Bootstrap false discovery for network creation

This code demonstrates the use of the Bootstrap false-discovery (BSFD) algorithm to mine interesting features from within a dataset of many features. The objective of this function is to generate association data that can be explored through network analyses using either the igraph and ggraph packages.

Read data

These data represent the abundances of more than 3,000 bacterial species associated with the roots of two wetland plants (48 samples). These data were obtained by extracting environmental DNA, amplifying (artificially replicating) a portion of a specific gene common to all bacteria (one that is particularly useful for taxonomic differentiation), and sequencing the amplified gene fragments using modern, next-generation sequencing technologies. The raw sequences were cleaned of sequencing errors and taxonomic classifications made using the bioinformatics pipeline mothur. `scan` is used because it can read in large matrices more efficiently than `read.csv` (which is more appropriate for data frames)

```r
dat <- scan("example_abundance_data.csv", sep = ",", what = integer(),
    skip = 1)
dat.names <- scan("example_abundance_data.csv", sep = ",", what = character(),
    nlines = 1)
dat <- matrix(dat, ncol = length(dat.names), byrow = T)
colnames(dat) <- dat.names
dim(dat)
```

```
## [1]   48 3145
```

```r
dat[1:10, 1:5]
```

```
##       species_1 species_2 species_3 species_4 species_5
##  [1,]       306        62       493       162        12
##  [2,]       159       114       425       121        30
##  [3,]       508       425       469       100        32
##  [4,]      1210       237       589       235        46
##  [5,]       152       410       184       168        35
##  [6,]       535       208       274       254         6
##  [7,]       363       210       250       259        37
##  [8,]        25        59       200         1         0
##  [9,]        34         6        28         7         2
## [10,]       234        36       109        78        14
```

Accompanying microbial species abundances are data pertaining to the habitat the samples were collected from. Samples were collected from three sites along a single wetland in Northwest Pennsylvania, were associated with either broadleaf cattail (*Typha latifolia*) or purple loosestrife (*Lythrum salicaria*), and were either found growing separately or together (bringing their root systems into direct contact).

```r
factors <- read.csv("example_sampling_data.csv")
factors[1:10, ]
```

```
##    names           site           species   occurrence
## 1   DL1S Dot Farm Marsh Lythrum salicaria     Separate
## 2   DL1C Dot Farm Marsh Lythrum salicaria Co-occurring
## 3   DL2S Dot Farm Marsh Lythrum salicaria     Separate
## 4   DL2C Dot Farm Marsh Lythrum salicaria Co-occurring
## 5   DL3S Dot Farm Marsh Lythrum salicaria     Separate
## 6   DL3C Dot Farm Marsh Lythrum salicaria Co-occurring
## 7   DL4S Dot Farm Marsh Lythrum salicaria     Separate
## 8   DL4C Dot Farm Marsh Lythrum salicaria Co-occurring
## 9   DT1S Dot Farm Marsh   Typha latifolia     Separate
## 10  DT1C Dot Farm Marsh   Typha latifolia Co-occurring
```

These factors can be used to separate the dataset to explore any differences in network structure. Since plant species are known to be important determinants of the community composition of root bacteria, the data will be subset by the `species` factor to compare the structures of these networks separately.

```
cattail <- dat[factors$species == "Typha latifolia", ]
loosestrife <- dat[factors$species == "Lythrum salicaria", ]
```

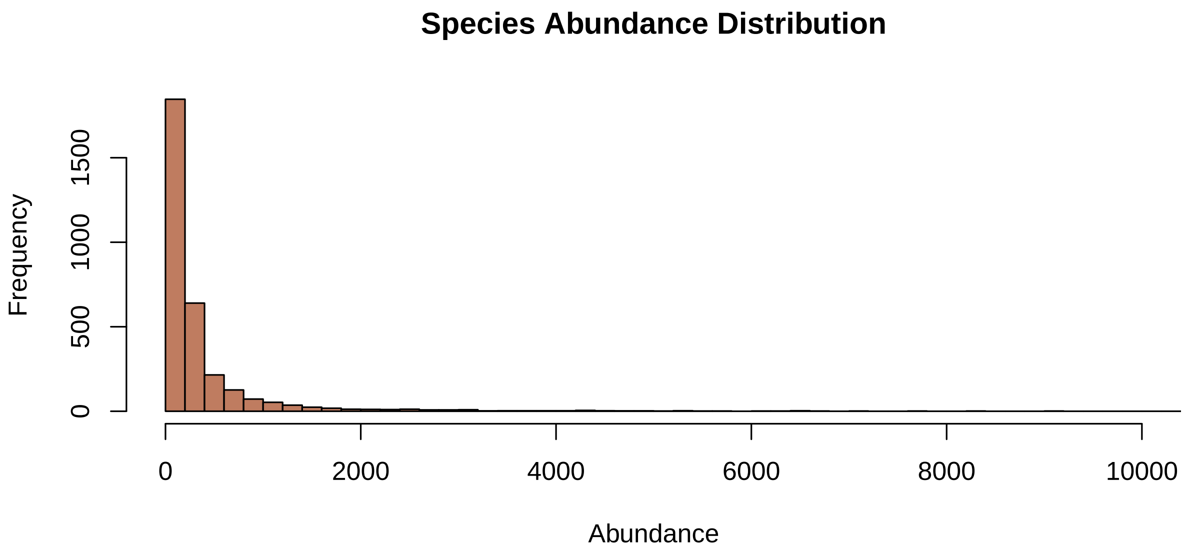
```
dim(cattail)
```

```
## [1]   24 3145
```

```
dim(loosestrife)
```

```
## [1]   24 3145
```

One of the few universal patterns in ecology is the numerical dominance of the community by a small proportion of overall diversity contrasted by the rarity (low numbers and sparse occurrence) of the vast majority of species. This can be represented by a *Species Abundance Distribution*. The figure below demonstrates that most species occur only a limited number of times. The main problem is that such rare species have an abundance of zero in many of the samples, but if two species occur together and are rare, they will demonstrate very high correlation. This will produce a network of poorly-connected peripherals and so species that occur infrequently should be curated. It is common to remove those features which occur in less than half of the samples, but being even more stringent may produce a more informative network. The code below converts the matrix into a presence-absence (0,1) matrix and subsets only those bacteria which occur more than 18 times.



```
incidence <- apply(cattail, 2, pmin, 1)
total.incidence <- colSums(incidence)
cattail <- cattail[, which(total.incidence > 18)]
ncol(cattail)
```

```
## [1] 727
```

```
incidence <- apply(loosestrife, 2, pmin, 1)
total.incidence <- colSums(incidence)
loosestrife <- loosestrife[, which(total.incidence > 18)]
ncol(loosestrife)
```

```
## [1] 783
```

There are 727 bacterial species that occur more than 18 times in cattail roots and 783 in loosestrife roots.

Determine significant associations between bacteria (features)

Here, the `bs_fdr` function will be used to generate an association matrix of the bacterial data, and apply a user-defined cutoff to keep only the associations with sufficiently strong (positive or negative) values. Before calling the function, it is important to consider several variables that will affect the final output:

- **Initial threshold value:** The minimum strength of the association that the researchers is interested in. Since many association values scale from 0 to 1 (or -1 to 1), this value is set to ±0.5. The BSFD algorithm will determine how much this initial value must be increased in order to satisfy the acceptable number of false discoveries as provided by the user.
- **False discovery rate (fdr):** The number of false positives (associations deemed significant due to chance, also known as Type I error) that are acceptable to the user. The default is 1.
- **Risk:** The probability that the number of false positives / false discoveries will exceed the rate specified by the user. A risk of 0.05 (the default) means there is a 5% chance the false discovery rate is higher than specified and, conversely, a 95% chance the the false discovery rate is equal to, or lower than, that specified.
- **Correlation method:** The measure used to calculate feature associations. Possible measures are: Spearman (default), Pearson, and Kendall (from `stats`) as well as several ecologically-relevant measures provided by the `vegdist` function from the vegan package: Manhattan, Euclidean, Canberra, Bray-Curtis, Kulczynski, Jaccard, Gower, alt-Gower, Morisita, Horn, Mountford, Raup-Crick, Binomial, Chao, Cao, Mahalanobis.

```
source("bs_fdr.R")
```

`bs_fdr` takes the resulting association matrix (of class `distance`) and makes assessments in 10,000 row blocks. This is to allow the function to compute large distance matrices generated from bacterial datasets often with tens of thousands of species across columns. In this example, the initial threshold of interest is 0.3, as the default results in only a few acceptable associations. Although 1,000 is the default number of bootstrapped iterations, 10,000 produces a more consistent selection. **Note:** These functions produce messages about the progress of the calculations (for larger datasets), significant features kept, and the total adjustment of the significance threshold. These messages were masked in the creation of this document.

```
cattail_edges <- bs_fdr(cattail, init.threshold = 0.3, fdr = 1,
    risk = 0.05, iters = 10000)
loosestrife_edges <- bs_fdr(loosestrife, init.threshold = 0.3,
    fdr = 1, risk = 0.5, iters = 10000)
```

This returns 66 associations from cattail-associated communities and 340 from loosestrife-associated communities. Next, a data frame of the species that these associations include needs to be constructed in order to plot the results in igraph. These will form the nodes, or vertices of the network.

```
construct_nodes <- function(x) {
    to <- as.character(unique(x$to))
    from <- as.character(unique(x$from))
    nodes <- unique(c(to, from))
    nodes <- data.frame(species = nodes)
}
```

```
cattail_nodes <- construct_nodes(cattail_edges)
cattail_abund <- colSums(cattail)
cattail_abund <- data.frame(species = attr(cattail_abund, "names"),
    abundance = cattail_abund)

cattail_nodes <- merge(cattail_nodes, cattail_abund)
nrow(cattail_nodes)
```

```
## [1] 72
```

```
cattail_nodes[1:6, ]
```

```
##        species abundance
## 1 species_1011       137
## 2 species_1015       106
## 3 species_1029       119
## 4 species_1049       127
## 5 species_1063       143
## 6 species_1069        91
```

```
loosestrife_nodes <- construct_nodes(loosestrife_edges)
loosestrife_abund <- colSums(loosestrife)
loosestrife_abund <- data.frame(species = attr(loosestrife_abund,
    "names"), abundance = loosestrife_abund)

loosestrife_nodes <- merge(loosestrife_nodes, loosestrife_abund)
nrow(loosestrife_nodes)
```

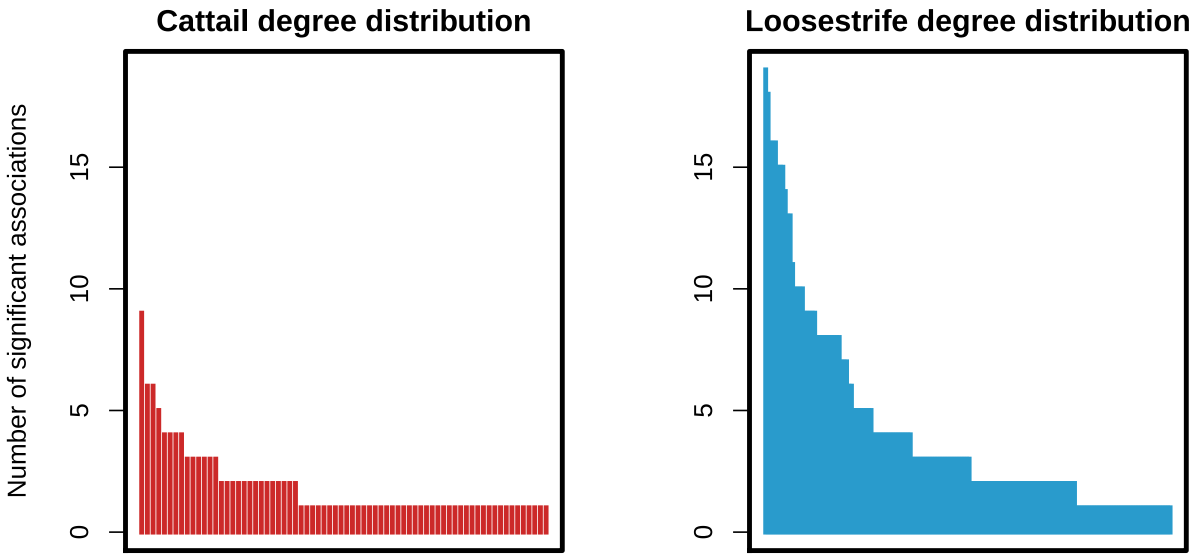
```
## [1] 166
```

```
loosestrife_nodes[1:6, ]
```

```
##        species abundance
## 1 species_1002       125
## 2 species_1003       126
## 3 species_1011       127
## 4 species_1012        71
## 5 species_1015       157
## 6 species_1016       210
```

We see that 72 bacterial species form important associations in bacterial associations in cattail roots compared to 166 bacterial species. From here, some basic network properties can be identified. One the most basic is *degree*. The degree of the network is simply the number of connections each species has in the network. Below the distribution of degree values is plotted for the nodes (sorted).

```
cattail_degree <- with(cattail_edges, {
    to <- as.character(to)
    from <- as.character(from)
    as.data.frame(table(c(to, from)))
})
loosestrife_degree <- with(loosestrife_edges, {
    to <- as.character(to)
    from <- as.character(from)
    as.data.frame(table(c(to, from)))
})
```



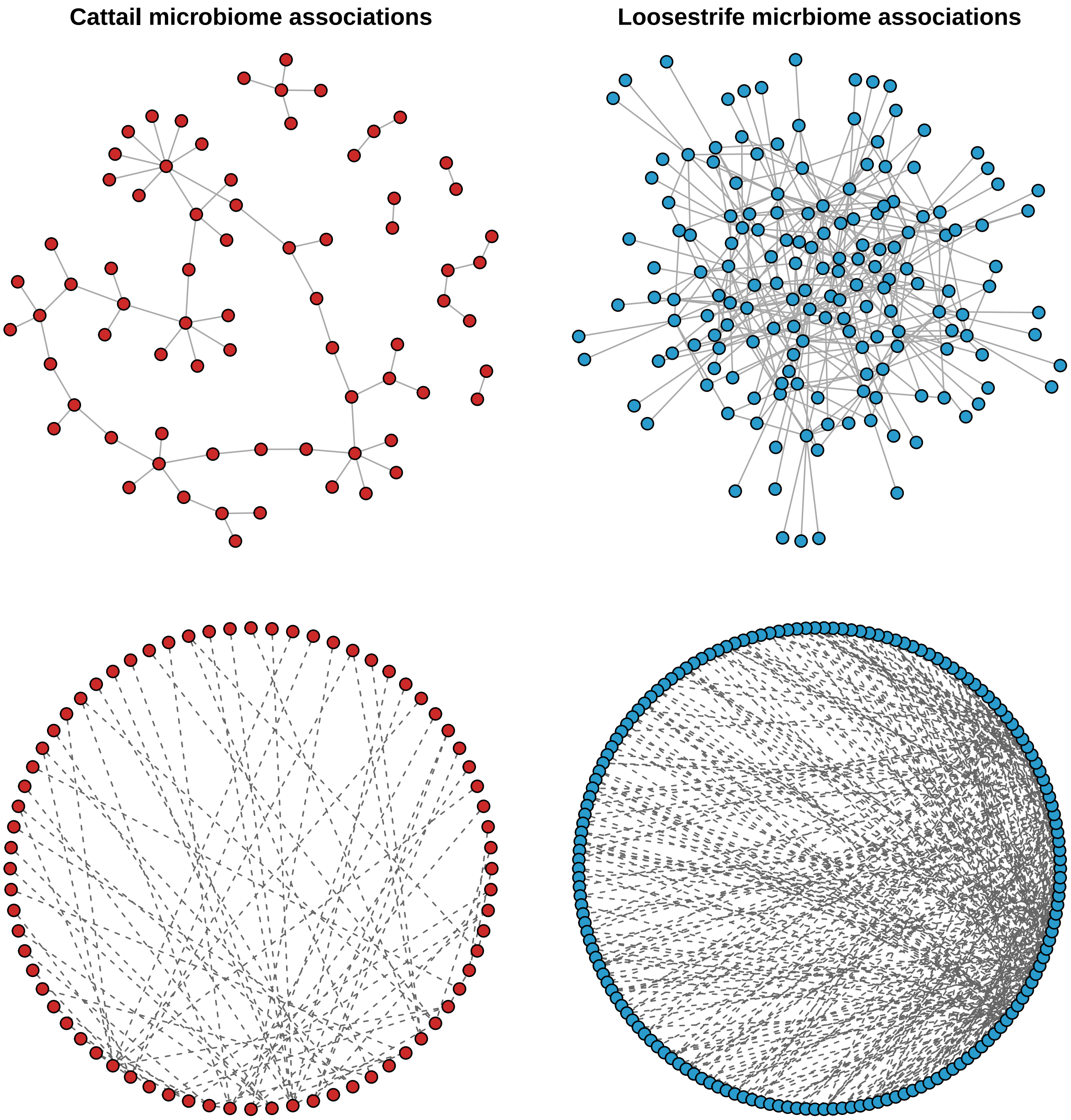
Graphing association networks

```r
library(igraph)
```

The igraph package can generate a network from either a symmetrical association matrix, or from data frames prepared beforehand. Here, cattail_edges and cattail_nodes will be combined into a single network while loosestrife_edges and loosestrife_nodes will be combined with the graph_from_data_frame function. The network can then be plotted.

```r
net_cat <- graph_from_data_frame(d = cattail_edges, vertices = cattail_nodes,
    directed = F)
net_loose <- graph_from_data_frame(d = loosestrife_edges, vertices = loosestrife_nodes,
    directed = F)
```

```r
par(mar = c(0, 0, 1.5, 0), mfrow = c(2, 2), byrow = T)
plot(net_cat, vertex.label = NA, vertex.size = 5, vertex.color = hsv(0,
    0.8, 0.8), main = "Cattail microbiome associations")
plot(net_loose, vertex.label = NA, vertex.size = 5, vertex.color = hsv(0.55,
    0.8, 0.8), main = "Loosestrife micrbiome associations")
plot(net_cat, layout = layout.circle(net_cat), vertex.label = NA,
    vertex.size = 5, edge.lty = 2, vertex.color = hsv(0, 0.8,
        0.8), edge.color = gray(0.4))
plot(net_loose, layout = layout.circle(net_loose), vertex.label = NA,
    vertex.size = 5, edge.lty = 2, vertex.color = hsv(0.55, 0.8,
        0.8), edge.color = gray(0.4))
```



Cattail microbiome associations

Loosestrife micrbiome associations

Conclusions

There were more bacterial associations in the loosestrife microbiome above the adjusted significance threshold given by the `bs_fdr` function. All interactions were positive, suggesting that competition between bacterial species is not an important influence in the bacterial community. For biological, ecological, and other sparse datasets (likely transactional information), care must be taken to reduce the influence of zero-count cells on the `bs_fdr` selection. Other options that exist include transforming the data or using correlation methods that take into account sparse occurrence. Several analytical pathways exist from this point by exploiting network properties, such as identifying network *hubs* (which in ecology represent potential keystone species).