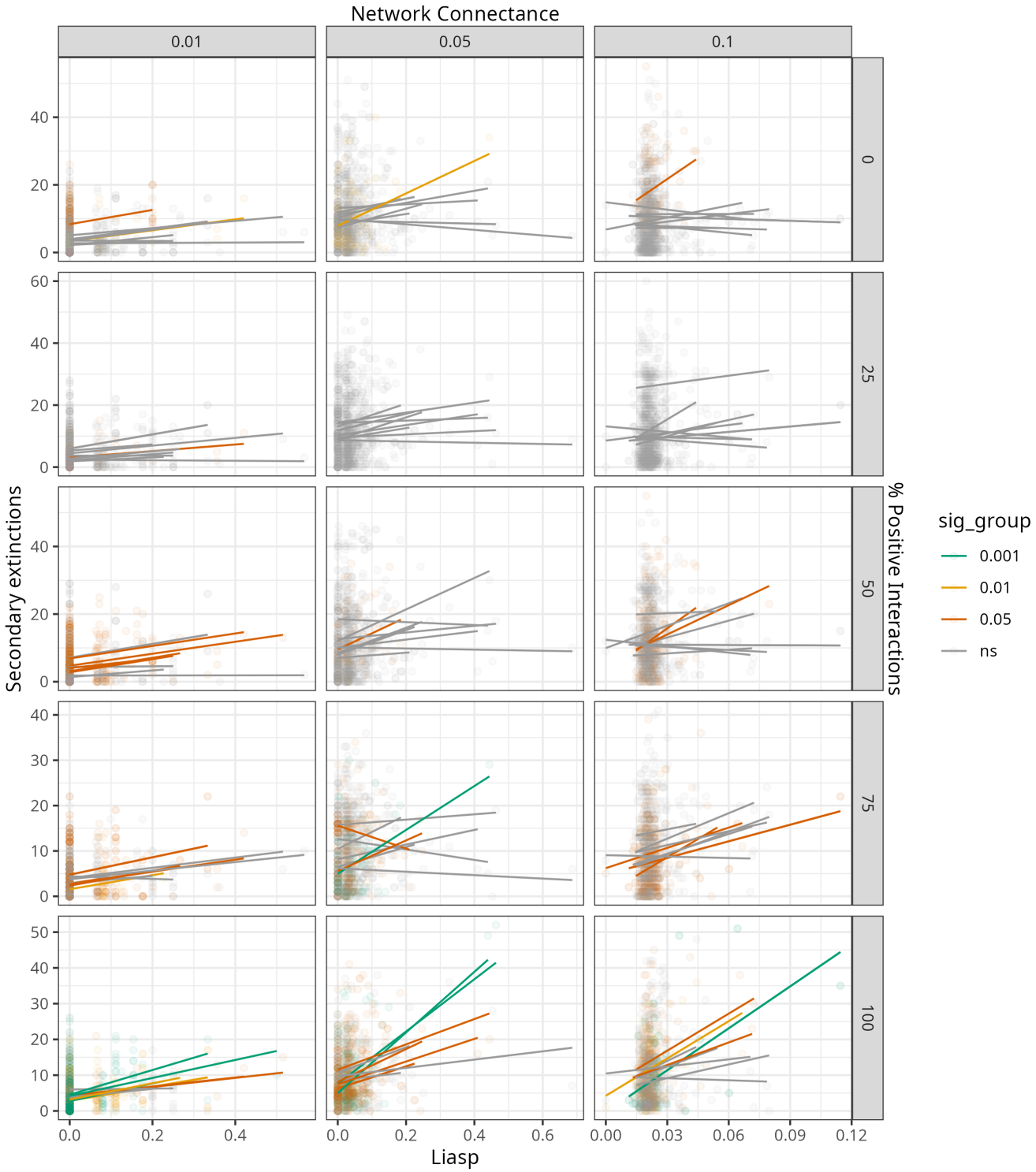


Figure 6: *The regressions between LIASP and secondary extinctions for each network. Each panel contains the information of the 10 replicates of the networks with the same proportion of positive interactions (rows) and network connectance (columns). Each point in a given panel refers to one population in one of the replicates network. The lines indicate the regression between LIASP and secondary extinction for each network. Points and lines in each panel are colored by the level of significance of the correlation.*

Number of Keystones

In Figure 7, we show a jittered scatter plot of the number of keystones identified by LIASP and by secondary extinctions metric for each network. We can observe that the number of keystones predicted by secondary extinctions metric does relatively homogeneous and larger than the number of keystones predicted by LIASP.

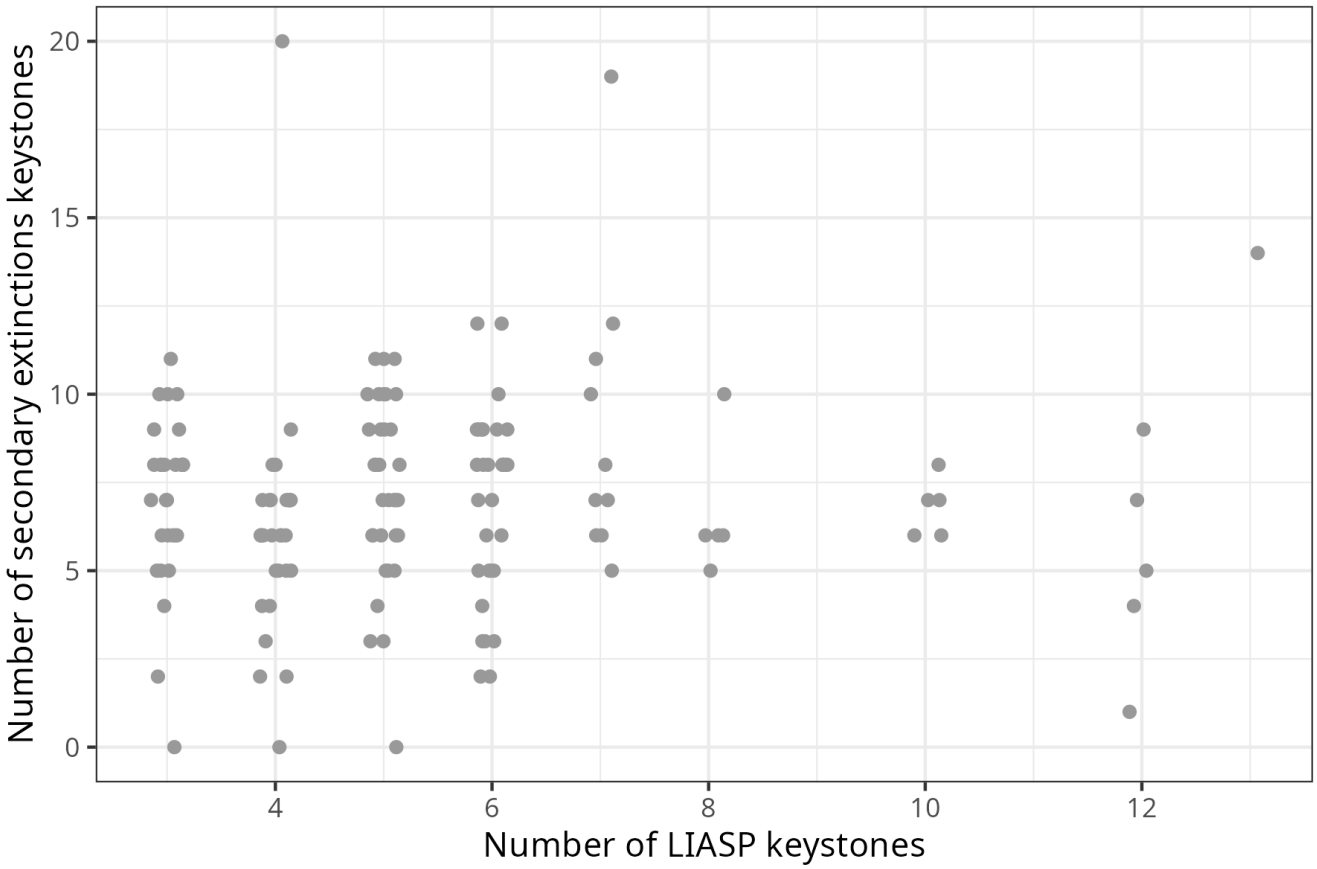


Figure 7: *The number of keystones predicted by secondary extinctions is plotted against the number of keystones predicted by LIASP. The points were jittered (0.15) in the x direction to aid in the visualization of clustered points.*

Proportion of Consensus Keystones (new*)

The Jaccard Similarity index is used to calculate the number of taxa identified as keystone by both metrics (PP = true positives) divided by the number of keystones identified only by LIASP (PN), only by Secondary extinctions (NP) and both (PP). In this sense we have a proportion of consensus keystones between both metrics relative to all keystones that were found for a given network/community. We calculate the index for each network, the index is bounded between 0 to 1, where 1 means that all taxa identified as a keystone for a given network/community was identified for both metrics and none of the metrics identified any other keystones where the other disagrees.

If $\mathcal{L}$ is the set of keystones identified by the LIASP metric and $\mathcal{S}$ is the set of keystones identified by the secondary extinctions metric, the Jaccard Similarity (J) is defined as

$$J = \frac{|\mathcal{L} \cap \mathcal{K}|}{|\mathcal{L} \cup \mathcal{K}|}$$

in which $|\cdot|$ denotes the size of the set.

Figure 22A shows the there is no clear relationship between Network connectance and the consensus between metrics. Figure 22B shows that the consensus between both metrics increase as the number of positive edges increase in the network. Considering the mean Jaccard Index for all 10 networks generated for each combination of connectance and % of positive edges, figure 22C shows that the conditions (network parametrization) promoting more consensus between the metrics are for intermediate connectance and 100% positive edges (mutualistic networks).

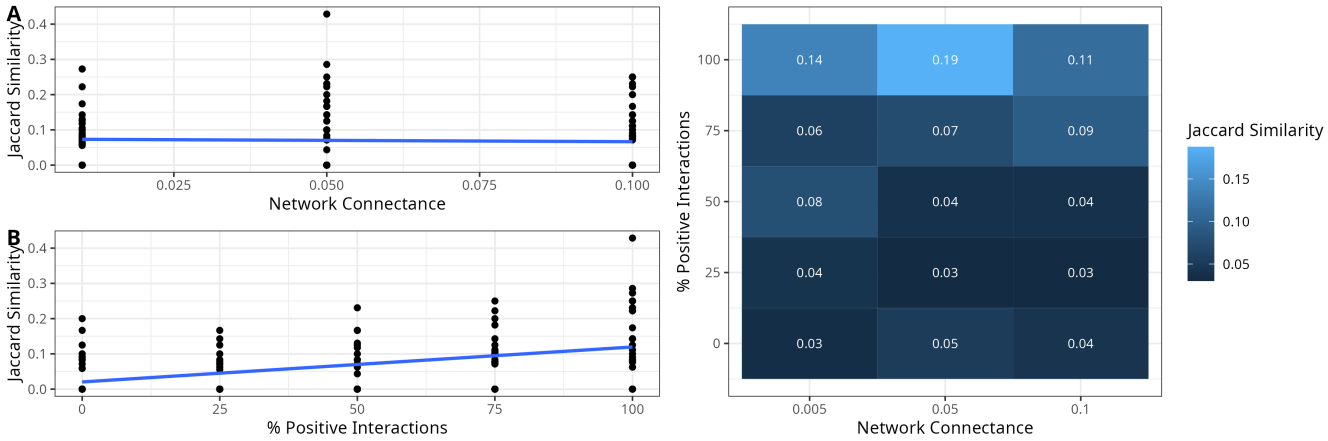


Figure 8: *Consensus between LIASP and Secondary Extinctions. A) shows the there is no clear relationship between Network connectance and the consensus keystones. B) shows that the consensus between both metrics increase as the number of positive edges increase in the network. C) shows that the conditions (network parametrization) promoting more consensus between the metrics are for intermdiate connectance and 100% positive edges (mutualistic networks).*

LIASP identifies keystones that are more distinct between keystones and non-keystones (new*)

For each node/taxa in a given network, we calculate the LIASP and Secondary Extinction index, hence, for each network simulated there is a distribution of LIASP and secondary extinction values. We found that LIASP distribution is more positively skewed in relation to the SE suggesting that the keystones identified by LIASP are more differentiable from non-keystones than the SE. The explanation is that the keystones are the outliers identified to the right of the distribution. A larger skewness indicates that these outliers are more distant to the center of mass of the distribution.

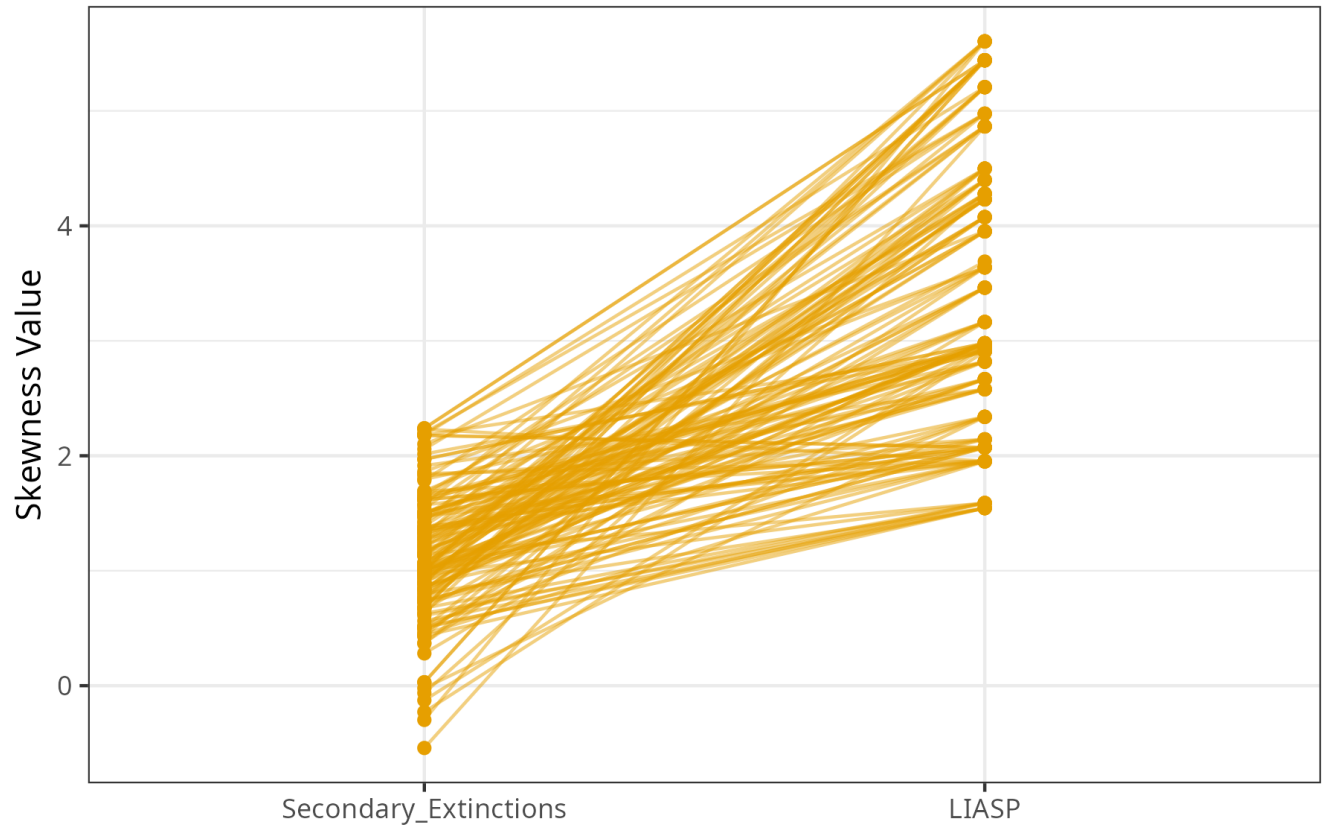


Figure 9: *Skewness of LIASP and SE distribution. Each line in the graph represents a single*

network/community and the circles in the extremes are the value of how skewed the distribution of LIASP and SE values are. Positive skews means that the tail of the distribution are in the right-hand side (larger values). Larger skew values suggest that the keystones of LIASP are more differentiable from non-keystones than the SE.

Percentage of identified keystones of each metrics (new*)

We used a modified Jaccard Similarity Index to calculate the proportion of keystones identified by each metric. To identify how much of the total keystone taxa identified was made by LIASP we used

$$J_1' = \frac{|\mathcal{L}|}{|\mathcal{L} \cup \mathcal{K}|} = \frac{PP + PN}{PP + PN + NP}$$

whereas, to identify how much of the total keystone taxa identified was made by LIASP we used

$$J_2' = \frac{|\mathcal{K}|}{|\mathcal{L} \cup \mathcal{K}|} = \frac{PP + NP}{PP + PN + NP}$$

Considering all networks simulated, LIASP identified the mean of 47.40% of all keystones:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max
0.1739	0.3588	0.4444	0.4740	0.5556	1.0000

Whereas, SE identified the mean of 59.59%

Min.	1st Qu.	Median	Mean	3rd Qu.	Max
0.0000	0.5000	0.6364	0.5959	0.7000	0.8889

Discussion

Lack of correlation between the two metrics

The low correlation between the LIASP and secondary extinction metrics in identifying keystone species can be due to their completely different approaches. The LIASP metric is a structural metric that uses a snapshot of the community in the form of an unweighted and undirected network to examine the impact of removing a particular population on the network through its (direct and indirect) effect on the shortest paths. Conversely, Secondary Extinction is a dynamical metric that requires an underlying dynamical model and a comprehensive knowledge of the community's initial state to predict the repercussion of a local population's extinction in a community. We observed that LIASP is more precise in predicting the keystones assigned by the secondary extinctions in mutualistic or commensalistic networks, which could suggest that, in these type of networks, the elimination of a topologically significant node leads to a more extensive extinction cascade in the community. However, the exact mechanism is still obscure.

Limitations of metrics based on community dynamics and the importance of topological metrics

One of the main challenges to study microbial community networks is the lack of full knowledge about the dynamics of the network/community, which hinders the prediction of consequences of a species extinction. Secondary extinctions, a dynamical metric, require an underlying model and full knowledge of the pristine state of the community. However, the usage of generalized Lotka-Volterra (gLV) models is an

approximation that might not reflect the complexity of interactions among microorganisms. For instance, such models assumes that the impact of one population upon others is proportional to their abundance, which may not hold true for species that secrete metabolites with catalytic effects. Moreover, the parameterization of even a simple model such as the gLV model requires knowledge of at least N^2 parameters, which may be impractical to estimate without extensive time series data of microbial communities for each environment. Therefore, the literature has shifted its focus to topological measures, such as degree and betweenness centrality, whose interpretation heavily relies on the network topology. In contrast to Secondary Extinctions, LIASP is a topologic-based metric that identifies organisms with a significantly large impact on the average shortest path of interactions in the community, which is determined from a mathematical standpoint. The increase in the shortest path of the community is directly related to the flow of information in the network.

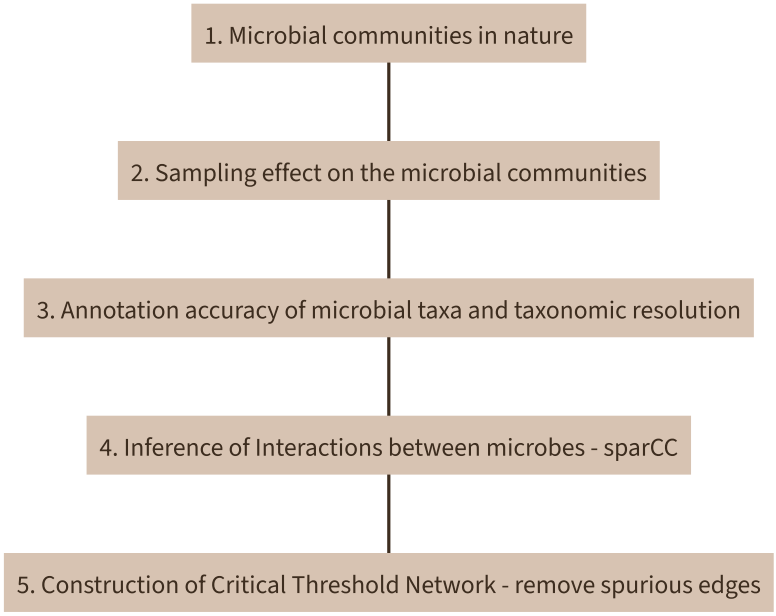
Data availability and the choice between dynamical and topological metrics

The insipient amount of time series datasets for microbial communities and the extensive community structure datasets of microbial communities make topological based metrics to estimate keystone taxa, such as LIASP, more feasable than dynamical based metrics, such as Secondary Extinctions. In this sense, this area of research has a historical focus on topological measures, such as node degree and node betweenness centrality, to estimate the importance of a node/species to the microbial community. However, the interpretation of these measures heavily depends on a priori knowledge of the topological features of the network, making the development of agnostic topological metrics like LIASP a valuable metric in microbial ecology.

While LIASP keystones may not match those predicted by a dynamical-based metric, the identification of keystones through LIASP can guide future research into the functional and ecological aspects related to the structural importance of organisms. LIASP provides crucial information by identifying structurally key species to guide experimentation and investigation of singular taxa, which can highlight the importance of specific organisms in maintaining ecosystem functioning.

Differences between the networks generated in this report and the networks reconstructed from data in the manuscript

In this study, we employed an artificial network construction method to investigate the performance of a topological and a dynamic metric in microbial ecology. While this approach represents a best-case scenario in which all real interactions in the community would be recovered (besides the growth rates and equilibrium abundances), it is important to note that the pipeline for constructing networks from natural communities involves several critical steps, including sampling, annotation, and inference of interactions from co-occurrence of taxa, as well as the removal of topologically spurious edges (steps 2 - 5 in the flowchart below). These methodological steps are necessary for generating accurate representations of real-world microbial networks, and the accuracy of the final outcome network is dependent on the accuracy of these intermediate steps, in addition to the accuracy of the metric analysis itself. Thus, while the artificial communities generated in this report differ from those constructed from data in the manuscript, our findings highlight the importance of careful consideration and evaluation of the methodological steps involved in network construction.



Conclusion

In conclusion, the lack of correlation between the LIASP and secondary extinction metrics can be attributed to their different approaches, with the former being a topological measure and the latter requiring a model and comprehensive knowledge of the ecological dynamics of community. Our findings suggest that the two metrics are more correlated in mutualistic or commensalistic networks, however, the exact mechanism is still obscure.

Topological measures like LIASP are crucial in microbial ecology, especially given the limitations of using community dynamics-based metrics. While LIASP keystones may not match those predicted by alternative metrics, they provide crucial information in identifying structurally key species to guide experimentation and investigation of singular clades.

Finally, it is important to note that the pipeline for reconstructing networks from natural communities involves several critical intermediary steps that determine the accuracy of the final network in addition to the accuracy of the metric analysis itself. Hence, careful consideration and evaluation of methodological steps involved in network reconstruction are essential.

References

Berry, David, and Stefanie Widder. “Deciphering Microbial Interactions and Detecting Keystone Species with Co-Occurrence Networks.” *Frontiers in Microbiology* 5 (2014). https://www.frontiersin.org/articles/10.3389/fmicb.2014.00219.

Faust, Karoline, Franziska Bauchinger, Béatrice Laroche, Sophie De Buyl, Leo Lahti, Alex D. Washburne, Didier Gonze, and Stefanie Widder. “Signatures of Ecological Processes in Microbial Community Time Series.” *Microbiome* 6, no. 1 (December 2018): 120. https://doi.org/10.1186/s40168-018-0496-2.

Klemm, Konstantin, and Víctor M. Eguíluz. “Growing Scale-Free Networks with Small-World Behavior.” *Physical Review E* 65, no. 5 (May 8, 2002): 057102. https://doi.org/10.1103/PhysRevE.65.057102.

Latora, Vito, and Massimo Marchiori. “Efficient Behavior of Small-World Networks.” *Physical Review Letters* 87, no. 19 (October 17, 2001): 198701. https://doi.org/10.1103/PhysRevLett.87.198701.