**Social Network Analysis**

This assignment is about Amazon’s co-purchase network; i.e., which products are purchased together on Amazon. There are two data files for this assignment: “*products.csv*” and “*copurchase.csv*”, both of which are provided as attachments on Canvas. “*products.csv*” contains the information about the products. “*copurchase.csv*” contains the co-purchase information (i.e., if people bought one product, then they also co-purchased the other). This is a *directed* graph, as explained below.

The file “*products.csv*” has the following variables:

* **Id:** Product id (number 1, 2, ..., 262109)
* **title:** Name/title of the product
* **group:** Product group (Book, DVD, Video or Music etc.), in this assignment, we only work with the books group.
* **salesrank:** Amazon [Salesrank](https://www.amazon.com/gp/help/customer/display.html?nodeId=525376) (click the link to see the description of Salesrank)
* **rating:** User rating on a scale of 1 to 5
* **total number of reviews:** Total number of product reviews available on Amazon

The file “*copurchase.csv*” has two variables, which are Product ID’s:

* **Source:** This is the focal product that was purchased
* **Target:** People who bought “Source” also purchased the Target product

Note that the edge between Source and Target is directed: Source 🡪 Target.

Please complete the following steps, based on the two data sets provided:

1. Delete products that are not books from “*products*” and “*copurchase*” files.

Note: In social network analysis, it important to define the boundary of your work; in other words, the boundary of the network.

With respect to the Social Network Team Assignment, please add the following sentence in part 1:

And then delete the books with salesrank>150,000 or salesrank = -1.

1. Create a variable named *in-degree*, to show how many “Source” products people who buy “Target” products buy; i.e. how many edges are *to* the focal product in “co-purchase” network.
2. Create a variable named *out-degree*, to show how many “Target” products people who buy “Source” product also buy; i.e., how many edges are *from* the focal product in “co-purchase” network.
3. Pick up one of the products (in case there are multiple) with highest *degree* (*in-degree* + *out-degree*), and find its *subcomponent*, i.e., all the products that are connected to this focal product. From this point on, you will work only on this subcomponent.
4. Visualize the subcomponent using iGraph, trying out different colors, node and edge sizes and layouts, so that the result is most appealing. Find the diameter, and color the nodes along the diameter. Provide your insights from the visualizations.
5. Compute various statistics about this network (i.e., subcomponent), including degree distribution, density, and centrality (degree centrality, closeness centrality and between centrality), hub/authority scores, etc. Interpret your results.
6. Create a group of variables containing the information of neighbors that “point to” focal products. The variables include:
   1. Neighbors’ mean rating (nghb\_mn\_rating),
   2. Neighbors’ mean salesrank (nghb\_mn\_salesrank),
   3. Neighbors’ mean number of reviews (nghb\_mn\_review\_cnt),

Note: you may recall the functions in “dplyr” such as group\_by, inner\_join, summarize, mean, etc.

1. Include the variables (taking logs where necessary) created in Parts 2-6 above into the “products” information and fit a Poisson regression to predict *salesrank* of all the books in this subcomponent using products’ own information and their neighbor’s information. Provide an interpretation of your results.

Note: Lower *salesrank* means higher sales. Data points in the network are related. The performance of one node is influenced by the performance of its neighbors. Also, it’s not necessary that all variables matter.