Behavioral Entropy and Profitability in Retail

Riccardo Guidotti[1,2], Michele Coscia[3], Dino Pedreschi[2], Diego Pennacchioli[1]

[1] KDDLab ISTI-CNR, Via G. Moruzzi, 1, Pisa, Italy, Email: {name.surname}@isti.cnr.it
[2] KDDLab University of Pisa, Largo B. Pontecorvo, 3, Pisa, Italy, Email: pedre@di.unipi.it
[3] CID - Harvard Kennedy School, 79 JFK Street, Cambridge, MA, US, Email: michele_coscia@hks.harvard.edu

***Abstract*—Human behavior is predictable in principle: people are systematic in their everyday choices. This predictability can be used to plan events and infrastructure, both for the public good and for private gains. In this paper we investigate the largely unexplored relationship between the systematic behavior of a customer and its profitability for a retail company. We estimate a customer's behavioral entropy over two dimensions: the basket entropy is the variety of what customers buy, and the spatio-temporal entropy is the spatial and temporal variety of their shopping sessions. To estimate the basket and the spatio-temporal entropy we use data mining and information theoretic techniques. We find that predictable systematic customers are more profitable for a supermarket: their average per capita expenditures are higher than non systematic customers and they visit the shops more often. However, this higher individual profitability is masked by its overall level. The highly systematic customers are a minority of the customer set. As a consequence, the total amount of revenues they generate is small. We suggest that favoring a systematic behavior in their customers might be a good strategy for supermarkets to increase revenue. These results are based on data coming from a large Italian supermarket chain, including more than 50 thousand customers visiting 23 shops to purchase more than 80 thousand distinct products.**

I. Introduction

To some extent, human behavior is predictable. Humans do not change their behavior randomly from one day to the other and their patterns usually follow a given routine. This is true at the crowd level: groups of humans flock together in predictable patterns. For instance, people are more mobile early in the morning and late in the afternoon, around the working day, creating an M-shaped pattern. But humans are also predictable at the individual level. Bursty patterns of activities have been observed and can be predicted, for instance in writing e-mails. Also individual mobility is predictable: most people will commute every working day between the same two points, and can be predicted to do so with very high accuracy [3], [25].

Predicting the behavior of a particular set of humans, customers, is of great value for a commercial enterprise. Knowing that each customer is systematic, or not, in her behavior might have important consequences for sales. Note that a shop can be interested in different dimensions of customer behavior. Up until now we have discussed examples of studies focused on spatio-temporal variables, but a retail shop has other information about customers, namely what they put in their baskets. Customers might not be predictable in the time of the day they visit the shop, but they might be highly predictable in that they always purchase the same products. If a systematic customer is more valuable because she spends more, then the shop might want to encourage more and more people to be systematic. On the other hand, if no link is found, then the shop is better off using classical marketing strategies.

In this paper, we find evidence suggesting that the former hypothesis might be true: systematic customers visit the shop more, buy more products and spend more money in the shop. Customers that can be classified as systematic, both in spatio-temporal and in basket composition variables, have higher average expenditure per capita when compared to customers that are not systematic, or systematic only in one dimension. However, we also find that the total volume of revenue that these systematic customers generate for the shop is not high. In fact, the truly systematic customers are a small minority of the entire customer base of the shop. In our data, they are barely composing less than 7% of the total customer set. We are also able to characterize their typical basket, which is composed mainly by fruit, vegetables and generally perishable products.

We estimate a customer's behavior using the information theoretic concept of entropy. Given a customer, we have the composition of all the baskets she purchased during our observation period. We are able to associate the basket to a specific basket pattern, in a one-to-one fashion: each basket has one classification. We then calculate the entropy of the baskets by looking at the probability of appearance of each pattern. This is a measure of how unpredictable a customer's basket is, and we call it Basket Revealed Entropy (BRE). Similarly, we can calculate a customer's spatio-temporal entropy. Every basket she purchased has been originated from a particular shop, in a particular day of the week and in a particular time. All these patterns have a probability of appearance for each customers, and therefore we can use again the information theory formulation of entropy to assess the customer's systematic behavior. In this case, we calculate how unexpected each customer shopping session is, and we call this measure the Spatio-Temporal Revealed Entropy (STRE).

To detect the patterns, needed for the basket classification in the BRE measure, we use the frequent itemset mining technique. Each basket is represented as an itemset, and the frequent products that co-appear in baskets are the patterns we use to label the basket. The basket is labeled with the largest possible frequent pattern that it contains. There is no need to apply itemset mining to the spatio temporal indicators, because they are always composed by three items, so we can simply calculate their relative frequency. In the paper, we employ other data mining techniques to classify the customers. Once we calculated the Basket Revealed Entropy and the Spatio-Temporal Revealed Entropy for all customers, we can classify them according to these values. We use the k-means clustering algorithm to group customers in this two dimensional space.

The analysis included in this paper are based on real world data. The provider of data is UniCoop Tirreno, which is one of the largest Italian supermarket chain. The chain has millions of customers and it sells hundreds of thousands of distinct products. In this work, we focus on a single Italian province, because in this area the company's market penetration can ensure a total representativeness of our sample. The company records the behavior of the customers that are identifiable because they associate each of their shopping session to their fidelity card. Therefore, given a customer, we can identify the time and place of each of her shopping sessions, and the composition of the basket she purchased. In the dataset of the selected Leghorn province, we have more than 50 thousand customers visiting 23 shops to purchase more than 80 thousand distinct products. We preserved the privacy of each customer by using anonymous IDs and by exclusively publishing only aggregate patterns that cannot be associated to any individual.

The rest of the paper is organized as follows. Section II discusses related literature to this paper. We present the details of the data used in Section III. The methodology is described in Section IV, including the formulation of the BRE and STRE measures, while Section V presents our case study. Discussion and future works are reported in Section VI.

II. Related Work

This paper is focused on the estimation of human predictability in a retail scenario. Human predictability is a vast research field, tackled with a number of approaches and for a number of different reasons. An accessible source for the general human behavior prediction task can be found in Barabasi's book "Bursts" [3]. In the book, Barabasi presents the theory that human behavior is bursty, i.e. humans have long inactivity intervals separated by moments of rapid activity. This theory has been tested repeatedly recently [2], [24].

As for the retail scenario, there are a number of works trying to predict customer behavior. One of the classic approach is to use data mining [1], [10], as it is very difficult to create a comprehensive model of overall customer behavior, as each single individual acts according to a very nuanced and personal utility function. Multiplex approaches are then used [4], [15]. However, recent research showed that it is possible to describe the retail market as a complex system [16].

These works focus on the detection of regularities in what customers buy. Another promising line of research investigates where customers go to buy what, i.e. how much the shops they visit are predictable [12] and how much they are willing to travel to satisfy their needs [7], [17]. In both cases, customers are shown to be rather predictable in their movements. In this work we improve over the state of the art discussed so far, by combining both dimensions: we evaluate customer predictability in what they buy and in where they buy it.

The mobility dimension is very important for two reasons. First, it is highly predictable [14], [21]. Second, it is intimately linked with the social dimension. It has been shown that it is possible to predict the places an individual will visit because we know that their friends visit them, and that social ties are more easily created among people who travel to the same places [5], [6], [23], [25]. The predictability of the creation of new social ties by an individual is a classic problem in social network analysis, even in isolation from mobility [19], [13]. This literature is relevant, because it is easy to imagine that the social connections play a not negligible role in influencing the individuals in buying new products. This social contagion effect has been studied in multiple scenarios [18], [9].

In this paper we choose an intermediate and innovative approach to the problem of exploiting human predictability for the retail market. Instead of using a pure data mining approach such as the ones presented in this section, that can be summarized with the OLAP framework [11], we use the hybrid approach combining data mining with a more systemic view. We do not use the mined patterns directly, but we use them to construct systemic measures estimating the degree of an individual's predictability. On top of these systemic measures, we apply again a data mining step, identifying the main customer classes based on their predictability.

III. Data

Our data come from one of the main retail supermarket chains in Italy. The chain serves several millions customers across the Italian territory. The chain operates three different tiers of shops according to their size: *Iper* shops are the largest, the Italian equivalent of a US mall; *Super* are the middle level, a large supermarket; and *Small* is the smallest shop type, whose size is comparable to a dollar store.

Customers of the retail chain can obtain a fidelity card. Through the card, customers can get a discount. The company is able to tie each shopping session to the card. For each shopping session, or basket, the company knows:

- Which customer made the purchase;
- All single items composing the basket;
- The time and the day of the shopping session;
- In which shop the transaction happened.

The dataset including this information is the one used in this paper. For simplicity and data cleaning purposes, we perform a series of filters on this dataset. First, we select a single year. The analysis performed in this paper is based on all observations recorded during 2012. Second, we focus on a narrow area of operation. The supermarket company was founded in Leghorn and we consider exclusively the shops that are in this Italian province. We do so because the market penetration of the company in this province is so high that we can effectively say that all inhabitants of Leghorn are represented in the data. Finally, we drop all customer who did not perform at least a shopping session per month. The area around Leghorn has an high influx of tourists from other areas of Tuscany, so supermarket customers from other provinces might sporadically use their card in shops in Leghorn province, thus introducing noise in our estimates.

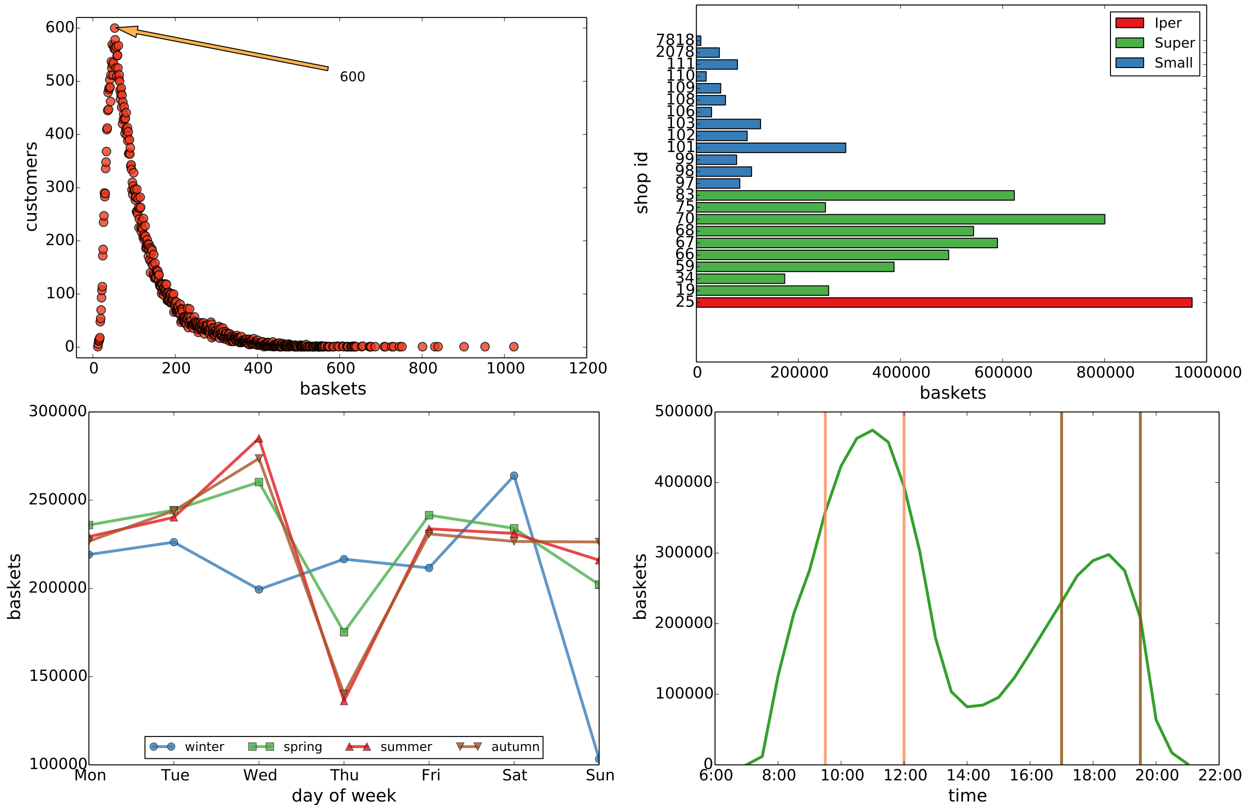

Fig. 1: Some facts customer behavior: distributions of baskets per customer (top left), distribution of baskets per shop (top right), distribution of baskets per weekday (bottom left) and distribution of baskets per time of the day (bottom right).

After this filter phase, we have 56,448 customers. Here, “customers” refers to customer cards. A card can be shared by an entire family. The province of Leghorn had a population of 343,003 in 2012. Assuming an average size of three people per household, we estimate that we cover at least 50% of the population. The total number of distinct products bought is 84,362. The total item scans in the dataset amounts to 71,172,672, and it has been generated from 23 shops.

Figure 1 depicts stylized facts about shopping sessions (baskets). Figure 1 top left: the number of baskets per customer. The mode is ~ 100, meaning that customers usually visit the shops around twice a week. The distribution does not follow the Zipf law because Leghorn does not have enough inhabitants to support it, since 50% or more of them are actually regulars. Figure 1 top right: the number of baskets per shop. Each of the 23 shops is represented here. There is a correlation between shop type and the number of customers it attracts. Figure 1 bottom left: the number of baskets per weekday. Customers have a remarkable preference for some days instead of others, also given the season. Fewer shopping sessions happen on Thursday, while Wednesday is the most popular day. Figure 1 bottom right: the number of baskets per time of the day. An M-shaped pattern appears: most shopping sessions happen in the morning or after working hours.

From Figure 1 we see that there are some general patterns in the customer behavior. Customers tend to shop twice a week, they are likely to be attracted to larger shops, they have favorite weekdays and time of the day to perform their shopping sessions. On these observations, we build our customer behavior entropy measures in the following section.

IV. Methodology

Our methodology aims at estimating the behavioral entropy of each customers. The two entropy measures are the Basket Revealed Entropy (BRE, Section IV-A) and the Shopping Time Revealed Entropy (STRE, Section IV-B). These measures tell us respectively how unpredictable is the basket composition and the visiting pattern of a given customer.

A. *Basket Revealed Entropy*

The objective of the mining step is to detect what are the behavioral patterns of a customer. There are two types of behavioral patterns in which we are interested: basket composition and spatio-temporal behavior. For the basket composition, we apply a frequent itemset mining algorithm [1]. For each customer, we apply the Apriori algorithm [22] on her baskets to detect her patterns. We drop the non-frequent patterns, i.e. the ones that are not present in at least $minsup$ baskets. Then, we assign each of her baskets to the largest pattern it contains. Note that each basket must be assigned to a pattern, and a pattern can classify multiple baskets.

To better understand the procedure, suppose that a customer visited the shop 5 times, purchasing these baskets:

1) {Cheese, Banana, Tomato, Bread}.
2) {Cheese, Banana, Tomato}.

3) {Cheese, Banana, Tomato, Coffee}.
4) {Cheese, Banana, Tomato, Bread}.
5) {Cheese, Meat, Shoes, Bread}.

For this example, we set the minimum support threshold to 3, i.e. each pattern has to be present in at least three baskets. Then, the mining algorithm will find the following patterns:

- Support = 5: {Cheese}.
- Support = 4: {Cheese, Banana, Tomato}, {Banana, Tomato}, {Cheese, Banana}, {Cheese, Tomato}, {Banana}, {Tomato}.
- Support = 3: {Bread, Cheese}, {Bread}.

We name those patterns *representative baskets*. In the following we use patterns and representative baskets as synonyms.

Finally, we classify baskets 1 to 4 with the {Cheese, Tomato, Banana} representative basket, because it is the longest pattern contained in them; and basket 5 with the representative basket {Cheese, Bread}. We now have a series of representative baskets with a given probability of appearance for our customer. The Basket Revealed Entropy (BRE) is calculated following the information-theoretic concept of entropy [20]:

$$BRE(RB) = \frac{-\sum_{i=1}^{n} \mathrm{f}(rb_i) \log \mathrm{f}(rb_i)}{\log n}$$

where RB is the set of representative baskets of our customer, rb_i is the i-th representative basket frequency (i.e. number of occurrences), $\mathrm{f}(rb_i)$ is the representative basket's relative frequency and $n = |RB|$ is the number of representative baskets. BRE takes values between 0 and 1, as it is normalized with $\log n$, that is the expected entropy of a fully random set of patterns. In our example we have only two representative baskets, with relative frequencies $4/5$ and $1/5$. Thus, the BRE of our hypothetical customer is ~ 0.72.

B. Spatio-Temporal Revealed Entropy

The calculation of the Spatio-Temporal Revealed Entropy (STRE) is similar in spirit to the procedure outlined in the previous section. However, the first computational step is easier. Here, we connect each basket to its spatio temporal characteristics. These characteristics are always represented by a tuple of three elements: the shop in which the basket was purchased (which provide the spatial dimension), the time of the day and the day of the week (the temporal dimension). Since all tuples always have three elements, we do not need to perform a mining step, and we can just count the relative frequency of each possible tuple. The relative frequencies are then fed into the information theoretic entropy formula.

Let us consider again a hypothetical customer. Her five shopping sessions happened in this order:

1) Shop 25, Weekend, Evening.
2) Shop 19, Weekday, Late Afternoon.
3) Shop 19, Weekday, Late Morning.
4) Shop 19, Weekday, Late Afternoon.
5) Shop 19, Weekday, Early Morning.

We have four patterns, three with probabilities $1/5$ and one with probability $2/5$, which results in an entropy ~ 0.96. Note that we aggregate days in two bins, weekday and weekends, as keeping days separate would generate too many fluctuations.

Algorithm 1 BRE($baskets$, $minsup$)

```
$IS \leftarrow$ getItemSet($baskets, minsup$)
$RB \leftarrow$ getReprBasketsFreq($baskets, IS$)
$bre \leftarrow -\sum_{rb \in RB} \mathrm{f}(rb_i) \log \mathrm{f}(rb_i) / \log(|RB|)$
return $bre$
```

Algorithm 2 getReprBasketsCount($baskets$, IS)

```
$RB \leftarrow \emptyset$
for $b \in baskets$ do
    $D \leftarrow \{rb \in IS \mid rb \subseteq b\}$
    if $D = \emptyset$ then
        $RB_b \leftarrow 1$
        continue
    $D' \leftarrow \mathrm{argmax}_{rb \in D} |rb \cap b|$
    if $|D'| = 1 \wedge D' = \{rb\}$ then
        $RB_{rb} \leftarrow RB_{rb} + 1$
        continue
    $D'' \leftarrow \mathrm{argmax}_{rb \in D'} \, sup(rb)$
    if $|D''| = 1 \wedge D'' = \{rb\}$ then
        $RB_{rb} \leftarrow RB_{rb} + 1$
        continue
    $D''' \leftarrow \mathrm{argmin}_{rb \in D''} \, lift(rb)$
    if $|D'''| = 1 \wedge D''' = \{rb\}$ then
        $RB_{rb} \leftarrow RB_{rb} + 1$
        continue
    for $rb \in D'''$ do
        $RB_{rb} \leftarrow RB_{rb} + \frac{1}{|D'''|}$
return $RB$
```

C. Pseudocode

In this section we provide and discuss the pseudocode of our analytic framework. We do it for two reasons: (i) to ensure a better understanding of our analytical workflow and (ii) sharing the pseudocode will favor modifying it for custom applications. We focus on pseudocode for BRE as it is the most complex. STRE computation does not require a mining step and every basket is already naturally associated with its own triple (shop, day-of-week, time-slot).

The full procedure has three logical steps that are reported in Algorithm 1. Step #1 is the detection of the frequent patterns from all the baskets of a customer. It can be implemented with any frequent itemset mining algorithm. In our experiments we implemented getItemSet($baskets, minsup$) with *Apriori* [22]. In the case of STRE, we simply calculate the relative frequencies of the triple (shop, day-of-week, time-slot). The set IS contains all frequent patterns appearing for the customer at least $minsup$ times. We want all the patterns returned by Apriori and not only maximal and closed patterns because otherwise we could not consider useful patterns (e.g. the pattern {Cheese} in the example). In our experiments we used $minsup$ as a relative frequency, i.e. $minsup = 24$ means that a certain pattern must be present in at least the 24% of the baskets of the customer analyzed and not in exactly 24 baskets.

Step #2 is the core of our contribution. It classifies each basket of the customer with the maximum matching frequent

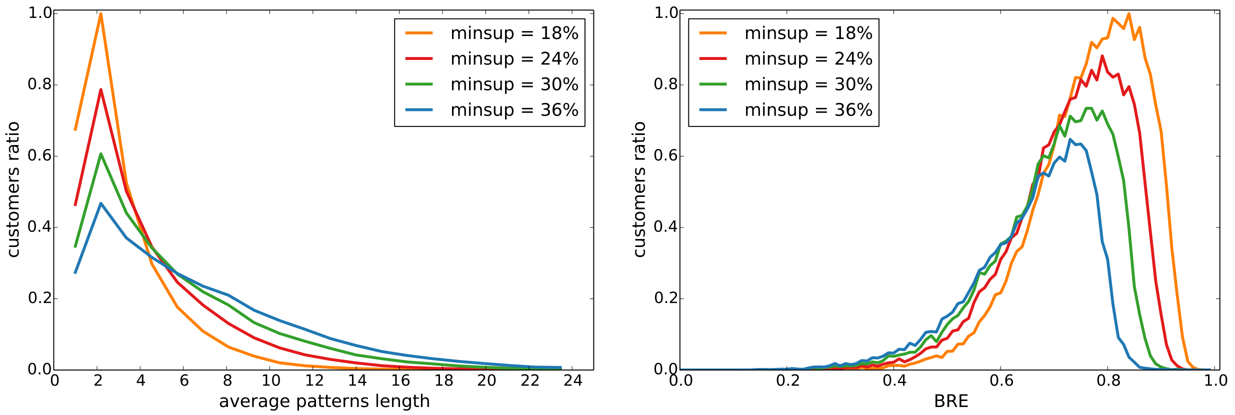

Fig. 2: The effect of $minsup$ on the length of the extracted patterns (left) and on the distribution of the BRE values (right).

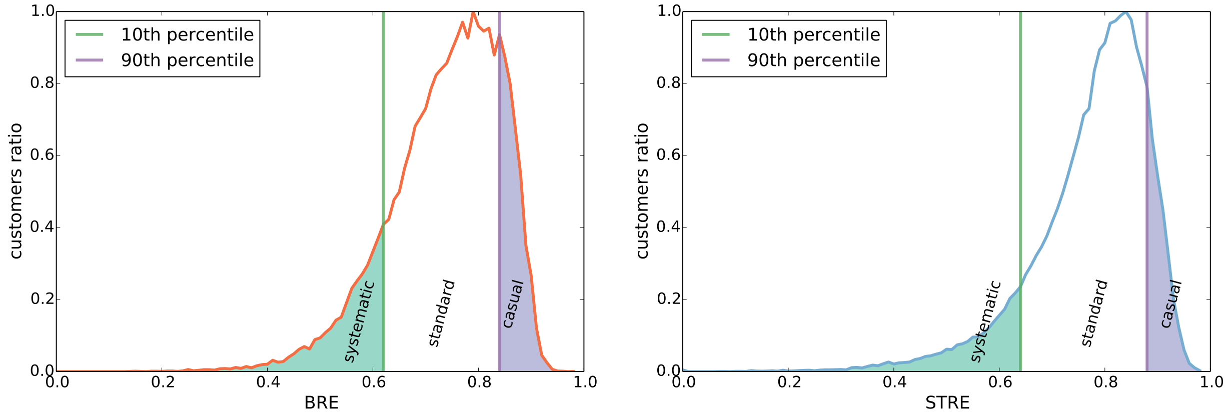

Fig. 3: Distributions of the BRE and STRE measures in our dataset, with highlighted 10th and 90th percentiles.

pattern. Its routine is expanded in Algorithm 2. For every basket, we define set D as the set of all IS patterns that are completely contained in the pattern. Note that there might be no patterns included, depending on the $minsup$ threshold choice. In this case we say that this basket can be only represented by itself, and its frequency is set to 1 (Steps #4-5).

Otherwise we have to extract from D the most significant representative basket that we use to classify the basket. The cascade of "if" conditions selects the representative basket rb as the most significant if: *(i)* there is one representative basket larger than any other representative basket in D (i.e. it contains the absolute highest number of elements, Step #8); *(ii)* there is one basket with the highest support (Step #12); *(iii)* there is one basket with the lowest $lift$ (Step #16). In data mining, $lift$ tells us how much more than expected a given customer purchased a given product: $lift > 1$ means higher than expected, $lift < 1$ means lower than expected. We use the lowest $lift$ because being the least unexpected means to be more representative. In all cases, the representative basket rb is found and its frequency (RB_{rb}) is increased by 1.

If it is impossible to reduce the set of included representative baskets to only one element, we classify the basket with all the remaining representative baskets, which are weighted one over the number of surviving representative baskets (Steps #19-20). Once we have the frequency of all representative baskets, we calculate BRE in Step #3 of Algorithm 1.

V. Case Study

In this section, we deploy our analytic framework to analyze the relationship between the behavior of customers and their shopping sessions[1]. We first take a look at what products can be found in the baskets of the systematic customers, in Section V-A. Then, we classify customers in five classes according to their behavioral entropy and we use these classes to understand the relationship between behavioral entropy and the customer profitability, in Section V-B. Finally, in Section V-C we validate and strengthen both results.

Before moving to the results, we provide our motivations for the required parameter $minsup$ (see Algorithm 1). $minsup$ is used for the frequent itemset mining and it is the minimum number of times a pattern has to appear to be considered frequent. We tested different values for this parameter, from 18% to 36%. This parameter influences the average pattern length we found. Higher $minsup$ generates shorter, and therefore less descriptive, patterns: the more elements a pattern has, the least likely it is to appear in full. Figure 2 (left) depicts this effect. $minsup$ also influences the distribution of BRE values. Higher $minsup$ generates less patterns and therefore BRE tends to take lower values, as each pattern is a new symbol and more symbols require more bits to be encoded. Figure 2 (right) depicts this effect. We chose $minsup = 24$ as a good

[1] A sample of the dataset used and the code to calculate BRE and STRE is available at https://goo.gl/UCqrUq

Product	Sup	Product	Sup
Bananas	82.44	Fresh Eggs	64.08
Vine Tomatoes	74.22	Parsley	62.71
Sugar	72.04	Nectarines	62.55
Fennels	69.12	Green Tomatoes	62.49
Dark Zucchini	67.80	Fresh Eggs (Organic)	62.23
Bright Zucchini	67.37	Roma Tomatoes	61.49
Cherry Tomatoes	65.52	Melons	61.17

TABLE I: The list of products of the systematic customers.

balance between the expressiveness of the detected patterns, and it does not skew the BRE distribution too much. Note that $minsup$ has no effect on STRE, as for STRE we consider all possible triples (shop, day-of-week, time-slot) and we do not use any frequent itemset mining technique. In particular, in our experiments we selected day-of-week in {weekday, weekend}, and time-slot in {07:00-9:30, 09:30-12:00, 12:00-17:00, 17:00-19:30, 19:30-21:00} according to Figure 1 bottom right.

We calculate the BRE for all customers included in the dataset using $minsup = 24\%$, Figure 3 (left) depicts the distribution: a skewed bell shape, peaking at 0.79; where 80% of the customers take values between 0.62 and 0.84. The 10th and 90th percentiles are highlighted in green and purple, respectively. Customers beyond the 90th percentile are "casual", customers below the 10th percentile are "systematic", and the remaining customers are "standard". Figure 3 (right) reports the STRE customer distribution. The distribution is a skewed bell shape, similar to the one observed for the BRE measure. The peak is now around 0.85; and 80% of the customers take values between 0.64 and 0.88. Also in this case, we report the 10th and 90th percentiles with green and purple lines.

A. Systematic Basket

In this section we look at the products purchased by the systematic customers. For this section, we define a systematic customer as a customer who is below the 10th percentile either for the BRE or for the STRE measure. Table I reports the list of the systematic products. To be a systematic product, a product must have been purchased at least ten times and it has to be purchased by at least 60% of the systematic customers. We dropped from this list all the meaningless products such as discount coupons and the plastic shopping bag.

From the list, we see that this selection includes mostly perishable products, from the fruit, vegetable and diary sectors. The only exception is sugar. It appears that the systematic customer's basket is characterized by very fresh products, that have a short shelf life and need to be purchased often.

B. Customer Classification

We now classify customers according to their observed BRE and STRE values. We represent each customer as a point in a two dimensional space. Her coordinates in this space are her BRE and STRE values. Then, we apply the k-means clustering algorithm [8] to detect clusters of customers in this space. The k-means algorithm requires to specify k, the number of clusters, or customer classes. The standard approach to determine k is to run k-means with varying ks (from 1 to 20 in our case), to calculate the sum of squared errors (SSE) for each k and choose the highest k beyond which SSE does not improve significantly. In our case, we have $k = 5$. Figure 4 depicts the evolution of the SSE values.

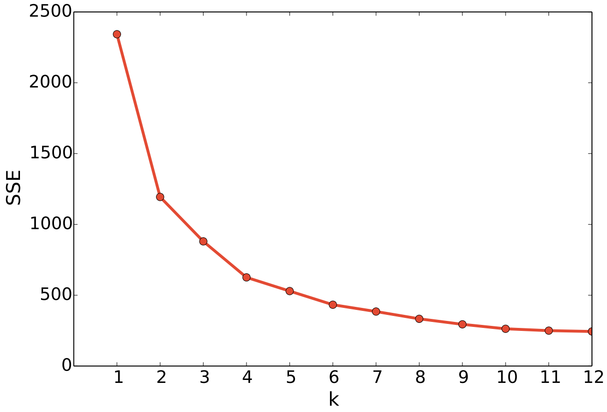

Fig. 4: The evolution of the sum of squared errors (SSE) for increasing ks in the k-means runs.

Cluster	Size	BRE	STRE
A	19.5%	0.45	0.57
B	17.3%	1.00	0.84
C	34.6%	0.50	1.00
D	21.7%	0.00	0.81
E	6.9%	0.17	0.00

TABLE II: Statistics of the detected customer clusters: relative size, and normalized average BRE and STRE scores.

For each detected cluster, k-means automatically detects the centroid, i.e. the most representative point of the cluster. If a point x belongs to cluster A, then the centroid of A is the closest centroid to x. The centroids are also representative of the cluster, as their BRE and STRE values are the average of the cluster. In Table II we report the statistics of the five detected clusters. We can see that the cluster sizes are well balanced, where three clusters contain around 20% of customers as expected. The exceptions are the larger C cluster and the smaller E cluster. We also report the normalized BRE and STRE values of each cluster's centroid. The normalization simply rescales BRE and STRE such that the minimum of all centroids equals to zero and the maximum equals to one.

From the reported values, we can easily characterize the five clusters. To aid the understanding of our interpretation, we display a simplified representation of the relative position

		STRE		
		Low	Medium	High
BRE	High			B
	Medium		A	C
	Low	E		D

TABLE III: The relative positions of the detected clusters in the BRE-STRE space. Note the triangular structure.

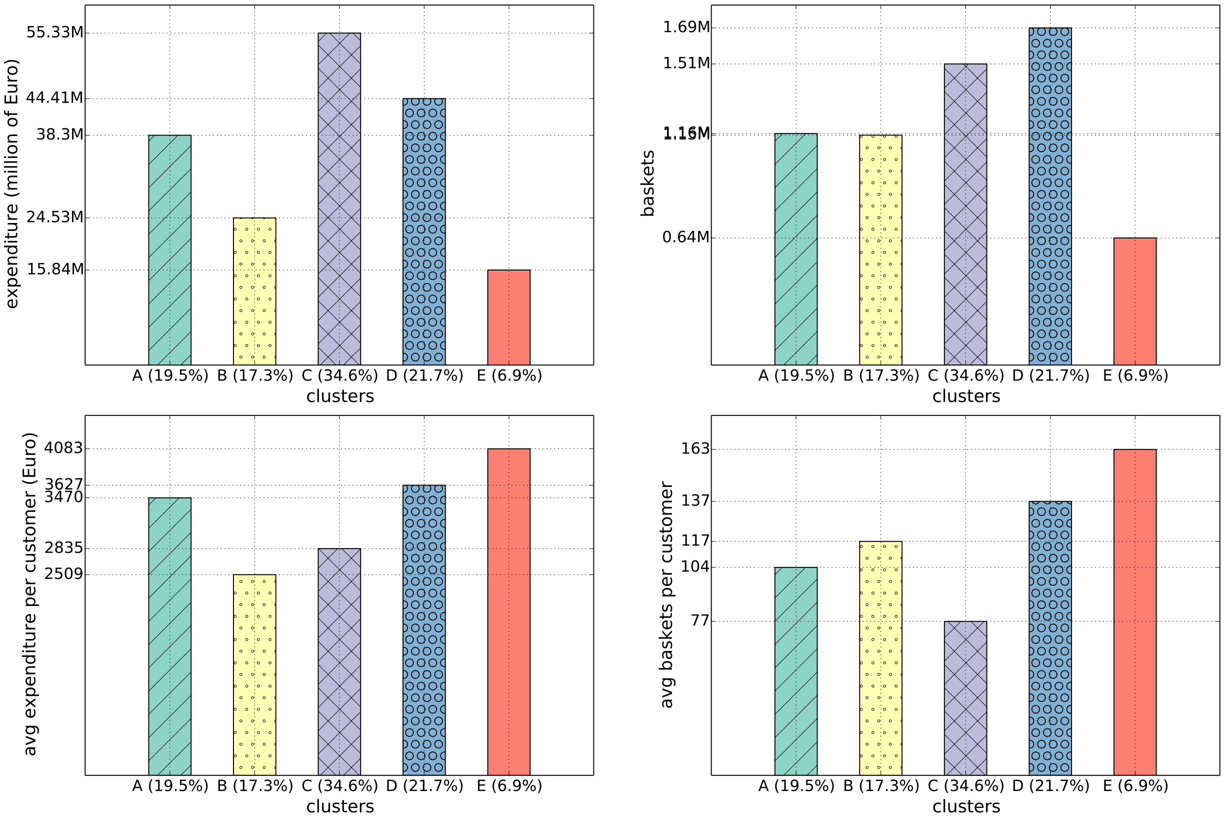

Fig. 5: The characteristics of customers belonging to the different clusters: on the left the customer expenditures (totals on top and per capita at the bottom), on the right the number of baskets (totals on top and per capita at the bottom).

of the centroids in the two dimensional space in Table III. We can see that each cluster can be characterized as follows according to BRE and STRE respectively: A medium BRE medium STRE; B high BRE high STRE; C medium BRE high STRE; D low BRE high STRE; E low BRE low STRE. From these results we can infer that the BRE-SPRE space has a triangular shape. Unpredictability in basket composition implies unpredictability also in the spatio-temporal dimension. On the other hand, unpredictability in the spatio-temporal dimension does not imply anything in the basket dimension (see cluster D for instance): the fact that we cannot predict when and where a customer will have a shopping session does not hinder us in predicting what products she is going to buy.

Before looking at more advanced statistics, we point out that the very regular customers, the ones characterized by both a low BRE and a low STRE, are the ones classified in cluster E. Cluster E is the smallest cluster, including the fewest number of customers, just below 7%. We conclude that the set of very regular customers is actually a large minority, at least in this supermarket chain. When projecting on one dimension, we see that the basket regular customers are less than 29% (clusters D and E), while the spatio-temporal regular are still just 7%, due to the triangular shape of our space (they can be found only in cluster E). In fact, we can conclude that most of the customers are spatio-temporal irregular, but somewhat basket regular. The two largest clusters are clusters C and D and while they both have high spatio-temporal entropy, they also have low or medium basket entropy. We can conclude that customers are more predictable in what they buy, rather than in when and where they perform their shopping sessions.

Once we detected our customer clusters, we can describe how the behavioral differences of the customers classified in them impact the profitability for the supermarket chain. For each cluster, we can calculate the total and per capita expenditures generated by customers classified in it, and we can calculate the total number of baskets and the per capita average. All these statistics are reported in Figure 5.

The most remarkable feature of the Figure is that it shows cluster E scoring the highest in expenditure per capita. We can calculate each cluster's leverage, that is the ratio between revenue share and relative size of the cluster. For E, since it includes 6.9% of customers and the total revenue is 178.41 million euros, the leverage equals to $(15.84/178.41)/0.069 = 1.29$. Second best is cluster D with a leverage of 1.15, while clusters C and B lag behind with a leverage of 0.9 and 0.79 respectively. This fact is hinting that the regularity of the customer might have a connection to her expenditure.

Looking at the broader picture, we see that there is a negative relationship between irregularity and per capita expenditure. We sort clusters from the most to the least irregular (by summing their BRE and STRE centroid values): $B \to C \to A \to D \to E$. We obtain a reverse order with the average per capita expenditure (see Figure 5): $E \to D \to A \to C \to B$.

We already saw that most customers are irregular in their spatio-temporal patterns. From the Figure, we also see that spatio-temporal irregular customers visit the stores more spo-

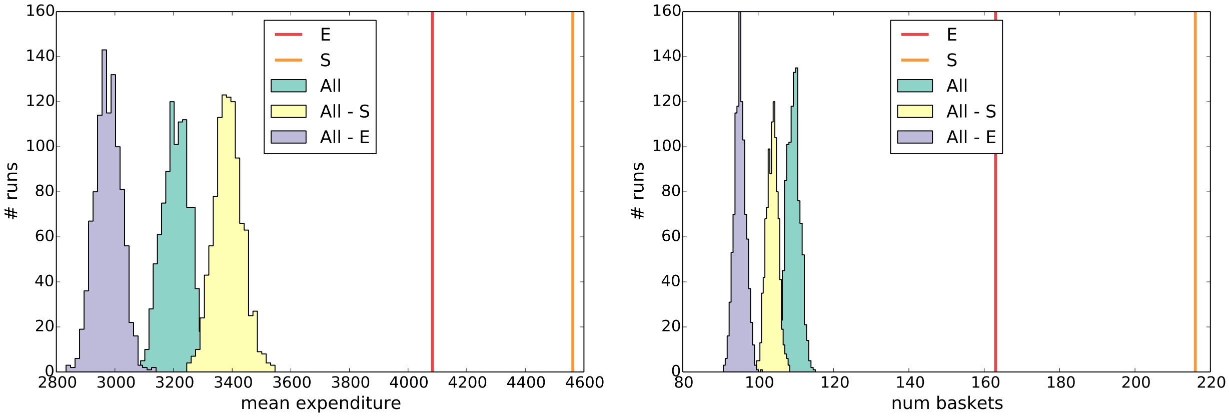

Fig. 6: The distributions of the average expenditure (left) and the average number of baskets (right) for the three null models "All", "All except systematic" ("All - S") and "All except E" ("All - E").

radically. The number of per capita baskets is lowest for cluster C, the most spatio-temporal irregular cluster, and generally low for most of the high STRE clusters.

The observed patterns have profound implications for the supermarket company. The average revenue generated per customer is higher for customers with low behavioral entropy. Cluster E is only two fifths in size of cluster B, but generates almost two thirds of cluster B's total revenues. However, the revenue from regular customers is low in absolute terms. An increase in customer's regularity could generate profits for the retail chain. On the basis of the data we gathered and the observations we just provide, our conclusion would be that the supermarket should encourage regularity to increase revenues. We now turn to some sanity checks to understand the significance of cluster E's observed profitability.

C. Validation

We start our result validation from the systematic basket composition. We use the $lift$ measure (see Section IV-C). For each customer, we calculate the $lift$ measure for the products in Table I. We then count how many systematic customers have $lift > 1$ for each of these products and we compare the three customer classes: systematic, standard and casual. We see that, on average, for each product there are 16% (st.dev. 5%) more systematic customers with $lift > 1$ than casual ones, and 9% (st. dev. 4%) more systematic than standard.

Now we focus on understanding if cluster E is really the most profitable per capita or if its expenditure level is not significantly different from a random occurrence. We perform two tests: a null model validation and a targeted model validation. Finally, we perform a last validation abstracting from the detected clusters and testing the direct connection between behavioral entropy and personal expenditure.

In the null model validation we want to explain the expenditure level and the number of baskets of the customers belonging to cluster E. We create some random E clusters with different characteristics and we observe their expected characteristics. We define three models called "All", "All except systematic" ("All - S") and "All except E" ("All - E"). We run each model a thousand times and we plot the distribution of their expenditure levels and number of baskets in Figure 6. The red band in Figure 6 is the observed E value.

The "All" model constructs a purely random E cluster. We extract uniformly at random 7% of the customers in our data and we calculate their average expenditure and their average number of baskets. Figure 6 reports that this model has an expected expenditure of 3,200 euros, that is slightly more than three quarters of the actual E expenditures.

The "All except systematic" model constructs a random E cluster by (randomly) selecting customers outside the "systematic" cut, i.e. all customers that have BRE and STRE values higher than the 10th percentile. By restricting to these customers we attempt to counter the argument that it is the BRE and STRE values driving the expenditure and not other common factors of customers included in cluster E. However, we obtain again a lower expected expenditure: 3,400 euros or just 83% of the actual expenditure of cluster E.

Finally, with the "All except E" we construct a random E cluster by selecting customers at random from the pool of customers that are NOT part of the original cluster E. In this model we investigate if it is likely to find a random composition of customers outside cluster E that are characterized by higher expenditure levels than the members of cluster E. This is the model that performs the worst, even worse than the "All" model, proving that "All" model's performance was actually driven by E cluster members. In "All except E", the expected expenditure is just below 3,000 euros. The number of baskets of cluster E members is impossible to match too. In this case, the "All" model performs better than "All except systematic", hinting that the number of baskets is more dependent on the behavioral entropy than the expenditure level.

Moving to targeted model validations, we define two: one based on expenditure and one based on the behavioral entropy. Differently from before, we are not composing a random E model, but we are sorting all customers in descending order of the chosen measure. Starting with expenditure, we collect the 7% top-spending customers and we count how many of them are classified in cluster E. The result is 14.66%, meaning that cluster E is represented in the top spending customers twice as much as its size would suggest. This confirms the strong relation between cluster E and high expenditure levels.

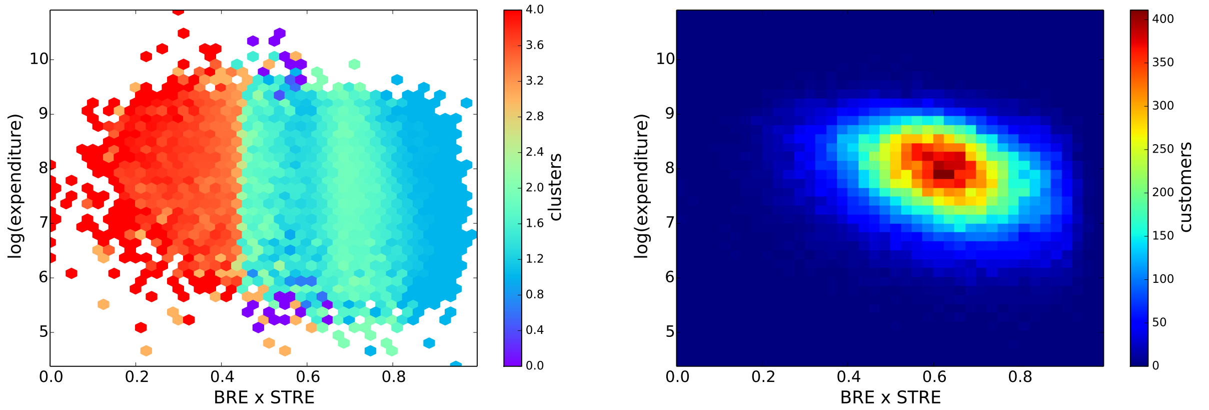

Fig. 7: Heatmaps depicting the relationship between the combined behavioral entropy (x axis) and the expenditure level (y axis, in log scale). We have both the average cluster composition of the cell (left, from 0 = A , to 4 = E) and the simple count of the number of customers (right).

	Dependent variable:					
	log(expenditure)			log(baskets)		
	(1)	(2)	(3)	(4)	(5)	(6)
BRE	-1.123^{***} (0.021)			-1.124^{***} (0.020)		
STRE		-1.152^{***} (0.026)			-1.738^{***} (0.024)	
BRE * STRE			-1.269^{***} (0.019)			-1.492^{***} (0.018)
constant	8.738^{***} (0.017)	8.754^{***} (0.020)	8.635^{***} (0.012)	5.360^{***} (0.016)	5.833^{***} (0.019)	5.394^{***} (0.011)
Observations	56,448	56,448	56,448	56,448	56,448	56,448
R^2	0.048	0.034	0.071	0.055	0.088	0.112
Adjusted R^2	0.048	0.034	0.071	0.055	0.088	0.112
Residual Std. Error	0.676	0.681	0.668	0.631	0.620	0.611
F Statistic	2,851.457***	1,970.123***	4,322.123***	3,277.555***	5,419.480***	7,131.251***
				Note: *p<0.1; **p<0.05; ***p<0.01		

TABLE IV: BRE and STRE effect in predicting the total expenditure level (models 1 to 3) and the number of baskets (models 4 to 6) of a customer for the year 2012. This is a standard OLS model. The R^2 can be interpreted as the square of the correlation coefficient of the variables.

If we include also the cluster characterized by the second most regular customers, cluster D, the share goes up to 38.7%.

The second targeted model involves selecting the customers in the "systematic" cut regardless the cluster in which they were classified. Their expenditure levels are very high, higher than the members of cluster E. This hypothetical super-systematic cluster has an expected expenditure level of almost 4,600 euros, as the orange band depicts it in Figure 6.

For our last validation step we abstract from the cluster division, to observe the direct relationship between a customer's behavioral entropy and her profitability for the retail company. This is done by plotting the behavioral entropy of a single customer against her expenditure level. Figure 7 shows two variants of this plot. In both cases we have a heatmap that groups the customers in a given interval of expenditures and of entropy. The x axis combines BRE and STRE by multiplying them. The y axis reports the logarithm of the expenditure level.

On the left of Figure 7, we have the average cluster composition of the cell. To calculate this average each cluster is mapped to an integer. The heatmap has a left to right gradient, where the lowest values on the x axis correspond to highest clusters (D and E). The heatmap contains a negative relationship between combined entropy and expenditure.

To better highlight this negative relationship, on the right of Figure 7 we use a different coloring logic for the heatmap. Instead of reporting the cluster composition, we color the cell according to its number of customers. Blue means few or no customer, red means a high concentration of customers. We can see that now the negative relationship is more clear: the densely populated cells show a downward pattern.

To quantify objectively the size of the effect depicted in Figure 7, we set up a model where we attempt to predict the logarithm of the customer's expenditure (or baskets) by using her BRE and STRE level. First we test the two measures separately, then we create a global measure by multiplying them. Table IV reports the result of this regression.

Both BRE and STRE have significant effects, with comparable levels. We are using a log-linear space, thus a coefficient of -1.123 means that increasing the entropy level by 1 is associated with an expected expenditure drop of almost a third ($e^{-1.123} \sim 0.325$)[2]. An hypothetical perfectly predictable customer (entropy = 0) would make three times as many profits for the company than a hypothetical completely unpredictable customer (entropy = 1). Combining BRE and STRE together, the effect almost reaches a fourfold increase ($e^{-1.269} \sim 0.281$). The effect is stronger if we predict the number of baskets instead of the expenditure level. The unit decrease in combined entropy is associated with almost a fivefold increase in number of baskets purchased ($e^{-1.492} \sim 0.225$).

VI. Conclusion

In this paper we investigated the effects of customer predictability in the retail market scenario. We estimated how much the behavior of a customer is predictable along two dimensions: basket composition, i.e. the items a customer purchases; and spatio-temporal, i.e. where and when a customer purchases the products she needs. This classification is done with a new framework that includes frequent itemset mining and information theory concepts such as information entropy. We showed that it is possible to divide customers into systematic and non-systematic, and even define five distinct classes. The systematic customers have been showed to be a minority in the supermarket's customer base, but their per capita expenditure and expected number of baskets is much higher than average. Our customer entropy measures have proved to be significant predictors of the value of a customer for the supermarket and point out that nudging customers to be regular could be an interesting strategy to increase revenues. It should be remarked here that we are addressing "individual" entropy, not "collective" entropy of the entire ecosystem of customers. Accordingly, high predictability of individual customers can coexist with a broad diversity of shopping behavior at collective scale.

There are a number of future works. We could use our behavioral entropy measures as additional features in a more comprehensive model for predicting a customer value, using other data mining tools such as decision trees. When combining our behavioral entropy measures with other customer features we will be able to evaluate if there are collinear factors or if our measures are mostly orthogonal to other characteristics like customer demographics. Finally, the relationship between our conclusions and other traditional marketing strategies could be tested: to which point is it beneficial to favor systematic behavior over the introduction of unexpected elements by the supermarket, such as rotation of product placements or special offers?

Acknowledgements. We thank the supermarket company UniCoop Tirreno for sharing the data with us and allowing us to analyze and to publish the results. We also thank Gian Felice Meloni for the preliminary analysis he made that lead to the creation of this paper.

[2] See http://www.ats.ucla.edu/stat/mult_pkg/faq/general/log_transformed_regression.htm for an explanation of our coefficient interpretation.

References

[1] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases," in *SIGMOD International Conference*, Washington, D.C., 1993, pp. 207–216.

[2] A.-L. Barabasi, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, no. 7039, pp. 207–211, 2005.

[3] A.-L. Barabási, *Bursts: the hidden patterns behind everything we do, from your e-mail to bloody crusades*. Penguin, 2010.

[4] Z.-Y. Chen and Z.-P. Fan, "Distributed customer behavior prediction using multiplex data: A collaborative mk-svm approach," *Knowledge-Based Systems*, vol. 35, pp. 111–119, 2012.

[5] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1082–1090.

[6] M. De Domenico, A. Lima, and M. Musolesi, "Interdependence and predictability of human mobility and social interactions," *Pervasive and Mobile Computing*, vol. 9, no. 6, pp. 798–807, 2013.

[7] R. R. Dholakia and N. Dholakia, "Mobility and markets: emerging outlines of m-commerce," *Journal of Business research*, vol. 57, no. 12, pp. 1391–1396, 2004.

[8] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Applied statistics*, pp. 100–108, 1979.

[9] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[10] 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.

[11] I. D. Kocakoç and S. Erdem, "Business intelligence applications in retail business: Olap, data mining & reporting services," *JIKM*, vol. 9, no. 2, pp. 171–181, 2010.

[12] C. Krumme, A. Llorente, M. Cebrian, E. Moro *et al.*, "The predictability of consumer visitation patterns," *Scientific reports*, vol. 3, 2013.

[13] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[14] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, "A tale of many cities: universal patterns in human urban mobility," *PloS one*, vol. 7, no. 5, p. e37027, 2012.

[15] D. Pennacchioli, M. Coscia, and D. Pedreschi, "Overlap versus partition: marketing classification and customer profiling in complex networks of products," in *Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on*. IEEE, 2014, pp. 103–110.

[16] D. Pennacchioli, M. Coscia, S. Rinzivillo, F. Giannotti, and D. Pedreschi, "The retail market as a complex system," *EPJ Data Science*, vol. 3, no. 1, pp. 1–27, 2014.

[17] D. Pennacchioli, M. Coscia, S. Rinzivillo, D. Pedreschi, and F. Giannotti, "Explaining the product range effect in purchase data," in *Big Data, 2013 IEEE International Conference on*. IEEE, 2013, pp. 648–656.

[18] D. Pennacchioli, G. Rossetti, L. Pappalardo, D. Pedreschi, F. Giannotti, and M. Coscia, "The three dimensions of social prominence," in *Social Informatics*. Springer, 2013, pp. 319–332.

[19] C. Scholz, M. Atzmueller, and G. Stumme, "On the predictability of human contacts: Influence factors and the strength of stronger ties," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 2012, pp. 312–321.

[20] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal, The*, vol. 27, no. 3, pp. 379–423, July 1948.

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

[22] P.-N. Tan, M. Steinbach, V. Kumar *et al.*, *Introduction to data mining*. Pearson Addison Wesley Boston, 2006, vol. 1.

[23] J. L. Toole, C. Herrera-Yaqüe, C. M. Schneider, and M. C. González, "Coupling human mobility and social ties," *Journal of The Royal Society Interface*, vol. 12, no. 105, p. 20141128, 2015.

[24] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási, "Modeling bursts and heavy tails in human dynamics," *Physical Review E*, vol. 73, no. 3, p. 036127, 2006.

[25] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1100–1108.