Tommaso Carraro
student number: 015147592

Big Data Management: Homework 3

1. The meaning of the requested layouts is provided below:
    a. Force Atlas: this layout is a classic force-directed approach that uses the principles of repulsion, attraction, and gravity to provide a high degree of accuracy for small to fairly large datasets. It is made to spatialize Small-World / Scale-free networks. It is focused on quality to allow a rigorous interpretation of the graph with the fewest biases possible, and a good readability even if it slow. With this layout it is possible to see how many clusters of nodes our network has. The main parameters of this layout that we can tweak are:
        i. repulsion: it defines how strongly does each node reject others. Higher levels force stronger rejection levels and tend to create a graph that has greater spacing between nodes;
        ii. attraction: it defines how strongly each pair of connected nodes attract each other. Higher levels of attraction strength will draw connected nodes together, leading to a network that is potentially more clustered, depending of course, on the underlying dataset;
        iii. gravity: it can be used to attract all nodes to the center to avoid dispersion of disconnected components. Lower settings can help the network to spread in cases where extreme crowding exists at the center of the network;
        iv. Adjust by Sizes: it is a check-box that tells the algorithm to avoid overlapping of nodes. This is highly useful when we have a network with large hubs that could easily land atop smaller nodes;
        v. Speed: it defines how fast the layout is applied to the network.
    b. Fruchterman-Reingold: this layout is a force-based algorithm that has different parameters respect to the previous one. It simulates the graph as a system of mass particles. The nodes are the mass particles and the edges are springs between the particles. The algorithms try to minimize the energy of this physical system. This method is very slow on big networks. With this layout it is possible to see which nodes have the biggest number of edges (tried on Gephi). The main parameters of this layout are:
        i. Graph size area: it is the parameter that substitute the repulsion and attraction settings of the previous layout. It acts as a surrogate for both, by spreading the network farther apart or by drawing it closer together;
        ii. The other parameters are Gravity and Speed that work like in the previos layout.
    c. Radial Axis: this algorithm positions nodes along radial axes using a predetermined number of radians. This method is not force-based, giving it a significant speed advantage. Instead, users specify how they wish to group nodes, how the nodes should be laid out, and several additional selections. It groups in axes radiating outwards from a central circle. Groups are generated using a metric or an attribute. It is possible to use this layout to study

homophily by showing distributions of nodes inside groups with their links. The main parameters of this layout are:

i. Scaling Width: it adjusts the size of the entire network;
ii. Resize Nodes and Node Size: these options allow to resize all nodes to a common value;
iii. The other parameters are used to sort and group the nodes.

d. Yifan Hu: this layout is a force-based algorithm that is designed to run more quickly than many of the other force-based algorithms while still providing a reasonably accurate result. It is a very fast algorithm with a good quality on large graphs and it focuses on attraction and repulsion at the neighborhood level. It combines a force-directed model with a graph coarsening technique to reduce the complexity. The repulsive forces on one node from a cluster of distant nodes are approximated by a Barnes-Hut calculation, which treats them as one super-node. It doesn't need to be stopped because it stops automatically. With this layout it is possible to see how many clusters of nodes our network has. I tried this on Gephi and the result was similar to the result of Force Atlas but it took less time for the computation. This method has the following important parameters:
    i. step ratio: it defines the ratio used to update the step size. It is possible to increase it reaching better quality but a slower computation. This parameter measures the relationship between repulsion and attraction levels;
    ii. optimal distance: if defines the natural length of the springs. It is possible to increase it to place nodes farther apart.
e. ARF: this is a force-based algorithm which allows users to adjust settings to better optimize the network display. I didn't try it on Gephi because it isn't included in the basic installation and it isn't possible to find it online. From other sources I've found that it provides a useful layout tool that affords considerable flexibility through attraction and repulsion settings. ARF outputs tend toward a more circular appearance than many of the other spring-based algorithms. The main parameters of this layout are:
    i. Neighbor attraction force: it is used to pull neighboring nodes closer together, or conversely, further apart;
    ii. General attraction force: it is applied to all nodes in a network, without regard for their neighbor status. This can be used to spread the network out or draw it closer together, independent of how the neighboring nodes relate to one another;
    iii. Repulsive force: it is another means to spread the network apart;
    iv. Precision: it is possible to make the graph as accurate as possible adjusting this setting to a higher level. High levels of precision will result in a longer running time and harder work for the layout algorithm, as it will try to maximize accuracy, and at some point, of maximum necessary precision there will be minimal benefit derived.

2. All correct except of the D
3. All correct except of the C

4. The requested queries are:
    a. SELECT Firstname, Lastname FROM Student
    b. SELECT * FROM Student WHERE Major <> "CS"
    c. SELECT * FROM Student WHERE Firstname LIKE 'J%' AND Lastname LIKE 'S%'
    d. UPDATE Student SET Major = "CS" WHERE StudentId = 101

5. The requested operations are present below in requested order:

A	B	R.C	S.C	D	E
a1	b1	c1	c1	d1	e1
a1	b1	c1	c2	d2	e2
a3	b3	c3	c1	d1	e1
a3	b3	c3	c2	d2	e2

A	B	R.C	S.C	D	E
a1	b1	c1	c1	d1	e1

A	B	C
a1	b1	c1

A	B	R.C	S.C	D	E
a1	b1	c1	c1	d1	e1
a3	b3	c3	null	null	null

A	B	R.C	S.C	D	E
a1	b1	c1	c1	d1	e1
a3	b3	c3	null	null	null
null	null	null	c2	d2	e2