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.*

```r
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

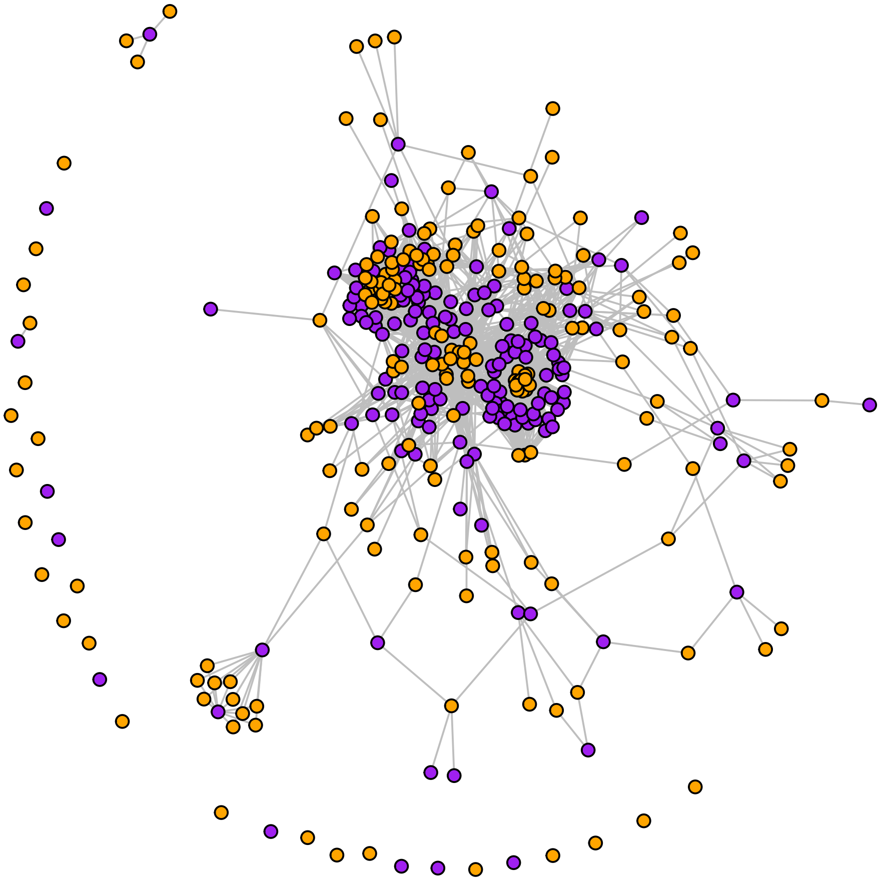
	Dataset statistics
Nodes	7341
Drug nodes	5017
Gene nodes	2324
Edges	15138
Nodes in largest SCC	6621
Fraction of nodes in largest SCC	0.901921
Edges in largest SCC	14581
Fraction of edges in largest SCC	0.963205
Diameter (longest shortest path)	18.000
90-percentile effective diameter	8.000

There are differences in the fraction of nodes in largest SCC, the longest shortest path and the 90-percentile effective diameter between the BioSNAP summary and the summary produced here. The differences are small, and the concrete statistics such as nodes and edges are the same, meaning the constructed graph is very similar but calculated differently. Most likely, the differences comes from choices of computation, such as the diameter being calculated with different algorithms which might handle edge cases in different ways. Also, the graph created here has been created as undirected, but perhaps extra information exists that could make it directed. The fraction of nodes in the largest SCC differ, while the fraction of edges in the largest SCC are identical, which means there is some difference in how the fraction is calculated in the BioSNAP database.

Plotting the full graph will not yield a meaningful visualization as there are too many nodes and edges, therefore we sample the most contributing SCC to show the most important genes.

```r
# find amount of edges per node
node_degrees <- degree(scc)

# sort descendingly and find names of the best, cutoff point is at 14 edges
best_nodes <- names(node_degrees[node_degrees > 14])

# create the subsampled graph
sampled_graph <- induced_subgraph(scc,best_nodes)

# set node types
V(sampled_graph)$type <- ifelse(V(sampled_graph)$name %in% data[,1], "Drug", "Gene")

# set colors for nodes
V(sampled_graph)$color <- ifelse(V(sampled_graph)$type == "Drug", "purple", "orange")

plot(
  sampled_graph,
  vertex.color = V(sampled_graph)$color,
  vertex.size = 3,
  vertex.label=NA,
  edge.color = "gray",
  main = "Network of drugs and genes"
)
```

Network of drugs and genes



The plotted graph above shows Drug nodes as purple and Gene nodes as orange. The cutoff for amount of edges per node required to be plotted was at 14 edges, which is the highest amount igraph could handle. This graph shows that there is a large cluster of drugs and genes that are closely connected, and a large half-circle of both type of nodes that are further away which doesn't have any visualized edges as the connected nodes are not plotted.

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.*