Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

Influence-based approach to market basket analysis

Ariel Monteserin, Marcelo G. Armentano*

ISISTAN Research Institute (CONICET/UNICEN), Campus Universitario Paraje Arroyo Seco, Tandil, Buenos Aires B7001BBO, Germany

ARTICLE INFO

Article history:
Received 15 April 2017
Revised 27 December 2017
Accepted 16 January 2018
Available online xxx

Keywords:
Social influence
Market basket analysis
Credit distribution model

ABSTRACT

In this article, we propose an approach to market basket analysis based on the notion of social influence. While traditional market basket analysis looks for combinations of products that frequently co-occur in transactions, we seek to find a set of influential products that, if bought by a customer, will increase the sales volume of the shop. We believe that customers who purchase influential products would also be influenced to purchase other products. We validated our approach with two real-world datasets collected from online shoppings and one dataset collected from a supermarket concluding that influential products identified by our approach increase the influence spread with respect to different baselines: best-selling, highest centrality, frequent sequence initiator, and most promoted products.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

In the last years, the interest on the analysis of social influence has increased significantly. Social influence occurs when one' s actions are affected by others. For example, user *A* exerts influence on user *B* when *B* buys a given product because *A* recommended it or bought it previously. Many applications exploit the concept of social influence. Particularly, viral marketing has found in social influence a key ally. One of the main problems related to social influence is the social influence maximization problem. The social influence maximization problem involves finding a set of users in a social network such that by targeting this set, the expected influence spread in the network is maximized [10,13]. From the point of view of viral marketing, by obtaining a set of influential users we can reach a high number of potential customers[1] with a minimum of effort [1,17–19].

All these works analyse how a user exerts influence on other users during the acquisition of a given product. However, we claim that another kind of influence can be observed: the influence that leads a customer to purchase a product because previously purchased another one. Drawing parallels, this represents the influence that a product exerts on other products from the point of view of a given customer. For example, a product *X* (a photo camera) exerts influence on a product *Y* (a wide-angle lens) when a customer *A* buys *Y* because she bought *X* previously. Thus, if we are able to discover the set of influential products, we can intensify the marketing on this set, under the assumption that the customers that bought any of those products will probably buy others. By looking at the social influence maximization problem from the products perspective, we can see some resemblance to the market basket analysis problem. Market Basket Analysis studies the composition of a shopping basket of products purchased during a single shopping event [20] with the aim of identifying what products tend to be purchased together.

In this article, we present an approach that allow us to identify the set of influential product from a complete set of products of a given seller or company. To do this, we apply a data-based approach to social influence maximization, named Credit Distribution (CD) model [10], that learns how influence flows in a network by directly leveraging available propagation traces. The main difference of our approach with the classical market basket analysis problem is that we consider *long-term* shopping baskets in which customers not necessarily buy items in the same transaction.

We validated our proposal with experiments performed with two real-world datasets extracted from Alibaba and Ponpare websites, and a database of Foodmart Supermarket. We compared the true influence spread produced by different seed sets. The true influence spread measures how many nodes (i.e. products) of the network will be activated (purchased by some customer) after the seed nodes (the potential influential products) are activated. We compared the true influence spread obtained from different seed sets: (a) the set of influential products obtained by our approach; (b) the set of best-selling products, (c) the set of products with highest centrality in a co-purchased network, (d) the set of frequent sequence initiator products extracted from sequential patterns mined with PrefixSpan [22], (e) the set of products with highest weight in the first component of V matrix, according with

* Corresponding author.
E-mail address: marcelo.armentano@isistan.unicen.edu.ar (M.G. Armentano).

[1] In the text, we refer to users and customers as synonyms, since with the term “users” we refer to users of online stores or e-commerce websites.

https://doi.org/10.1016/j.is.2018.01.008
0306-4379/© 2018 Elsevier Ltd. All rights reserved.

Singular Value Decomposition (SVD) technique [8], (f) the set of most promoted products (only for Foodmart dataset, which counts with this information) and (g) random sets of products. In the domain of this work, the value of the influence spread is related to the number of products that will be potentially sold after selling the influential products (seeds). The results clearly revealed that the number of potential sales is increased when the set of influential products discovered with our approach is activated. Additionally, we run a test on the datasets to determine the existence of influence among the products [3,7], and to distinguish correlation from causality.

The article is organized as follows. In Section 2, we present some background on market basket analysis and different approaches to the problem found in the literature. In Section 3, we present our approach for mining influential products. Section 4 presents a test to distinguish influence from correlation. Next, in Section 5, we present the experiments performed to validate our approach. Finally, Section 6 exposes our conclusions and future work.

2. Background

In market basket analysis, association rules [2] have been intensively explored as a means for finding products that are bought together with other products. Generating association rules involves looking for frequent itemsets. Then, since all subsets of a frequent set are also frequent, for each frequent set X, each subset $Y \subset X$ is tested to verify whether $X - \{Y\} \Rightarrow Y$ has a confidence value greater than the minimum confidence allowed. Association rules have been successfully used for product assortment decisions [6], credit card business [29], profit mining [27,28], market basket analysis [16], among other fields. The main difference of our work with respect to frequent-itemset approaches is in the data available for mining. Frequent itemset approaches work with a transactional database in which each transaction consists of a set of products that the customer bought together and probably some additional information such as the benefits from the transaction. In our approach, transactions are individual, that is, triples (*I, C, T*) meaning that item *I* (a given product) was bought by customer *C* on time *T*. This implies that we consider items bought by the same customer in different times.

Tan et al. [24] studied the problem from another perspective. Instead of using association rules, they approach the problem of finding similar time series of product sales in transactional data. Authors state that the use of quantity and time information yield to richer and more insightful results.

Li et al. [15] approached the problem of helping manufactures position their products in the market, based on dominant relationship queries: Linear Optimization Queries, Subspace Analysis Queries and Comparative Dominant Queries. Authors propose the construction of a data cube to facilitate more advanced data mining, by using the relationships of dominated/dominating customers and products as a basis for decision-making. Similarly, Vlachou et al. [26], proposed an algorithm for identifying the most influential products by means of reverse top-k queries. Given a product, reverse top-k queries retrieve the customers to whom the product belongs to their top-k result set. The problem of finding the most influential products is then expressed as a query that finds the k products with the highest influence score. In Vlachou et al. [26] approach, a product is considered influential if it is appealing to many users. Differently, in our approach, we consider that a product is influential if it motivates users to buy other products.

An alternative approach for finding influential products is by means of network analysis in search of relationships rather than associations among products. Raeder and Chawla [23], for example, use community detection algorithms to detect strong relationships among products ("communities" of products). Kim et al. [14] used two networks structures for market basket analysis: (1) Market basket network (MBN), in which there exist a link between two products if there is an pre-computed association rule that relate both products; the strength of the link is given by the support of the association rules; and (2) Co-purchase product network (CPN), in which there is a link between two products if they were bought by the same customer; the strength of the link in this network is given by the frequency of two products purchased together. Differently to our approach, the work by Kim et al. does not consider purchase time-stamps. They found that top products in terms of *centrality* in MBN correspond mainly to daily necessity products while in CPN are living necessities for ordinary families. They also conclude that products with high centrality can both increase sale volumes and be effective in promotion or cross selling. In the experiments presented in Section 4, we show that our approach is able to increase sale volumes with respect to consider the products with higher centrality as the seed set. In the same way, Videla-Cavieres and Ríos [25] generate a product network based on transactions where each product is linked to others because they appear in the same transaction from the same customer and then apply a temporary set of filters to check quality and stability of the communities found. Finally, overlapping community detection algorithms are used to discover frequent item set.

All the approaches presented above identify products that are usually bought together, in the same transaction, by the same user or any combination of these restrictions. These approaches do not consider the order in which products are purchased and then it is difficult to identify which product(s) in the frequent itemsets is the influential product. Differently, the approach presented in this article is able to identify which is the set of products that are assumed to motivate the purchase of other products, what we call *influential products*. We seek to find the set of products that, when bought by customers, will maximizes the sale volume of the shop.

3. Mining influential products

Kempe et al. [13] formalized the influence maximization problem as following: given a directed graph $G = (V, E, p)$, where nodes are users and edges are labeled with influence probabilities among users, the influence maximization problem looks for a set of seeds (users) that maximizes the expected spread of influence in the social network under a given propagation model. A propagation model indicates how influence propagates through the network. Kempe et al. proposed two propagation models: the Independent Cascade (IC) and the Linear Threshold (LT) models. In both models, each node can be either active or inactive at a given moment. Moreover, the tendency of each node to become active increases monotonically as more of its neighbors become active.

Given a propagation model m (for example, IC or LT) and an initial seed set $S \subseteq V$, the expected number of active nodes at the end of the process is the expected (influence) spread, denoted by $\sigma_m(S)$ [10]. Then, the influence maximization problem is defined as follows: given a directed and edge-weighted social graph $G = (V, E, p)$ (where nodes are users and edges are labeled with influence probabilities among users), a propagation model m, and a number $k \leq |V|$, find a set $S \subseteq V$, $|S| = k$, such that $\sigma_m(S)$ is maximum. Several approaches have been developed to solve this problem. Despite the fact that this problem is NP-hard under both the IC and LT propagation models, some characteristics of the function $\sigma_m(S)$ (monotonicity and submodularity, see [13] for further details) made it possible to develop a greedy algorithm to solve the problem.

One of the limitations of the IC and LT propagation models is that the edge-weighted social graph is assumed as input to the problem, without addressing the question of how the probabili-

Please cite this article as: A. Monteserin, M.G. Armentano, Influence-based approach to market basket analysis, Information Systems (2018), https://doi.org/10.1016/j.is.2018.01.008

ties are obtained [9]. For this reason, Goyal et al. [10] proposed the Credit Distribution (CD) model, which directly estimates influence spread by exploiting historical data. In this context, the influence maximization problem to be solved under the CD model is reformulated as follows: given a directed social graph $G = (V, E)$, an action log L, and an integer $k \leq |V|$, find a set $S \subseteq V$, $|S| = k$, such that $\sigma_{cd}(S)$ is maximum. Under the CD model, $\sigma_{cd}(S)$ is defined as $\sigma_{cd}(S) = \sum_{u \in V} \kappa_{S,u}$, where $\kappa_{S,\, u}$ represents the total credit given to S for influencing u for all actions. For a pair of users v and u, the average credit given to v for influencing u, over all actions that u performs is denoted by Eq. (1).

$$\kappa_{v,u} = \frac{1}{A_u} \sum_{a \in A} \Gamma_{v,u}(a) \tag{1}$$

In Eq. (1), A_u is the number of actions performed by u, A is the set of actions to be propagated. Here, $\Gamma_{v,\, u}(a)$ represents the total credit given to v for influencing u on action a, corresponding to:

$$\Gamma_{v,u}(a) = \sum_{w \in N_{in}(u,a)} \Gamma_{v,w}(a) \cdot \gamma_{w,u}(a) \tag{2}$$

where $\gamma_{w,\, u}(a)$ indicates the direct credit given by u to a neighbor w for action a and $N_{in}(u,\, a)$ is the set of neighbors of u which activated on action a before (potential influencers on action a). The direct credits are computed by Eq. (2).

$$\gamma_{v,u}(a) = \frac{infl(u)}{|N_{in}(u,a)|} \cdot \exp\left(-\frac{t(u,a) - t(v,a)}{\tau_{v,u}}\right) \tag{3}$$

In Eq. (2), $infl(u)$ denotes the user influenceability and is defined as the rate of actions that u performs under the influence of at least one of its neighbors; $\tau_{v,\, u}$ is the average time taken for actions to propagate from user v to user u; and $t(x,\, a)$ is the time in which user x performed action a. Thus, the direct credit represents the idea that influence decay over time in a exponential fashion and that some users are more influenceable than others [9].

To solve the problem of social influence maximization, Goyal et al. developed an algorithm that works under the CD model by scanning the action log L to learn the influence probabilities in the social network and computing “influenceability scores” for the users. An action log is a set of triples (u, a, t) which say user u performed action a at time t. Then, the seed set is selected under the CD model by using a greedy algorithm with CELF optimization [11] according to a set of training actions. Finally, the true influence spread is computed by taking into account a set of action for testing. See [10] for further details on algorithm implementation.

Our approach consists of varying the definition of the influence maximization problem. First, since we want to identify influential products, we define the directed graph G as a graph of products where nodes are products and edges are some kind of relationship among products [14,23,25]. The relationships among products can be varied. For example, two products can be related if they have the same category, if they are manufactured by the same company or if they are sold in the same shop, among other kinds of relationships. Fig. 1 shows an example of a product graph in which the products are related by category and brand.[2] Thus, cameras are related to lens, smartphones to headphones, and *Smartphone Y* to *Camera Y* (where *Y* is the brand of both products).

Second, the action log L is completed with real sales, but varying the order of the data. In the traditional problem, a sale is represented in the action log as a set of triples (u, p, t) which indicates that user u (a customer) bought product p at time t. This is because the goal of the traditional problem is to determine the influential users. Thus, users receive credits for influencing other users to purchase the same products. In contrast, in our approach, we represent a sale as a triple (p, u, t) indicating that product p was bought by user u at time t. Thus, the products receive credits for influencing other products to be purchased by the same users. Following the example of Fig. 1, Table 1 shows an example of an action log.

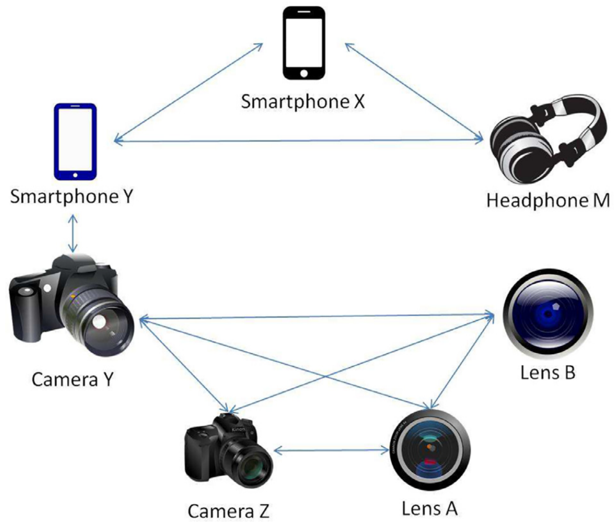

Fig. 1. Example graph of products.

Table 1
Example action log for the network in Fig. 1.

Product	User	Timestamp
Smartphone Y (*sy*)	1	1
Camera Y (*cy*)	1	5
Lens A (*la*)	1	8
Smartphone Y	2	5
Headphone M (*hm*)	2	10
Lens B (*lb*)	3	3
Smartphone X (*sx*)	3	15
Camera Y	4	12
Smartphone X	5	8

Taking into account the graph of Fig. 1 and the action log of Table 1, our approach indicates that *Smartphone Y* is the most influential product with a marginal gain of 1.527. This value is obtained by computing the total credit given to $S = \{sy\}$ for influencing all the products for all training actions. Eq. (4) shows how this value is computed.

$$\begin{aligned}\sigma_{cd}(sy) &= \sum_{u \in V} \kappa_{sy,u} \\ &= \kappa_{sy,sy} + \kappa_{sy,cy} + \kappa_{sy,la} + \kappa_{sy,hm} + \kappa_{sy,lb} + \kappa_{sy,sx} \\ &= 1 + 0.092 + 0.067 + 0.368 + 0 + 0 = 1.527\end{aligned} \tag{4}$$

$\kappa_{sy,\, sy}$ is 1, since the base of the recursion in Eq. (5) is $\Gamma_{v,v}(a) = 1$. Moreover, $\kappa_{sy,\, lb}$ and $\kappa_{sy,\, sx}$ are 0 because neither *lb* nor *sx* were acquired by users who had purchased *sy*. In contrast, *cy, la* and *hm* give to *sy* real credit. Eq. (5) exemplifies how $\kappa_{sy,\, cy}$ is computed.

$$\begin{aligned}\kappa_{sy,cy} &= \frac{1}{A_{cy}} \sum_{a \in A} \Gamma_{sy,cy}(a) \\ &= \frac{1}{2}(\Gamma_{sy,cy}(1) + \Gamma_{sy,cy}(2) + \Gamma_{sy,cy}(3) + \Gamma_{sy,cy}(4) + \Gamma_{sy,cy}(5)) \\ &= \frac{1}{2}(0.184 + 0 + 0 + 0 + 0) = 0.092\end{aligned} \tag{5}$$

where $A_{cy} = 2$ (users *1* and *4*) and $\Gamma_{sy,\, cy}(2)$, $\Gamma_{sy,\, cy}(3)$, $\Gamma_{sy,\, cy}(4)$ and $\Gamma_{sy,\, cy}(5)$ are 0 because *sy* and *cy* did not purchased together by users *2, 3, 4* or *5*. However, *sy* receives credit from *cy* because user *1* bought *cy* after buying *sy*. Eq. (6) shows how this value is com-

[2] Notice that other kind of relationships among products can be considered.

Please cite this article as: A. Monteserin, M.G. Armentano, Influence-based approach to market basket analysis, Information Systems (2018), https://doi.org/10.1016/j.is.2018.01.008

puted by using Eqs. (2) and (3).

$$\Gamma_{sy,cy}(1) = \sum_{w \in N_{in}(cy,1)} \Gamma_{sy,sy}(1) \cdot \gamma_{sy,cy}(1) = 1 \cdot \left(\frac{0.5}{1} \cdot \exp\left(-\frac{5-1}{4} \right) \right) = 0.184 \quad (6)$$

where $N_{in}(cy, 1) = \{sy\}$ and $infl(cy) = 0.5$ since cy was purchased after sy by user 1, but in first place by user 4.

4. Influence vs. correlation

In the literature on social influence, differentiating influence from correlation is known to be a difficult task. This is because there are factors as homophily or unobserved confounding variables that can induce statistical correlation between actions of users in a social network [3]. Thus, distinguishing influence from these factors is the problem of distinguishing correlation from causality. This fact can also be present when we analyze influential products.

Anagnostopoulos et al. [3] identified three causes of correlation in social networks:

- Influence: this occurs when an action of a user is triggered by one of his/her friend's recent actions. In the context of this work, this occurs when the purchase of a product is triggered by a recent purchase of other product.
- Homophily: this means that individuals often befriend others who are similar to them, and hence they perform similar actions. In this work, this means that similar products are related; hence, the same customers may purchase them.
- Environment (confounding factors or external influence): this occurs when external factors are correlated both with the event that two individuals become friends and with their actions. These external factors can also be correlated with the products purchased by a given customer.

The main assumption to distinguish social influence from other types of correlation is that under the other types of correlation the probability that an individual is active can be affected by *whether* their friends become active, but not by the time *when* they become active. Taking into account this assumption, Anagnostopoulos et al. [3] proposed the *shuffle test* for identifying social influence. This test is based on the idea that if influence does not play a role between two nodes in the network, even though a node's probability of activation could depend on her friends, the timing of such activation should be independent of timing of other nodes [3]. The shuffle test proposes to compare (a) the social correlation of a given social network, a set of nodes and a set of actions executed by the nodes in a specific time; and (b) the social correlation of the same social network, nodes and actions, but shuffling the time when the node executed the actions. Then, the shuffle test declares that the model exhibits no social influence if the values of social correlation are close to each other. Recently, Chen et al. [7] applied an adaptation of the shuffle test to prove the existence of influence on interactions of Twitter users. They constructed a shuffled dataset by permuting the sequence of tweets for each user randomly.

Similarly, to prove the existence of influence among products, we propose to compare the influence spread obtained by the influence seed set in a real action log and in a shuffled one. To generate the shuffled action log, for each product p_t purchased by user u in different time-stamps $t \in \{1, \ldots, l\}$, we pick a random permutation π of $\{1, \ldots, l\}$, and set the new action log entries as $(p_t, u, \pi(t))$ for each $t \in \{1, \ldots, l\}$. Table 2 shows the shuffled action log generated from the example presented in Table 1.

Table 2
Shuffled action log.

Product	User	Timestamp
Smartphone Y (*sy*)	1	8
Camera Y (*cy*)	1	1
Lens A (*la*)	1	5
Smartphone Y	2	10
Headphone M (*hm*)	2	5
Lens B (*lb*)	3	3
Smartphone X (*sx*)	3	15
Camera Y	4	12
Smartphone X	5	8

5. Experimental evaluation

5.1. Experimental settings

To evaluate our approach, we ran experiments on real-world datasets extracted from Ponpare[3] and Alibaba[4] websites, and a database of Foodmart Supermarket grocery store. For each of these datasets, we compared the true influence spread obtained by different seed sets: (a) the set of influential products obtained by our approach; (b) the set of best-selling products; (c) the set of products with highest centrality; (d) the set of products appearing in the first elements of a frequent sequential pattern; (e) the set of products with highest weight in the first component of V matrix, according with Singular Value Decomposition (SVD) technique; and (f) random sets of products. Additionally, since the Foodmart dataset includes information about product promotions, we also added to the comparison the seed set representing the most promoted products. The true influence spread represents the number of active nodes after the nodes of the seed set become active. In the context of our proposal, this spread represents how many products (nodes) could be sold (become active) after the products of the seed sets are sold (become active).

We describe each dataset in the following Sections.

5.1.1. Ponpare dataset

Ponpare is Japan's leading joint coupon site. This dataset[5] provides a year of transactional data for 22,873 users on Ponpare site. Each transaction contains a user, a coupon (i.e. a discount price for a given product) and a time-stamp. The dataset consists of 168,996 transactions, 22,873 users and 19,413 coupons. Note that in this dataset the coupons represent the products of the shop.

5.1.2. Alibaba dataset

This dataset[6] was obtained from Alibaba Tianchi Contest, which was carried out during the 2nd International Workshop on Social Influence Analysis [4], co-located with the International Joint Conference on Artificial Intelligence (IJCAI 2016). The dataset consists of different online sales achieved by a set of different sellers between July 1st, 2015 and November 30th, 2015. Each sale consisted of a user, a seller, a product, and a timestamp. In total, the dataset consists of 9,348,756 sales, 885,759 users, 1,144,124 products and 9997 sellers.

5.1.3. Foodmart dataset

This dataset[7] consists of different transactions performed by customers of a supermarket chain during two years (1997 and

[3] https://ponpare.jp/.

[4] http://www.alibaba.com/.

[5] Available at: https://www.kaggle.com/c/coupon-purchase-prediction.

[6] Available at: https://tianchi.aliyun.com/competition/introduction.htm?spm=5176.100066.333.7.i39riv&raceId=231532.

[7] Available at: https://sites.google.com/a/dlpage.phi-integration.com/pentaho/mondrian/mysql-foodmart-database.

Please cite this article as: A. Monteserin, M.G. Armentano, Influence-based approach to market basket analysis, Information Systems (2018), https://doi.org/10.1016/j.is.2018.01.008

1998). This is a real world dataset that has been used in different works [5,12,21]. The dataset offers information about sales, products, and promotions, among other features. The dataset includes 251,357 transactions representing purchases of 1559 products by 8736 customers.

5.2. Procedure

Since the information about the products was scarce in the datasets, we decided to build the product graph as a complete graph. That is, there exist an undirected edge between every pair of products in the graph. Thus, we eliminated any constraint among products, which allows us to find associations between any pair of them. Although constructing a complete graph when the number of nodes is large is an expensive task, the product graph can be built only one time. To exemplify, the execution time taken to build the largest graph (from Ponpare dataset) was 120.9 s.[8]

After building the graph, we divided the users in a *training set*, by taking the 70% of users with transactions with earliest timestamp, and a *testing set*, by taking the rest of the users. Then, we obtained a seed set of 100 influential products under the CD model by running the greedy algorithm with CELF optimization. Finally, we computed the true influence spread achieved by the seed set according to the testing set of users.

Moreover, to compare the influence spread achieved by the set of influential products, we generated other sets of products following different criteria:

- Best-selling products: this set was composed of the 100 products that appeared in more transactions in the log.
- Highest centrality products: according to Kim et al. [14], this set consisted of the 100 products with more connections to other products, taking into account a co-purchased product network.
- Frequent sequence initiator products: to build this set, we ran the well-known sequential pattern mining algorithm *PrefixSpan* [22]. A sequential pattern-mining algorithm allows us to find all the frequent sub-sequences from a set of sequences, where each sequence consists of a list of elements and each element consists in a set of items (products). Here, each sequence represents all the products purchased by a single customer and each element represents the set of products purchased in the same transaction. Thus, this set consisted of the 100 products that compound the first element of the most frequent sub-sequences.
- SVD products: this set was obtained by applying the *Singular Value Decomposition* (*SVD*[9]) technique on the matrix of customers and products. The singular value decomposition of a matrix A is the factorization of A into the product of three matrices $A = U\Sigma V^T$, where U and V are a matrix whose columns contain the left and right singular vectors of X, respectively. In addition, Σ is a diagonal matrix that contains the singular values of X. Thus, the set of products obtained with this technique was composed of the 100 products with highest score in the right singular vector with highest weight in Σ.
- Most promoted products: this set was composed by the products that were promoted during more days. The promotions include campaigns in daily papers, radio and TV; mail campaigns, and discount coupons, among others. This set was only built for the Foodmart dataset.
- Random products: as a baseline, we decided to generate 10 different sets of 100 random products. The reported results will be an average of these sets.

Then, we computed the true influence spread achieved for each of these sets and compared the results. The true influence spread was computed under the CD model, since the actual spread is best approximated using this model [10]. Basically, this spread indicates how much credits a seed set will receive taking into account the testing users (users that were not taken into account to compute the seed set). From the point of view of our domain, the influence spread represents how much a product in the seed set influences the purchase of other products by taking into account the testing customers (users).

Since the Alibaba dataset contains information about sellers, we followed the same procedure but the experiments were run independently for each seller. We also considered for the experiments sellers with more than 500 products appearing in the transactions set. As a result, the experiments were run for 185 different sellers. To carry out the experiments, we divided the sellers in two groups. First group included 144 sellers that sold between 500 and 1000 products, and second group included 41 sellers with more than 1000. We decided to discard the sellers with less than 500 products because we think that a small number of products and sales could produce not generalizable results.

5.3. Results

5.3.1. Ponpare dataset

Fig. 2 shows a comparison among the true influential spread achieved by the influential, best-selling, highest centrality, frequent sequence initiator and random seed sets according to the seed set size. As we can see, the influential spread achieved by the set of influential products clearly overcame the spread obtained by the other seed sets. Taking into account 100 seeds, this spread was 24.77% higher than the spread obtained by the seed set composed of the frequent sequence initiators. These differences are statistically significant according to the different seed set size ($p < .0001$). It is worth noticing that in this dataset the most influential product (first element of the influential seed set) also was the best sold product and the product with the highest centrality in the co-purchase product network. For this reason, the three influential spreads started in the same point in the plot. However, we can see how the spread achieved by the influential products overcomes the spread of the other seed sets from seed sets size greater than 2. As expected, the average of the random seed sets obtained a really poor performance.

5.3.2. Alibaba dataset

Since we divided the experiments with the Alibaba dataset in two parts, we first present the results disaggregated by seller in Table 3. The first six columns of this table shows the seller ID, the total number of sales (transactions) carried out by the seller (S), the number of different products sold by the sellers (P), the number of different customers who bought products from the seller (C), the sales/product (S/P) and sales/customer (S/C) rates. Se remaining columns show the true influence spread achieved by the influential (I), best-selling (BS), highest centrality (HC), frequent sequence patterns (SP) and SVD sets of products. We think that it is important to show these results disaggregated since there are several differences among the sellers regarding to the information about sales as well as the results obtained.

Additionally, Figs. 3 and 4 show the comparison of the true influence spread for each different seller. As we can see, for 39 out of 41 sellers, the influence spread achieved by the influential seed set overcame the other spreads. Only in two cases, seller 977 and seller 6045, the best-selling and highest centrality set slightly overcame the influential seed set, respectively. Notice that some influence spreads occasionally decreased as the seed set size increased. This occurred when the following situation arose: we analyze a

[8] In a server with two Intel Xeon E5620 at 2.40 Ghz and 32Gb of RAM.

[9] We used Smile (https://haifengl.github.io/smile/) to compute SVD technique.

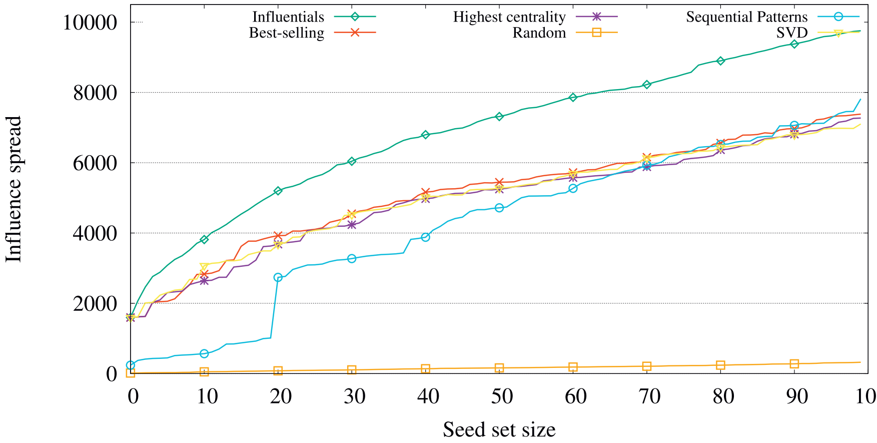

Fig. 2. Influence spread comparison for Ponpare dataset.

Table 3
Statistics about different sellers and true influence spread obtained with different seed sets, highlighting best results for each seller. Results show that the influential seeds in the most of the cases obtained the best influence spreads.

Seller ID	S	P	C	S/P	S/C	I	BS	HC	SP	SVD
504	20,156	1092	5760	18.46	3.50	**158.779**	130.453	133.577	130.712	126.639
528	115,151	1473	12,820	78.17	8.98	**2220.97**	1915.48	1934.18	1894.84	1905.45
529	11,379	1675	6573	6.79	1.73	**158.72**	121.534	124.871	72.5893	115.375
641	4201	1404	2177	2.99	1.93	**293.702**	190.405	161.415	188.531	123.409
935	5090	1297	1693	3.92	3.01	**2056.36**	1400.3	1988.99	1572.78	636.853
977	14,422	1555	2588	9.27	5.57	562.448	**562.918**	537.133	535.599	521.664
1225	15,224	1091	4324	13.95	3.52	**161.513**	137.977	139.95	140.193	132.508
1828	9416	1052	3791	8.95	2.48	**202.151**	173.184	179.194	175.482	168.216
1887	19,344	1223	3358	15.82	5.76	**309.584**	242.524	274.262	242.349	247.86
2154	3991	1019	1825	3.92	2.19	**189.052**	144.534	125.038	145.541	112.592
2468	5698	1108	1700	5.14	3.35	**157.546**	126.31	133.251	127.143	121.308
2920	5797	2061	2820	2.81	2.06	**246.69**	145.584	145.364	40.8915	111.134
2973	13,538	1232	1729	10.99	7.83	**394.041**	343.002	341.957	340.29	352.812
3845	24,665	1419	2992	17.38	8.24	**220.073**	162.684	170.416	157.96	170.794
4230	4516	1219	2041	3.70	2.21	**169.22**	132.476	136.496	131.532	128.187
4455	4894	1123	1439	4.36	3.40	**8941.62**	3646.15	4811.42	4496.33	1448.47
4557	3925	1156	1684	3.40	2.33	**206.728**	133.899	135.755	128.937	113.745
4562	3588	1310	1546	2.74	2.32	**245.016**	164.506	174.957	166.511	139.889
4913	38,593	2399	14,294	16.09	2.70	**222.243**	190.53	198.901	186.42	184.385
5085	3516	1082	1829	3.25	1.92	**243.462**	190.233	143.167	190.421	120.094
5216	14,377	1113	3305	12.92	4.35	**214.892**	177.026	169.907	177.32	172.85
5218	5268	1127	632	4.67	8.34	**197.93**	130.976	141.387	132.689	141.749
5414	8320	1752	1395	4.75	5.96	**1428.29**	835.963	975.332	851.258	792.403
5468	4051	1870	1924	2.17	2.11	**210.388**	127.508	120.283	39.7769	101.513
5524	1295	1013	42	1.28	30.83	**5350.89**	1497.69	1606.23	1761.32	1483.48
5542	3468	1203	1163	2.88	2.98	**477.362**	399.879	237.675	409.389	248.358
6045	11,566	2390	4134	4.84	2.80	368.663	199.942	**380.419**	200.49	159.225
6360	5051	1058	2324	4.77	2.17	**269.016**	182.32	183.029	182.038	140.019
6433	64,121	1110	9181	57.77	6.98	**167.775**	158.037	157.291	159.428	155.704
6545	3613	1142	1398	3.16	2.58	**162.377**	127.551	125.938	127.441	118.075
6792	3981	1362	1849	2.92	2.15	**202.274**	136.653	113.315	138.635	106.9
6883	6791	2304	2665	2.95	2.55	**181.98**	139.464	142.32	138.915	111.764
7065	37,526	1429	6654	26.26	5.64	**218.801**	160.672	165.337	159.701	148.315
7823	5291	3268	1987	1.62	2.66	**256.506**	130.325	75.6982	26.7966	97.834
8007	10,505	1412	2798	7.44	3.75	**356.599**	299.908	319.918	303.918	280.469
8417	4806	2975	2726	1.62	1.76	**364.335**	126.026	92.6549	9.7184	72.299
8743	5778	1615	2628	3.58	2.20	**148.03**	112.613	115.784	112.558	102.408
9093	11,295	9213	5184	1.23	2.18	**503.712**	140.081	119.75	0.0	88.504
9410	8075	1010	4193	8.00	1.93	**143.271**	124.123	125.708	123.913	111.564
9480	4628	1041	896	4.45	5.17	**408.985**	282.106	297.139	285.666	273.109
9987	7151	1430	5001	5.00	1.43	**140.798**	115.493	108.679	50.5189	105.101

S: sales. P: products. C: customers. S/C: transactions per customer. S/P: transactions per product. I: influentials. BS: best-selling. HC: highest centrality. SP: sequential patterns. SVD: singular value decomposition.

seed set with size i and there is a node x that occupies position $i+1$ of the seed set who performed few actions in the action log; then, if the total credits given to the first i nodes of the seed set is higher than the number of actions performed by node x, the influence spread decreases when we analyze the seed set with size $i+1$. This situation is given because, according to the Credit Distribution model, the total credit given to a set of nodes S for influencing a user u on action a when u belongs to S is 1 [10]. Formally, $\Gamma_{S,u}(a) = 1$ if $u \in S$.

In addition, we show the average results for sellers with less and more than 1000 products sold. Figs. 5 and 6 show a comparison among the average true influential spread achieved by the different seed sets for each selected seller with less or more than 1000 products sold, respectively. In both cases, the influen-

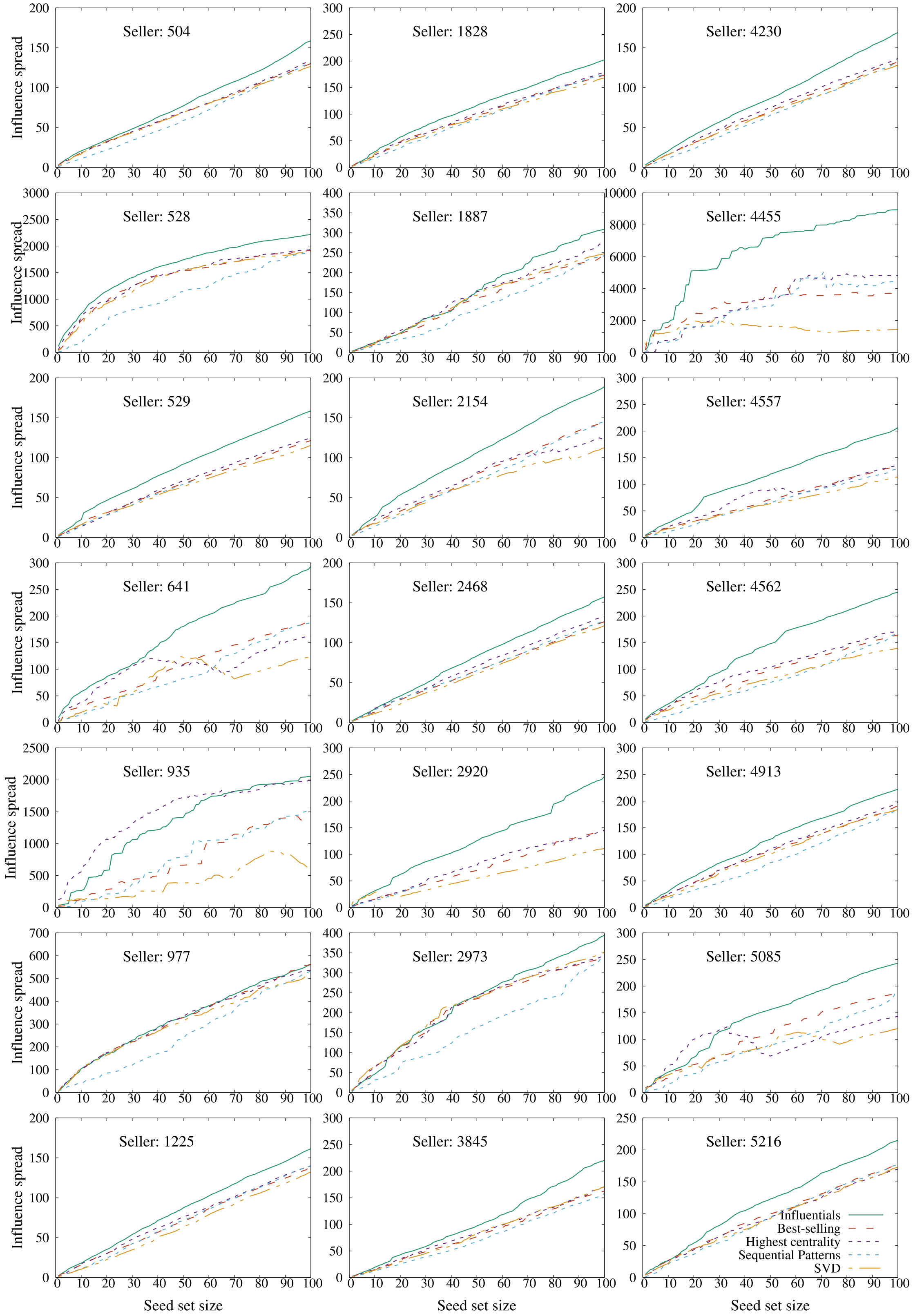

Fig. 3. Influence spread by seller (part 1).

tial spread achieved by our approach clearly overcomes the spread achieved by the other seed sets. As in the Ponpare dataset, the differences of the influential spreads according to the seed set size were statistically significant ($p < .0001$).

5.3.3. *Foodmart dataset*

This dataset showed similar results to the previous case study (Fig. 7). We can observe that the influential seed set, even when comparing it with the spread obtained from the Most Promoted seed set, also obtained the highest influence spread. Moreover, we

Please cite this article as: A. Monteserin, M.G. Armentano, Influence-based approach to market basket analysis, Information Systems (2018), https://doi.org/10.1016/j.is.2018.01.008

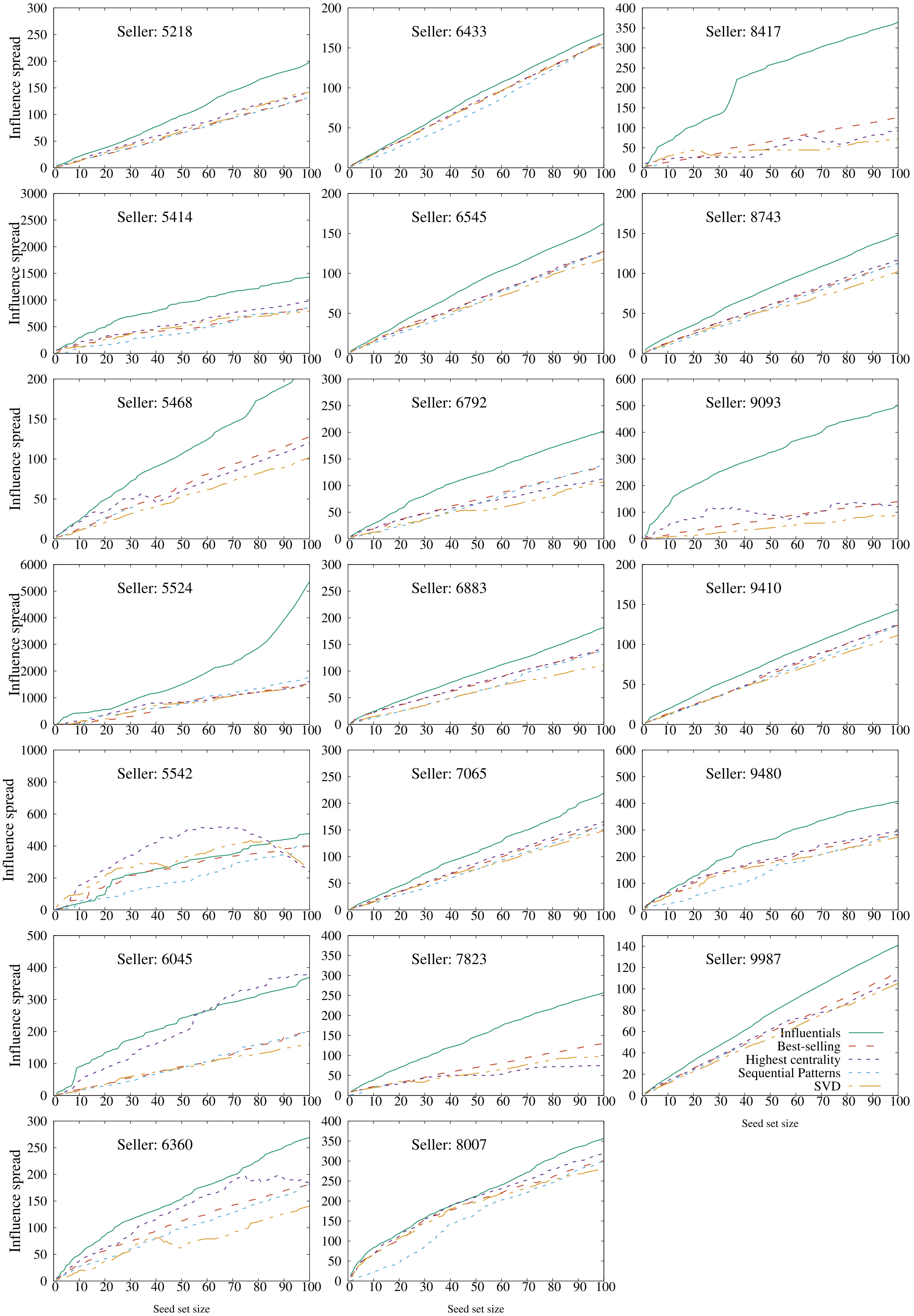

Fig. 4. Influence spread by seller (part 2).

can also note that the difference with the random seed set was smaller. We think that this could indicate a greater parity between the influence power of the products of the dataset.

In addition, since the Foodmart dataset includes information about the products, Table 4 shows the products influenced by the top influencer in this dataset, where the “Total credit” column indicates the total credit given to the top influencer for influencing the influenced product for all customers. The top influencer was the product “Dried Mushrooms” of brand “Ebony”. This product belongs to the “Fresh Vegetables” subcategory of the “Vegeta-

Please cite this article as: A. Monteserin, M.G. Armentano, Influence-based approach to market basket analysis, Information Systems (2018), https://doi.org/10.1016/j.is.2018.01.008

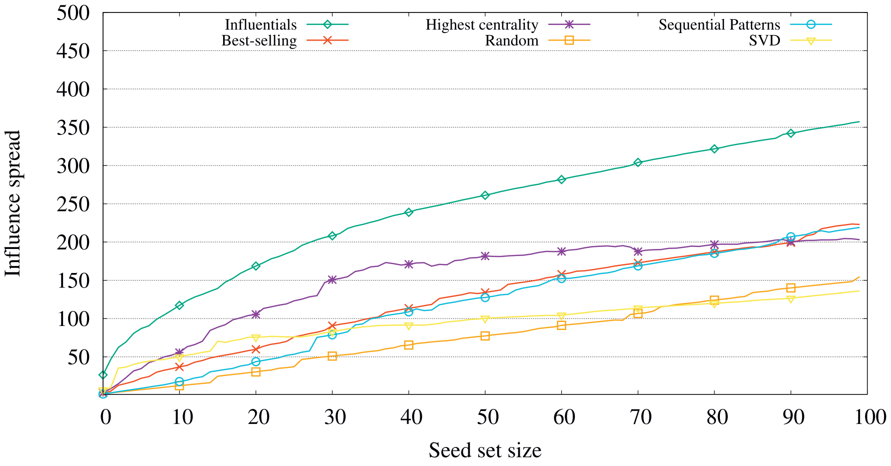

Fig. 5. Average influence spread for Alibaba dataset for sellers with less than 1000 products.

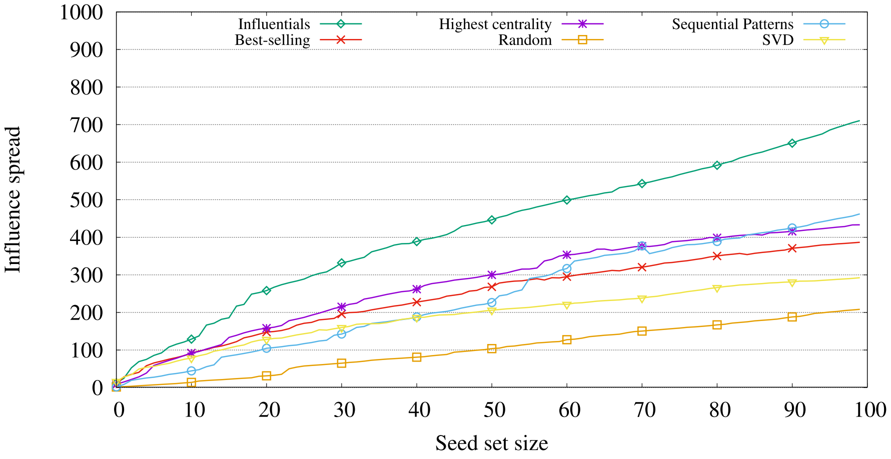

Fig. 6. Average influence spread for Alibaba dataset for sellers with more than 1000 products.

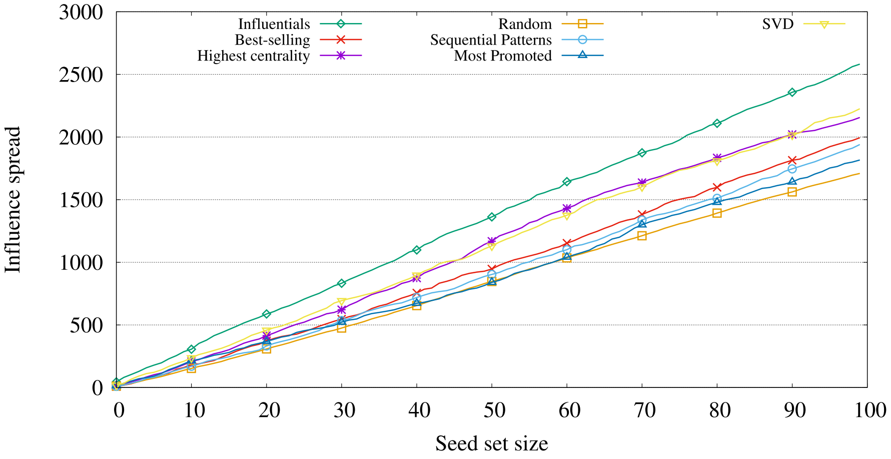

Fig. 7. Influence spread comparison for Foodmart dataset.

bles” category and “Produce” department. As we can see, there is no evident relationship between the influencer and the influenced products: only one influenced products belongs to the same brand (Ebony) that the influencer and two products belongs to the same subcategory (Fresh Vegetables).

5.3.4. Shuffled test: influence vs. correlation

As explained in Section 4, we need to prove the existence of influence among products to validate our results. To do this, we applied the *shuffle* test presented in Section 4 to each dataset used in the experiments. Fig. 8 shows the influential spreads obtained by the influential seed set in the original action log (*Influentials*) and the shuffled one (*Shuffled*). As we can observe, the influential spread obtained with the shuffled action log stands significantly below the spread obtained with the original action log, proving the existence of influence as expected [3,7]. However, we can also observe some differences among the results obtained with

Table 4
Products influenced by the top influencer.

Brand	Product	Subcategory	Category	Department	Total credit
Johnson	Corn Puffs	Cookies	Snack Foods	Snack Foods	1.235
Carrington	Home Style French Fries	Frozen Chicken	Meat	Frozen Foods	1.074
Tell Tale	Sweet Peas	Fresh Vegetables	Vegetables	Produce	1.008
Big Time	Frozen Chicken Breast	Dried Fruit	Snack Foods	Snack Foods	0.944
Horatio	Low Fat Popcorn	Fresh Vegetables	Vegetables	Produce	0.906
Just Right	Fancy Canned Anchovies	Anchovies	Canned Anchovies	Canned Foods	0.889
Fast	Dried Dates	Popcorn	Snack Foods	Snack Foods	0.885
Best Choice	Lemon Cookies	Cereal	Breakfast Foods	Breakfast Foods	0.869
Ebony	Green Pepper	French Fries	Vegetables	Frozen Foods	0.863
CDR	Low Fat Apple Butter	Preserves	Jams and Jellies	Baking Goods	0.861

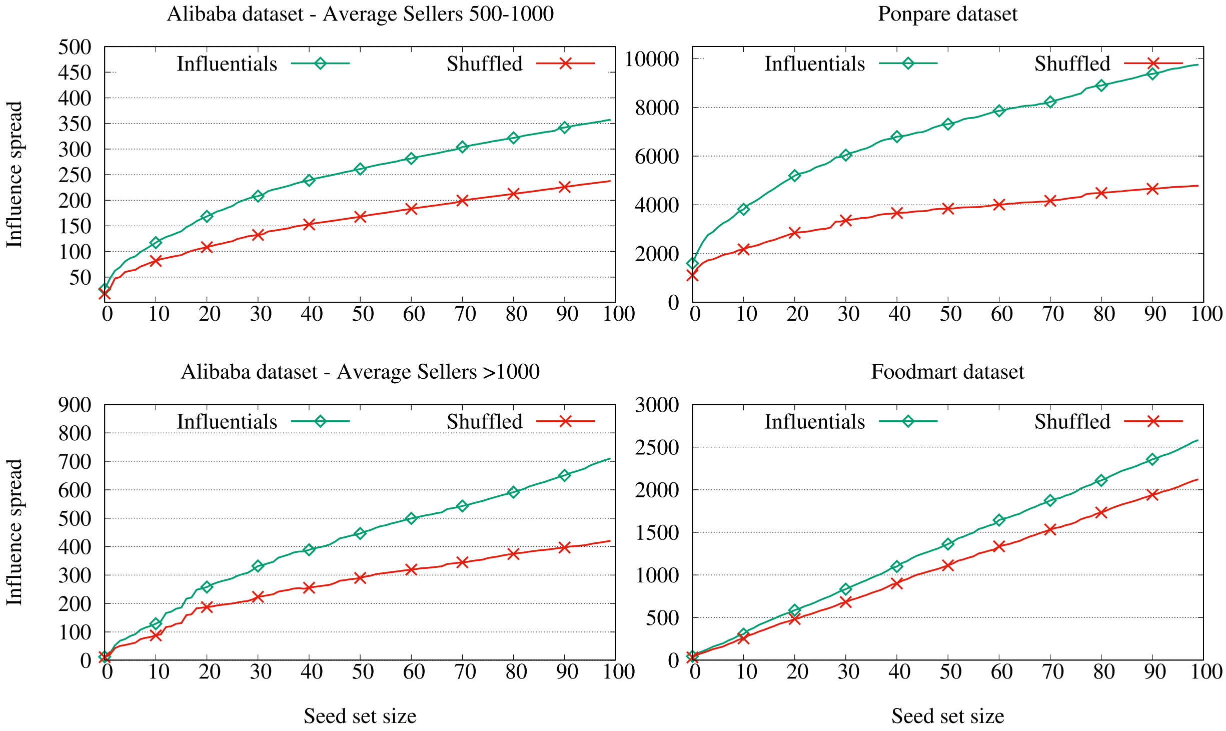

Fig. 8. Results of shuffle test.

different datasets. The highest difference was obtained with the Ponpare dataset (103.88%). In contrast, Foodmart dataset exhibited the lowest difference (21.66%). Moreover, we noted similarities between these differences and the differences between the influential spread obtained by the influential seed sets and the random seed sets. We believe that both comparisons, against the shuffle action log and the random seed set, are indicators of the existence and the strength of the influence among the products in the dataset.

5.3.5. Discussion

In the three datasets analyzed, the experiments clearly showed how the influential products obtained by our approach achieved a highest influence spread. This implies that if we promote the purchase of the influential products, the customers who purchase them would also be influenced to purchase other products. We claim that the results presented demonstrated that the social influence maximization problem better captures the dynamics of the sequence of transactions in the datasets analyzed. Although several works have shown that there are products more influential than others are, these approaches fail to consider two important facts: (1) the influence decay over time and (2) some products are sequentially purchased by a same user in different transactions. This is the case of co-purchase product networks from which previous approaches obtained the highest centrality products as the most influential products. Similarly, the sequential pattern algorithm finds all the frequent sub-sequences from a set of product sequences without considering the time and the length of the sequences. On the other hand, the poor performance of random sets also showed the importance of identify influential products. Furthermore, we observed that the difference between the influence spread of the seed sets and the random set varied depending on the datasets. This could be due to the differences between the influence power of the products of each particular dataset. This fact was also confirmed by the *shuffled* test, which clearly allow us to distinguish between influence and other factors of correlation such as homophily and confounding variables. Moreover, the differences between the influence spread obtained in each particular dataset were due to several intrinsic characteristics of them (i.e. number of transactions, average time between transactions, number of users and products, among others).

Since there is no information about the costs of products in the datasets, an analysis of the *profit* obtained by the potential sales could not be carried out. However, we claim that this information could be taken into account by our approach by modifying the way in which the direct credit is computed (Eq. (3)). Thus, the credit distributed by the algorithm can be directly related to the profit of selling the product (the more profit, the more credit).

Please cite this article as: A. Monteserin, M.G. Armentano, Influence-based approach to market basket analysis, Information Systems (2018), https://doi.org/10.1016/j.is.2018.01.008

The experiments also showed that our approach was able to find influenced products that are not related by brand or category to the influencer. This is an interesting finding since it is well-known in the marketing literature that a product stimulates purchases of other products in the same brand or category. Thus, our approach could give additional information to a marketing analyst.

Finally, our approach could be combined with the traditional approach for finding influential users. Thus, direct marketing could be performed on the influential users by suggesting purchasing the influential products. To do this, it is worth noticing that we need, in addition, a social network of customers.

6. Conclusions and future work

In this article, we presented an approach for mining influential products under the hypothesis that if customer buys these products the total sales volume of the shop will increase. Our approach is based on the concept of social influence, particularly applied to marketing, which is known as social influence marketing or influencer marketing. Differently to influencer marketing, which aims at targeting users with some kind of social influence, our approach aims at targeting products with some kind of *influence* on other products.

The approach presented in this article has several implications for both shops and e-shops. First, it can be used for target marketing, by designing campaigns that send customers who bought influential products discounts or offers on products that she might be also interested in buying. Second, online recommendation engines can be enhanced with our approach by providing suggestions such as “customers who bought this product also were interested in these other products”. Finally, our approach can help e-shop designers, to find ways to display the products related to influential products in such a way that their visibility is highlighted or made more accessible.

As we discussed in the previous section, future works will focus on enriching the direct credit computation with additional information on the products. In addition, we are interested in the analysis of the effects of the structure of the product graph when this is not a complete one. Moreover, we will work in the combination of influential users and influential products. Furthermore, we plan to analyze the use of the proposed approach for personalized product recommendation.

Acknowledgment

This research was partially supported by ANPCyT through PICT project no. 2014-2750.

References

[1] N.S.N. Abadi, M.R. Khayyambashi, Influence maximization in viral marketing with expert and influential leader discovery approach, in: 8th International Conference on e-Commerce in Developing Countries: With Focus on e-Trust., 2014, pp. 1–8.

[2] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. SIGMOD 93., ACM, New York, NY, USA, 1993, pp. 207–216.

[3] A. Anagnostopoulos, R. Kumar, M. Mahdian, Influence and correlation in social networks, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2008, pp. 7–15.

[4] M.G. Armentano, A. Monteserin, J. Tang, V. Yannibelli, Preface, in: Proceedings of the 2nd International Workshop on Social Influence Analysis co-located with 25th International Joint Conference on Artificial Intelligence (IJCAI 2016). Vol. 1622. CEUR Workshop Proceedings, 2016.

[5] A. Augustin, P. Vince, V.G. Nair, High utility itemset mining with top-k CHUD (TCHUD) algorithm, Int. J. Comput. Appl. 165 (3) (2017).

[6] T. Brijs, G. Swinnen, K. Vanhoof, G. Wets, Using association rules for product assortment decisions: a case study, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '99, ACM, New York, NY, USA, 1999, pp. 254–260, doi:10.1145/312129.312241.

[7] C. Chen, W. Li, D. Gao, Y. Hou, Exploring interpersonal influence by tracking user dynamic interactions, IEEE Intell. Syst. 32 (3) (2017) 28–35.

[8] G.H. Golub, C. Reinsch, Singular value decomposition and least squares solutions, Numer. Math. 14 (5) (1970) 403–420.

[9] A. Goyal, F. Bonchi, L.V. Lakshmanan, Learning influence probabilities in social networks, in: Proceedings of the Third ACM International Conference on Web Search and Data Mining. WSDM '10, ACM, New York, NY, USA, 2010, pp. 241–250.

[10] A. Goyal, F. Bonchi, L.V.S. Lakshmanan, A data-based approach to social influence maximization, PVLDB 5 (1) (2011) 73–84.

[11] A. Goyal, W. Lu, L.V. Lakshmanan, Celf++: optimizing the greedy algorithm for influence maximization in social networks, in: Proceedings of WWW 2011 (20th International World Wide Web Conference, 2011, pp. 47–48.

[12] R. Henriques, C. Antunes, Generative modeling of itemset sequences derived from real databases, in: ICEIS, 1, 2014, pp. 264–272.

[13] D. Kempe, J. Kleinberg, E. 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. KDD '03, ACM, New York, NY, USA, 2003, pp. 137–146.

[14] H.K. Kim, J.K. Kim, Q.Y. Chen, A product network analysis for extending the market basket analysis, Expert Syst. Appl. 39 (8) (2012) 7403–7410. http://www.sciencedirect.com/science/article/pii/S0957417412000796.

[15] C. Li, B.C. Ooi, A.K.H. Tung, S. Wang, Dada: a data cube for dominant relationship analysis, in: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. SIGMOD '06, ACM, New York, NY, USA,, 2006, pp. 659–670. http://doi.acm.org/10.1145/1142473.1142547.

[16] G.S. Linoff, M.J.A. Berry, Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 3rd, Wiley, 2011.

[17] C. Long, R.C.W. Wong, Viral marketing for dedicated customers, Inf. Syst. 46 (2014) 1–23. http://www.sciencedirect.com/science/article/pii/S0306437914000751.

[18] S.A. Naik, Q. Yu, Maximizing influence of viral marketing via evolutionary user selection, in: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013, 2013, pp. 1435–1436.

[19] S.A. Naik, Q. Yu, Evolutionary Influence Maximization in Viral Marketing, Springer International Publishing, Cham, 2015, pp. 217–247.

[20] D.L. Olson, Market Basket Analysis, Springer Singapore, Singapore, 2017, pp. 29–41.

[21] A. Oroojlooyjadid, L.V. Snyder, M. Takác, Applying deep learning to the newsvendor problem, CoRR (2016). arXiv:1607.02177. http://arxiv.org/abs/1607.02177.

[22] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, M.C. Hsu, Mining sequential patterns by pattern-growth: the prefixspan approach, IEEE Trans. Knowl. Data Eng. 16 (11) (2004) 1424–1440.

[23] T. Raeder, N.V. Chawla, Modeling a store's product space as a social network, in: Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining. ASONAM '09. IEEE Computer Society, Washington, DC, USA, 2009, pp. 164–169. https://doi.org/10.1109/ASONAM.2009.53.

[24] S.C. Tan, P.S. Lau, X. Yu, Finding Similar Time Series in Sales Transaction Data, Springer International Publishing, Cham, 2015, pp. 645–654.

[25] I.F. Videla-Cavieres, S.A. Ríos, Extending market basket analysis with graph mining techniques: a real case, Expert Syst. Appl. 41 (4) (2014) 1928–1936. http://www.sciencedirect.com/science/article/pii/S0957417413007094.

[26] A. Vlachou, C. Doulkeridis, K. Nørvåag, Y. Kotidis, Identifying the most influential data objects with reverse top-k queries, Proc. VLDB Endow. 3 (1–2) (2010) 364–372. https://doi.org/10.14778/1920841.1920890.

[27] K. Wang, S. Zhou, J. Han, Profit Mining: From Patterns to Actions, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 70–87.

[28] R.C.W. Wong, A.W.C. Fu, K. Wang, MPIS: maximal-profit item selection with cross-selling considerations, in: Third IEEE International Conference on Data Mining, 2003, pp. 371–378.

[29] R.C. Wu, R.S. Chen, C.C. Chang, J.Y. Chen, Data mining application in customer relationship management of credit card business, IEEE Comput. Soc. (2005) 39–40. Proceedings of the 29th Annual International Conference on Computer Software and Applications Conference. COMPSAC-W'05 Washington, DC, USA http://dl.acm.org/citation.cfm?id=1890517.1890537.

Please cite this article as: A. Monteserin, M.G. Armentano, Influence-based approach to market basket analysis, Information Systems (2018), https://doi.org/10.1016/j.is.2018.01.008