A Multistage Unsupervised Framework for Determining Product Substitutes

Bodhisattwa Prasad Majumder
Walmart Labs

Arunita Das
Walmart Labs

Amlan Das
Walmart Labs

Subhasish Misra
Walmart Labs

bodhisattwapm2017@email.iimcal.ac.in

***Abstract*— For a retailer, it is imperative to identify the substitution relationship between products since this aids in efficient assortment decision. Quirks of consumer behaviors result in two distinct substitution behaviors – traditional and variety substitution. We propose a multistage unsupervised framework to determine three mutually exclusive and exhaustive product pairs sets – traditional substitutes, variety substitutes and non-substitutes. We combine both behavior-based and content-based product attribution along with demographic and price information to exhaustively represent a product pair. Furthermore, the unsupervised nature of the framework makes even more acceptable when it comes to translating the result into other categories, even in the absence of any tagged data. We finally rank the substitutes given a product via an affinity based scoring mechanism. The ranked retrieval of substitutes not only improves the results but also enriches the product understandings of categories with no tagged data. Our final findings of categorization of product pairs achieve a high sensitivity and specificity irrespective of a very skewed minority population. Substitution being a rare phenomenon is rightly captured by our framework compared to supervised or heuristic based baselines.**

Keywords—substitutes, attributes, transaction metrics, clustering, affinity scores

I. Introduction

Shelf space in a brick and mortar store is constrained; necessitating selection of the most optimal products only. One way to do it is to delist products with low sales. But, if such a product has no substitute then we may lose the customers who come exclusively for it. This happened to a US retailer in 2009, where crude methods in product deletion resulted in an estimated loss of $2 Billion.

It is imperative then, to identify *substitutes* (and *non-substitute*) product pairs for a retailer. Within substitutes, consumer behavior quirks result in two classes of substitutes: ***traditional substitutes*** which satisfy the same consumers need state (diet Pepsi v/s diet Coke in the carbonated soft drink category) and ***variety substitutes*** which satisfy variety seeking tendencies. For e.g. buying a Pepsi (cola) & also, a Fanta (orangeade) in the same transaction. Fig. 1. presents an example of both type of substitutes in the category, pizza. Where '*DIGIRNO Original Thin Crust 7 Cheese*' and '*RED BARON Thin & Crispy Five Cheese*' are perceived as traditional substitutes in the sense that one can be replaced by other, '*RED BARON Thin & Crispy Five Cheese*' and '*RED BARON Thin & Crispy Pepperoni*' together serve to the consumers' variety seeking behavior.

Differentiating between traditional and variety is important. Variety substitutes aid in basket building behavior and hence bring in more sales. Ideally then, a product can be deleted only if it has poor sales performance, has traditional substitutes and is not a variety substitute to many products. On the other hand, we would be cautious of removing a product with poor performance, few/no traditional substitute, but which is a strong variety substitute (to some or many products).

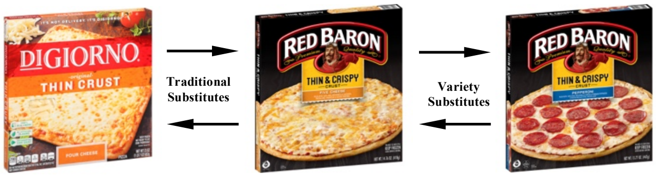

Fig. 1. A representative example of two classes of substitutes.

Our goal was to help optimize assortment by classifying any given product pair (for a given category) into any of the three mutually exclusive and exhaustive classes – non-substitutes, traditional substitute or variety substitute. While, this in the surface may appear to be multiclass classification problem, there are obstacles going that route. Multiclass classification ostensibly requires tagged data which may be difficult to get to since it requires intensive inputs on the nature of the substitute from a business subject matter expert. This means that doing this across different categories will be an issue. Another option could be to use managerial heuristics – rough rule of thumb approaches to get to classifying between different substitute types. These approaches, however, usually yield sub-par accuracies. In the finality, we will show how we circumvented these limitations and came up with an unsupervised and scalable way to identify both kinds of substitutes. Here, we list the major contributions of the paper:

a) We combine behaviour-based features and content-based features to represent a product pair. The combined representation also includes demographic and price information of the respective products which makes the representation more exhaustive to capture the regularities of the various kinds of substitution phenomena. The generic nature of these features makes the approach perfectly scalable.
b) We propose a completely unsupervised framework with multistage clustering to filter out substitutes, and then finally separate out both the classes of substitutes. To the authors' best knowledge, it is the first attempt to a scalable unsupervised approach for solving the fuzzy problem of identifying traditional and variety substitutes on large-scale data.
c) We also propose an affinity score based ranking mechanism to list out the most influential traditional and variety substitutes given a product.

Our approach perfectly scales in an unsupervised way across categories without the availability of tagged data which make the approach unique, robust and efficient.

The paper is further organized as follows: Section II captures the previous attempts to solve this problem. Section III outlines the methodology along with feature creation and the unsupervised framework. Section IV presents the details of the dataset used and the comparison of proposed method with various baselines. Finally, section V concludes the article with the mentions of future work.

II. Previous Works

In a typical micro-economic setup, the phenomenon of substitution is inbuilt in the utility maximization by each consumer. Also, the current literature on substitution address the problem from choice theoretic perspective [1]. However, none of these theories are self-sufficient to explain the phenomena of decision making over a wide choice set. These theories complement the understanding of the choice making, but in practical scenarios, where the choice universe is large enough to perform surveys, typical choice theoretic experiments fail. Besides, these experiments are too costly and difficult to perform in a practical scenario. There are few efforts made to address the problem using a content based approach. In one of these, the problem was looked at as a product network connected with edges representing association between the products learned from the buying pattern [2]. In this work, the major focus was unveiling the inter-category joint purchase behavior, the concept of intra-category substitution behavior didn't get much attention in this piece. In another effort, the authors employed a variant of PageRank algorithm to find valuable products that reside in the network. Even if the network shows the competitiveness among products it fails to give right direction to the phenomena of substitution [3].

In some of the works [4], cross-price elasticity has been proposed as an identifier for substitution behavior. For a product pair, Cross price elasticity is the percentage change in demand for one product divided by the percentage change in price of the second product (given that all other factor affecting demand remains the same) [5]. A positive cross-price elasticity indicates a product pair to be substitutes whereas a negative cross-price elasticity indicates complementarity. Cross-price elasticity is very infrequently used to identify substitution behaviour in retail because of multiple reasons. Day, shocker and Srivastava [6] argue that the measure is static and doesn't incorporate price matching behaviour by other firms in lieu of price change by one firm.

Lattin et. al. [7] describes a variety seeking model of identifying the asymmetric relationship between product pair in terms of substitutability and complementarity but the proposed method has been modelled in an individual household level. In the case of retail chain having millions of customers, insights at the individual level does not help the store authorities to choose a right mix in the modular for a particular category. Also, non-traceable customers do not add any value in the study at individual level. McAuley et. al. presents *SCEPTRE* – a recommendation engine [8] which identifies the substitutes and complements product given a query product. they attempted to understand the semantic meaning of relatedness and employed customer reviews, product description as content based features to predict links (of substitutability and complementarity) in a product graph.

III. Problem Formulation

In this section, we formally define the problem of identifying substitute product pairs. The section defines the vocabulary for the paper and builds the methodology on it.

A. Vocabulary

Table I contains the problem specific symbols along with their definitions those are used in this paper.

TABLE I. Vocabulary

Symbols	**Definitions**
$\mathfrak{p}$	A *product* – generally sold at physical stores
$\mathbb{P}$	*Product Universe* – the collection of all products $\mathfrak{p}$
$\mathfrak{c}$	A *category* – a discrete collection of similar products $\mathfrak{p}$
$\mathbb{C}$	*Category Universe* – collection of all categories $\mathfrak{c}$
$\mathbb{T}$	A set of product pairs known as *Traditional Substitute*
$\mathbb{V}$	A set of product pairs known as *Variety Substitute*
$\mathbb{S}$	A set of product pairs known as *Substitutes*
$\mathbb{NS}$	A set of product pairs known as *Non-Substitutes*
$\mathfrak{h}$	A *household* – the smallest unit of the consumers
$\mathfrak{b}$	A *basket* – a collection of products from various category purchased by a single household $\mathfrak{h}$
x	A *feature* – a quantitative form of the factors those define the interrelationship in a product pair
$\mathfrak{B}$	*Brand* Universe – the collection of all possible product brands from a category
$\mathfrak{D}$	*Description* Universe – the collection of product descriptions from a category
$\mathfrak{P}$	Set of all possible *pack-size* buckets
$\mathfrak{R}$	Set of all possible *price* buckets

B. Problem Formulation

We define a product as $\mathfrak{p}$ from a product universe $\mathbb{P}$, which belong to a category $\mathfrak{c}$ from a category universe $\mathbb{C}$. In this paper, we aim to study the interrelationship of a product pair $(\mathfrak{p}_i, \mathfrak{p}_j)_{i \neq j} \in \mathbb{P} \times \mathbb{P}$ when both $\mathfrak{p}_i$ and $\mathfrak{p}_j$ come from same category, $\mathfrak{c}$. It is understood that a product pair may or may not share a competitive relationship among them. Thus, we define three mutually exclusive and exhaustive sets of product pairs which are defined based on various kind of competitive relationship of a product pair.

Definition 1. Traditional Substitute: A set $\mathbb{T}$ where $\mathbb{T} = \{(\mathfrak{p}_i, \mathfrak{p}_j)_{i \neq j} \in \mathbb{P} \times \mathbb{P}$ of category $\mathfrak{c} \in \mathbb{C} : \mathfrak{p}_i$ is bought in place of $\mathfrak{p}_j$, or vice-versa$\}$.

Definition 2. Variety Substitute: A set $\mathbb{V}$ where $\mathbb{V} = \{(\mathfrak{p}_i, \mathfrak{p}_j)_{i \neq j} \in \mathbb{P} \times \mathbb{P}$ of category $\mathfrak{c} \in \mathbb{C} :$ $\mathfrak{p}_i$ and $\mathfrak{p}_j$ are bought into same basket due to consumers' variety seeking behavior [2]$\}$.

Definition 3. Non-Substitute: A set $\mathbb{NS}$ where $\mathbb{NS} = \{(\mathfrak{p}_i, \mathfrak{p}_j)_{i \neq j} \in \mathbb{P} \times \mathbb{P}$ of category $\mathfrak{c} \in \mathbb{C} :$ $\mathfrak{p}_i$ and $\mathfrak{p}_j$ do not share any competitive relationship between them. In other words, product pairs those do not belong to $\mathbb{T}$ or $\mathbb{V}$, fall into $\mathbb{NS}$}.

It is also important to mention that $\mathbb{T} \cup \mathbb{V} = \mathbb{S}$, where $\mathbb{S}$ is a set called Substitute containing product pairs those are some kind of substitutes. We further define a product pair $(\mathfrak{p}_i, \mathfrak{p}_j)$ (dropping the notation $i \neq j$ for simplicity) as a set of different factors which in turn will describe their interrelationship. We define:

$$(\mathfrak{p}_i, \mathfrak{p}_j) = \{\textit{Attribute, Demographic, Consumption, Price}\}$$

Attribute signifies the relatedness in the intrinsic information of the products. For example, are the *brands* of the products in pair are same or not. The most obvious and direct way that consumers asociate a product is via product attributes.

Demographic signifes the comparison of the consumer population who usually (formally defined in *Feature Creation*) consume the product. We hypothesise that difference in population's demographic affects the purchase behavior.

Consumption denotes the consumption pattern of the products in a pair, in an aggregate level. How two products are associated based on their consumption pattern defines their competitive relatipnship.

Price is an important factor based on which the consumer perception varies from product to product. Given a pair, this factor captures the (dis)similarity in prices for both the products which tells how related are the products are.

These factors are quantitatively defined via popular metrics of defining them which finally constitute a d-dimensional vector derived for each product pair. This can be restated as

$$\overrightarrow{(\mathfrak{p}_\iota, \mathfrak{p}_\jmath)} = [x_1, x_2, \dots, x_i, \dots, x_d]$$

where x_i is a specific quantitative representation of any of the above factors. Finally, this vector for each product pair has been modelled to estimate the relationship between the products. Hence, if we seek to identify a model $\mathcal{M}$, which takes the input as a d-dimensional vector and finally generates an assignment $\mathfrak{a}_k$ where $\mathfrak{a}_k$ belongs to any of the symbols of {$\mathbb{T}$, $\mathbb{V}$, $\mathbb{NS}$}. This is to clarify that the model outputs the resulting set where the product pair should belong. There is no fuzzy assignment involved, i.e. one product pair can belong to only one set of {$\mathbb{T}$, $\mathbb{V}$, $\mathbb{NS}$}. Mathmatically,

$$\mathcal{M}\left(\overrightarrow{(\mathfrak{p}_\iota, \mathfrak{p}_\jmath)}\right) = \mathfrak{a}_k$$

$$\mathfrak{a}_k \in symbol\,\{\mathbb{T}, \mathbb{V}, \mathbb{NS}\}$$

C. Methodology

1) ***Feature Creation:*** We formally define all the features which in turn get used in multiple stages of clustering. Along with the definitions, we also describe why intuitively these features are important. Here, we capture both behavior-based and content-based characteristics of product pairs.

Behavior based features –

a. ***Proportion of households buying the product pair in different basket (%dfb):*** This feature captures the consumers' buying pattern to reflect the substitution behavior. Given any product pair $(\mathfrak{p}_i, \mathfrak{p}_j)$ we compute the proportion of households who bought both the products but not in the same basket-

$$\delta_{ij} = \frac{\left\{\sum \mathfrak{h} \middle|\ n\left(\mathfrak{b}_\mathfrak{h}(\mathfrak{p}_i, \mathfrak{p}_j)\right) = 0\,, n\left(\mathfrak{b}_\mathfrak{h}(\mathfrak{p}_i)\right) \geq 1, n\left(\mathfrak{b}_\mathfrak{h}(\mathfrak{p}_j)\right) \geq 1\right\}}{\left\{\sum \mathfrak{h} \middle|\ \left[n\left(\mathfrak{b}_\mathfrak{h}(\mathfrak{p}_i, \mathfrak{p}_j)\right) + n\left(\mathfrak{b}_\mathfrak{h}(\ \mathfrak{p}_j)\right) + n\left(\mathfrak{b}_\mathfrak{h}(\mathfrak{p}_i)\right)\right] \geq 1\right\}}$$

where $\mathfrak{b}_\mathfrak{h}(\mathfrak{p}_i, \mathfrak{p}_j)$ indicates a basket purchased by household $\mathfrak{h}$ which contained $\mathfrak{p}_i, \mathfrak{p}_j$.
Traditional substitutes are supposed to be purchased in lieu of the other, chances are households will buy them in different transactions - it could be non-availability of the product in the shelf leading to substitution. So, if there is a relatively large % of the population who have bought both the products but not in the same basket that can be due to the fact that the pair appears as a traditional substitute.

b. ***Proportion of household purchasing the pair in the same basket (%smb)****:* Given a product pair $(\mathfrak{p}_i, \mathfrak{p}_j)$ this feature computes the proportion of households $\mathfrak{h}$ who bought them together in the same basket at least once. Mathematically-

$$\beta_{ij} = \frac{\left\{\sum \mathfrak{h} \middle|\ n\left(\mathfrak{b}_\mathfrak{h}(\mathfrak{p}_i, \mathfrak{p}_j)\right) \geq 1\right\}}{\left\{\sum \mathfrak{h} \middle|\ \left[n\left(\mathfrak{b}_\mathfrak{h}(\mathfrak{p}_i, \mathfrak{p}_j)\right) + n\left(\mathfrak{b}_\mathfrak{h}(\ \mathfrak{p}_j)\right) + n\left(\mathfrak{b}_\mathfrak{h}(\mathfrak{p}_i)\right)\right] \geq 1\right\}}$$

where $\mathfrak{b}_\mathfrak{h}(\mathfrak{p}_i, \mathfrak{p}_j)$ indicates a basket purchased by household $\mathfrak{h}$ which contained $\mathfrak{p}_i, \mathfrak{p}_j$.

For any product pair $(\mathfrak{p}_i, \mathfrak{p}_j)$ if this percentage is high enough that might be indicative of variety seeking behavior and the product pair in consideration could be potential candidate for variety substitute. In contrast, if the proportion of such household is really small, it might be the indication that product pair serves the same need state and households substitute them in a traditional sense.

c. ***Lift****:* Given a product pair $(\mathfrak{p}_i, \mathfrak{p}_j)$, lift is the ratio of the probability of two products being bought together in the same basket to the probability of these two products being bought individually [9] -

$$\mathcal{L}_{ij} = \frac{\mathrm{P}\,(\mathfrak{p}_i \cap \mathfrak{p}_j)}{\mathrm{P}(\mathfrak{p}_i) \times \mathrm{P}(\,\mathfrak{p}_j)}$$

$$\mathcal{L}_{ij} = 1 \iff \mathrm{P}\,(\mathfrak{p}_i \cap \mathfrak{p}_j) = \mathrm{P}(\mathfrak{p}_i\,) \times \mathrm{P}(\mathfrak{p}_j)$$

⇒ purchase of $\mathfrak{p}_i$ and $\mathfrak{p}_j$ are independent event

$$\mathcal{L}_{ij} > 1 \iff \mathrm{P}\,(\mathfrak{p}_i \cap \mathfrak{p}_j) > \mathrm{P}(\mathfrak{p}_i\,) \times \mathrm{P}(\mathfrak{p}_j)$$

⇒ $\mathfrak{p}_i$ and $\mathfrak{p}_j$ are likely to be bought together in the same basket than being bought separately.

Thus, variety seeking behavior results in higher lift for such product pairs, on the other hand for traditional products purchase of one product nullifies the purchase probability of the other since they serve the same need, hence lift for a *traditiona*l pair is likely to be lower. Other metrics for market basket analysis - confidence and support are omitted due to confidence's asymmetric nature and support's tendency to be deflated in large scale transactions.

Content based features –

d. ***Brand Similarity****:* It is a mapping $\mathcal{F}: \mathfrak{B} \times \mathfrak{B} \rightarrow \mathbb{R}\,[0,1]$. If the product pairs have the same brand then brand similarity is one. Instead of using a discrete scale (0 or 1) we have computed the similarity in a continuous scale. A fuzzy matching takes care of the cases of 1) brand names written in abbreviations, 2) spelling mistakes etc. We apply character-level bigrams [10] to tokenize the brand name and then use Jaccard Similarity [11] to finally achieve the similarity score.

Brand loyal households are more likely to substitute one product for the other if they are from the same brand. Higher value for brand similarity indicates that the pair is likely to be a substitute pair.

e. *Pack-size Similarity*: For a product $\mathfrak{p}_i \in \mathbb{P}$ pack-size indicates number of units of the product $\mathfrak{p}_i$ being sold as one single product - for e.g. single product pack, double product pack etc. Bundling products into different pack sizes is prevalent in food and consumables category. Different pack-sizes cater to different household segments with varying household size and other demographic features. Accordingly, substitution is mainly possible within similar pack-sizes. Essentially, a multi-pack (pack of 12 or pack of 14) is less likely to be substituted with a single pack product.

Consider the function $\mu: \mathbb{I} \rightarrow \mathfrak{P}$ is a mapping from the set of intergers to set of all possible pack-size buckets, such that depending on the distribution of pack counts, μ assigns the product one pack-size bucket.

Then for a given pair $(\mathfrak{p}_i, \mathfrak{p}_j)_{i \neq j}$ we can define pack-size similarity index (ρ_{ij}) as –

$$\rho_{ij} = 1 \; if \;\; \mu(\mathcal{s}_{\mathfrak{p}_i}) = \mu\left(\mathcal{s}_{\mathfrak{p}_j}\right), \\ \textit{0 otherwise.}$$

A value of 1 for this feature indicates both the products fall into a similar pack-size bucket and more likely to be substitute.

f. *Product Description Similarity*: It is a mapping $\omega_{ij}: \mathfrak{D} \times \mathfrak{D} \rightarrow \mathbb{R}\,[0,1]$. Product description consists of explicit or implicit mentions of various attributes like flavor, gender and other functional forms except brand. Products having very similar descriptions are potential to be substitutes. Again, traditional substitutes are likely to be more similar in terms of product description as compared to variety substitutes pairs. We first create a word-similarity metric → $\mathbb{R}\,[0,1]$ considering the first and last letter match. We also consider all other common letters to be present in the smaller word. Furthermore, we create a normalised description similarity via taking a sum over all word similarities and finally preform an affine transformation [16] to determine its range correctly.

Other features –

g. *Demographic Similarity*: The framework captures the customer segments that purchase product from the category and taps their demographic information. Following table shows the information about groups and intuition behind using them. Each household $\mathfrak{h}$ can be represented as a collection of following factors –

$\mathfrak{h}$ = {Education, Ethnicity, Adult Quantity, Children Quantity, Income, Marital Status, Age}

We calculate the percentage count of all these variables for a product which brings down the values in a range of [0, 1]. But, difference in certain specific variable can get attenuated by other variables where difference is less while comparing two demographic vectors. To mitigate this, we performed Principal Component Analysis (PCA) to obtain the linear combinations of all the variables, doing in such a way so that the maximum variance of the data can be captured. We further use Kaiser criteria [12] to select the number of Principle Components to be considered. Finally, the cosine similarity between two vectors of reduced dimensions gets calculated. This similarity measure captures the demographic relatedness of customer segments for both the products in a product pair. If the final demographic representation of each households belongs to a space $\mathcal{R}_d$ with reduced dimension than original, then we define the similarity function as a mapping $\phi: \mathcal{R}_d \times \mathcal{R}_d \rightarrow \mathbb{R}\,[0,1]$, where ϕ is a cosine similarity function [13].

h. *Price Similarity:* Given a product pair $(\mathfrak{p}_i, \mathfrak{p}_j)$, price similarity index calculates how close the product pair is in terms of price point. It is only a plausible assumption that households look for the products in the same price bucket while substituting. Households who tend to buy products from a lower price bracket - essentially value products, are less likely to substitute their product of choice by some other product from a very high price bucket. We consider a transformation $\varphi(.)$, such that, given the price points $R_{\mathfrak{p}_i} \in \mathbb{R} \; \forall \; \mathfrak{p}_i \; \in \mathbb{P}, \varphi(R_{\mathfrak{p}_i}) \in \mathfrak{R}$. Essentially, $\varphi(.)$ is a function which takes each price point and assigns them to a suitable price bucket. Mathematically –

$$\textit{Price Similarity } (\sigma_{ij}) = \textit{1 if } \varphi\left(R_{\mathfrak{p}_i}\right) = \varphi\left(R_{\mathfrak{p}_j}\right), \\ \textit{0 otherwise.}$$

Now, we will substantiate our intuitions via exploratory analysis performed on our dataset (details in subsection *Data*).

*2) **Exploratory Analysis:*** We here present the exploratory analysis to bolster our motivation behind the methodology. The relative directions of the explanatory power of each variable conditional to identifying different group of substitute is presented here. It is seemed to hold across categories.

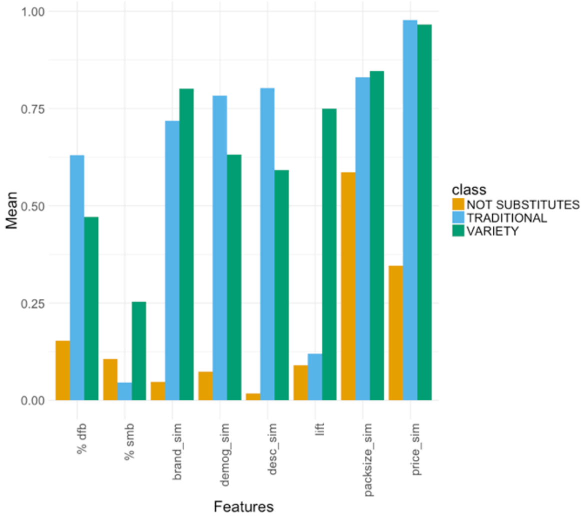

Fig. 2. Feature means conditional to different set of product pairs

Fig. 2. shows the normalised means of each variable conditional to three group of product pairs. %dfb exhibits major differences while constrasting substitutes and non-substitutes. Similarly, Brand Similarity Index, Pack-size Similarity, Price Similarity show similar behavior i.e. reveal significant differences while distigushing substitutes and non-substitutes. In contrast, Lift and %smb differentiate well the traditional substitutes and variety substitutes. High values of these these two features indicate variety substitution behavior. It is also interesting to notice that Description Similairity and Demographic Similarity show differences for all kind of product pairs.

*3) **Multistage Clustering:*** We propose an unsupervised multistage framework to estimate the assignment $\mathfrak{a}_k$ for a product pair. We seek to cluster the product pairs in to semantically meaningful groups, here which are $\mathbb{T}$, $\mathbb{V}$ or $\mathbb{NS}$. Since, $\mathbb{T}$ and $\mathbb{V}$ constitute to Substitute set $\mathbb{S}$ - our framework tries to cluster the data points into two mutually exclusive sets $\mathbb{S}$ and $\mathbb{NS}$. Furthemore, as it is observed that not all the features are well distinguishers for the separation when it comes to sets $\mathbb{S}$ and $\mathbb{NS}$, the framework considers specific set of features in different stages which we will elaborate here.

Stage 1:

Product pairs those belong to set $\mathbb{S}$ are perceived differently by the consumers compared to the product pairs belonging to set $\mathbb{NS}$. Also, from what we see empirically, the cardinality of set $\mathbb{NS}$ is way too high than the number of elements in set $\mathbb{S}$. These are the main reasons why the problem has to be broken into multiple stages. Specifically, in stage 1, we focus on distinguishing sets $\mathbb{S}$ and $\mathbb{NS}$.

Since attributes are the primary anchor by which consumers associate themselves with the products, content based features play the key role when it comes to differentiating *substitues* and *non-subtitutes*. Backed by our emperical observations from Fig. 2. we finally arrive at the set of variables that significantly distinguish or more specifically group the product pairs into two sets $\mathbb{S}$ and $\mathbb{NS}$:

%dfb, *Brand Similarity*, *Demographic Similarity*, *Pack-size Similarity*, *Price Similarity* and *Description Similarity*

We perform k-means [14] on $|\mathbb{P}\times\mathbb{P}|$ (excluding $(\mathfrak{p}_i, \mathfrak{p}_i)$ pairs) number of product pairs where each pair is represented by $[x_1, x_2, \dots, x_6]$. k-means divides the obeservation space into k clusters such that each observation belongs to the group with nearest '*mean*'. Here, the product pairs are represented via such variables which aid the simple clustering method to split the space into two groups which are semantically meaningful. The next step of assignning the cluster members is to find the right representations for both of the clusters. As explained, all the features used in this stage ideally take high values for pairs which belong to *substitute* set. We compute the central values of all the features for all observations by clusters and combine them via a convex combination. For simplicity, we assume all features has equal weightage and hence the combined representation is nothing but the empirical average of the mean of all features for a cluster. The cluster which has a greater combined value, thus obtained, will be tagged as a *subtitute* cluster. The other cluster is assigned as *non-subtitute* cluster.

Filtration Stage:

The preponderance of non-substitutes in any category may necessitate further fine-tuning of the substitute cluster. This is crucial because assigning non-substitute pair as substitite is costly. But this stage can be skipped if it meets following criteria:

- The cardinality of the substitute set assigned by the model meets the business benchmark for the substitutes in the category. Either prior guess of the category experts or the statistical prior estimate (via mean of samples) using multiple small samples (not more than 100) of the whole data can be taken into cinsideration.
- If the combined mean of substitute cluster assgined by the model is significantly larger than the non-substititute cluster.

If the first stage could not generate clusters those meet the above criteria, this filtration stage takes care of eliminating the non-substitute pairs from the substitute clusters. In this stage, we fine tune the substitute cluster by further forming two groups, one with the impurities (non-substitutes) and other with the finer set of substitutes. The system uses the product description similarity as the feature to capture the difference in perception of the consumers on these two groups. We perform k-means in similar fashion to first stage for this filtration process. It is intuitive that the substitute pairs will be more similar in terms of the product description as compared to non-substitutes and hence for substitutes, product description similarity will have higher values. After formation of the two clusters, assignment is achieved via product description similarity values in order to detect the cluster with the finer set of substitutes. The framework automatically decides to perform this stage based on the given prior proportions.

Stage 2:

In the second stage, the system takes the finer set of substitutes from the filtration stage and forms two cluster among them – one of traditional substitutes, another of variety substitutes. In this stage, the system takes the following set of features:

Lift, *%smb*, *Product Description Similarity*, *Demographic Similarity*

to form different clusters for traditional and variety substitutes. Higher values for lift for a pair represents variety behavior, since traditional substitutes are not bought in the same transaction basket. Similar intuition follows for the feature %smb - variety pairs will have higher values for this variable as compared to traditional product pairs. Traditional substitutes are very similar in terms of the attributes, whereas the variety differ in one or more attributes - traditional will be more similar in terms of product description as compared to varieties.

The framework handles this third stage differently from the previous two stages, precisely for the following two reasons – 1) Lesser number of pairs in this stage makes the task of identifying the pattern trickier. 2) Across categories the extent of traditional substitution and variety substitution could be different.

We employ a k-medoids [15] clustering which is similar to k-means but instead of 'mean', it assigns each observation to the group with nearest '*medoid*'. It is robust in presence of outlier observations and more useful than k-means when we have comparatively less observations [15]. The framework passes different combinations of the mentioned four variables to form the clusters. Since behaviour based features are most important in determining these two kinds of substitutes, the framework preserves Lift and %smb for all combinations and further takes one or both or none of Product Description Similarity and Demographic Similarity at a time and performs the clustering. After each run, the system assigns the cluster tag from the combined mean of lift and %smb. Thus, for each of the different feature combinations, it produces two clusters - one of traditional substitutes and another of variety substitutes. Finally, we use a voting mechanism to get to final class assignment for each of the pairs. In case of a tie, the system uses the combined score of Lift and %smb as a tie-breaker and assigns the pair in either of the two classes.

Affinity Score Creation:

Even though the framework return the final assignements, it is also pivotal to understand the depth or magnitude of such assignment as one product may have multiple traditional or variety substitute.

We created a score for each substitutable pairs. Define, θ_{ij}: Propensity score of Product pair $(\mathfrak{p}_i, \mathfrak{p}_j) \in \mathbb{S}$ that shows the strength of the substitutable relationship among the pairs. So if the pair be traditional substitutes i.e. $(\mathfrak{p}_i, \mathfrak{p}_j) \in \mathbb{T}$ then it shows how well they are traditional substitutes and if it's variety substitutes i.e. $(\mathfrak{p}_i, \mathfrak{p}_j) \in \mathbb{V}$, then it shows how strong variety substitute they are. Higher the value, stronger the relationship between the pairs. The value is normalized to 0 to 1 with a linear scale.

The score is created separately for the two cases of substitutions. It's thus best not to compare across the two scenarios. The scores for the two cases are as follows:

$$\theta_{ij} = \mathcal{F}(\delta_{ij} + \rho_{ij} + \sigma_{ij} + \omega_{ij} + \beta_{ij}) \,\forall\, i,j : (\mathfrak{p}_i, \mathfrak{p}_j) \in \mathbb{T}$$

$$\theta_{ij} = \mathcal{F}(\mathcal{L}_{ij} + 1 - \beta_{ij}) \,\forall\, i,j : (\mathfrak{p}_i, \mathfrak{p}_j) \in \mathbb{V}$$

Here, $\mathcal{F}(.)$ is a real valued function. $\mathcal{F}: \mathfrak{R} \rightarrow [0,1]$ and

$$\mathcal{F}(x_{ij}) = \frac{\left(x_{ij} - MIN_{\,i,j:(\mathfrak{p}_i,\mathfrak{p}_j)\in P}\,(x_{ij})\right)}{MAX_{\,i,j:(\mathfrak{p}_i,\mathfrak{p}_j)\in P}(x_{ij})} \,\forall\, x_{ij} \epsilon\, \mathfrak{R}, P = \mathbb{V} \text{ or } \mathbb{T}$$

The final score is used to retrieve the substitutes in a ranked manner. This consolidates the mechanism of identifying substitutes with ranked retrieval.

IV. Experiments

In this section, we evaluate our multistage clustering model. Prior to that we describe the data, the experiments, baseline and then move to results. Finally, a discussion on baseline and proposed model is also provided. We explain the imapct of the ranked retrieval of the substitutes for various query products. Further we explain the consumption of the result in a time agnostic manner. The framework also has been extended to multiple categories at the absence of any tagged data and validated by category experts.

A. Data

We used data from the *cold cereal* category of an US based retailer. The pair level data was manually tagged by category experts in three mutually exclusive and exhaustive classes – non-substitutes, traditional substitutes and variety substitutes. The class distribution is given in Table II below-

TABLE II. SUMMARY OF DATA

Relation	***Non- substitutes***	***Traditional***	***Variety***
No. of Pairs	34831	262	118
% of Pairs	98.9%	0.7%	0.3%

All the pairs bought by a minimum of 1000 households are considered. The features are generated at the pair level from three main data sources - point of sales data (~ 18 M) over the 1period of 52 weeks from the cold cereal category, household level demographic data and product attribute data.

B. Baselines

*1) **Supervised Framework - Classification tree (CT):*** If tagged data is available, a classifier can be constructed to classify the product pairs into traditional, variety and not-substitute classes. The available tagged data (from Table III) was divided into training (70%), test (20%) and validation set (10%). The same feature set has been used in building the classifier. A grid search carried out on the training set and the optimal value for the parameter got selected depending on the accuracy and sensitivity measures from both the training and test set.

*2) **Managerial Heuristics (MH):*** Prior study shows that there are certain managerial heuristics in place to identify the non-substitutes, traditional and variety substitutes. As the last section explains, non-availability of training data is detrimental to traditional way of modelling this kind of problem, since it will not scale up across categories. We have experimented with the following heuristics based results as baselines for our model. The motivation for these heuristics are rooted in the way how category managers and domain experts deal with the problem and mostly based out of their understanding and experience in the field.

***Measure of Association**:* Traditionally category managers look at some measure of association to understand the substitution behavior. Yules Q [17] is one such metric. A threshold of .6 is used as determinant of Substitutes, so pairs with Yules Q > = .6 is considered Substitute pairs and pairs with Yules Q less than .6 are considered as not-substitute pair.

Proportion of household purchasing the pair in the same basket: Once the substitute pairs are identified, among them the pairs having at least 50% of household purchasing the pair in the same basket are considered as variety substitutes and the remaining are considered as traditional substitutes.

C. Experimental Settings

We bucket Pack-size similarity, into $\mathfrak{P}$ = {single packs, small, medium, large, very large}. This bucketing is data driven and may vary across categories. For the feature Price Similarity also, the set of all possible price buckets are $\mathfrak{R}$ = {low, medium, high}, which is done based on the distribution of price points within the category. Also, since Lift can take very large values, we have used winsorization [17] to this feature.

For the first stage of clustering, we have used PCA in the selected feature set, the number of principle components to be used chosen by Kaiser's criterion [12]. Then the selected Principal components are passed into the k-means clustering algorithm. The first stage involved severe class imbalance problem, use of principal components helped accentuate the difference and formation of better clusters. Finally, we use k = 2, in all the stages for both k-means and k-medoids.

D. Results and Comparison

Before reporting the exhaustive comparison of our method and the baselines, we present the results of gradual improvements from each of the stages of our multi-stage approach.

1) ***Stage 1- Substitutes vs Non-Substitutes:*** In this stage, k-means identifies two broad clusters - one of substitutes and another of non-substitutes. The performance metrics are reported in the Table III.A. Sensitivity is computed with respect to the substitute (minority) class -

TABLE III.A STAGE 1: PERFORMANCE METRICS

Accuracy	Sensitivity	Specificity	Precision
96.98%	90.79%	97.05%	25.13%

After this first stage, we also observe 1028 non-substitutes pair entering the substitute cluster making precision very low, which necessitates further stage of filtration.

2) ***Filtration Stage - Substitutes vs Non-Substitutes:*** In this stage, we fine-tune the substitute the cluster identified from the first stage. In the last stage, the substitute cluster consisted of 1373 pairs. In this stage, the framework separates out the finer set of substitutes pair from these 1373 pairs. The performance, after this filtration stage has been summarized in the Table III.B. After this stage, we get a cleaner substitute cluster of 342 pairs. The improvement in precision is notable, which validates the necessity of the filtration stage.

TABLE III.B. FILTRATION STAGE: PERFORMANCE METRICS

Accuracy	Sensitivity	Specificity	Precision
99.63%	77.63%	99.87%	86.26%

3) ***Stage 2 - Traditional Substitutes vs Variety Substitutes:*** In these stage, the algorithm works on the refined set of substitutes, from the filtration stage and differentiates the traditional substitute pairs from the variety substitute pairs. This is the final stage, and the confusion matrix, taking into account all three sets, is summarized in Table III.C.

4) ***Comparison****:* In the Table III.D, we have consolidated the results from the baselines and proposed model. In terms of accuracy all the three models fair well, obtaining comparable accuracy – this is because of huge class imabalance.

In terms of sensitivity, our model performs better than the managerial heuristic based method or the classification tree method for the variety class. Identifying the variety pairs correctly are crucial, since they contribute in basket building behavior and are of high importance for retailers.

In terms of scaling up, the classification model trained only on the cold cereal category will superimpose the same boundaries learned on this category to new categories and may not perform well. Managerial Heuristics are essentially fixed thresholds and will have very poor sensitivity in other categories as well. However, our model is unsupervised, does not need tagged data, and will learn the category nuances from the features itself. Thus, it is more likely to give better results in terms of scaling up across categories.

As we can see in the Table III.D, classification tree performed quite well in identifying the non-substitutes pair correctly, but the sensitivity of the variety classes is not good enough. This is attributable to the class imbalance problem in the data. We have experimented with some over-sampling techniques but that resulted in negligible increase in sensitivity measures. Also, scarcity of tagged data makes the use of classification model less scalable.

TABLE III.C. FINAL STAGE: CONFUSION MATRIX

		Predicted			
	Class	***Non-substitutes***	***Traditional***	***Variety***	***Total***
Actual	***Non-substitutes***	34784	0	47	34831
	Traditional	46	163	53	262
	Variety	39	1	78	118
	Grand Total	34869	164	178	35211

TABLE III.D. FINAL PERFORMANCE METRICS

Model	Sensitivity			Accuracy	Scalability
	Non-substitutes	***Traditional***	***Variety***		
CT	99.90	67.39	50.00	99.58	N
MH	97.10	62.59	14.40	96.57	N
Our Model	**99.12**	**62.21**	**66.10**	**99.47**	**Y**

5) ***Ranked Retrieval:*** Furthermore, one we obtain the clusters, we derive the affinity scores for all pairs and present the final retrieval of both kind of substitutes given a query product. Given top retrievals are obtained for all cases, it is observed that the ground truth appears on the toppest retrieval most of the time, which can also be seen from the above sensitivity measures. Furthermore, considering a ranked retrieval improve the performance more due to the appearance of the ground truth, even if not at the topmost but in the top retrievals.

TABLE IV. RANKED RETRIEVAL FOR SUBSTITUTES

Query	Traditional	Variety
Kellogg's Special K Chocolate Protein Shake	**Kellogg's To-Go Chocolate Protein Shake**	**Special K Protein Strawberry Shake**
	Kellogg's Special K Chocolatey Delight	Kellogg's Special K Protein Shakes, Raspberry Cheesecake
	Kelloggs Special K Choco Almond	Kellogg's Special K Red Berries

Table IV shows an illustrative example of the retrieval. The ground truth are both retrieved at the first response. Moreover, the other retrieved results also exhibit traditional substitution behavior. Similarly, all of the products reported here are true variety substitutes of the query products. The rank was given by the affinity score to indicate the magnitude of the assignment by our unsupervised framework. This ranking further improves the specificity of substitute classes via multiple retrieval with their magnitude of assignments.

E. Time Agnostic Modifications

In retail, to keep up with the customer requirements and change in shopping patterns, categories i.e. $\mathbb{c}$ ($\in \mathbb{C}$) do undergo significant changes every year or even more frequent than that. Change in categories can be in many ways – new products are launched, same products are modified slightly i.e. new packaging, changed price etc. This means that over the course the product universe $\mathbb{P}$ gets changed. To tackle this, $\mathbb{P}_\tau$, a time-stamped product universe should be considered at time point τ. Even for an unchanged $\mathbb{P}$, there can be instances which change the feature representations for certain $\mathbb{p}$ ($\in \mathbb{P}$). Our approach makes provisions that helps to solve for such scenarios efficiently. In other words, given a solution on $\mathbb{P}_\tau$ of category $\mathbb{c} \in \mathbb{C}$, we can obtain solution for any time point τ' where $\tau' > \tau$. This can be achieved at any time point $\tau'(> \tau)$ accordingly. Once the solution for time point $\tau\,(< \tau')$ is available, the tagging for all the pairs $(\mathbb{p}_i, \mathbb{p}_j) \in \mathbb{P}_\tau \times \mathbb{P}_\tau$ of category $\mathbb{c} \in \mathbb{C}$ can be generated. Now at time τ', we can create the features for $\forall\, \mathbb{p} \in \mathbb{P}_{\tau'}$, and generate the distance from the representatives (preferably mean represenation) of each cluster. Also, it is possible to build a supervised model with tags generated on $\mathbb{P}_\tau$ and then apply it on $\mathbb{P}_{\tau'}$. Further to this, the affinity scores can be generated with newer assignments and further can be retrieved with new set of products. This makes our solution time agnostic which can accomodate the ever-changing buying patterns.

F. Application to other categories & Impact

To check the robustness, this methodology has been tested out for quite a few food and consumables categories (for e.g. milk, juice, meat etc.) as well as toiletries (for e.g. fragrance, soap, shampoo etc.). The results have been vetted by business domain experts as giving more accurate reads than any of the status-quo solutions they have used. The improved detection of variety substitutes will aid in basket building and correct detection of traditional substitutes will help in freeing the occupied shelf space to provide the room newer products. It is to be emphasised that, without tagged data in these above categories, the framework perfectly scales and achieves gold standard performance when validated by category experts. Substitution is a rare phenomenon and availability of tagged data is extremely scarce. Our unsupervised framework hence plays a crucial role to capture the interrelationships of the products by mining large-scale data.

V. Conclusion

The problem of appropriate substitute identification is an important one for any retailer. Misjudgements here may lead to severe repercussions – including, alienating the customer base. In this work, we present an automated framework to identify product substitutes into three mutually exclusive and exhaustive components – 1) non-substitute 2) traditional substitute and 3) variety substitute. We do this using a wide set of behaviour based and content based features along with demographic & price information. While the most obvious way to solve this problem would be to use multiclass classification techniques, these have issues in not being easy to scale up. For this we have designed a multistage clustering approach which in the finality is scalable as well as has better sensitivity than classification (or, the simpler managerial heuristic) based model. Beyond this, our methodology also provides an affinity score for a given pair to be a traditional or a variety substitute.

We intend to take this framework forward by using this to solve analytical problems in diverse retail domains. One such is in pricing and promotions. A product on promotion tends to cannibalize on the sales of a like product (traditional substitute), whereas has a positive effect on the sales of another, which is often bought together with it (variety substitute). To tease out the full impact of a promotion understanding these effects, our framework will be brought into play.

References

[1] S. Silberberg (2001), “The Structure of Economics, A Mathematical Analysis”. *McGraw-Hill*

[2] A. D. Shocker, B. L. Bayus, & N. Kim, “Product Complements and Substitutes in the Real World:The Relevance of Other Products”, *Journal of Marketing* Vol. 68 (January 2004), 28–40

[3] J. Singer, B. Libai, L. Sivan, E. Carmi, & Ohad Yassin, “The Network Value of Products”, *Journal of Marketing*,Volume 77 (May 2013)

[4] R. Bordley, "Relating Cross-Elasticities to First Choice/Second Choice Data", *Journal of Business and Economic Statistics*, (1985)

[5] S Hal R. Varian, “Microeconomic Analysis”

[6] Day, S. George and S. Allan D., “Identifying Competitive Product Market Boundaries: Strategic and Analytical Issues,” *Marketing Science Institute Report* No. 76 -112 (1976), Cambridge, MA

[7] J. M. Lattin and L. McAlister, “Using a Variety-Seeking Model to Identify Substitute and Complementary Relationships among Competing Products”, *Journal of Marketing Research*, Vol. 22, No. 3 (Aug., 1985)

[8] J. McAuley, R. Pandey, J. Lescovec, “Inferring Networks of Substitutable and Complementary Products,” *KDD* ’15 August 11 - 14, 2015

[9] Rakesh Agrawal, Ramakrishnan Srikant, “Fast Algorithms for Mining Association Rules”, *VLDB* 1994

[10] Michael Collins. “*A new statistical parser based on bigram lexical dependencies”. 34th Annual Meeting of the ACL*, CA. 1996. pp.184-191

[11] P-N. Tan,, M. Steinbach, V. Kumar, “Introduction to Data Mining”, 2005

[12] Braeken J, van Assen MA., “An Empirical Kaiser Criterion, Psychol Methods”. 2016 Mar 31

[13] Singhal, Amit (2001). "Modern Information Retrieval: A Brief Overview". *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24*

[14] Lloyd., S. P. (1982). "Least squares quantization in PCM" (PDF). *IEEE Transactions on Information Theory*. 28 (2): 129–137.

[15] L Kaufman. and P.J Rousseeuw, (1987), “Clustering by means of Medoids, in Statistical Data Analysis Based on the L1–Norm and Related Methods”

[16] Zwillinger, D. (Ed.). "Affine Transformations." 4*.3.2 in CRC Standard Mathematical Tables and Formulae. Boca Raton*, FL: CRC Press, pp. 265-266, 1995.

[17] H astings, Jr., Cecil; Mosteller, Frederick; Tukey, John W.; Winsor, Charles P. (1947). "Low moments for small samples: a comparative study of order statistics"

[18] George Udny Yule, “On the association of attributes in statistics”, Phil.Trans.A, 194, 257--319, 1900