Bundle entropy as an optimized measure of consumers' systematic product choice combinations in mass transactional data

Roberto Mansilla
N/LAB
Nottingham University Business School
Nottingham, UK
roberto.mansilla@nottingham.ac.uk

Gavin Smith
N/LAB
Nottingham University Business School
Nottingham, UK
gavin.smith@nottingham.ac.uk

Andrew Smith
N/LAB
Nottingham University Business School
Nottingham, UK
andrew.p.smith@nottingham.ac.uk

James Goulding
N/LAB
Nottingham University Business School
Nottingham, UK
james.goulding@nottingham.ac.uk

***Abstract*—Understanding and measuring the predictability of consumer purchasing (basket) behaviour is of significant value. While predictability measures such as entropy have been well studied and leveraged in other sectors, their development and application to very large multi-dimensional data sets present in the retailing sector are less common. While a small number of methods exist, we demonstrate they fail to accord with intuition, leading to the potential for misunderstandings between those who conduct the analysis and those who act on the insights. We delineate the requirements for such a measure in this domain to demonstrate these issues in context. A novel measure is then developed based on entropy to directly measure the predictability of basket composition. The measure is designated as *bundle entropy* (zero denotes a bundle's total predictability, one the total unpredictability). We empirically compare the proposed bundle entropy against existing measures using two large-scale real-world transactional data sets, each including more than 2,000 households (frequent shoppers) over two years. First, we demonstrate how the proposed measure is the only measure that behaves according to the desired properties. Second, we show empirically that bundle entropy differs noticeably from the other measures. Finally, we consider some use case analyses and discuss the utility of the proposed measure in practice.**

***Index Terms*—bundle entropy, systematic choices, systematic customers, consumer behaviour, basket analysis**

I. Introduction

Measuring and understanding the predictability of human behaviour is of growing value to scholars and commercial or policy decision makers alike. Measures of predictability, and their importance to both policy and the economy, are well studied within domains such as human mobility [1], [2], with a wide range of application areas from advertising [3] and service provision [4] to intelligent agents [5]. However, the study of predictability to understand and leverage regularity in other behaviours, such as purchasing patterns, is comparatively limited. This is despite large-scale transactional data sets now being routinely collected as part of our digital footprint, and processed as part of loyalty programmes and online purchase platforms.

The potential to leverage behavioural big data for practical purposehighs, as well as academic ones, is self-evident [6], [7]. For instance, systematic or predictable consumers can be more readily provided with relevant offers or new products and services with significant efficiency (and financial) gains for the retailer and increased utility for the household/consumer. Conversely, unpredictable households and consumers might be appropriate for targeting with innovations / new products and/or provided with more varied direct offers. The ability to assign a household or consumer with a predictability score, that can be incorporated into retail segmentation, descriptive, and predictive analytics [8], creates greater opportunities for personalising responses and offers. This accompanies the fact that the framing and messaging of direct-to-consumer marketing communications are increasingly informed by behavioural and propensity scores, ensuring that communication is congruent with consumer needs.

Motivated by the potential value for behavioural academics, retailers, and policymakers we focused on measuring the predictability of basket purchases from transactional big data. We acknowledge that this has been the focus of prior work in [9], and to a lesser extent, work on related measurements of variety, and diversity [10]–[12]. However, as will be demonstrated below, existing measures do not accord with an intuitive definition of basket predictability or are parameter dependent/unstable. This directly affects their actionability. While this is discussed in detail in §IV, the examples presented in Figure 1 illustrate both this work's goal and current approaches' drawbacks.

We note that measuring predictability at the basket level is

The purchasing patterns of five customers are listed below. Each basket (set) in a purchase history represents a distinct shopping visit, featuring one or more of the following items: *milk: m, bread: b, paper: p, sandwich: s, coffee: c, jam: j*:

$C1 = \{m, b, j\}, \{m, b, j\}, \{m, b, j\}, \{m, b, j\}, \{m, b, j\}$
$C2 = \{m, b, p\}, \{m, b, p\}, \{m, b, p\}, \{m, b\}, \{m, b\}$
$C3 = \{m\}, \{m, b\}, \{m, p\}, \{m, s\}, \{m, c\}$
$C4 = \{m, b\}, \{b, p\}, \{p, s\}, \{s, c\}$
$C5 = \{m\}, \{b\}, \{p\}, \{s\}, \{c\}$

Intuitively the behaviour of customer $C1$ is the most predictable, given their purchasing of the same items each visit. $C2$ remains relatively predictable, given they always purchase milk and bread, but only occasionally a paper. The basket composition of $C3$ is more unpredictable yet again, with only milk being consistently purchased. $C4$ and $C5$ are, however, the hardest to predict, although $C4$ at least demonstrates some commonality across baskets. For a measure which quantified predictability across these customers, we would therefore wish it to produce the following ordering:

	Low ← Unpredictability → High				
Expected:	C1	C2	C3	C4	C5

Yet existing measures one might apply do not match this intuition. Symbol (item) entropy fails to reflect the expected ordering altogether, whereas both Joint (Basket Level) entropy and Guidotti et al's measure [9] fail to distinguish the fact that C1 < C2 < C3 < C4 < C5 (even across various parameterizations).

Item Entropy	1.0	0.98	0.81	0.97	1.0
Basket Entropy	0.0	0.97	1.0	1.0	1.0
[9] (low param)	0.0	0.97	1.0	1.0	1.0
[9] (med param)	0.0	0.97	0.0	1.0	1.0
[9] (high param)	0.0	0.0	0.0	1.0	1.0

The inability, even in toy examples, of existing measures to match applied intuitions has motivated this work.

Fig. 1: Examples of consumer purchasing behaviour where the effectiveness of current approaches to measure basket predictability are insufficient.

also a goal of [9]. Specifically, we wish to consider someone as predictable based on the extent to which it is possible to predict the composition of either their basket or a sub-basket. This reflects the utility that can be derived from having certainty that a customer's next basket will contain certain items, with more utility being gained from being able to be certain about more and more items. It also reflects the real-world conditions where repeated baskets with the same content are uncommon due to variety seeking, cross-retailer shopping, group purchasing, and other factors. Failure to consider regularity due to a small number of random items in a basket risks labelling many customers as being unpredictable - when in fact the opposite is true.

This paper is organised as follows. In §II we briefly review related measures that attempt to quantify the predictability of purchases from transactional data, highlighting the different definitions of predictability these measures encode and postulate that none correctly quantifies the predictability of basket purchases. Subsequently, in §III, we motivate and list the properties that such a measure would need to follow. In §IV, §V and §VI we theoretically detail the shortcomings of existing approaches against these, propose a new measure *Bundle Entropy (BE)* and then empirically demonstrate its utility compared to existing approaches on two real-world data sets. A conclusion is provided in §VII.

II. Related Work

While many studies focus on explaining (often sequential) product choice through fitting latent models (e.g [13]–[24]) or building predictive models to predict subsequent visit behaviour (e.g [25]–[27]), limited work exists on directly quantifying the predictability of human (basket) purchases from transactional data to identify people for actionable interventions. The first attempt at this could perhaps be considered an attempt at quantifying variety via simple counts of distinct products [28]. Later the concept of entropy, more directly measuring predictability via the notion of uncertainty [2], [29], was proposed to consider the variety of products within a single group [30]. While other methods of variety have been proposed, i.e., the Hirschman-Herfindahl and the Gini coefficient, entropy's link with prediction provides a stronger actionable link in this context, with [10] additionally noting the utility of entropy measures in encoding desirable aspects related to the distribution, rareness, and commonness of the products contained in the group compared to these other measures.

Defined for a single group, however, entropy describes the difficulty (uncertainty) one faces in predicting a single item observed randomly from this group (say, the next item an individual might add to their basket). As such, it does not quantify the predictability of baskets (groups of items). A very simple example of this can be seen by considering an individual who always (say over 3 visits) buys a basket consisting of milk (m), coffee (c) and a sandwich (s) leading to a purchase sequence of $[\{m, c, s\}, \{s, c, m\}, \{c, s, m\}]$. Viewing this as a single unordered group ($[m, c, s, s, c, m, c, s, m]$) predicting what a random item would be next is almost impossible, and entropy captures this by reporting maximum entropy. However, considering these purchases at the basket level, the same basket is always purchased (m, c, s) and is 100% predictable - i.e., it should have reported zero uncertainty.

Motivated by this [9] develop a measure they term Basket Revealed Entropy (BRE) to provide a direct, individual-level measure of "how unpredictable ... [an individual's] basket composition [is]". By mining frequent patterns within customers' baskets (with some additional steps) and then computing the entropy based on these *common sub-baskets*, the authors implicitly indicate their view that the direct application of entropy is not appropriate. Notably, the reasons one would reject the direct application of entropy in this context are directly related to what is meant by *predictability* (i.e., the predictability of what?), which directly impacts the ability to formulate and take any subsequent action in a business context. Specifically, BRE considers predictability as the task of exactly predicting sub-baskets (the set of which is defined algorithmically and will be discussed shortly), ignoring any extra context (items in the basket). This contrasts the predictability defined as the task of exactly predicting an individual's complete basket composition; the task encoded if one was to consider the entropy based on a symbol set where each symbol represented a distinct basket. We subsequently refer to this as Basket Level Entropy (BLE)[1]. This is different from the aforementioned application of entropy at an item level (losing the attribution of items to baskets), where a symbol is then an item, and the prediction task is predicting which item I will select to put in my basket at any arbitrary time point.

The utility of these different measures of "predictability" must then be considered in context. Similarly to [9] in this work, we are interested in understanding customers to positively affect KPIs within the retail sector. Specifically, our discussion focuses on the fast-moving consumer goods sector, though the discussion is likely more generalizable. Considering such a sector, while [9] provides a good first attempt, we argue that their measure (1) does not accord to the most actionable properties and (2) lacks clarity in its definition of predictability due to its algorithmic definition and parametrization.

III. Measuring the predictability of basket composition

To underpin a measure of purchasing behaviour, that captures the predictability of basket composition while corresponding to intuition and the needs of real-world applications (see Fig 1), we propose the following three properties. Let $\mathcal{B} = [b_0, b_1, \ldots, b_n]$ be an individual's list of baskets, where baskets are sets of distinct items $b_x = \{\gamma_0, \gamma_1, \ldots\}$, and $\mathcal{M}(\mathcal{B})$ the value of the measure assigned to a given $\mathcal{B}$. Then:

P0: Sequences of baskets where all baskets contain the same items should have a score of *zero*; Sequences, were no basket shares any items with any other, should have a score of *one* (or the maximal value if not normalised - see later discussion). Thus:
$\mathcal{M}(\mathcal{B}) = 0 \quad \text{if } b_0 = b_1 = \ldots = b_n$
$\mathcal{M}(\mathcal{B}) = 1 \quad \text{if } b_0 \cap b_1 \cap \ldots \cap b_n = \emptyset$

P1: Increased (decreased) presence of common sub-baskets in a purchasing sequence should result in a lower (higher) score. Formally, if Γ is any arbitrary combination of items that was not previously present in every basket of $\mathcal{B}$ (i.e. $\exists b_k : \Gamma \not\subset b_k$), and we add Γ to each basket to produce $\mathcal{B}'$, then so long as $\mathcal{B}$ wasn't already totally predictable:
$\mathcal{M}(\mathcal{B}') < \mathcal{M}(\mathcal{B}) \quad \text{if } \mathcal{B}' = [b_0 \cup \Gamma, b_1 \cup \Gamma, \ldots]$

P2: Sequences with larger systematic sub-baskets should result in a lower score than sequences with smaller sub-baskets (relative to any basket size) unless the sequence is already fully predictable meeting $P0$.
$\mathcal{M}(\mathcal{B}^*) < \mathcal{M}(\mathcal{B}')$ if
$\mathcal{B}' = [b_0 \cup \Gamma, b_1 \cup \Gamma, \ldots,]$
$\mathcal{B}^* = [b_0 \cup \Gamma \cup \Gamma', b_1 \cup \Gamma \cup \Gamma', \ldots,]$
where Γ and Γ' are (different) arbitrary combination of items and $\mathcal{B}$ does not meet $P0$.

These properties are adapted from those expected with regard to BLE. $P0$ defines the expected behaviour in the extremes. If all baskets are identical, then the prediction of a (random next) basket is trivial - there is no uncertainty and the measure should equal zero. Conversely, if no baskets share any items, then there is maximal uncertainty as to what a subsequent basket should be, and the measure should reflect this. Such properties accord to those encoded by BLE as the task degrades into predicting a set of identical baskets (symbols/composition unimportant) or unique baskets (symbols/composition unimportant).

In contrast, $P1$ encodes a relaxed view on uncertainty concerning individuals with repeated sub-baskets. This contrasts to the use of joint entropy (BLE) which would not consider the predictability gains accrued when the baskets observed in a consumer's purchase history contain repeated sub-baskets. BLE takes an 'all-or-nothing' perspective to the prediction task, where sub-basket regularity is orthogonal to the sole task of predicting the entire basket. Relaxing this all-or-nothing view, $P1$ encodes the fact that being able to predict a sub-basket holds significant utility to a decision-maker. Furthermore, $P2$ acknowledges the fact that having a larger predictable sub-basket provides greater insights into real-world applications. The larger the size of predictable components within any purchase history, the more certain one can be about the composition of that customer's future purchasing behaviour (and likely their overall spending). $P2$ therefore further formalises this relaxation.

Finally, we note that entropy-based measures of predictability are often normalised. This acknowledges that the complexity of a prediction task increases as the number of output possibilities (the overall number of symbols in the case of entropy) increases. As such non-normalised versions of entropy-based measures, when used to compare individuals, conflate a measure of the individuals' uncertainty within a bounded set of choices (relative predictability across what they have access to) and measures of access to larger choice sets. Concerning the use case of consumer goods purchasing, the latter is often driven and/or constrained by factors such as household size and income [31], [32]. As such, it is generally desirable for a measure to remain independent of these covariates, often normalising against some measure of choice set size (i.e., the number of unique baskets or number of unique items). Following [9] we normalise our measure by dividing by the

[1] This is equivalent to the Joint Entropy of baskets with all items represented by indicator binary variables indicating if the item was in the basket or not.

number of unique baskets[2]. We note that, if desired by a practitioner, the non-normalised version of the measure we propose could be used, maintaining all motivating properties and with their associated proofs still holding (see Appendix A).

IV. Failure of existing methods

To consider the measures mentioned above (entropy at the basket level, entropy at the item level, BRE) in more depth, we first introduce some notation.

Let $\mathcal{B} = [b_0, b_1, \ldots, b_n]$ be an individual's list of baskets, where baskets are sets of distinct items $b_x = \{\gamma_0, \gamma_1, \ldots\}$. Let $p(\gamma)$ be the probability of γ appearing in a basket in $\mathcal{B}$. Further, let $p(b)$ be the probability of b, an observed basket in $\mathcal{B}$; and B be the set of distinct baskets in $\mathcal{B}$. Finally let $I = \bigcup_{b \in \mathcal{B}}$ be the set of distinct items purchased (and in all cases $0 \log_2 0$ is taken to be 0 as per convention)

Normalized Entropy at the item level is then defined as:

$$IE(\mathcal{B}) = -\frac{1}{\log_2 |I|} \sum_{\gamma \in I} p(\gamma) \log_2 p(\gamma) \qquad (1)$$

Normalized entropy at basket level (which we refer to as BLE) can be defined as:

$$BLE(\mathcal{B}) = -\frac{1}{\log_2 |B|} \sum_{b \in B} p(b) \log_2 p(b) \qquad (2)$$

In contrast, BRE, as proposed by Guidotti et al. [9], is defined by first constructing a new list of baskets, $\mathcal{B}' = [b'_0, b'_1, \ldots, b'_n]$, which replaces each basket $b \in \mathcal{B}$ with a *common sub-basket* b' according to the following algorithm:

1) first identifying a set of candidate *common sub-baskets* via the Apriori algorithm [33] with a user-defined minimum support parameter
2) replacing each basket with a single *common sub-basket* based on the following rules (expanding the set of *common sub-baskets* as required):
    a) the longest *common sub-basket* contained in the basket currently under consideration [3] [RULE 1]
    b) if no *common sub-basket* is contained, which may occur depending on the minimum support value defined by the user, then the full basket being considered, and this new symbol added to the *common sub-basket* list [RULE 2].

Let B' be the set of distinct baskets in $\mathcal{B}'$.
BRE is then defined as:

$$BRE(\mathcal{B}') = -\frac{1}{\log_2 |B'|} \sum_{b' \in B'} p(b') \log_2 p(b') \qquad (3)$$

Framing BRE as a measure of basket entropy of assigned *common sub-baskets* and considering its behaviour as its parameterization, the minimum support (*minsup*) for the Apriori algorithm is varied and highlights two key points: (1) When *minsup* tends towards zero then all baskets become part of the candidate *common sub-basket* set and all $b'_x = b_x$ (due to RULE 1); (2) When *minsup* tends towards 1 then it depends on the data. If a *common sub-basket* exists in all baskets (i.e., $\{m, b, p\}, \{m, s, c\}, \{m, b, j\}$), then RULE 1 will apply, leading to all b'_x being all the same (m in the example), though in most real-world cases this will not be true (i.e., $\{m, b, p\}, \{m, s, c\}, \{s, c, j\}$). In this case the candidate *common sub-basket* set will contain no candidates (as *minsup* is close to 1) and all $b'_x = b_x$ via RULE 2. Given that the entropy is then computed, considering each unique *common sub-basket* as a symbol, for the most common two of the three cases, the BRE degenerates to entropy at a basket level (BLE).

A. Property violations in existing measures

We now assess where existing predictability measures do not accord to the intuitive properties laid out in §III. First, theoretically item level entropy (IE) does not accord to $P0$, while BLE and BRE do. The violation of this property by IE can be highlighted by considering the trivial example from Figure 1 with an equivalent discussed in §II. BLE accords as when all baskets are identical $B = \{b\}$, $p(b) = 1$ and therefore BLE equals zero. Equally when all baskets are unique BLE is trivially maximised with all $b \in B$ having $p(b) = \frac{1}{|B|}$.

BRE accords, under any parameterization, as when all baskets share the same items under any parameterization, the entire repeated basket composition is always mined as a *common sub-basket* by the Apriori algorithm. Equally, when all baskets are unique, then there are no possible *common sub-baskets*, and by RULE 2, the original basket is always used. In both cases, the computation proceeds with the same input as BLE.

Further, Item Entropy (IE) violates $P1$, with the inclusion of systematic behaviour at the basket level having the potential to lead to an increase in item entropy in some cases. Consider the example of $\{m, b\}, \{b, p\}, \{b, p\}, \{b, p\}$ vs $\{m, b\}, \{m, b, p\}, \{m, b, p\}, \{m, b, p\}$, a clear increase the presence of systematic sub-baskets. Item Entropy in the first case (4x b, 3x p, 1x m) is 0.887 compared to after adding the systematic sub-basket: 0.992 (4x b, 3x p, 4x m).

BLE also violates $P1$. Consider another trivial example. Let $\{m, b, p\}, \{m, b, s\}, \{m, b, c\}, \{m, b, j\}$ denote a basket sequence. Increasing the presence of systematic sub-baskets ($P1$) will not change the fact that at a basket level all baskets will stay unique resulting in a maximal BLE score.

For BRE, the exact behaviour is dependent on the chosen parameter. Consider the case of

[2] We also note that other versions of normalisation could be included to achieve invariance to slightly different aspects or definitions of choice group sets. The definition in this work is chosen to align with that used in BLE and BRE to enable better evaluation of the proposed measure with regard to properties (P0-P2). Exploration of other variants is left as future work.

[3] additional rules exist for tie-breaking, see [9]

$\mathcal{B} = \{p, s\}, \{p, s\}, \{c, j\}, \{c, j\}$ vs. the case where $\{m, b\}$ has been systematically added to each basket to form $\mathcal{B}' = \{p, s, m, b\}, \{p, s, m, b\}, \{c, j, m, b\}, \{c, j, m, b\}$. When the *minsup* is set such that both $\{p, s\}$ and $\{c, j\}$ are mined as frequent patterns then in $\mathcal{B}$: $\{p, s\}$ becomes a distinct symbol *common sub-basket 1* (X), and $\{c, j\}$ becomes a distinct symbol *common sub-basket 2* (Y) and by RULE 1 and BRE is evaluated as the BLE of $\{X, X, Y, Y\}$. For the same *minsup* threshold for $\mathcal{B}'$, many two-item frequent patterns exist, but so do the longer *common sub-baskets* $\{p, s, m, b\}$ and $\{c, j, m, b\}$. By RULE1 it is these, and only these, that will then be selected to represent the baskets (as distinct symbols) and again BRE is evaluated as the BLE of $\{X, X, Y, Y\}$. This clearly violates $P1$ with no decrease reported by the measure.

Finally, considering $P2$. $P2$ is an extension of $P1$, effectively clarifying the expected behaviour based on the repeated application of $P1$. Measures failing $P1$ inherently cannot meet $P2$. As an immediate consequence Item Entropy, BLE and BRE cannot fully accord to $P2$.

V. Bundle Entropy

Having discussed the failure of Item Entropy, BLE and BRE to meet the properties we seek to practically describe human predictability, we now propose a new method, *Bundle Entropy* (BE), that does accord to properties $P0$ to $P2$. Bundle Entropy is realised as a conceptual extension of BLE. We define bundles as a collection (set) of products bought simultaneously. To expand on the definition of BLE we recast BLE's formulation as the (normalised) mean information for all baskets:
Let $\mathcal{B} = [b_0, b_1, \ldots, b_n]$ be the list of baskets for an individual, where baskets are sets of distinct items.
Let $B = set(\mathcal{B})$ and $p(b_k)$ denote the empirical probability of basket b_k given $\mathcal{B}$.
Then:

$$I(b_k) = -\log_2(p(b_k)) \tag{4}$$

Where $I(b_k)$ is the well known measure of self-information, measuring the amount of *surprise* we receive when b_k is observed given we expected b_k with probability $p(b_k)$. Given $I(b_k)$, BLE is then:

$$\begin{aligned} BLE(\mathcal{B}) &= \frac{1}{\log_2 |B|} \sum_{b \in B} p(b) I(b) \\ &= \frac{1}{\log_2 |B|} \times \frac{\sum_{b \in \mathcal{B}} I(b)}{|\mathcal{B}|} \end{aligned} \tag{5}$$

Note the distinction between B and $\mathcal{B}$ in the above. Ignoring the normalisation term, the final line highlights that non-normalised BLE represents the average amount of self-information over the observed data (typically assumed to be representative of population statistics).

Returning to the definition of self-information, we note that, given $p(b_k)$ is computed from an empirical probability, an alternative way of considering $I(b_k)$ is how surprised (and unhappy if one had taken action on) one might be if one predicted b_k and then observed the set of baskets ($\mathcal{B}$), getting the predictions (*exactly*) correct only $p(b_k) \times |\mathcal{B}|$ times[4]. So formally, the computation of $I(b_k)$ is based on the empirical probability, which can be written as:

$$p(b_k) = \frac{\sum_{b_q \in \mathcal{B}} \mathbb{1}(b_k = b_q)}{|\mathcal{B}|} \tag{6}$$

Leveraging this prediction point of view in the context of the desired concept of Basket Entropy, we note that in wanting to capture the predictability of sub-baskets, we wish to incorporate the fact that we would be happy with a prediction even if it is not *exactly* correct, by measuring the predictions partial worth based on some measure of expected utility. Assuming utility is gained from the correct prediction of sub-baskets, with utility increased proportionally to the relative size of the sub-basket correctly predicted, and given we consider baskets as sets of items, such a measure corresponds to a set similarity measure such as Jaccard [34] or Overlap [35]. Notably, these measures match the *exact match similarity function* in the extremes (zero: no partial match, one: exact match). In this work, we propose the use of a variant of the Overlap coefficient, specifically:

$$\mathcal{S}(b_k, b_q) = \frac{|b_k \cap b_q|}{max(|b_k|, |b_q|)} \tag{7}$$

Similarly to Jaccard and Overlap the measure is defined as the proportion of items shared between the predicted and truth sets with the numerator being the number of shared items between these sets. Differing is the denominator - Overlap is the proportion of the smaller basket ($\frac{|b_k \cap b_q|}{min(|b_k|, |b_q|)}$) and fails to penalise, for instance, over-predictions. In contrast, Jaccard sets the denominator to be the number of matched plus unmatched items between the prediction and truth ($\frac{|b_k \cap b_q|}{|b_k \cup b_q|}$). These double counts the incorrect prediction of an item as the incorrectly predicted symbol and the true symbol are both unmatched. Ensuring over-predictions are penalised while only counting incorrect predictions once results in the overlap variant in Equation 7.

Substituting this measure (Equation 7) into Equation 4 via Equation 6 by replacing the exact match indicator function ($\mathbb{1}(b_k = b_q)$) with the similarity function ($\frac{|b_k \cap b_q|}{max(|b_k|, |b_q|)}$) we get our alternative definition of bundle self-information, information like measure which we term *regret* ($R(b_k)$). The interpretation of this measure is no longer one of how surprised one is when b_k is observed, but a measure quantifying how much regret one might feel if they had assumed it was going to be b_k.

$$\mathcal{R}(b_k) = -\log_2 \left(\frac{\sum_{b_q \in \mathcal{B}} \frac{|b_k \cap b_q|}{max(|b_k|, |b_q|)}}{|\mathcal{B}|} \right)$$

$$\mathcal{R}(b_k) = -\log_2 \left(\sum_{b_q \in B} p(b_q) \frac{|b_k \cap b_q|}{max(|b_k|, |b_q|)} \right) \tag{8}$$

[4] Alternatively if one assumes the empirical $p(\cdot)$ is the true generative distribution then $I(b_k)$ is how unhappy one would be on average if one predicted b_k over an infinite amount of time.

The proposed regret based bundle entropy is therefore defined as:

$$BE(\mathcal{B}) = \frac{1}{\log_2 |B|} \times \sum_{b_k \in B} p(b_k)R(b_k) \qquad (9)$$

The measure meets all properties, $P0 - P2$. Proofs are provided in Appendix A. As $BE(\mathcal{B})$ accords to these properties it performs as expected in the aforementioned examples, specifically as shown in Figure **??**

VI. Evaluation

The section is divided into three parts. The first illustrates empirically how bundle entropy meets the desirable properties described in §III) while the others do not. The second demonstrates empirically that the proposed measure differs noticeably from each other, highlighting that the choice of measure will lead to differences in conclusions when used by practitioners. The final part considers some exemplar analyses, demonstrating the utility of the measure and replicating and comparing to the relevant parts of a case study from [9].

Each part compares and contrasts bundle entropy with the related measures of consumers' buying behaviour predictability previously identified: Item Level Entropy (IE), Bundle Level Entropy (BLE), and Bundle Revealed Entropy (BRE). We used three different parametrizations of *minsup* for BRE, 10%, 24%, and 70%. This value directly affects the set of *common sub-baskets* that are used to represent the purchase history (see §IV for more information). A *Minsup* of 24% is selected as it is the value that [9] recommends. However, since our data is different in size and context, we also test BRE with *minsup* of 10% and 70%.

The evaluations are based on two different, real-world, mass transactional data sets. The first is *Dunnhumby - The complete Journey* a freely available[5] data set. The data set includes grocery purchases at a household level over two years from 2,500 frequent shoppers, providing a cohort for tracking systematic choices over time. The data set contains over 2.5 million records of 'what', 'how much', 'where', and 'when' each transaction was made by each household. All code to replicate experiments presented on this data set has been made available[6]. The second data set is a large transactional data set from 2,181 loyalty card holders over 20 months (between 2014 and 2016) from a large UK grocery retailer. Similarly to the Dunnhumby data set, the data set records of 'what', 'how much', 'where', and 'when' each transaction with transactions linked to a customer via their loyalty card[7].

A. Quasi-synthetic Data

This section empirically considers how BE, BLE, BRE, and IE accord to the desired properties P1 and P2 listed in §III. P0 is not considered as it stipulates desired edge cases to which IE clearly does not meet as it measures predictability of a different conceptual level (items rather than baskets), with the remaining measures all meeting the property.

P1 states that baskets with systematic sub-baskets should result in a lower score than those without. We investigate the performance of the measures based on this by adding systematic sub-baskets to each basket in each household from the Dunnhumby data set. For measures that accord to property 1, these values will consistently strictly[8] lower bound the measure computed from the original basket set. Overall results are shown in Table I, indicating the percentage of households, per measure, accorded to this property. As expected, the proposed bundle entropy measures always and accords while item entropy typically accords with minor exceptions. In contrast, BLE never accords due to a failure with respect to $P1$, with the addition of the systematic baskets not altering the number of unique baskets and hence not lowering the BLE score. As discussed in §III, the violation of $P1$ by BRE is data and threshold dependent, and the results in Table I column 2 show that violations are not uncommon in practice, though the violations only appear to occur at lower *minsup* levels. Investigating this behaviour further, scores for individual households were considered and three are shown as illustrative examples in Figure 2. The results are generally as expected. An exception is the scores when adding systematic sub-baskets (lower orange dots) for BRE with a *minsup*=70% (and household 2 in BRE with a *minsup*=24%). In these cases, we see the introduction of systematic sub-baskets incorrectly, causing the household to be considered 100% predictable. This is due to the systematic sub-basket (that was added to each original basket) representing all baskets in the BRE algorithm. As such, while the measure strictly holds to the $P1$ property it does so in a degenerate way. To quantify the extent of this effect we compute the number of times this occurs as a percentage of all households for each measure. This is shown in column three of Table I. The results clearly highlight that as the *minsup* of BRE is increased helping it to meet $P1$ it overwhelmingly does so in this degenerate way significantly degrading the utility of the measure.

TABLE I: Measures vs. Properties 0 & 1 and the percentage of households considered as fully predictable.

	Property accorded to		% Households measure considered fully predictable
Measures	**P0**	**P1 (% Households)**	
Bundle entropy	✓	100.0	0.0
Entropy	✗	99.0	0.0
BLE	✓	0.0	0.0
BRE 10%	✓	70.9	5.2
BRE 24%	✓	63.2	5.1
BRE 70%	✓	99.8	98.8

P2 states that sequences with larger systematic sub-baskets should result in a lower score than smaller sub-baskets relative to basket size (and vice versa). To investigate the empirical

[5] https://www.dunnhumby.com/source-files/

[6] https://github.com/nlab-admin/bundle-entropy

[7] This second data set is unavailable for public release

[8] Unless an individual's basket sequence is already maximally predictable meeting $P0$, something that did not occur in this data set. See the definition of $P1$ in §III for more information.

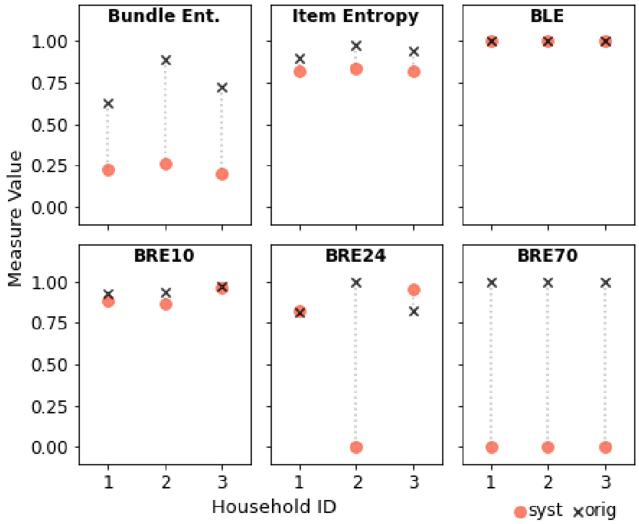


Fig. 2: Illustrative examples of three household's scores for the evaluated measures when adding systematic bundles to the household's purchases.

performance of the measures regarding this property, we selected all the baskets from a random sample of 1,000 households[9] and incrementally added systematic bundles of different sizes (from one to ten items) to each household's baskets. After every iteration, we computed the mean score per measure across all households. These results are shown in Figure 3. Once again, as expected, bundle entropy accords to P2 decreasing as the size of the systematic bundle added to each household's basket is increased.

In contrast, BLE, BRE10, BRE24, and BRE70 remain indifferent to the added bundles at different score levels. This indifference can be explained by further considering how the *minsup* affects the mining of common sub-baskets considering the new sub-basket component that has to be introduced to all baskets at each (x-axis) point. In the case of BRE70 the threshold is high enough that no sub-baskets are being found in the original data and as soon as the sub-basket of length 1 is introduced this is almost invariably mined as the common basket. Subsequently, all baskets are then represented by this making the behaviour appear almost completely predictable. Conversely when a *minsup* of 24% or 10% is set other sub-baskets that were already considered common are extended with these new sub-basket components. These extended sub-baskets are now typically longer than the original but, given they exist within the same proportion of baskets, have identical support. This results in no changes to the symbol set and the baskets are then mapped to and subsequently, no change to the score computed in the subsequent entropy calculation. Interestingly, entropy is the only one with similar behaviour to bundle entropy. However, this is a consequence of its item-level aggregate approach, where adding larger systematic bundles also increases the overall probabilities of each item added, lowering the overall entropy. We reiterate that conceptually in this case what is being measured is significantly different.

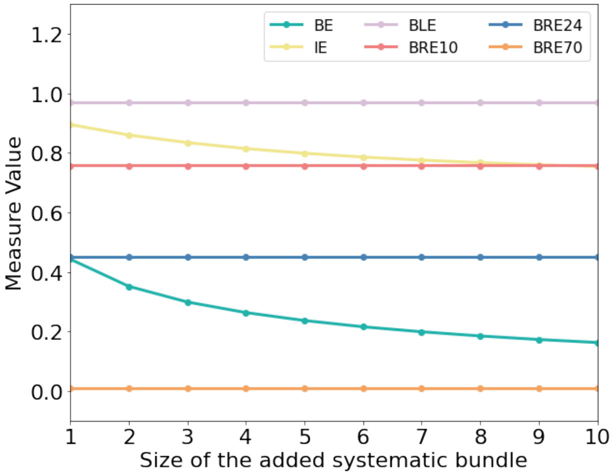


Fig. 3: Comparing measures by increasing the size of systematic bundles added to all baskets.

B. Rank similarities in practice

Having shown that the proposed measure accords in theory and practice with the desired properties while other measures do not, we now demonstrate that the selection of BE over the other measures will have a notable real-world impact on analysis and subsequent actions. To demonstrate this, we consider how similarly the measures rank households and customers in our two real-world data sets. This is achieved by computing all previously discussed measures/measure variants (BE, IE, BLE, BRE10, BRE24, BRE75) for all households (the Dunnhumby data set) and customers (the Large UK grocery retailer data set). For each measure/variant and data set pair, a ranked list of the households/customers according to the measure is then generated. For each data set, we then make pairwise comparisons of all lists computing the Kendall Tau Rank Agreement and the Mean Rank Difference.

The Kendall Tau Rank Agreement indicates the difference between the probability that pairs of households (customers) will be in the same rank order according to both measures and the probability that the pairs will have a different rank order [36]. The Mean Rank Difference provides a similar, more conceptually straightforward indication of rank similarity, measured by (1) matching the two lists by households (customers), (2) taking the differences in rank before (3) computing the mean of these differences. The results are shown in Figure 4.

The results, as expected, show the measures are all related to some degree, though the measures are noticeably data set dependent. Also, as expected (see IV), BRE with a high *minsup* (70%) is highly correlated with the BLE measure. While the measures show notable agreement, they also highlight non-trivial differences, with the proposed BE measure differing in mean rank difference by 194 to 560 (out of 2,213 households) for Dunnumby and 287 to 411 (out of 2,181 customers)

[9] A large sample was taken due to computational costs.

for the large UK grocery retailer. This evidences that the measures' different properties lead to differences in practice, with the potential to arrive at different conclusions within any analytics based on them. The results highlight differences between the proposed measure, with its clear interpretation and theoretical properties, and the others across the different parameterizations of BRE. This indicates the sensitivity of the BRE measure to the *minsup* parameter, with different choices able to directly influence the outcome of any analytics, making its selection and motivation crucial to any interpretation and/or subsequent action.

C. Case Study / Exemplar Analysis

Finally, we consider the utility of the proposed measure in practice. The utility of a measure for understanding the predictability (or systematic nature) of an individual's baskets can be grouped into two main groups. The first is its use as an explanatory variable to describe an individual and/or segment. For instance, if one knew that an individual (or segment) was highly predictable, then one may appeal to this regularity in the wording of any communication. This requires the measure to match the practitioner's understanding/intuition of the measure and has motivated the evaluation of the previous sections.

The second group of analysis uses such measures to evidence a relationship between the measure, potentially in conjunction with other measures, and an output variable within a predictive framework. This could be driven by commercial or other imperatives (e.g. social good, consumer welfare). Such a use reflects the use case explored in [9] which analysed the relationship between systematic customer behaviour[10] and profitability within the supermarket retail setting.

To consider the relationship between basket predictability and profitability, [9] consider fit a single variable linear regression model for two variables (average basket spend and the number of visits[11]) and report the equivalent of a Pearson Correlation of -0.3253 and -0.3249. Based on this, the authors conclude that "predictable systematic customers are more profitable for a supermarket: their average per capita expenditures are higher than non-systematic customers". In order to demonstrate the similar utility of the proposed BE measure and compare the results of its use against the BRE measure, we compute the Pearson Correlation for all measures and average basket spend and the number of visits for both the Dunnhumby and UK grocery retail data sets. In addition, we consider the correlation of the measures with another measure of performance relevant to the fast-moving consumer goods industry - individual's average spend per month (an indication of potential lifetime value). The results are shown in Table II.

The results of BE lend weight to the conclusions of [9] with negative correlations observed for both *mean basket spend* and *number of visits* for both data sets. Depending on the parameterization, BRE indicates positive or negative relationships between the parameterised BRE measure for *mean basket spend* for both data sets and *number of visits* for the Dunnhumby data set. This once again highlights that the interpretation/meaning of the BRE measure is bound to the *minsup* threshold. We highlight again that in addition to complicating any interpretation, setting this parameter is non-trivial. Considering both item level and basket level entropy, we see that they clearly measure conceptually different concepts with results indicating a non-significant correlation with *mean basket spend* for Dunnhumby and only a small positive correlation for BLE for the second data set.

Perhaps notably, all measures for all parameterizations have a negative relationship with *mean spend per month*, although the parameterization of BRE alters the relative strength of the relationships inconsistently across the data sets. This, even if BRE is ignored, indicates that *mean spend per month*, an indicator of lifetime value, is linked to basket/item predictability more generally, both sharing information regarding and across the item/basket levels an observation that has the potential to inform future work.

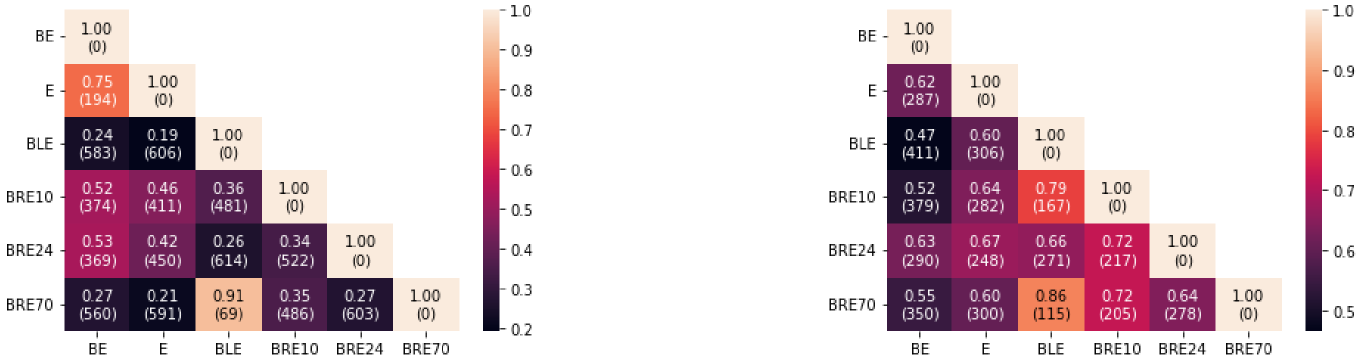
Differing purchase behaviour types across the data are clear from the differences in the two data sets, confirming the need for the measures to have clear interpretations and theoretical underpinnings. Concerning the proposed measure, BE, the precise interpretation of basket predictability, and its stable theoretical underpinnings provide a reasonable basis for future work to further consider its use as either an explanatory variable or predictor for all three factors considered within the supermarket domain, given the consistency of the results. The results are consistent across both data sets (significant negative correlations), and the relative magnitude is consistent with respect to the factors in both data sets. Given the factors are likely related, this potentially speaks to the different target markets of the two supermarkets. This is in comparison to either item or basket level entropy, which either shows limited / non-significant correlation (IE / BLE), inconsistent relationships, or inconsistent relative magnitudes depending on the parameterization (BRE). Examples of the latter include BRE24 showing a positive relationship with *number of visits* for the Dunnhumby data set but a negative correlation in the second data set and for all other *minsup* parameters in both.

VII. Conclusion

In this work, we address quantifying the predictability of human purchasing behaviour by introducing a novel measure, *Bundle Entropy*. Motivated by the failure of existing methods to accord with intuition on simple examples, the work developed a set of simple properties such measures should meet, noting the failure of existing methods to meet them theoretically and empirically. Bundle Entropy was then developed to meet these properties. The new measure was then compared empirically to real-world, large-scale grocery store transactional data sets. The results demonstrated that (1) that the proposed measure accords, in theory, and practice to the desired properties while the others do not, (2) the measures

[10] In this work we replicate the evaluation concerning their proposed BRE measure, omitting evaluation which relates to the complementary Spatio-temporal measure that could be equally used in-conjunction with Bundle Entropy as proposed in this work.

[11] [9] refer to this as expenditure and baskets respectively in their figures.



(a) Dunnhumby

(b) Large UK grocery retailer

Fig. 4: Kendall Tau Rank Agreement (Mean Rank Difference) of relative household/customer *predictability* for pairs of measures.

	Correlation with:					
	Dunnhumby			UK grocery retailer		
	Mean Basket Spend	Mean Spend per Month	Number of Visits	Mean Basket Spend	Mean Spend per Month	Number of Visits
BE	-0.187^*	-0.475^*	-0.374^*	-0.215^*	-0.401^*	-0.371^*
BRE 10%	0.009	-0.494^*	-0.439^*	0.066^*	-0.226^*	-0.365^*
BRE 24%	-0.290^*	-0.340^*	-0.108^*	-0.001	-0.262^*	-0.342^*
BRE 70%	-0.134^*	-0.268^*	-0.182^*	0.002	-0.152^*	-0.223^*
Item Entropy	0.027	-0.380^*	-0.456^*	0.000	-0.236^*	-0.315^*
BLE	-0.034	-0.268^*	-0.323^*	0.088^*	-0.138^*	-0.270^*

TABLE II: Pearson Correlation between the measures and spending and visiting factors. * denotes statistical significance. $p - values$ were adjusted using the Benjamini–Hochberg false discovery procedure with a q-value of 0.05 [37].

are notably different and should not be used interchangeably, and (3) the proposed measure has higher utility in practice, providing a consistent, parameter-less measure that accords to well-defined, intuitive properties allowing practitioners to efficiently and correctly interpret and action the outcome of analytics and insights based on the measure.

Appendix

A. Proofs that bundle entropy meets properties P0-P2.

Given $BE(\mathcal{B})$ is defined as:

$$\frac{1}{\log_2|B|} \times \sum_{b_k\in B} p(b_k)\left[-\log_2\left(\sum_{b_q\in B} p(b_q)\frac{|b_k\cap b_q|}{max(|b_k|,|b_q|)}\right)\right]$$

P0.a: When $b_0 = b_1 = \ldots = b_n$ then $\frac{|b_k\cap b_q|}{max(|b_k|,|b_1|)} = 1$ resulting in $BE(\mathcal{B}) = 0$.

P0.b: When $b_0 \cap b_1 \cap \ldots \cap b_n = \emptyset$ then:

$$BE(\mathcal{B}) = \sum_{b_k\in B} p(b_k)[-log_2(p(b_k))]$$

As $|b_k \cap b_q| = 0$ except when $b = q$. As each b_k is unique with a probability of $\frac{1}{|B|}$, then the term excluding the normalisation term sums to $log_2|B|$ resulting in a value of 1.

P1: When $\mathcal{B}' = [b_0 \cup \Gamma, b_1 \cup \Gamma, \ldots,] = [b'_0, b'_1, \ldots,]$ where Γ is any non-zero arbitrary combination of items and $\exists b_k : \Gamma \not\subset b_k$.

Given: $p(b_k) = p(b'_k)$
And:

$$\frac{|b'_k \cap b'_q|}{max(|b'_k|,|b'_q|)} = \frac{|(b_k\cup\Gamma)\cap(b_q\cup\Gamma)|}{max(|b_k\cup\Gamma|,|b_q\cup\Gamma|)} \geq \frac{|b_k\cap b_q|}{max(|b_k|,|b_q|)}$$

With the inequality strict when $\Gamma \not\subset b_k$ and $b_k \neq b_q$ which by definition must be true at least once or all baskets are the same resulting in P0.a.
Via the summation and negative log and since $|B'| \leq |B|$ as baskets are represented as sets and adding identical sets to all sets in an existing collection of sets can only reduce the number of distinct sets in the collection, then $BE(\mathcal{B}') < BE(\mathcal{B})$.

P2: Holds via the $P1$ proof, mapping $b_0 \cup \Gamma$ to b_0 & Γ' to Γ.

Acknowledgement

This research was supported by the EPSRC project "Neo-demographics: Opening Developing World Markets by Using Personal Data and Collaboration" EP/L021080/1, the EPSRC project "Trusted Data-Driven Products" EP /T022493/1 and the EPSRC project "The CIVIC Project: A Sustainable Platform for COVID-19 syndromic-surveillance via Health, Deprivation and Mass Loyalty-Card Datasets" EP/V053922/1.

References

[1] C. Song, Z. Qu, N. Blumm, and A. L. Barabási, “Limits of predictability in human mobility,” *Science*, vol. 327, no. 5968, pp. 1018–1021, 2 2010.

[2] G. Smith, R. Wieser, J. Goulding, and D. Barrack, “A refined limit on the predictability of human mobility,” *2014 IEEE International Conference on Pervasive Computing and Communications, PerCom 2014*, pp. 88–94, 2014.

[3] J. Krumm, “Ubiquitous advertising: The killer application for the 21st century,” *IEEE Pervasive Computing*, vol. 10, no. 1, pp. 66–73, 2010.

[4] B. Y. Jung, M. S. Choi, H. Y. Youn, and O. Song, “Vertical handover based on the prediction of mobility of mobile node,” in *2010 8th IEEE International Conference on Pervasive Computing and Communications Workshops, PERCOM Workshops 2010*, 2010, pp. 534–539.

[5] J. Froehlich and J. Krumm, “Route Prediction from Trip Observations,” in *Society of Automotive Engineers (SAE) 2008 World Congress, April 2008*, 2008.

[6] M. A. Hossain, S. Akter, and V. Yanamandram, “Revisiting customer analytics capability for data-driven retailing,” *Journal of Retailing and Consumer Services*, vol. 56, 9 2020.

[7] G. R. Foxall, “Foundations of consumer behaviour analysis,” *Marketing Theory*, vol. 1, no. 2, pp. 165–199, 2001.

[8] Y. T. Wen, P. W. Yeh, T. H. Tsai, W. C. Peng, and H. H. Shuai, “Customer purchase behavior prediction from payment datasets,” *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pp. 628–636, 2018.

[9] R. Guidotti, M. Coscia, D. Pedreschi, and D. Pennacchioli, “Behavioral entropy and profitability in retail,” *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015*, pp. 1–10, 2015.

[10] S. M. Straathof, “Shannon’s entropy as an index of product variety,” *Economics Letters*, vol. 94, no. 2, pp. 297–303, 2007.

[11] L. Jost, “Entropy and diversity,” *Oikos*, vol. 113, no. 2, pp. 363–375, 2006.

[12] D. V. Budescu and M. Budescu, “How to measure diversity when you must,” *Psychological Methods*, vol. 17, no. 2, pp. 215–227, 2012.

[13] A. S. Ehrenberg, “Repeat-Buying,” in *Repeat-Buying: facts, theory and applications*, 1988, pp. 31–78.

[14] G. A. Frisbie, “Ehrenberg’s Negative Binomial Model Applied to Grocery Store Trips,” *Journal of Marketing Research*, vol. 17, no. 3, pp. 385–930, 1980.

[15] G. Goodhardt, A. Ehrenberg, and C. Chatfield, “The Dirichlet : A Comprehensive Model of Buying Behaviour,” *Journal of the Royal Statistical Society. Series A (General)*, vol. 147, no. 5, pp. 621–655, 1984.

[16] G. M. Allenby and P. J. Lenk, “Modeling household purchase behavior with logistic normal regression,” *Journal of the American Statistical Association*, vol. 89, no. 428, pp. 1218–1231, 1994.

[17] P. S. Fader and D. C. Schmittlein, “Excess Behavioral Loyalty for High-Share Brands: Deviations from the Dirichlet Model for Repeat Purchasing,” *Journal of Marketing Research*, vol. 30, no. 4, pp. 478–493, 1993.

[18] M. Uncles, A. Ehrenberg, and K. Hammond, “Patterns of Buyer Behavior : Regularities , Models , and Extensions,” *Marketing Science*, vol. 14, no. 3, pp. G71–G78, 1995.

[19] M. Uncles and K. Hammond, “Grocery store patronage,” *The International Review of Retail, Distribution and Consumer Research*, vol. 5, no. 3, pp. 287–302, 1995.

[20] G. J. Russell and W. A. Kamakura, “Modeling multiple category brand preference with household basket data,” *Journal of Retailing*, vol. 73, no. 4, pp. 439–461, 1997.

[21] C. B. Bhattacharya, “Is your brand’s loyalty too much, too little, or just right?: Explaining deviations in loyalty from the Dirichlet norm,” *International Journal of Research in Marketing*, vol. 14, no. 5, pp. 421–435, 1997.

[22] B. Sharp and A. Sharp, “Loyalty programs and their impact on repeat-purchase loyalty patterns,” *International Journal of Research in Marketing*, vol. 14, no. 5, pp. 473–486, 1997.

[23] G. J. Russell and A. Petersen, “Analysis of cross category dependence in market basket selection,” *Journal of Retailing*, vol. 76, no. 3, pp. 367–392, 2000.

[24] B. Sharp, M. Wright, J. Dawes, C. Driesener, L. Meyer-Waarden, L. Stocchi, and P. Stern, “It’s a dirichlet world: Modeling individuals’ loyalties reveals how brands compete, grow, and decline,” *Journal of Advertising Research*, vol. 52, no. 2, pp. 203–213, 2012.

[25] E. Kim, W. Kim, and Y. Lee, “Combination of multiple classifiers for the customer’s purchase behavior prediction,” *Decision Support Systems*, vol. 34, no. 2, pp. 167–175, 2003.

[26] D. Van Den Poel and W. Buckinx, “Predicting online-purchasing behaviour,” *European Journal of Operational Research*, vol. 166, no. 2, pp. 557–575, 10 2005.

[27] C. Lo, D. Frankowski, and J. Leskovec, “Understanding behaviors that lead to purchasing: A case study of pinterest,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016. Association for Computing Machinery, 8 2016, pp. 531–540.

[28] B. E. Kahn and D. R. Lehmann, “Modeling Choice Among Assortments,” *Journal of Retailing*, vol. 67, no. 3, pp. 274–299, 1991.

[29] H. Akaika, *Prediction and entropy*. Springer US, 1985.

[30] P. J. Alexander, “Product variety and market structure: A new measure and a simple test,” *Journal of Economic Behavior and Organization*, vol. 32, no. 2, pp. 207–214, 1997.

[31] H. Li and M. G. Russell, “The Impact of Perceived Channel Utilities, Shopping Orientations, and Demographics on the Consumer’s Online Buying Behavior,” *Journal of Computer-Mediated Communication*, vol. 5, no. 2, 1999.

[32] S. Bellman, G. L. Lohse, and E. J. Johnson, “Predictors of online buying behavior,” *Communications of the ACM*, vol. 42, no. 12, pp. 32–38, 1999. [Online]. Available: www.gvu.gatech.edu/

[33] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules,” in *Proc. of 20th International Conference on Very Large Data Bases, VLDB.*, 1994, pp. 487–499. [Online]. Available: citeseer.ist.psu.edu/agrawal94fast.html

[34] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, “Using of jaccard coefficient for keywords similarity,” *Lecture Notes in Engineering and Computer Science*, vol. 2202, pp. 380–384, 2013.

[35] L. R. Lawlor, “Overlap, similarity, and competition coefficients,” *Ecology*, vol. 61, no. 2, pp. 245–251, 1980.

[36] H. Abdi, “Kendall Rank Correlation Coefficient,” *The Concise Encyclopedia of Statistics*, vol. 2, pp. 508–510, 2007.

[37] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Source: Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.