Uncover viral sharing through global srtucure of host-virus meta-network

Mathis Gheno [1,2,‡], Timothée Poisot [2,3,‡]

[1] Univeristé de Rennes; [2] Université de Montréal; [3] Québec Centre for Biodiversity Sciences
[‡] Equal contributions

Correspondance to:
Timothée Poisot — timothee.poisot@umontreal.ca

Purpose: This template provides a series of scripts to render a markdown document into an interactive website and a series of PDFs.
Internals: GitHub actions and a series of python scritpts. The markdown is handled with `pandoc`.
Motivation: It makes collaborating on text with GitHub easier, and means that we never need to think about the output.

Keywords:
pandoc
pandoc-crossref
github actions

0.1. Introduction

1. Global health needs more ecology in viral forecasting

   Predicting viral spillover for future pandemic prevention + importance of ecology in the context of global health (Belay *et al.* 2017; Carroll *et al.* 2018) (Plowright *et al.* 2017). Works have been done to predict viral hotspot according to climate change (Carlson *et al.* 2022), future host distribution (Morales-Castilla *et al.* 2021) and potential viral sharing among host (Albery *et al.* 2020). Bring informations on potential spillover path. Albery *et al.* (2020) viral sharing is based on phylogenetic and geographic overlap, they show that phylogenetic similarity is much more decisive on viral sharing than geographical overlapping. Knowing that species migration is at is highest, there is a need to develop models that predict potential viral sharing of species even if they don't overlap. Using phylogenetic seams to be conclusive. We want to bring an other approach that use global structure Metaweb. A metaweb is a network that resume potential interactions… Metawebs contains precious ecological information (**Morales-Castilla2015Inferrina?**)

   Poisot *et al.* (2023) have develop a link prediction method for host virus metaweb. Showing that unrealised or non discover associations can be infer using a network dimension reduction method. Enable a better prediction of viral infection on human. This imputed metawab has no geographical constrain. A question can arise, can we use the imputed metaweb structure to make prediction of most likely viral spillover ? and top viral sharing species with human (or any other) ? (*I'm not sure what can communicability add to the embedding, because embedding already take into account the global structure of the network. A prediction of link isn't enough ? Plus, the embedding already give a probability of observing an association. I still need to think about it.*)

   We want to propose an approach that take the big picture as input (the network as a hole) and provide information at the lower level (the link). For this we used a matrix function introduce by Estrada & Hatano (2008)

2. Communicability is really flexible, as shown in Benzi & Boito (2020) it already has application in multiple subject such as neurology, cancer detection, economic… and is used for community detection, spread of information/contagion. This matrix function barely (or not) used on ecological network.

3. Using communicability we try to extract informations on metaweb with predicted association. We are exploring the community detection method to extract group of host with similar viral sharing, group of virus with similar host sharing.

   By using global structure of a network it is possible with to deduce small scale properties such as viral sharing potential between to species, importance of species or importance of association in the network.

4. Explaining why we think that communicability is a proxy for viral sharing. Host host part of A^2 is known to be direct viral sharing, aka the exact number of virus share between 2 host (same apply for virus-virus, the the number of host share between two). Host host part of A^4 is the number of virus share between the 2 host plus the number of virus share with any other intermediary host, etc… . Communicability quantify this but add a penalization each time we add intermediary. This penalization enable to simulate the "difficulty" for a virus to transit from one host to an other if they are "far away" in the network.

5. Communicability (G) is a quantification closely related to total sharing viral diversity (A^2). G will be strong when one or both host have a high viral diversity. It can be simply explain by the fact that a hight diversity mean in general more path to connect both nodes, so the communicability between those is better. The clustering (ΔG, or viral spectrum similarity) is not exactly the same. It will not by highly positive when both have a hight viral diversity or highly negative when both have a low viral diversity. No! it will depend on the similarity of both node, the quantity is not taken in consideration. For example, lets suppose that one part of a pair have a hight viral diversity and the other less, also that the hight viral diverse one have most virus of the second one. G will be strong between the two host because of the one highly connected. But if we look at the clustering, the highly connected host is considered as a "generalist host", it share virus with a lot of other host. Even if the less connected one does share almost all of is virus with the other the "virus spectrum" of both host is completely different, so the clustering between those two will be weak.

   Next: make an host host matrix with percentage of same virus shared

0.2. Methods Communicability quantify how well information transit between two nodes by considering all possible path in a network and penalizing longer ones. It is compute with the exponent of the adjacency matrix of the network.

$$G = \sum_{k=0}^{\infty} \frac{(\mathbf{A}^k)}{k!} = e^{\mathbf{A}}$$

where G is the communicability matrix, A the adjacency matrix and k is used as a penalizing term. It is possible to compute the exponential of a matrix with the graph spectrum :

$$G = \sum_{j=1}^{n} \varphi_j \varphi_j^T e^{\lambda_j}$$

where φ_j and λ_j are respectively the j^{th} eigenvectors and eigenvalues of the matrix A. Obtain cluster with Communicability is done only by removing the fist dimension in the computation of the previous equation.

$$\Delta G = \sum_{j=1}^{n} \varphi_j \varphi_j^T e^{\lambda_j} - \varphi_1 \varphi_1^T e^{\lambda_1}$$

In this equation we are removing to global sum the dot production of the first eigenvector(and eigenvalue) to find the clustering matrix ΔG. Let's keep in mind this final result and explain why this work for clustering.

We start with the spectral form of G that can be decompose by the following way :

$$G = \varphi_1 \varphi_1^T e^{\lambda_1} + \sum_{j=2}^{n} \varphi_j^+ \varphi_j^{+T} e^{\lambda_j} + \sum_{j=2}^{n} \varphi_j^- \varphi_j^{-T} e^{\lambda_j} + \sum_{j=2}^{n} \varphi_j^- \varphi_j^{+T} e^{\lambda_j} \tag{1}$$

where φ_j^+ or φ_j^- indicate respectively all the positives or negatives values of the j^{th} eigenvector. A way to think about it would be that when φ_j^+ all negative values are set to 0 and when when φ_j^- all positive values are

set to 0. Estrada & Hatano (2008) explain that "two nodes have the same sign in an eigenvector if they can be considered as being in the same partition of the network, while those pairs having different signs correspond to nodes in different partitions.".

$$\sum_{j=2}^{intracluster} \varphi_j \varphi_j^T e^{\lambda_j} = \sum_{j=2}^{n} \varphi_j^+ \varphi_j^{+T} e^{\lambda_j} + \sum_{j=2}^{n} \varphi_j^- \varphi_j^{-T} e^{\lambda_j}$$

and

$$\sum_{j=2}^{intercluster} \varphi_j \varphi_j^T e^{\lambda_j} = \sum_{j=2}^{n} \varphi_j^- \varphi_j^{+T} e^{\lambda_j}$$

so the clustering matrix is obtain with

$$\Delta G = \sum_{j=2}^{intracluster} \varphi_j \varphi_j^T e^{\lambda_j} - \left| \sum_{j=2}^{intercluster} \varphi_j \varphi_j^T e^{\lambda_j} \right|$$

The absolute operator is not useful, it is just here to remind that all inter cluster values are negatives. The dot product of $\varphi_j \varphi_j^T$ produce a matrix with positive and negative sign depending on the sign of $\varphi_j(p)$ and $\varphi_j(q)$. The first is not include because all the values the eigenvector the same sign, so it is not really informative (In fact it consider the hole network as one cluster thus it does not bring interesting information on the clustering).

In short we can rewrite ΔG as follow :

$$\Delta G = G - \varphi_1 \varphi_1^T e^{\lambda_1}$$

which was the form in which we introduce it.

0.2.0.0.1 Remark 1 Each dimension of the spectra form the 2nd to the last one is a clustering configuration of the network. Cluster identified by dim 3 are not necessarily independent form those in dim 2 or 4 (or all others). Although the cluster form by dim 2 are "stronger" then 3, 4, 5 … and so on till the last one. So the dim 2 is the one that contribute the most for ΔG (we can see it in the following example).

0.2.0.0.2 Remark 2 In a graph nodes can be consider as diffuser or receiver. The communicability matrix can be use to quantify the capacity of nodes to communicate when they act like a diffuser a receiver. For example host-host or virus-virus parts of G give information on the capacity of respectively host or virus to communicate when they are playing the same role in he network. see Benzi & Boito (2020) for better explaination

```
library(tidyverse)
library(lattice)
library(igraph)
library(colorRamps)
A  = matrix(c(0,1,0,1,1,0,0,0,0,0,0,
              1,0,1,1,1,0,0,0,0,0,0,
              0,1,0,1,1,0,0,0,0,0,0,
              1,1,1,0,1,0,0,0,0,0,0,
              1,1,1,1,0,1,0,0,0,0,0,
              0,0,0,0,1,0,1,0,0,0,0,
              0,0,0,0,0,1,0,1,1,0,1,
              0,0,0,0,0,0,1,0,1,1,1,
              0,0,0,0,0,0,1,1,0,1,1,
              0,0,0,0,0,0,0,1,1,0,1,
              0,0,0,0,0,0,1,1,1,1,0), nrow =11, ncol =11)
grap = graph_from_adjacency_matrix(A, mode = "undirected")
plot(grap)
```

0.2.0.1 Example A graph with 11 nodes and 2 distinct group. First we need to compute the graph spectrum

```
spectra = eigen(A)
levelplot(spectra$vectors, ylab ="j th position" , xlab ="eigenvectors")
```

The above plot just represent the 11 eigenvectors. We can see that the fisrt one is the only full of same sign value.

Now let's take the 2^{nd} dimension as an example.

```
##
G_dim2 = spectra$vectors[,2]%*%t(spectra$vectors[,2])*exp(spectra$values[2])
levelplot(G_dim2, ylab = "node", xlab ="node",col.regions = rev(matlab.like(16)))
```

And that it ! The second dimension of the graph communicability identify 2 cluster (in blue).

We can compute for the third dimension

```
G_dim3 = spectra$vectors[,3]%*%t(spectra$vectors[,3])*exp(spectra$values[3])
levelplot(G_dim3, ylab = "node", xlab ="node",col.regions = rev(matlab.like(16)))
```

Which identify clusters between 5:6 and 6:7. The cluster of the third dimension are less "obvious" than those from the second dimension

Now if we want to use both dimension in the clustering, we just have to add

```
levelplot(G_dim2+G_dim3, ylab = "node", xlab ="node",col.regions = rev(matlab.like(16)))
```

We could continue like that till the last dimension (11th), but it was for the explanation. So now we can compute directly ΔG by adding dimension from 2 to 11 (or subtraction of the first dim which is exactly the same)

```
delta_G = matrix(0, nrow =11, ncol =11)
for(dim in 2:11){
  delta_G = delta_G + spectra$vectors[,dim]%*%t(spectra$vectors[,dim])*exp(spectra$values[dim])
}
levelplot(delta_G, ylab = "node", xlab ="node",col.regions = rev(matlab.like(16)))
```

0.3. Results

0.4. Conclusion

Albery, G.F., Eskew, E.A., Ross, N. & Olival, K.J. (2020). Predicting the global mammalian viral sharing network using phylogeography. *Nature Communications*, 11, 2260.

Belay, E.D., Kile, J.C., Hall, A.J., Barton-Behravesh, C., Parsons, M.B., Salyer, S., *et al.* (2017). Zoonotic Disease Programs for Enhancing Global Health Security. *Emerging Infectious Diseases*, 23, S65–70.

Benzi, M. & Boito, P. (2020). Matrix functions in network analysis. *GAMM-Mitteilungen*, 43, e202000012.

Carlson, C.J., Albery, G.F., Merow, C., Trisos, C.H., Zipfel, C.M., Eskew, E.A., *et al.* (2022). Climate change increases cross-species viral transmission risk. *Nature*, 607, 555–562.

Carroll, D., Daszak, P., Wolfe, N.D., Gao, G.F., Morel, C.M., Morzaria, S., *et al.* (2018). The Global Virome Project. *Science*, 359, 872–874.

Estrada, E. & Hatano, N. (2008). Communicability in complex networks. *Physical Review E*, 77, 036111.

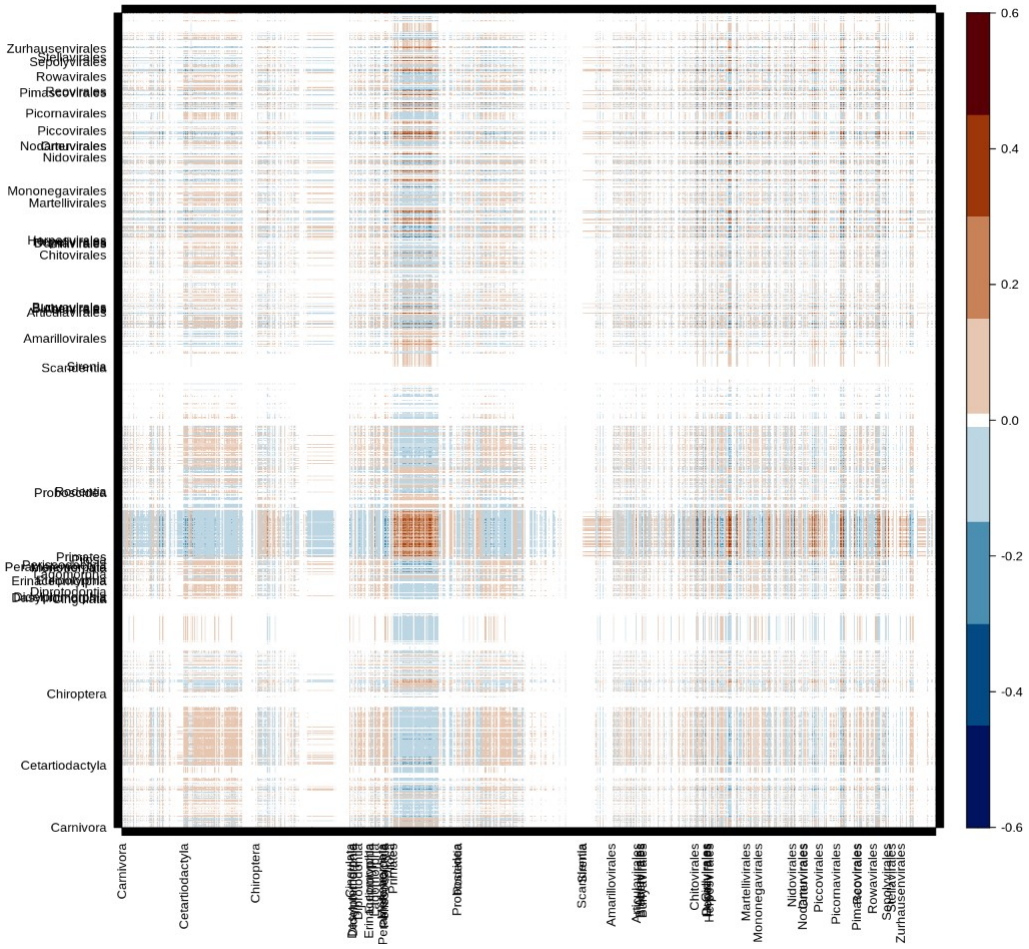

Figure 1 Global matrix of clustering communicability. Positive values indicate species in same cluster, negative value species in "opposite" cluster. Only host order and virus order names are display on x and y.

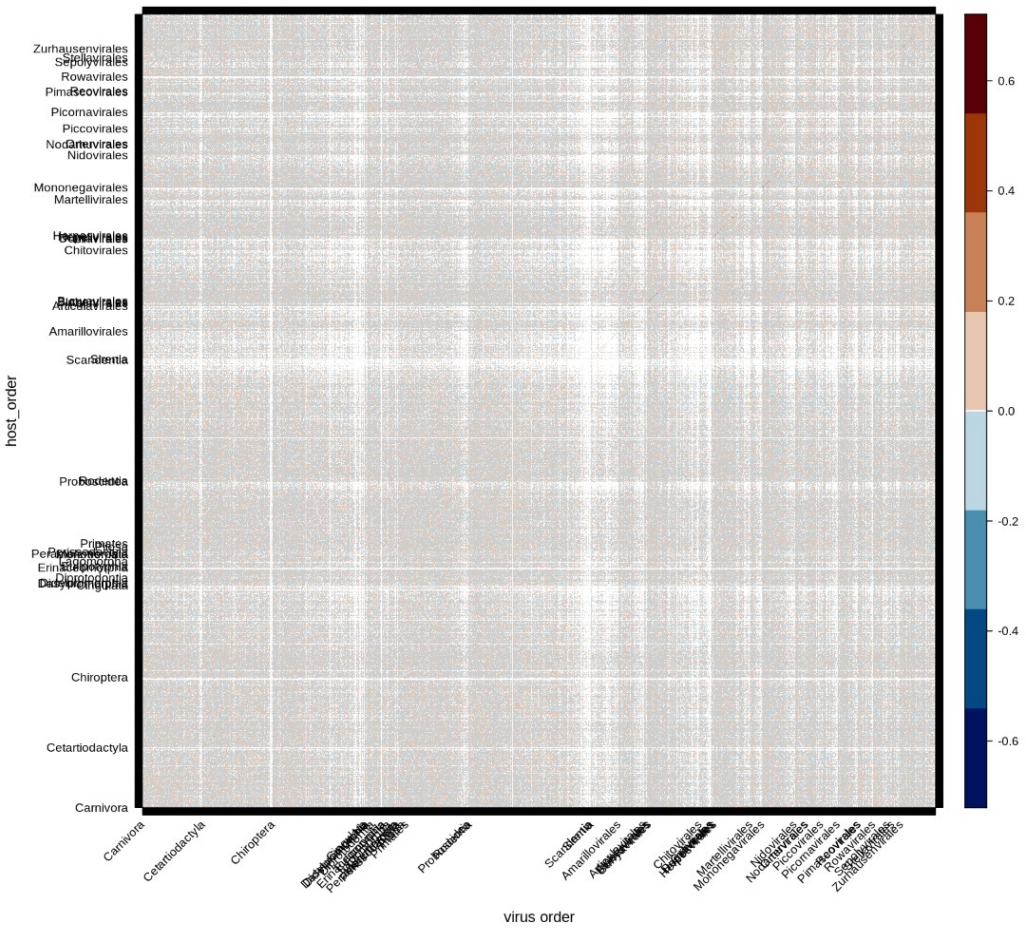

Figure 2 Global matrix of clustering communicability, with random edges connection. Number of edges for each nodes (Host and virus) is conserved. X and Y are ordered the same way as fig. 1

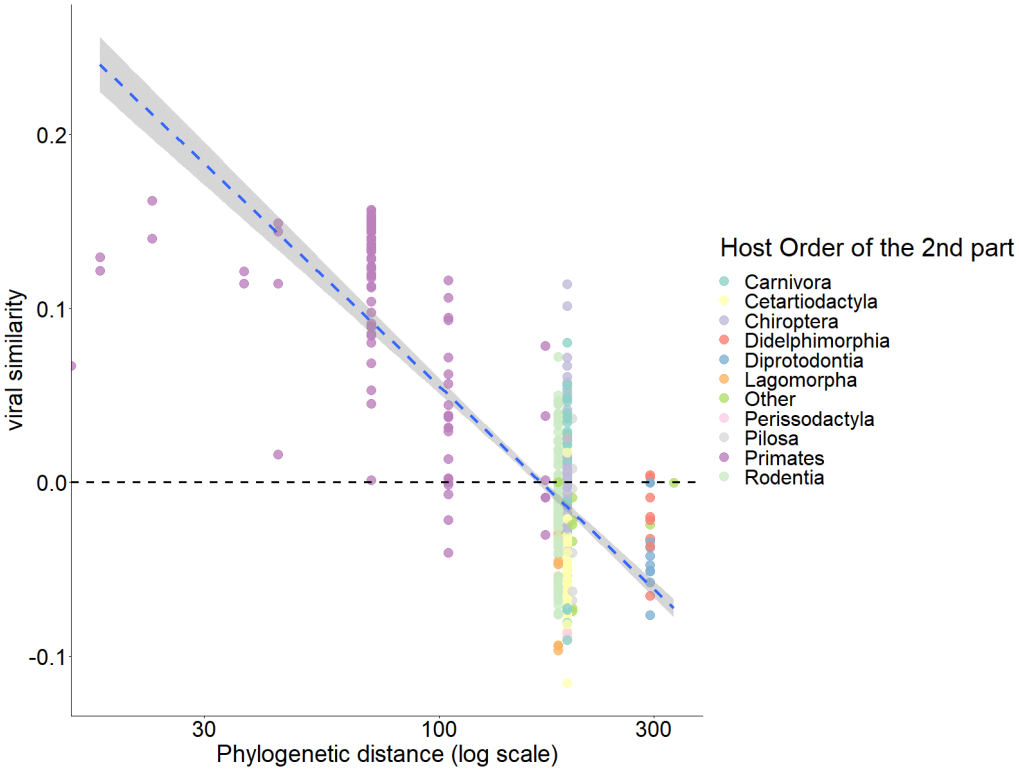

Figure 3 Phylogenetic distance explain sharing similarity of Homo sapiens (with almost all trefle host)

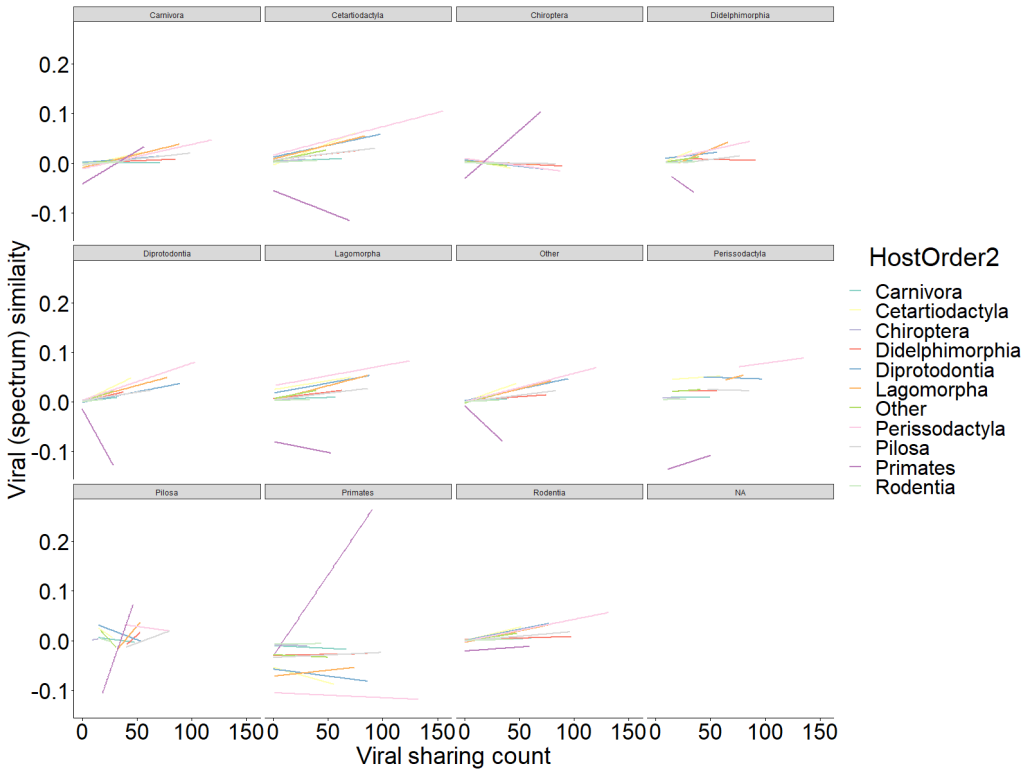

Figure 4 Viral sharing count between host isn't always correlated with sharing similarity

Morales-Castilla, I., Pappalardo, P., Farrell, M.J., Aguirre, A.A., Huang, S., Gehman, A.-L.M., *et al.* (2021). Forecasting parasite sharing under climate change. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 376, 20200360.

Plowright, R.K., Parrish, C.R., McCallum, H., Hudson, P.J., Ko, A.I., Graham, A.L., *et al.* (2017). Pathways to zoonotic spillover. *Nature Reviews Microbiology*, 15, 502–510.

Poisot, T., Ouellet, M.-A., Mollentze, N., Farrell, M.J., Becker, D.J., Brierley, L., *et al.* (2023). Network embedding unveils the hidden interactions in the mammalian virome. *Patterns*, 100738.

0.5. Usefull fig

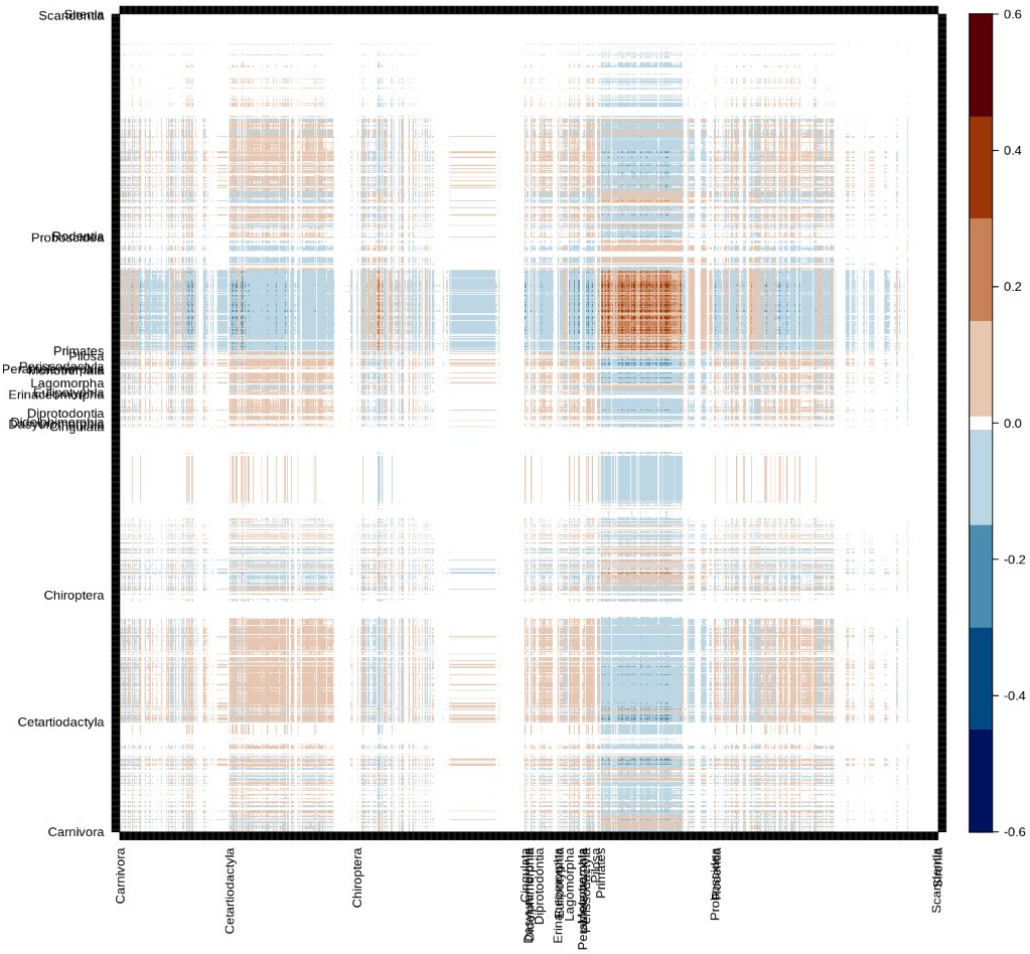

Figure 5 Host-Host cluster matrix, can be consider as viral sharing similarity for host pairs

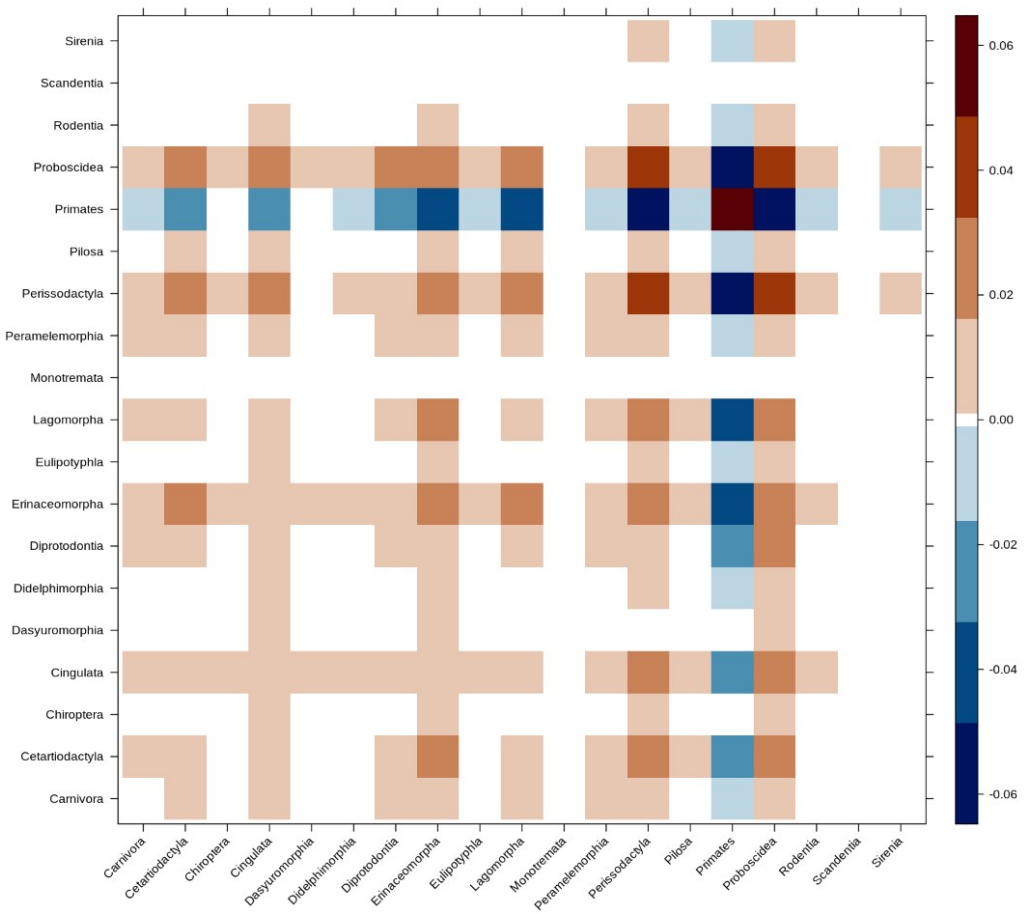

Figure 6 Host-Host recap (with median of clustering value for each order)

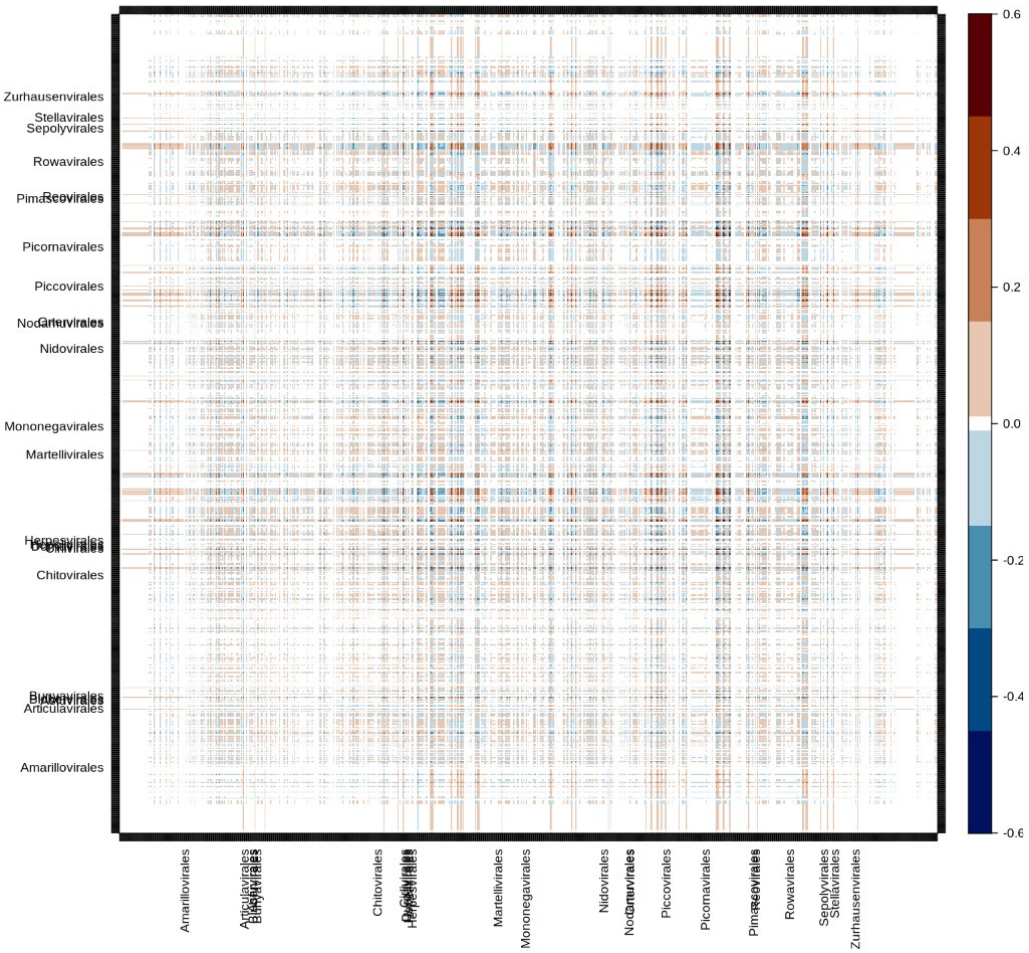

Figure 7 Virus-Virus cluster matrix, can be consider as host sharing similarity for virus pairs

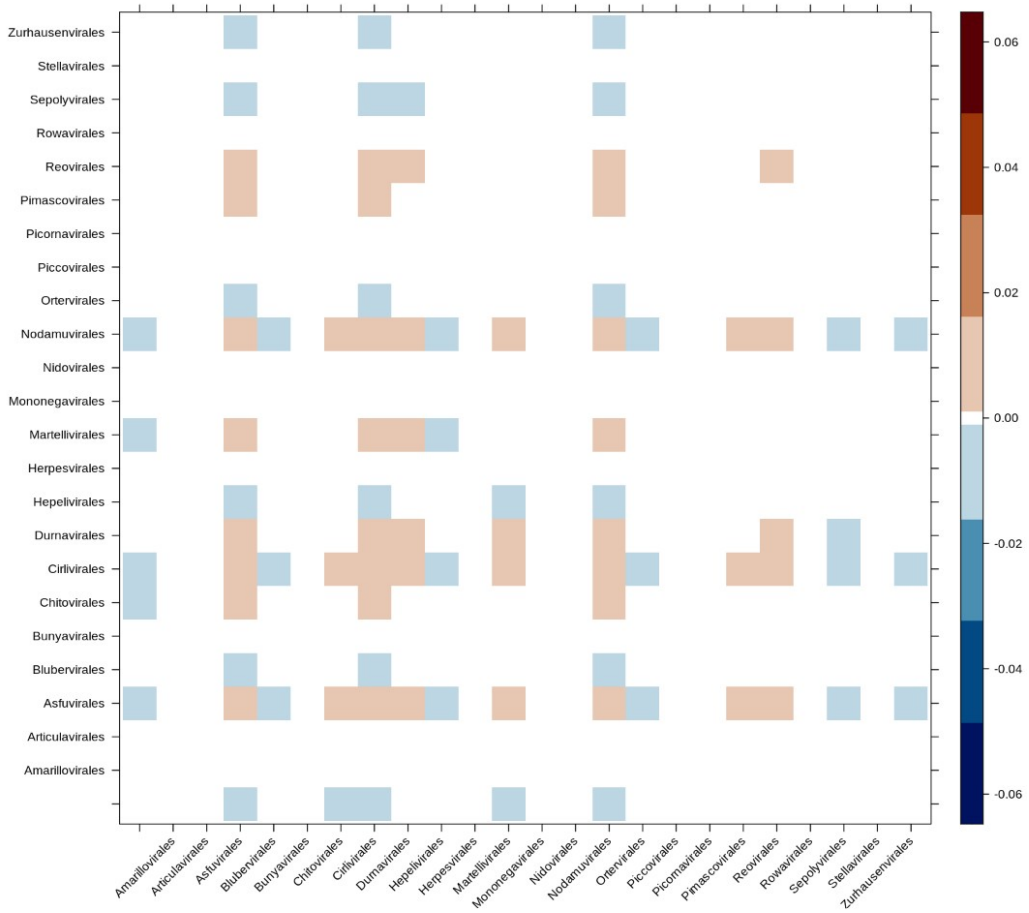

Figure 8 Virus-Virus recap (with median of cluster value for each order)

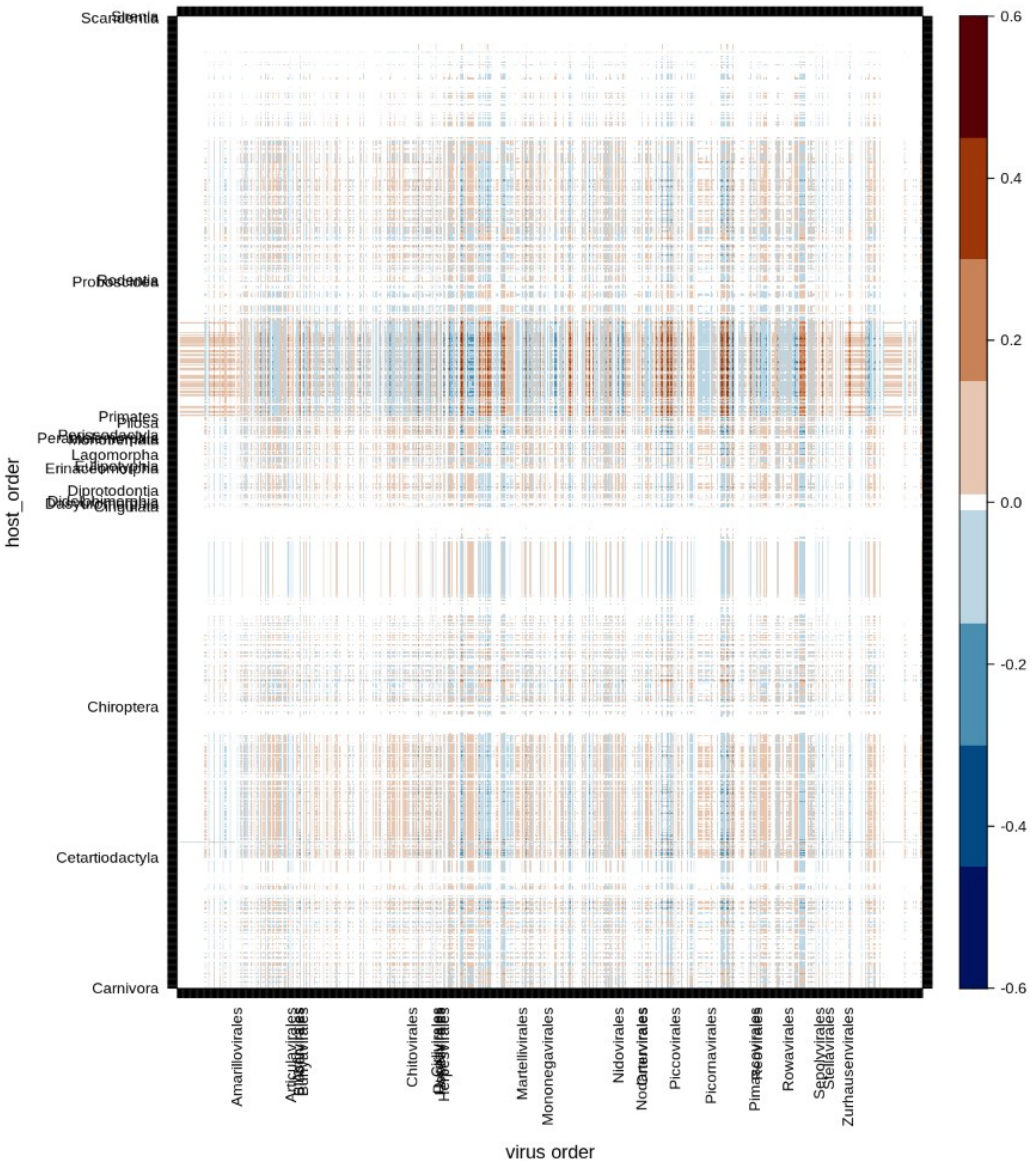

Figure 9 Host-Virus cluster matrix, can be seen as a quantification of the expect association occurrence for host-virus couples according to the global struct of the network

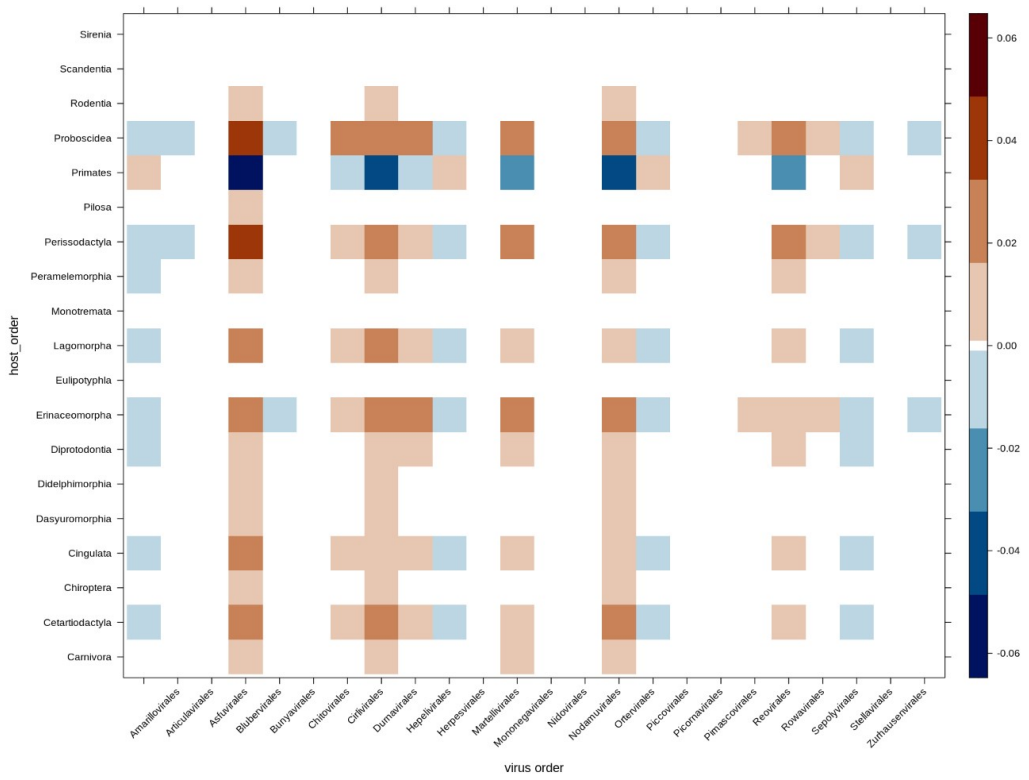

Figure 10 Host-Virus recap (with median of cluster value for each order)

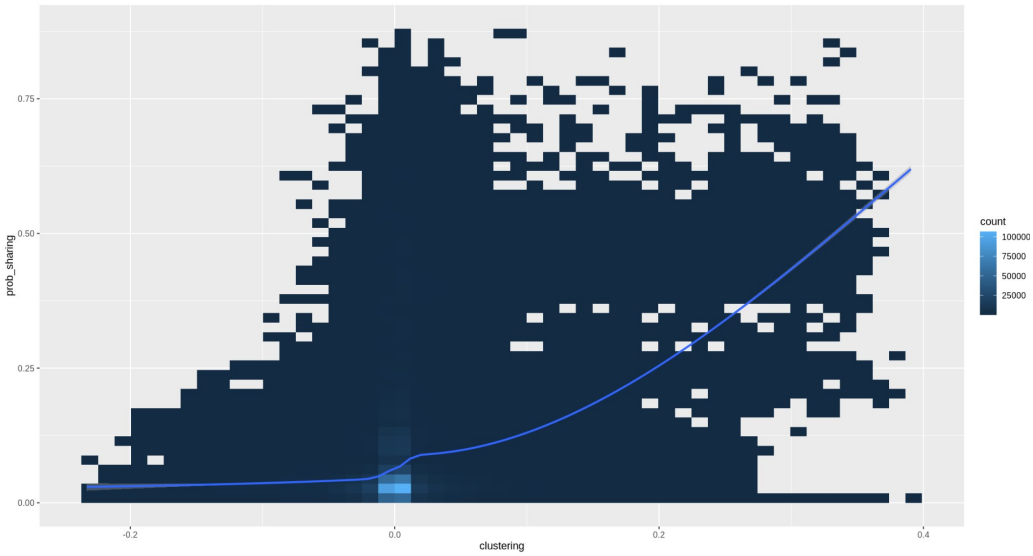

Figure 11 comparison with Albery sharing metric.

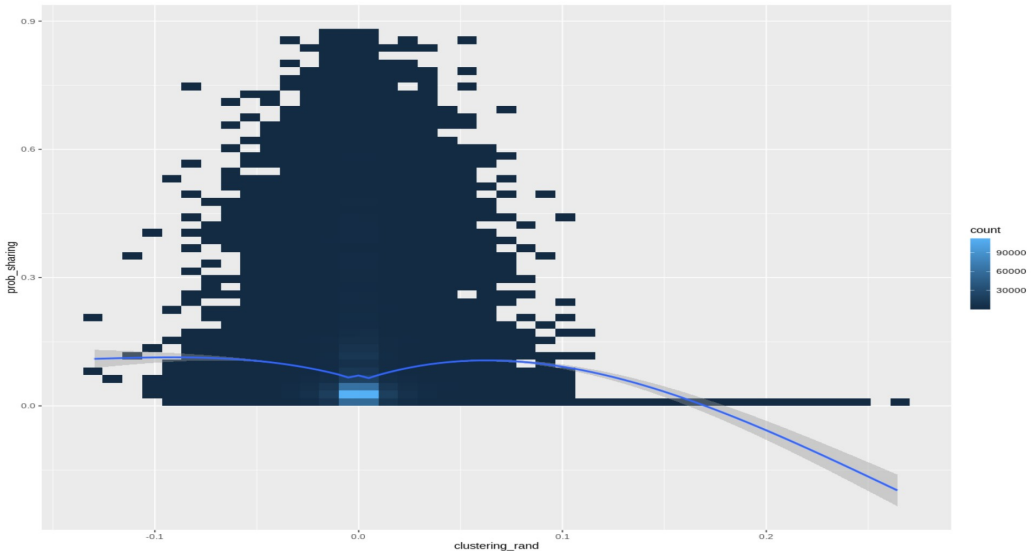

Figure 12 comparison with Albery sharing metric for a random network