Alpha多样性箱线图

徐锐

2020/6/17

### 全网最简单的网络图画法，小白福音包学包会

徐锐（助理研究员），广东省生态环境技术研究所，土壤微生物与宏基因组方向

版本1.0.1，更新日期：2020年6月23日

本项目永久地址： <https://github.com/YongxinLiu/MicrobiomeStatPlot> ，本节目录 212RareCurve，包含R markdown(\*.Rmd)、Word(\*.docx)文档、测试数据和结果图表，欢迎广大同行帮忙审核校对、并提修改意见。

#### 背景知识

本教程使用方法

1. 将251NetworkXuRui.zip解压至自己喜欢的目录即可（脚本部分修改成对应的工作目录）。本文件夹包含的子文件夹路径、名字不建议修改，脚本容易报错。
2. 部分脚本可取消#注释，供有需求时使用
3. 默认脚本与案例介绍均采用“微生物OTU丰度-环境理化Ev”的网络关系，可自行调整为“OTU-OTU”模式

##### 重要术语

**节点**（node/）：基因、物种OTU、环境因子等对象。若为有向网络，则可细分为源节点（Source）和目标节点（Target）。无向网络则不区分。

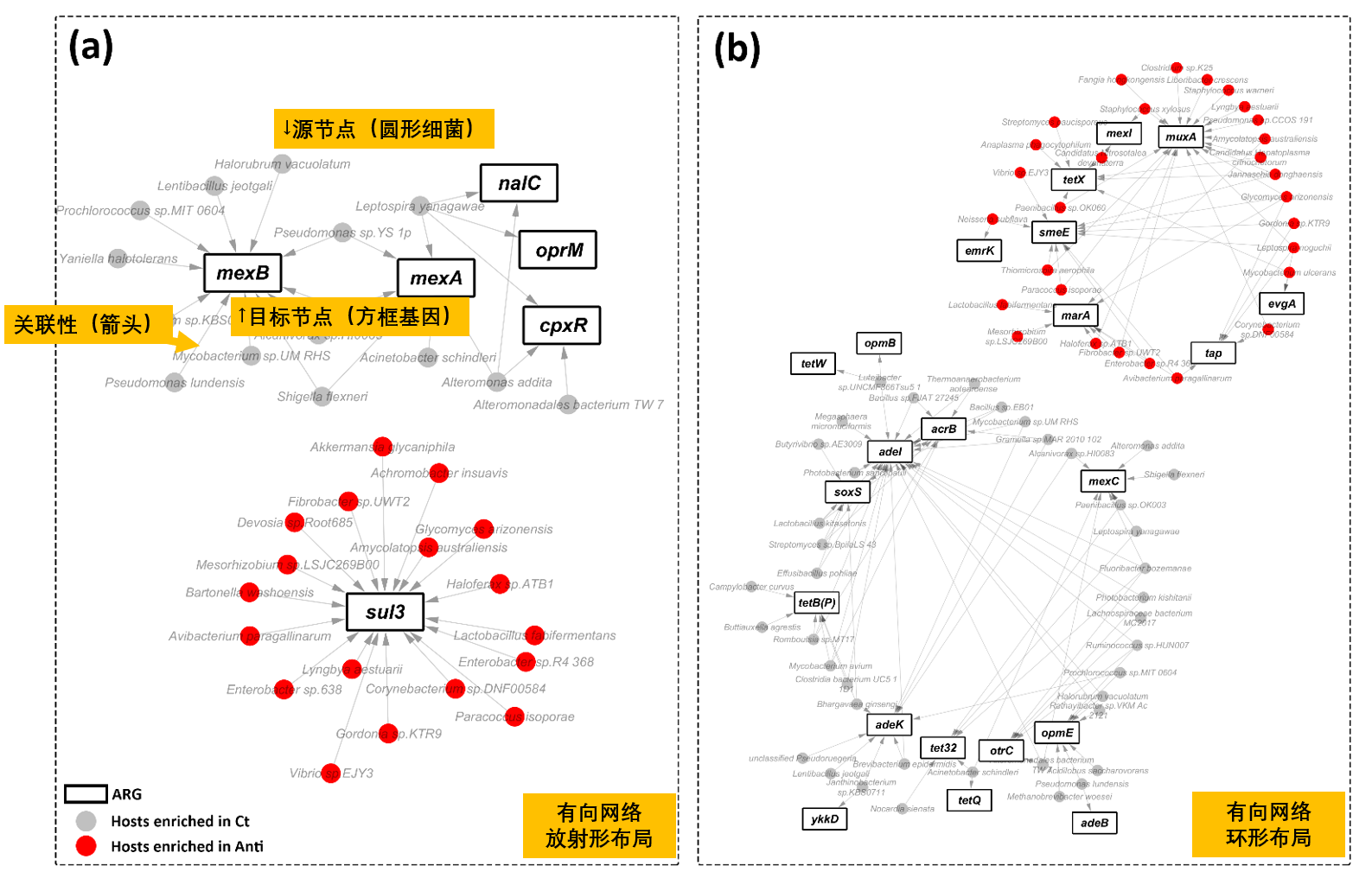
**边**（edge/link）：两个节点之间的连线，通常映射为相关性系数（r-value）。还可定义成正/负相关、有/无相关等。

**相关性**（correlation）：两个节点相关系数的计算结果，根据数据不同可采用不同的模型方法，如常见的Pearson、Spearman、SparCC等。

**显著性**（significance）：验证相关性是否具有显著性（p-value）。

**布局**（layout）：节点及边的分布形式，如常见的球形、圆环形、放射形等。

**拓扑特性**（topological property）：描述网络特征的数学参数，如连接度、中心性、模块数等



网络示意图

##### 简介

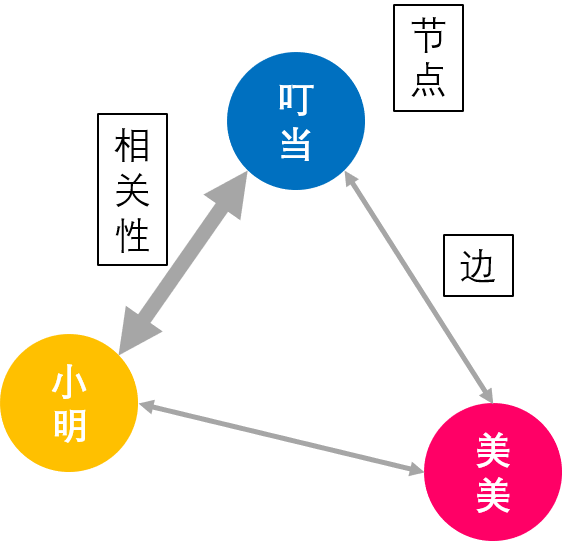
网络图（俗称毛线球）是一种广泛用于复杂、高维度数据的分析与可视化方法，常用于表现成百上千个微生物(菌群)、基因、蛋白等对象（统称为特征，feature）之间的关联形式。在网络图中，各个特征被定义为”节点“，而两两特征之间的关系被定义为“相关性”，并以“边”的形式将两个节点进行连接。

##### 简例

幼儿园小朋友们给自己最喜爱的食物打分：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 小明 | 叮当 | 美美 |
| 鸡腿 | 89 | 91 | 20 |
| 牛排 | 95 | 85 | 22 |
| 花椰菜 | 10 | 12 | 84 |

简单观察得分表后可以发现，**小明**与**叮当**都比较爱吃肉，而**美美**则是素食爱好者。那么**小明**和**叮当**成为好朋友的可能性可能要比**美美**更高，用网络图表示就是：



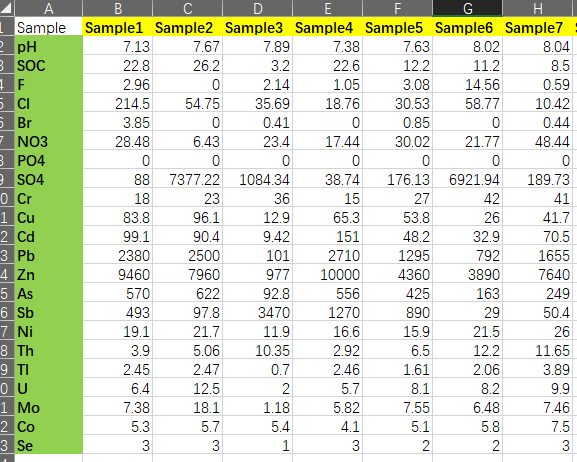
图中：三名小朋友就是**节点**，对食物的喜爱偏好就是**相关性**，并且小明和叮当之间的偏好要**强于**美美，故**边**也更宽。当然，仅仅根据食物偏好去判断是否能够成为好朋友，可信度不高。因此哪怕相关性很高（边更宽），但其背后的统计意义却不显著，也无法下结论。此外，本例不区分节点的方向，属于无向网络，故三名小朋友既是源节点，也是目标节点。最后，图片的整体分布特征（三角形）就是**布局**。

#### 实战范例

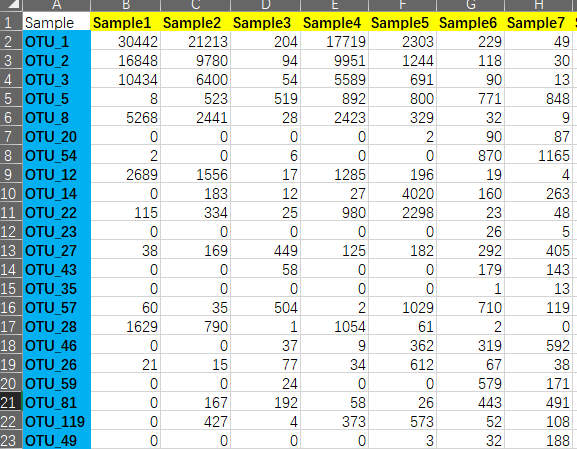
上述简例为抛砖引玉，实际的数据分析过程远比简例复杂。继续以最常见的微生物菌群丰度（OTU）与环境理化因子（Ev）的网络分析为例进一步说明：

##### 数据准备

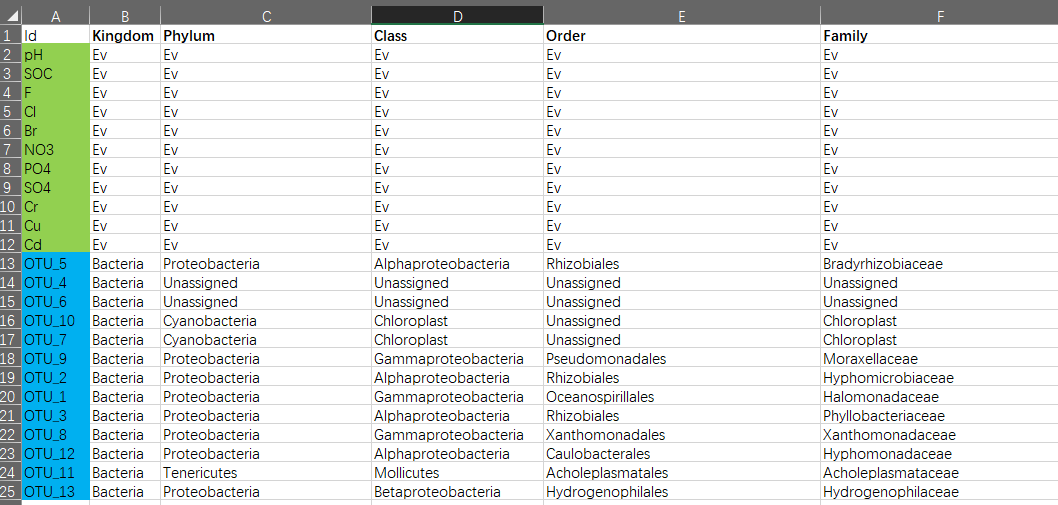
1. 环境理化因子数据（data/data\_Ev.txt）：行为理化指标，列为各样品



1. 微生物菌群OTU的丰度数据（data/data\_OTU.txt）：行为OTU丰度，列为各样品



1. Ev和OTU的补充注释信息，需手动整理，推荐合成一个表（data/taxonomy.txt）：



**注意**！

* Ev表和OTU表的样品要对应关系要一致
* 尽量避免行、列名使用非法字符，如#、！、数字开头、空格等，并预先剔除全为0的行/列，常常因为格式不对而报错！

##### 网络的计算

准备工作，安装R包，读取输入文件

# 检测包，是则跳过，没有则安装  
if (!requireNamespace("psych", quietly=TRUE))  
 install.packages("psych")  
if (!requireNamespace("reshape2", quietly=TRUE))  
 install.packages("reshape2")  
  
# 加载包  
library(psych)  
library(reshape2)  
  
# 导入数据(txt)，可在RStudio右上角手动Import Dataset，设置如下  
# OTU/Ev文件：heading=Yes, Row names=first column  
# Taxonomy文件：heading=Yes, Row names=automatic  
  
# 或使用如下命令导入数据：  
Ev <- read.table("data/Ev.txt", sep="\t", header=T, row.names=1)  
OTU <- read.table("data/OTU.txt", sep="\t", header=T, row.names=1)  
  
# 导入节点注释文件  
tax <- read.table("data/taxonomy.txt", sep="\t", header=T)  
names(tax)[1] <- "Id"

数据预处理

# 转置数据格式  
  
# # 情形1（默认）：两数据Ev-OTU表格时:  
Ev=t(Ev)  
OTU=t(OTU)  
  
# # 情形2：单数据OTU-OTU表格时：  
# OTU=t(OTU)

设定分析阈值

结果不理想时可反复修改这些阈值

# 若OTU数目太多，极大影响计算速度，而且结果不具有可读性  
# 按丰度值的百分比进行筛选, 默认保留相对丰度>0.05%的OTU  
abundance=0.05  
  
# 筛选  
OTU <- OTU[,colSums(OTU)/sum(OTU)>=(abundance/100)]  
  
# 网络分析的关联阈值  
r.cutoff=0.6  
p.cutoff=0.05

开始计算，不用修改

全选脚本后一键Enter~等待自动生成结果吧！超级爽

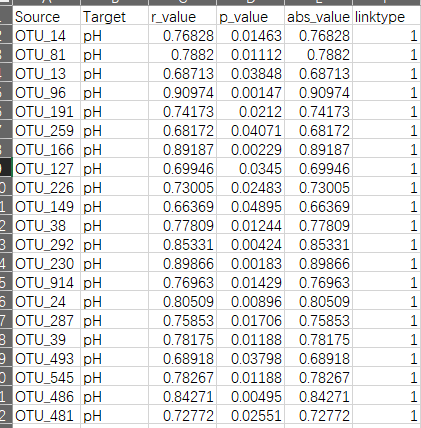
# 计算r、p  
  
# 情形1：两数据Ev-OTU表格时，默认  
occor=corr.test(OTU, Ev,  
 use="pairwise",  
 method="spearman", # 可选pearson/kendall  
 adjust="fdr",  
 alpha=0.05)  
  
# 情形2：单OTU-OTU  
# occor=corr.test(OTU,  
# use="pairwise",  
# method="spearman",  
# adjust="fdr",  
# alpha=0.05)  
  
# 获取相关矩阵及边数据  
  
# 提取相关性矩阵的r、p值  
r\_matrix=occor$r  
p\_matrix=occor$p  
  
# 确定物种间存在相互作用关系的阈值，将相关性R矩阵内不符合的数据转换为0  
r\_matrix[p\_matrix>p.cutoff|abs(r\_matrix)<r.cutoff]=0  
  
# 转换数据为长格式形式，方便下游分析  
p\_value=melt(p\_matrix)  
r\_value=melt(r\_matrix)  
  
#将r、p两表合并  
r\_value=cbind(r\_value, p\_value$value)  
  
# 删除含r\_value=0的行  
r\_value=subset(r\_value, r\_value[,3]!=0)  
  
# 删除含r\_value=NA的行  
r\_value=na.omit(r\_value)  
  
# 对r表格增补绝对值、正负型等信息  
abs=abs(r\_value$value)  
  
linktype=r\_value$value  
linktype[linktype>0]=1  
linktype[linktype<0]=-1  
  
r\_value=cbind(r\_value, abs, linktype)  
  
# 重命名r、p表头  
names(r\_value) <- c("Source","Target","r\_value","p\_value", "abs\_value", "linktype")  
names(p\_value) <- c("Source","Target","p\_value")  
  
# 输出结果为csv文件  
write.csv(r\_value,file="result/1.边数据.csv", row.names=FALSE)  
write.csv(r\_matrix, file="result/4.corr\_matrix.csv")  
write.csv(r\_value,file="result/5.r\_value.csv", row.names=FALSE)  
write.csv(p\_value,file="result/6.p\_value.csv", row.names=FALSE)  
  
  
# 获取节点数据  
# 从边文件提取节点并去除重复  
node\_OTU <- as.data.frame(as.data.frame(r\_value[,1])[!duplicated(as.data.frame(r\_value[,1])), ])  
node\_Ev <- as.data.frame(as.data.frame(r\_value[,2])[!duplicated(as.data.frame(r\_value[,2])), ])  
  
names(node\_OTU)="Id"  
names(node\_Ev)="Id"  
  
# OTU ID和Ev ID合并成节点索引表，用于检索注释信息  
list <- rbind(node\_Ev, node\_OTU)  
write.csv(list,file="result/3.node\_list.csv", row.names=FALSE)  
  
# 筛选节点对应的注释信息  
list=subset(tax,Id %in% list$Id)  
  
# 复制一列当节点Label  
list$Label <- list$Id  
  
# 输出结果为csv文件  
write.csv(list,file="result/2.节点数据.csv", row.names=FALSE)

查验结果

在result文件夹中查验生成的表格结果，主要使用**1.边数据.csv**和**2.节点数据.csv**两个。

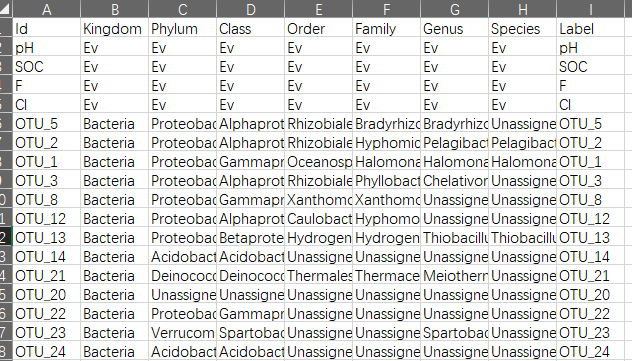
**1.边数据：**

说明：当区分有向、无向网络时，Source和Target节点才有区别。r-value表示符合网络阈值的相关性数值，p-value供说明相关性的显著程度。abs-value是将可正可负的r-value取绝对值，用于画图时表示连线的宽度（关联强度）。linktype表示正（1）、负（-1）相关性，可在画图时用于指定连线的颜色（红=正相关，蓝=负相关）。



**2.节点数据：**

说明：Id表示边数据中的Source、Target节点，后续几列为节点的注释信息，如分类水平、性别（如有）、采样点等。最后一列Label用于指定画图时节点显示的标签字符，可手动删除不想显示的内容。



**疑问？**：为什么需要生成节点数据？因为如果直接使用最开始的taxonomy.txt注释文件（总表）画图，会存留许多非网络节点的节点（冗余）。因此需要根据边数据中保留下的节点（符合r/p网络阈值的），从taxonomy总表中挑选出来制作画图用的节点数据（子表）。

##### 网络的可视化

数据在手，天下我有！只要有脚本生成的、或不怕麻烦自己excel手动整理的**1.边数据.csv**和**2.节点数据.csv**就可以进行网络图的可视化啦~推荐使用Cytoscape或者Gephi两个软件。以Cytoscape为例：

**Cytoscape安装及下载**

下载页：<https://cytoscape.org/download.html>

下载最新版本的Cytoscape和对应的Java环境版本

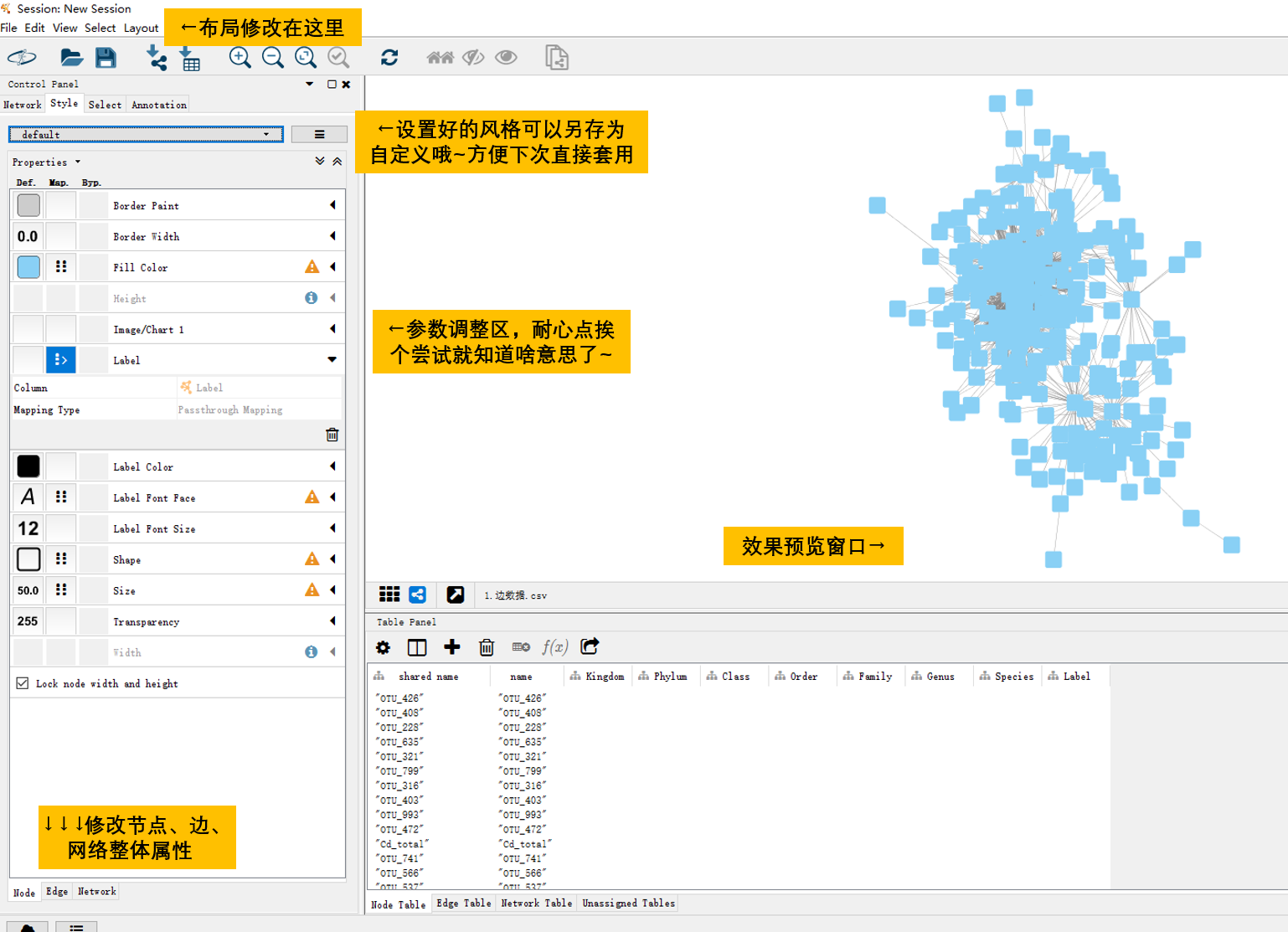
* 导入数据

**边数据**：【File】【Import】【Network from file】【1.边数据.csv】

**节点数据**：【File】【Import】【Table from file】【2.节点数据.csv】

* 美化

Cytoscape中可以修改几乎所有能够想到的网络图属性，节点、连线的颜色、粗细、透明度当然不在话下，还可以修改多种布局、标签显示方式，甚至还能计算网络的拓扑参数。由于这部分不是本文重点，不再赘述，具体可参见其他指导手册~



**CytoScape网络可视化相关资源**：

* [操作界面介绍](https://mp.weixin.qq.com/s/ZSoW7-qWs3BuSB7bkDnfmA)
* [一网打尽](https://mp.weixin.qq.com/s/fbiMeE1QDMQn3HSgusxh1Q)
* [轻松实现](https://mp.weixin.qq.com/s/tPOO-9aTorMPeXyUQn696Q)

**Gephi网络可视化相关资源**：

* [Gephi方法](https://mp.weixin.qq.com/s/FwktSxYbiN0-ilIyceDAcQ)

**学术论文案例**

* 厌氧消化反应器中的微生物网络: <https://www.sciencedirect.com/science/article/abs/pii/S0960852418306060> “Organic loading rate and hydraulic retention time shape distinct ecological networks of anaerobic digestion related microbiome”
* 抗生素抗性基因与潜在宿主网络: <https://www.sciencedirect.com/science/article/abs/pii/S0960852419303281> “Metagenomic analysis reveals the effects of long-term antibiotic pressure on sludge anaerobic digestion and antimicrobial resistance risk”
* 酸性矿山废水侵蚀下的微生物网络: <https://www.sciencedirect.com/science/article/pii/S0269749119368253> “Uncovering microbial responses to sharp geochemical gradients in a terrace contaminated by acid mine drainage”

责编：刘永鑫，中科院遗传发育所

版本1.0.0，网络基本讲解和网络文件准备 版本1.0.1，改写为Rmd版本，建议作者增加实战讲解和点评、绘图实战的典型操作和经验。