Lab 5_simon

Simon Jorstedt

2024-12-17

```r
# Dependencies
library(magrittr)
library(igraph)
```

Question 1

Go to the webpage http://snap.stanford.edu/biodata/ and choose one of the provided datasets. Download it and reproduce the statistics concerning the graph. If you obtain different values, then discuss this in your report. Visualize the graph. The next step is to try to identify some clusters (communities in the graph). You can follow the tutorial at https://psych-networks.com/r-tutorial-identify-communities-items-networks/ to achieve this. Once you have found some clusters, identify the elements in it and try to find information on this cluster. Is it related to some known biological phenomena? If you do not find anything, then document your search attempts. If it will not be possible to do this question on the whole downloaded graph, then you may take some sub-graph of it.

From the provided website we obtain the “SS-Butterfly_weights.tsv.gz” dataset. In the chunk below, we obtain some properties/statistics of the dataset. We find that the graph as provided is fully connected, directed, and non-cyclic. However, it is implied that the direction describes the nature of the similarity (captured by edge weight). Therefore we will treat all edges as bidirectional. Similarly to the website, we find in the chunk below that the number of nodes and edges are 832 and 86528. There is one strongly connected component, containing all nodes. Below are the properties according to the website.

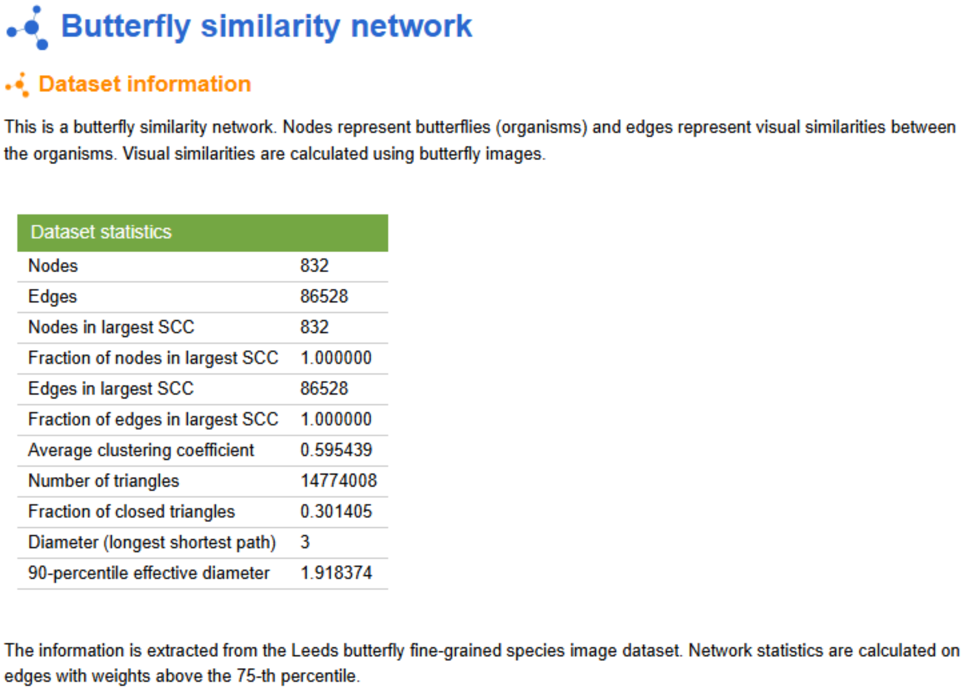
```r
# Load data
butterfly_weights <- read.table(gzfile("SS-Butterfly_weights.tsv.gz"))
butterfly_names <- read.table(gzfile("SS-Butterfly_labels.tsv.gz"))

# Number of nodes
c(butterfly_weights$V1, butterfly_weights$V2) %>% unique() %>% length()
```

```
## [1] 832
```

```r
# Number of edges
nrow(butterfly_weights)
```

```
## [1] 86528
```

```r
# Finding the number of strongly connected components
graph_1 <- igraph::graph_from_data_frame(butterfly_weights[,c(1,2)], directed = FALSE)
igraph::components(graph_1, mode="strong")$membership %>% unique() %>% length()
```



Butterfly similarity network

Dataset information

This is a butterfly similarity network. Nodes represent butterflies (organisms) and edges represent visual similarities between the organisms. Visual similarities are calculated using butterfly images.

Dataset statistics	
Nodes	832
Edges	86528
Nodes in largest SCC	832
Fraction of nodes in largest SCC	1.000000
Edges in largest SCC	86528
Fraction of edges in largest SCC	1.000000
Average clustering coefficient	0.595439
Number of triangles	14774008
Fraction of closed triangles	0.301405
Diameter (longest shortest path)	3
90-percentile effective diameter	1.918374

The information is extracted from the Leeds butterfly fine-grained species image dataset. Network statistics are calculated on edges with weights above the 75-th percentile.

Figure 1: Apparently true properties

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

Now we obtain the average clustering coefficient (which matches with the website), and the number of triangles, which does not match. The website suggests that there are over 14 million triangles, but even if we count all the triangles for every node and sum them up (triple-counting all triangles), we only achieve about 13 million triangles/triples. In turn, we achieve a different fraction of closed triangles. Also the diameter is different. We also find that all nodes are connected by a path with a most 4 nodes, and so it makes sense that the 90-percentile effective diameter is very low, although the corresponding value on the website is slightly lower (1.918374). Regarding the "number of triangles", we will conclude that the website is too unspecific to warrant further investigation.

```r
# Average clustering coefficient
igraph::transitivity(graph_1, type="average")
```

```
## [1] 0.5954394
```

```r
# Number of triangles
igraph::count_triangles(graph_1) %>% sum()
```

```
## [1] 13358898
```

```r
# Fraction of closed triangles
igraph::transitivity(graph_1, type="global")
```

```
## [1] 0.5641441
```

```
# Diameter
igraph::diameter(graph_1)
```

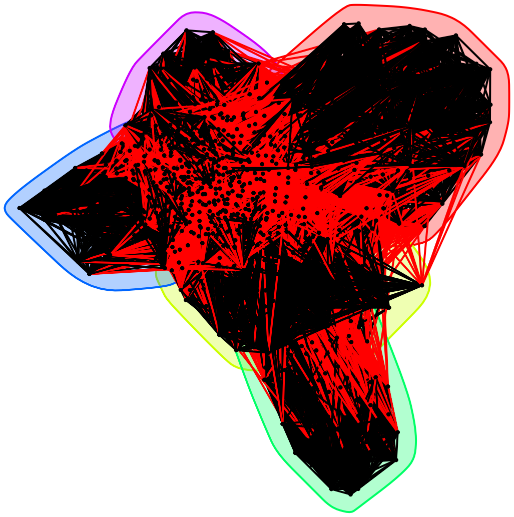
```
## [1] 4
```

```
# 90 percentile effective diameter
sp_lengths <- igraph::distances(graph_1)
quantile(sp_lengths, 0.9)
```

```
## 90%
##   2
```

Now we perform Louvain clustering, and plot the result in Figure 2. This not very informative, since there are many nodes that are all quite tightly connected. In Figure 3 below we see the separation of butterflies into the four clusters generated by the louvain clustering. We tried multiple values for the resolutions parameter and found that a value of 1 or 1.1 achieved best clustering. Note that some jitter has been added to differentiate overlapping points. The data only contains integer representation of the species, which we must assume is to properly anonymize the butterflies identities. In Figure 3, we see clearly that some species were easier to differentiate than others, but no one species is completely separated from the others. For example, species 6 and 9 are both clustered together with few butterflies from the other species, while species 2, 3, 7 are clustered together, along with a not insignificant amount of butterflies from species 9 and 10.

```
# Perform louvain clustering and then plot the graph
louvain <- igraph::cluster_louvain(graph_1, resolution = 1.1)
plot(louvain, graph_1, vertex.size = 1, vertex.label = NA,
     layout = layout_with_fr(graph_1),
     main="Fig. 1: Clustered graph")
```



```
plot(y = jitter(louvain$membership), x=jitter(butterfly_names$V2),
     main="Fig. 2: Inferred vs true species.",
     xlab="True species",
     ylab="Inferred species")
```