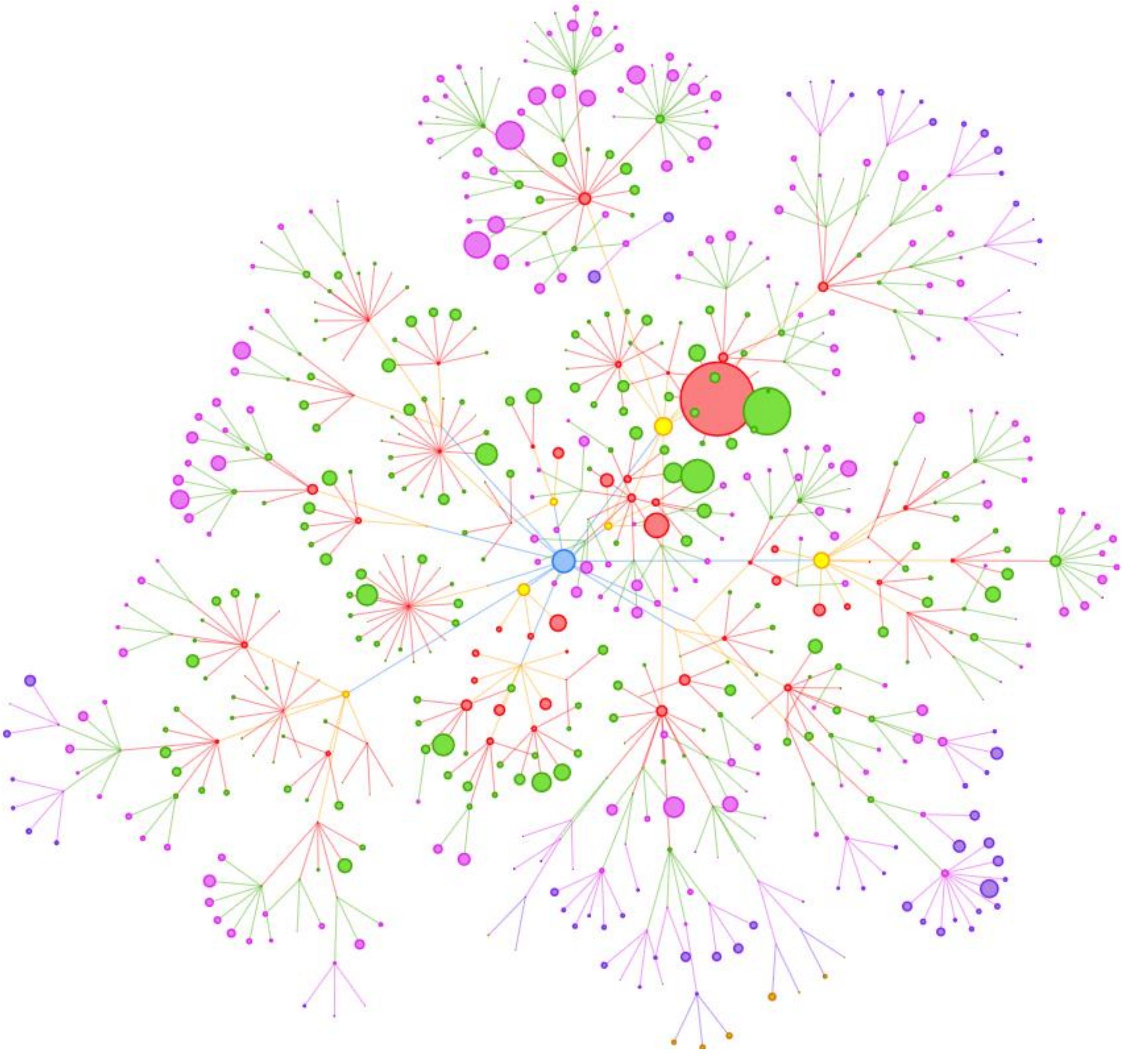
DC3: The Tree of Stuff

Jiun-Ting Chen

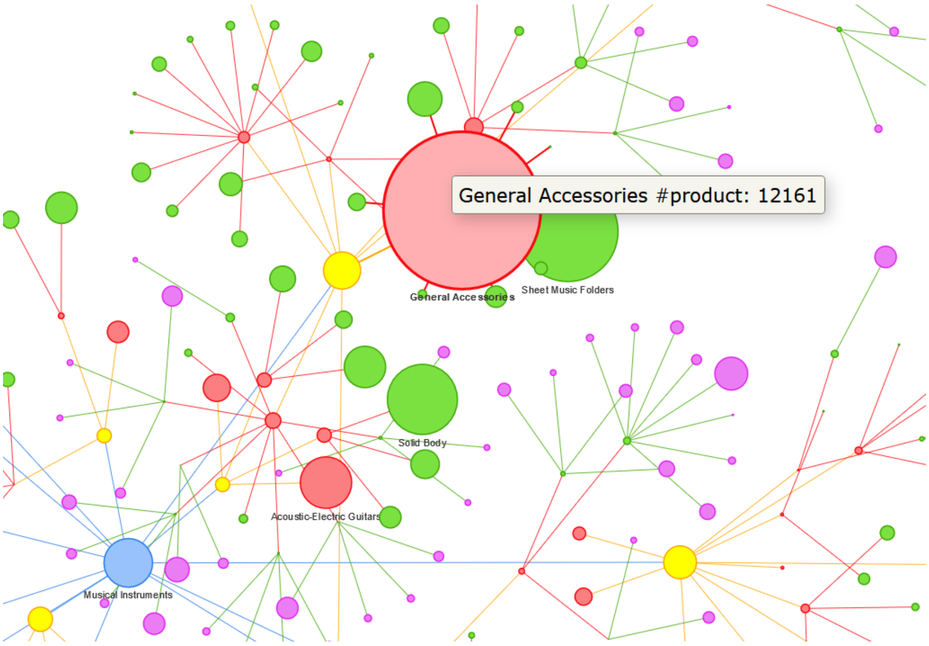
Designs and Intents



The visualization shows the distribution of product amount for each subcategory, where the node size stands for their product amount (products of their subcategories are not included), while the colors stand for tree levels.

The main goal of the task is to understand if a main category is skewed, where the products concentrate in certain subgroups; Or if some categories have an extreme long path. Once we find out such case with the graph, we could evaluate the existing category design in a clear fashion and gain more insignt to improve.

The above graph is an example from *Musical Instruments*, a main category of the dataset. With this graph, we see that the product amount of each subcategory resides in a limited range. Besides, it seems that all paths end in around 5 levels, which is quite balance and thus a good property. However, we could also see that the second level nodes (the yellow ones) tend to have fewer products. And there exist two subcategories with relatively large scale (the green and the red node on the upper-right quadrant). To improve the category design, we may want to know more about the large subcategories and their product amount.



My visualization design supports the following three main interactive functions, which aims at handling such requirement.

1. **Zoom in**: User could scroll with mouse to enlarge/shrink the graph.
2. **Express label for important nodes**: The name of the important nodes would also appear when a user zooms in to some certain extent. In order to let users to get the crucial node at a glance, the graph only shows the name of the top 1% nodes (when the number of its product is more than 99% of the nodes in that main subcategory.)
3. **Detail information pop-out**: Some information would pop-out once a user hovers over a node. My current design shows the name of the subcategory and its product amount.

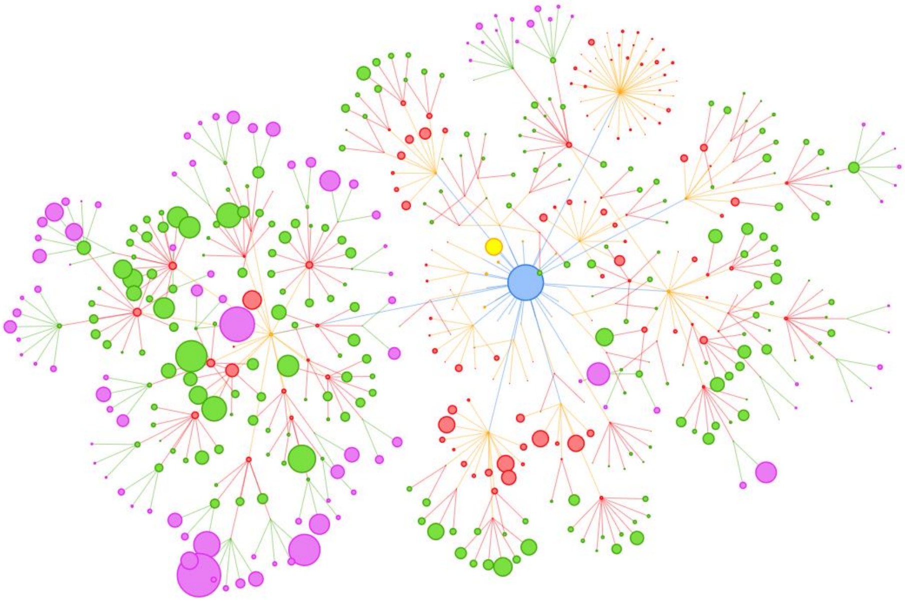
In this example, we see that one of the large nodes is *general accessories*. Hence, we may want to divide it to enable user to find out certain product easier.

Discussion of Data and Findings

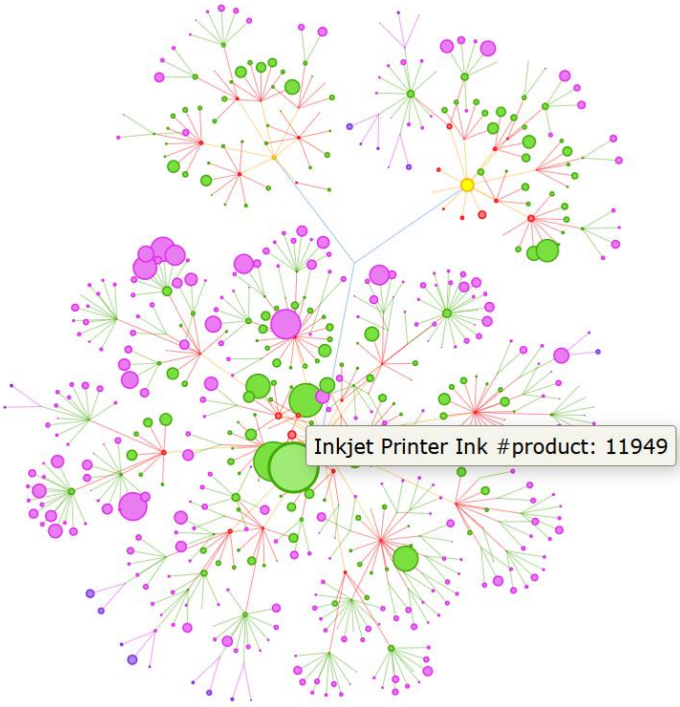
In addition to *Musical Instruments,* I also explored other main categories and found that their distributions vary. The following shows three different scales of categories and some findings of the result.

1. Small:

Pet Supplies: 110707

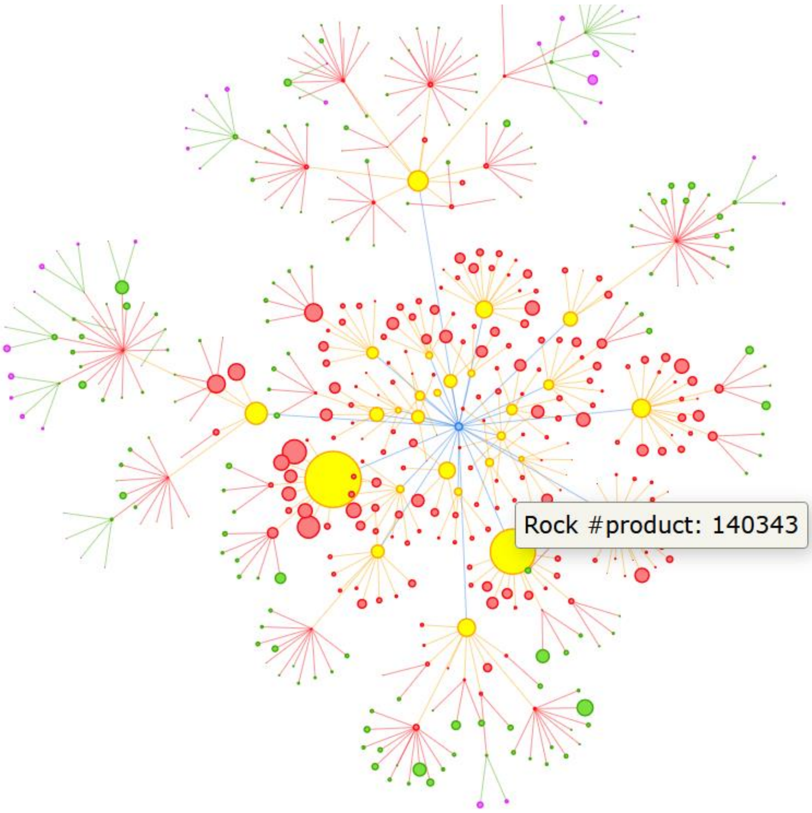


Office Products: 134838

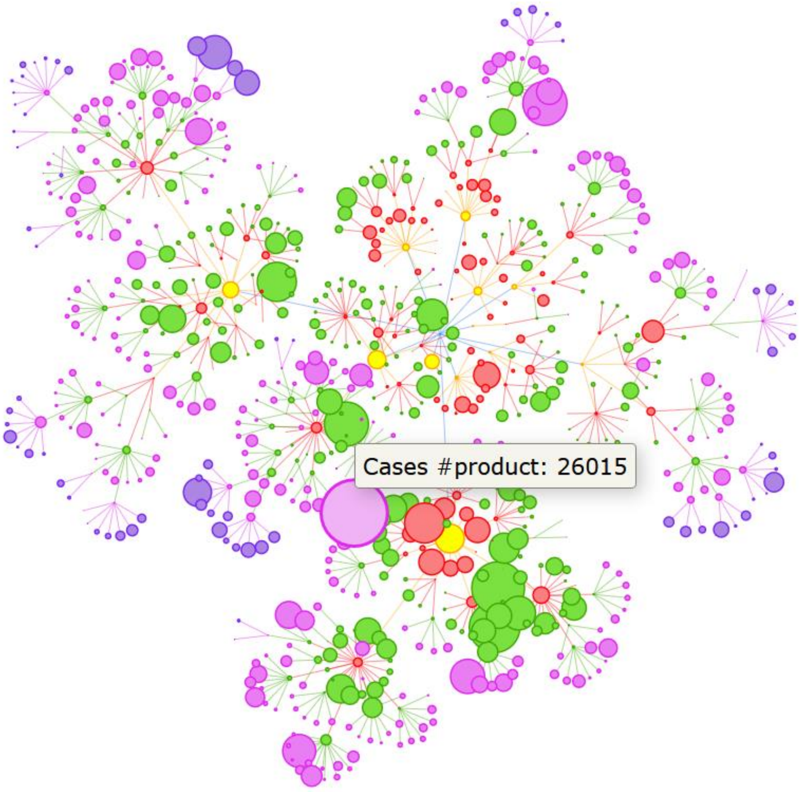


2. Medium:

CD&Vinyl-492799

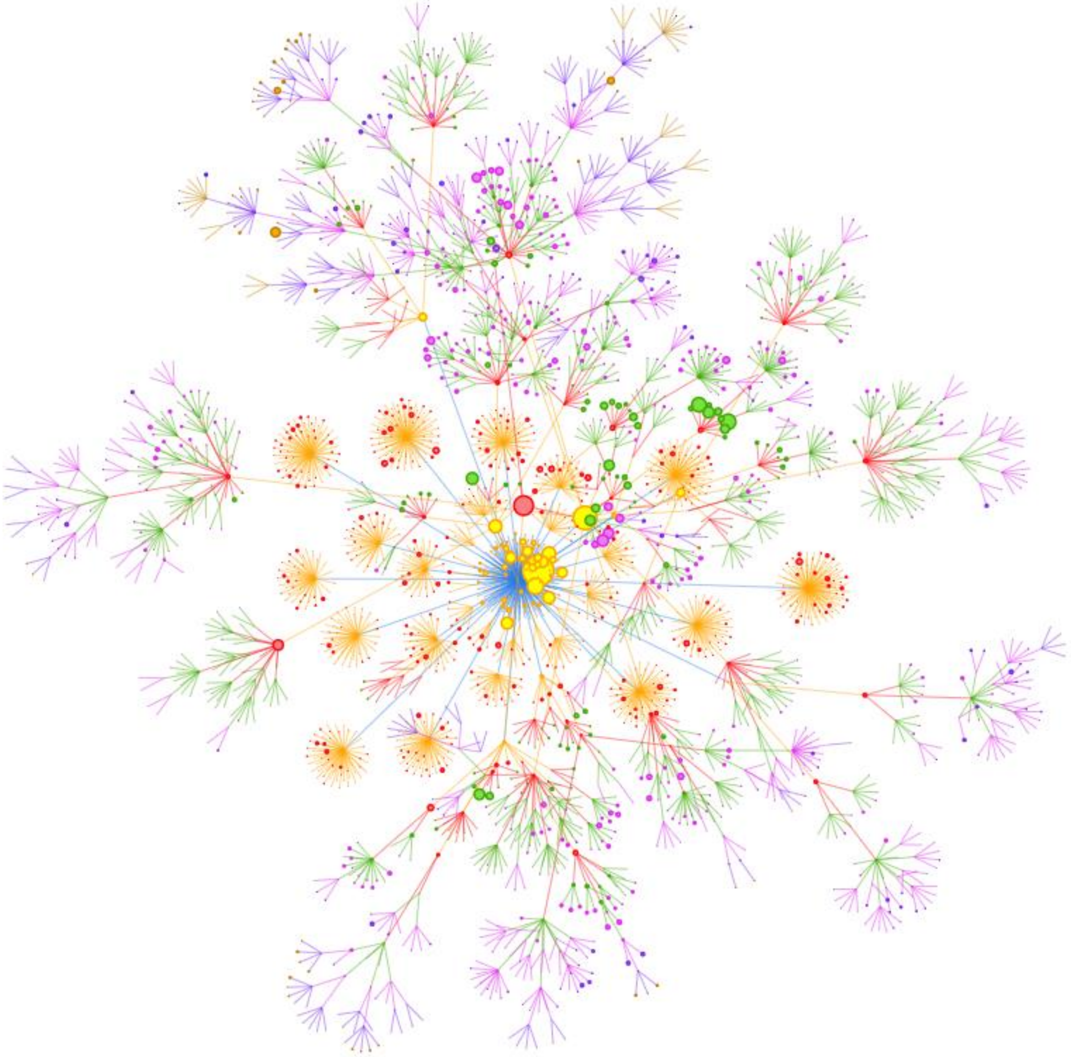


Electronics: 498196

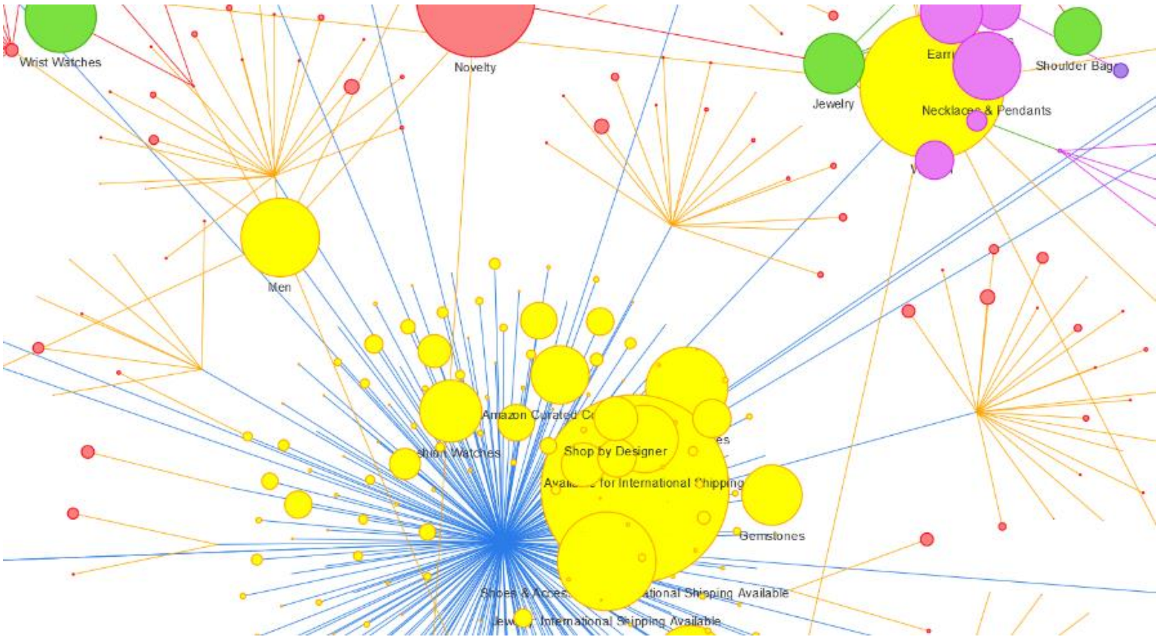


3. Large:

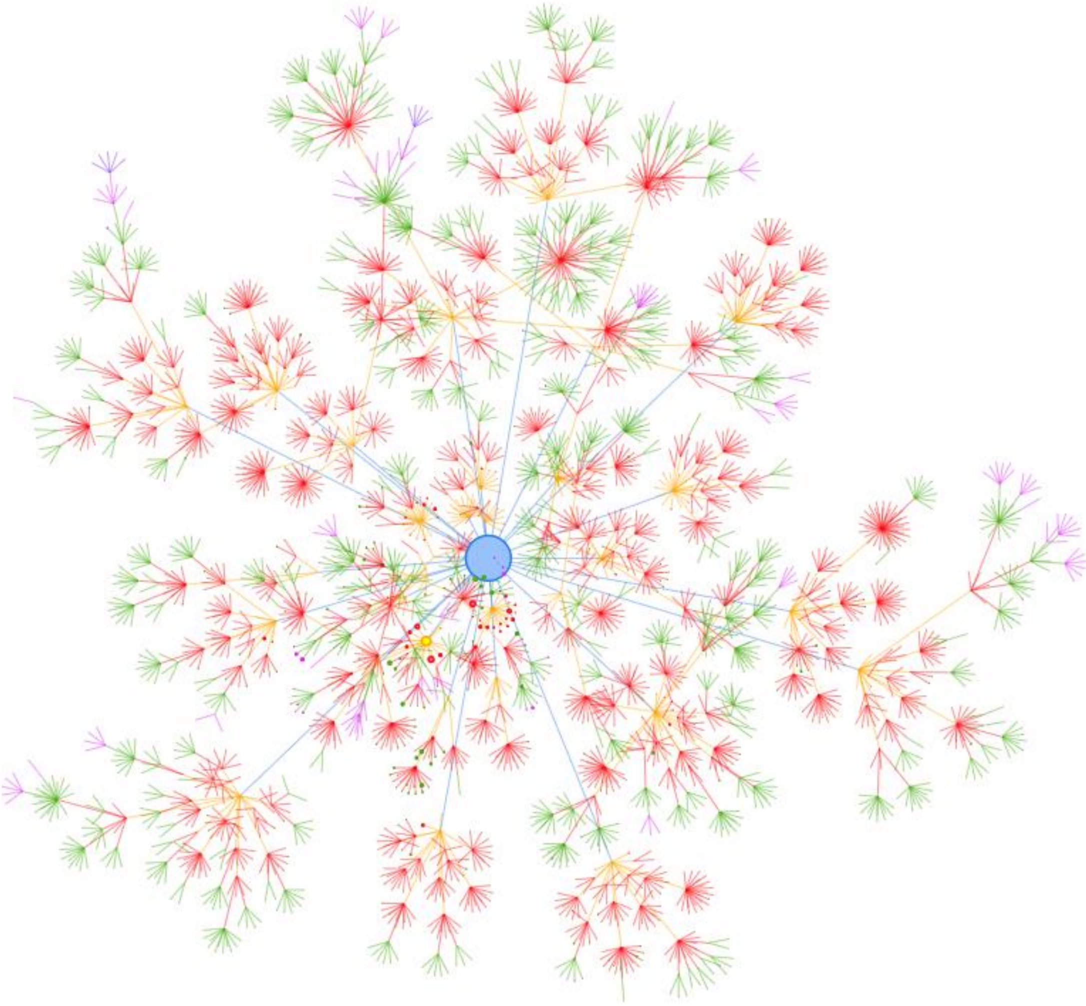
Clothing, Shoes & Jewelry-1503384

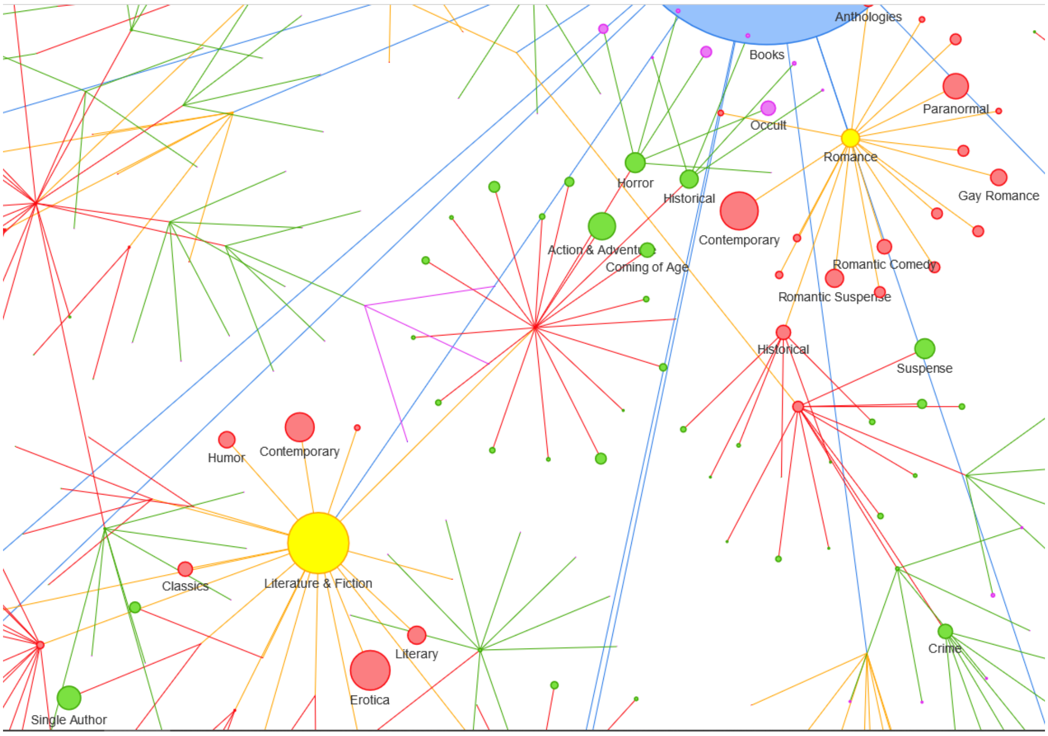


,



Book- 2370585





For *Pet supplies*, the size of each node is in a certain range as well. But we can find out some nodes on the lefthand side is a bit larger than other. It may not be a big issue but the database designer could also check this nodes to see if it's necessary to divide them. Regarding *Office Products*, some paths end with a large green node, which is one level upper than the purple ones. As the graph shows, one of them is Inkjet Printer Ink. I think dividing such nodes by their size or other specification to would be a workable way to make the users easier to find out their target.

As for the medium size of categories, there are more nodes in the graphs but they are still clean enough to show us some valuable information. In *CD & Vinyl*, some path has huge second level (yellow) or third level (red) nodes. Since the products are categorized by the music type, and thus people tend to search with these keywords, dividing them with other rare terms may not be quite useful. And for the Electronics, the graph is still neat even though more nodes appeared. Besides, we could also point out some possible nodes to improve.

The processing time for the large dataset becomes much longer, which takes about 3 minutes to generate a graph. Moreover, as the size of some nodes gets larger, we need to rescale the graph to keep it readable. Fortunately, the processed graphs are also good enough to help us understand their subcategories. Both two graphs show a quite symmetric snowflake-like shape, which means that the number of branches is well-designed and no path is with extremely long path. Furthermore, there some large second level nodes in the clothing and the root of Books is huge, which may be resulted from the fact that many products are also labelled as these categories. It's quite interesting that only these large categories have such property.

Self-Assessment

This assignment is implemented by python and mainly based on the tree structure provided by the professor and two graph packages (networkx and pyvis).

I did not have experience in drawing decent graphs so it took me much time to finish this assignment. First, I used the provided csv files and wrote some BFS code to build graphs. However, I realized that some python scripts were provided. Hence, I read the tree structure inside and implemented the tree-traversal algorithm on it. For graph construction, I tried networkx from the beginning and found it quite useful. And for the visualization, I planned to represent a static version with matplotlib since I was not pretty sure how to build an interactive one. Finally, after trying many packages, I found pyvis meet my requirement, which outputs a decent interactive graph given some large dataset.

I think my work still has room to improve but I cannot make it due to the limited time in the late semester. In short, the work I have done so far merely shows an idea on building a graph with large dataset. I expected to construct a website, which offers an interface for user to choose any main category. Besides, it would be great if I have time to modify pyvis so that some more statistic information could pop-out after the user click a certain node.