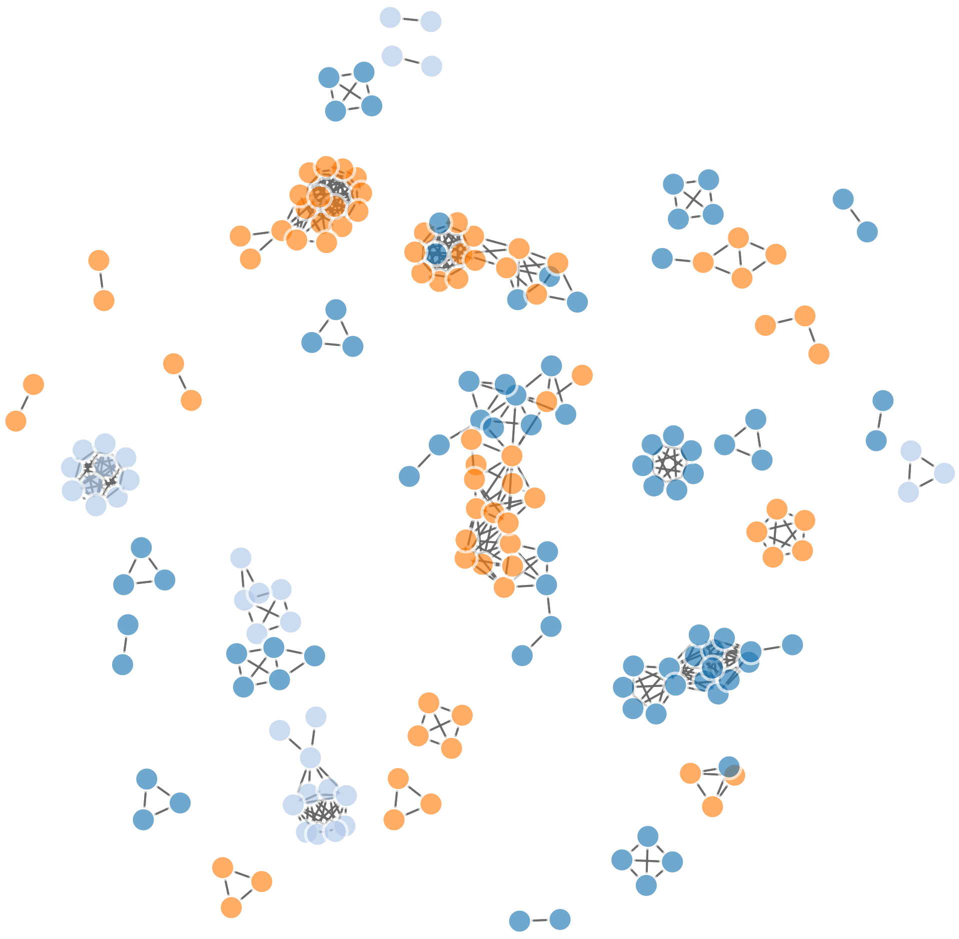
8 Models Clustered by Tag Similarity

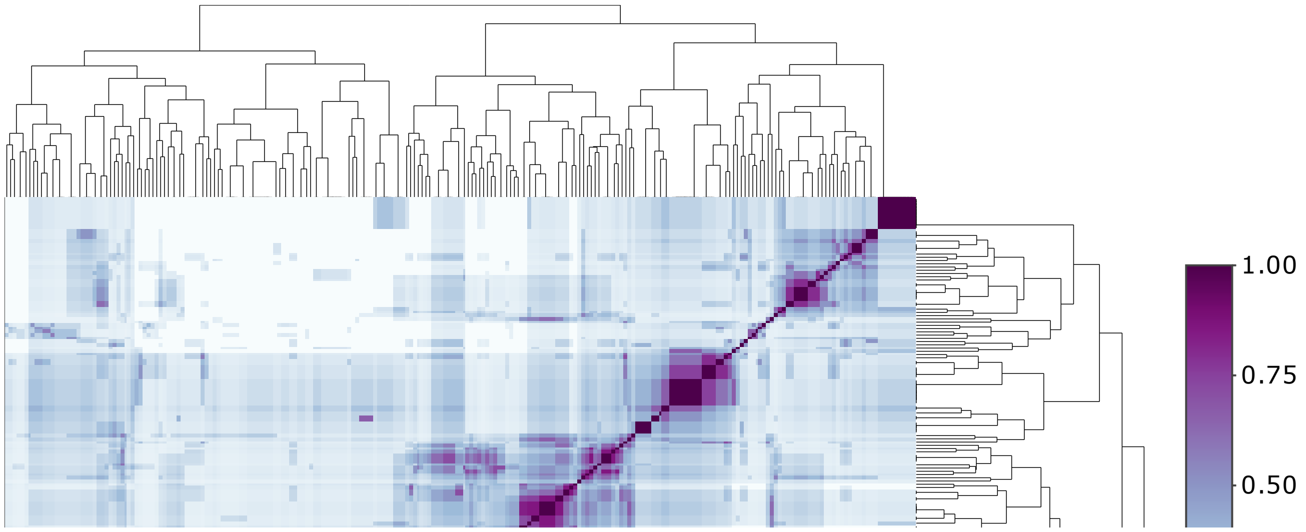
This page shows a network diagram of all the models that can be accessed by `train`. See the Revolutions blog for details about how this visualization was made (and this page has updated code using the `networkD3` package). In summary, the package annotates each model by a set of tags (e.g. “Bagging”, “L1 Regularization” etc.). Using this information we can cluster models that are similar to each other.

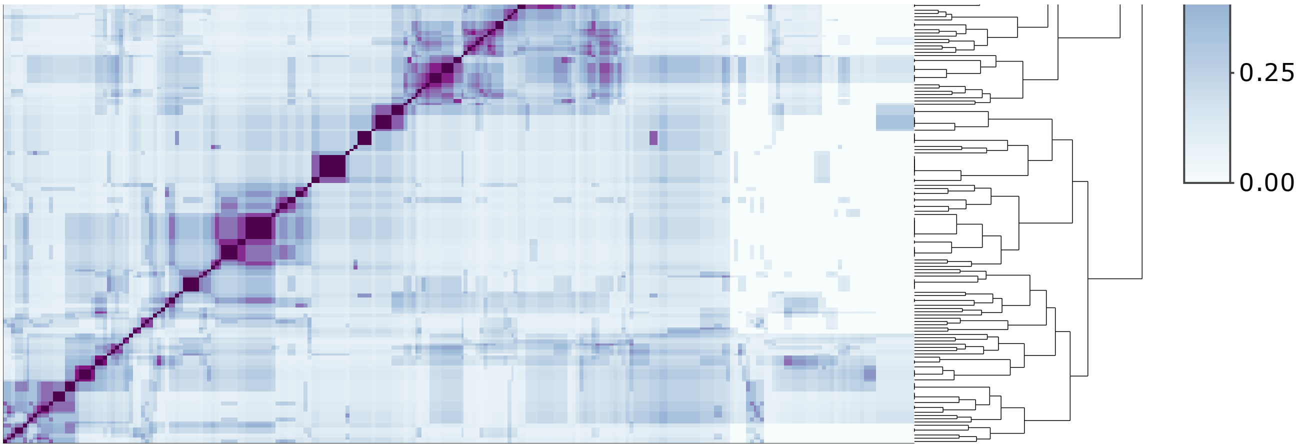
Green circles are models only used for regression, blue is classification only and orange is “dual use”. Hover over a circle to get the model name and the model code used by the `caret` package and refreshing the screen will re-configure the layout. You may need to move a node to the left to see the whole name. 43 models without connections are not shown in the graph.



The data used to create this graph can be found here.

The plot below shows the similarity matrix. Hover over a cell to see the pair of models and their Jaccard similarity. Darker colors indicate similar models.





You can also use it along with maximum *dissimilarity* sampling to pick out a diverse set of models. Suppose you would like to use a SVM model with a radial basis function on some regression data. Based on these tags, what other four models would constitute the most diverse set?

```
tag <- read.csv("tag_data.csv", row.names = 1)
tag <- as.matrix(tag)

## Select only models for regression
regModels <- tag[tag[,"Regression"] == 1,]

all <- 1:nrow(regModels)
## Seed the analysis with the SVM model
start <- grep("(svmRadial)", rownames(regModels), fixed = TRUE)
pool <- all[all != start]
```

grep返回rownames是regModels的indices
start & pool都作为indices
maxDissim取pool中与start最大差异的，nextMods总数为4

```
## Select 4 model models by maximizing the Jaccard
## dissimilarity between sets of models
nextMods <- maxDissim(regModels[start,,drop = FALSE],
                      regModels[pool, ],
                      method = "Jaccard",
                      n = 4)

rownames(regModels)[c(start, nextMods)]
```

```
## [1] "Support Vector Machines with Radial Basis Function Kernel
## [2] "Cubist (cubist)"
## [3] "Bayesian Regularized Neural Networks (brnn)"
## [4] "Negative Binomial Generalized Linear Model (glm.nb)"
## [5] "Logic Regression (logreg)"
```