# CIS 509 – Mining Assignment 3 - Strategy

When a user purchases a product, there could be three types of recommendations of a new product.

1. **Collectibles**: These are products which are frequently co-purchased. These products may not be similar, but have a very strong dependency or relationship.
2. **Similar items**: These are products which fall in the same category or have a high similarity in terms of the categories and are co-purchased.
3. **Popular items:** These are products which were co-purchased by the product under consideration.

The strategy would be collect these items with the order of priorities as: Collectibles, Similar items and most popular items. The intention is users purchase items in this order. For example, if a user had already purchased similar items and collectibles, then it would be meaningful to recommend other popular items. Different metrics need to be considered to **rank** the nodes. In the calculation, **Lower** the rank of a node, **higher** the recommendation of the product associated with the node.

**Collectibles**: To find collectibles or most often purchased items along with the product under consideration, **cliques** are used. A **clique** is a complete subgraph. Set of cliques which contain the product under consideration is computed. Therefore, these cliques capture products which were co-purchased at some point of time. A limitation with the available data is that there is no information about how often two products are purchased **together**. So, to find the most often co-purchased items, **SalesRank** metric is used. For each clique, let **A, average SalesRank** for the clique is measured. As nodes, can be part of many cliques, **Rank** of a node is calculated as using **F,** frequency of node appearance in a clique, **S,** size of the clique **as A/(S\*F)**. Thus, lower Sales Rank, higher frequency and higher size of clique it appears improves the **Rank.**

**Similar items:** To find similar items, **similarity edge weight**, is used. In the first step, **Islands,** are used to narrow down on the nodes. The threshold is calculated as the **median** value of the ego network of product under consideration. The reason for **median** is because of the **skewness** in the weights. Once the islands are formed. Let **S** be the **SalesRank** of the node and **E** be the **edge weight,** then **rank** is **S/E.** Thus, lower the edge weight, lesser the similarity and higher the **Rank.**

**Popular items:** To find popular items, **clustering coefficient,** is used. First, the ego network is used to select the neighbors of the product under consideration. Then **clustering coefficient (coeff)** is computed by **removing** the product under consideration from the graph. This step will ensure that the product is popular **irrespective** of its association to the product under consideration. Then **Rank** is calculated as **S/coeff** where **S** is **SalesRank.** Thus, if clustering coefficient is high, the Rank gets improved.

**Three lists** are formed and sorted by their **Rank.** These lists are merged and the final list is generated in the order: **Collectibles, Similar items and Popular items.**