COMP 767 - Assignment 1

Yi Tian Xu

September 15, 2016

1 Vintage Luxury Trends

One dataset that can interesting and challenging to look at comes from the company that I worked with in the past year. The company, LXR&CO, is a vintage luxury item company with its main focus on second-hand high end bags. Currently, they have over 20 retail stores around the world and an e-commercial website open for North America customers. As the company plan to expand in size, a pertinent aspect to consider is the sale trend of their products at different location. Such analysis can potentially help them to understand the behaviour of buyers, the performance of physical stores and online platform across time, and help them in distributing their products to the best location.

1.1 Data Source and Network Construction

The data gathered is from the company's database, containing all the sold and return items since the beginning of the year 2016, totally to more than 48,000 entries. Each item has an associated store name or shipping address down to the city level, a price and discount values in a particular currency, the brand and the category name. If the item is associated with a physical store, information about whether this store's item can be accessible on the e-commerce website is also retrieved. The list of items is queried using MySQL and exported as a CSV file.

The network construction will depend on the exact problem that we want to study. For example, if we want to know how to efficiently distribute the inventory, we can possibly construct bipartite graphs, linking available items to corresponding locations by edges and find the graph that maximize a

"profit score". We can define the weight of an item vertex to be a combination of the item's attributes (price, discount, brand, category, refunded or sold, etc.) and infer the optimal weight space for each location based on past data to design the "profit score".

2 Phylogenetic Tree

Motived by my interest in evolution theory, I choose the phylogenetic tree for my second network. Defining species as vertices and evolutionary relationships as edges, we can obtain a directed graph. We call a species u a parent of species v if species v is evolved from species u and no other species is in any path between u and v. This relationship can be marked by a directed edge from child to parent. If there are vertices in any path between u and v, we call v an ancestor of u. We also call a path between two vertices a evolutionary path between two species. Some potentially interesting exercises that can be done with this network are visualization and clustering (e.g.: classifying species into families).

2.1 Data Source and Gathering Method

The data I gathered is from the Ensembl Rest API (https://rest.ensembl.org). Their taxonomic classification endpoint returns a list of species in the evolutionary path between the queried species and an ancestor. Each returned species contains the parent and a list of children species.

The method that I used is a script that recursively queries for the children and parent until all the species are collected. The scripts saves a list of edges (child to parent) in a CSV file. As each species in Ensembl Rest API has one parent, the number of vertices should be equal to the number of edges.