732A51 Bioinformatics

Lab 5

Johannes Hedström, Mikael Montén

STIMA
Institutionen för datavetenskap
Linköpings universitet

2024-12-11

Contents

1 Question 1

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.*

```
data <- read.table("ChG-Miner_miner-chem-gene.tsv", header = FALSE)
colnames(data) <- c("drug","gene")
# create graph
graph <- graph_from_data_frame(data, directed = FALSE)

total_nodes <- vcount(graph)
drug_nodes <- as.numeric(length(unique(data$drug)))
gene_nodes <- as.numeric(length(unique(data$gene)))
edges <- ecount(graph)

# strongest connected components
scc <- largest_component(graph)
nodes_scc <- vcount(scc)
nodes_fraction <- nodes_scc/total_nodes # not identical to bioSNAP
edges_scc <- ecount(scc)
edges_fraction <- edges_scc/edges

graph_diameter <- diameter(graph, directed = FALSE) # not identical to bioSNAP
graph_dist <- distances(graph) # get all distances
graph_dist <- as.vector(graph_dist[graph_dist < Inf]) # remove unconnected nodes
effective_diameter <- quantile(graph_dist, probs = 0.9) # not identical to bioSNAP

summary_stats <- rbind(total_nodes, drug_nodes, gene_nodes, edges, nodes_scc,
                       nodes_fraction, edges_scc, edges_fraction,
                       graph_diameter, effective_diameter)
```

Table 1: Summary statistics for chosen dataset

	Dataset statistics
Nodes	7341
Drug nodes	5017
Gene nodes	2324
Edges	15138
Nodes in largest SCC	6621
Fraction of nodes in largest SCC	0.901921
Edges in largest SCC	14581
Fraction of edges in largest SCC	0.963205
Diameter (longest shortest path)	18.000
90-percentile effective diameter	8.000

2. *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.