Marvasti, Neda B.; Huhtala, Juho-Petteri; Yousefi, Zeinab R.; Vaniala, Iiro; Upreti, Bikesh; Malo, Pekka; Kaski, Samuel; Tikkanen, Henrikki

Is this company a lead customer? Estimating stages of B2B buying journey

Research paper

Is this company a lead customer? Estimating stages of B2B buying journey

Neda B. Marvasti [a],[1],[*], Juho-Petteri Huhtala [b],[1],[*], Zeinab R. Yousefi [a], Iiro Vaniala [b], Bikesh Upreti [c],[d], Pekka Malo [c], Samuel Kaski [a], Henrikki Tikkanen [b]

[a] Departement of Computer Science, Aalto University School of Science, Espoo, Finland
[b] Departement of Marketing, Aalto University School of Business
[c] Department of Information and Service Economy, Aalto University School of Business
[d] UQ Business School, The University of Queensland, St Lucia, QLD 4072, Australia

A B S T R A C T

By advancement in digital marketing, business-to-business (B2B) buyers carry out over half of the buying process through digital touchpoints before they establish any significant contact with the B2B seller. Knowing the buying stage of a potential buyer can bring a substantial advantage to the B2B seller given the complexity of the transaction and the associated value. In this paper, the authors propose a machine learning approach to infer the stages of the B2B buying journey by observing the online browsing behavior of buyer companies. It is shown that observing the buyer's online behavior effectively allows us to estimate the buying stages with high accuracy by utilizing the hidden Markov models. Managers in B2B seller companies may use these techniques for adjusting their marketing efforts to improve the fit with the information demands of the B2B buyer prospects along with their buying journey, and thus, improve the hit rate of marketing and sales activities.

1. Introduction

Due to the development of digital information technologies (DIT) and a huge increase in the amount of digital information is being generated, stored, and made available for analysis, business-to-business (B2B) marketing has become more digital and buyer-driven than before. For example, according to a survey by Schwartz & Kim, 2012, more than 70% of buyers kickstart their buying journey digitally with a Google search to improve their understanding of the market and products. Further, as the buying journey progresses, digitally-driven information search constitutes a major portion of the B2B buying journey as buyers spend a considerable amount of time searching and analyzing digitally available information. Today, a tremendous amount of digitally available information has reduced the buyers' dependency on sellers for the relevant information and buyers can access a wide variety of information without involving the sellers in the process. As a result, it is estimated that when a buyer approaches a complex B2B sale, roughly 60% of the buying journey is already completed even before the potential buyer comes in contact with sellers' representatives (Grewal et al.,

2015).

B2B buyers are continuously seeking information that can be used to support their reasoning along the buying journey (Steward, Narus, & Roehm, 2018; Steward, Narus, Roehm, & Ritz, 2019). Today, these potential customers are leaving vast amounts of digital traces from their information-seeking behavior. Simultaneously, the B2B sellers also have a massive amount of behavioral customer data at their disposal, arising from the digital information search, which they are often unable to utilize (Kwon, Lee, & Shin, 2014). While analyzing behavioral customer data can potentially enable B2B sellers to improve their understanding of the B2B buying journey and decision-making, such an endeavor entail significant challenges related to processing and synthesizing behavioral browsing data.

As Steward et al. (2019) noted, compared to their B2C counterparts, B2B sellers are currently lagging in the effective use of customer data. The underlying problem is that virtually all academic research in analytics is focused on business-to-consumer markets. To make a significant leap of improvement in B2B analytics, academic research should urgently pay more attention to the rhythm and flow of the B2B buying

* Corresponding authors.
E-mail addresses: neda.marvasti@aalto.fi (N.B. Marvasti), juho-petteri.huhtala@aalto.fi (J.-P. Huhtala), zeinab.rezaeiyousefi@aalto.fi (Z.R. Yousefi), iiro.vaniala@aalto.fi (I. Vaniala), bikesh.upreti@aalto.fi (B. Upreti), pekka.malo@aalto.fi (P. Malo), samuel.kaski@aalto.fi (S. Kaski), henrikki.tikkanen@aalto.fi (H. Tikkanen).
[1] Equally contributing authors.

journey and provide sophisticated methods to track B2B buying journey without making the tracking process costly and more complicated (Steward et al., 2019). Our paper aims to provide the means to start building data and a fact-based understanding of the B2B buying journey.

B2B analytics is no B2C analytics and it has its own distinctive characteristics which, to a large extent, makes the application of existing customer analytics solutions to the B2B market nearly useless (Lilien, 2016). Further, Lilien (2016) also suggests that while developing a B2B analytics solution, researchers should embrace the distinctive factors of B2B markets, such as complexity and length of the buying process. Particularly, to know user transitions during the buying process, B2B sellers need to gain insights about buyers' information search, browsing behavior, and content-related needs and interests. Despite the complexity of B2B buying (Lilien, 2016), the underlying notion has been that the B2B seller still can widely control the B2B buying journey. Notably, it has been argued that B2B sellers may strategically manage when, where, and how their customers are engaged with them (Brennan & Croft, 2012). This view has been emphasized in many empirical and conceptual studies focused on determining the activities and sequence of actions performed in various B2B buying stages (see, e.g., Johnston & Lewin, 1996). The recent transition from seller-directed to buyer-directed approach has pushed B2B sellers to rethink the dynamism of B2B buying behavior process fundamentally, and precisely, how marketing and sales activities should be aligned to better match with the actual phase of B2B buying journey. The development of DIT can be considered as a double-edged sword for B2B sellers. Although it has made the buying process more buyer-driven, it has also brought numerous opportunities to understand the B2B buying behavior per se better. Seminal understanding of the B2B buying process is offered by research that is mostly either qualitative or conceptual (Johnston & Lewin, 1996). As information-seeking behavior occurs in real-time in the digital environment, B2B sellers may now develop tools and techniques to systematically collect and analyze large sets of behavioral data and gain objective insights from the actual B2B buying process. This knowledge can be used to adjust marketing and sales activities better and eventually increase the efficiency and hit rate of these activities (Kiang & Chi, 2001).

However, many firms are still struggling with these techniques, due to 1) lack of experience and understanding of data analysis and 2) working with inconsistent datasets (Kwon et al., 2014). Moreover, tightening data protection regulations, such as the General Data Protection Regulation of the EU (GDPR), is already making it even more challenging to obtain and use behavioral data of customers in the future. There are currently available numerous digital technologies that claim to help the B2B sellers to utilize their big data and map the buying journey with the aid of Artificial Intelligence (Paschen, Pitt, & Kietzmann, 2020). Deployment of these technologies requires substantial investments in the technological infrastructure, which has made the implementation of these systems less appealing from the B2B seller point-of-view (Dwivedi et al., 2019; Mikalef & Gupta, 2021).

In this paper, we aim to develop a model that can predict the stage of the buying journey based on the browsing buying behavior. The bearing idea is to develop a model that 1) uses data that is accessible to all B2B sellers with digital presence 2) can be implemented to B2B sales without heavy and costly investments in technological infrastructure. In line with Information Foraging Theory (IFT) as a basis for understanding B2B buying behavior, we use the browsing activities as a proxy of behavior to capture the latent states in the buying journey. IFT assumes that customers will "modify their strategies or the structure of the environment to maximize their rate of gaining valuable information" (Pirolli & Card, 1999, p. 643). Thus, browsing data can be used to proxy latent stages of the B2B buying process (see, e.g., D'Agostino, Gasparetti, Micarelli, & Sansonetti, 2016; Siriaraya et al., 2017).

However, the environment has often "patchy" structure – For instance, content relevant to a specific B2B buyer may reside in piles of multiple sources, such as in various content categories (Pirolli & Card, 1999). To capture these valuable patches empirically, extant studies have been using cluster analysis (see, e.g., Lawrance, Bellamy, & Burnett, 2007; Pirolli & Card, 1999). Following the same path, we categorized the web pages of seller companies into eight different groups and studied the online browsing behavior of customers in these categories. We recognize that almost every buying decision is group-based. Although the decision is group-based, it is influenced by individual browsing and individual actions, which was the seminal argument of behavioral organizational studies (Cyert & March, 1963). Thus, we operate with the browsing data collected from individuals.

In this paper, we propose an approach of modeling the sequential online browsing behavior of customers with a hidden Markov model (HMM) to estimate the corresponding latent B2B buying stage of each observed day. To test the performance of our proposed model, we used both simulated and real data. To the best of our knowledge, this is the first academic study that focuses on predicting the stages of the B2B buying journey rather than conceptualizing the stages of buying. This information helps B2B seller companies to adjust their marketing efforts to fit them with the information needs of potential customers, which typically increases the efficiency and hit rate of the marketing efforts (Kiang & Chi, 2001).

This paper is organized as follows: section Background discusses the theory behind our modeling, and the technical part of our proposed machine learning model is implemented in the section Technical Implementation. Simulation and real experiments are presented and discussed over in sections Simulation Study and Real-world Dataset Experiments, respectively. Finally, the results and impact of our work are discussed and concluded in sections Discussion and Conclusions.

Note that throughout this manuscript, bold lowercase letters (e.g., **a**) and bold capital letters (e.g., **A**) denote the vector quantities and matrices, respectively. Elements of matrices and vectors are shown in lowercase letters, e.g., a_{ij} denotes the element in the ith row and jth column of the corresponding matrix.

2. Background

As the B2B buying process differs from the B2C buying, consequently, it is also widely recognized that the B2B buying behavior differs vastly from that of the consumers (Grewal et al., 2015). For instance, it is more likely that in B2B context more than one person is involved in the buying decision process in various kinds of roles (Grewal et al., 2015). Also, according to Lilien, 2016 and Wouters, 2004 B2B buying is more heterogeneous than the buying behavior of consumers as the buying process varies a lot among firms and situations.

Although every B2B buying process is claimed to be unique (Woodside, 1996), there have been many attempts to generalize and conceptualize patterns in B2B buying (see Chavan, Chaudhuri, & Johnston, 2019; Steward et al., 2019). Often B2B buying process is divided into multiple stages, states, or steps that present linearly progressive information-seeking and/or decision-making behavior of B2B buyers and spans both ends of before and after the actual purchase (see, e.g., Hutt & Speh, 2007; Johnston & Lewin, 1996). B2B buying begins from a business problem that the B2B buyer firm needs to solve and can involve a group of decision-makers that dictate the buying behavior throughout the buying process (Grewal et al., 2015). Thus, conceptualizations of the B2B buying process often include and start with an awareness stage where the actual business problem is recognized and further clarified. After gaining a better understanding of the essential need, B2B buyer moves to an evaluation stage where it explores and searches for possible solutions and assesses the strengths and weaknesses of each potential option. Then in a decision stage, B2B buyer attempts to decrease the number of options, commits to a specific set of options (for instance, a shortlist of 2–3 suppliers that could help the company to solve the problem), and finally validates and justifies one of the options.

In this paper, rather than using a fine-grained conceptualization of

the B2B buying process (see, e.g., Hutt & Speh, 2007), we assumed that during the B2B buying process, the buyers move from one stage to another over time. These buying stages are hidden from the sellers, and the critical data to proxy these stages constitutes the browsing behavior of potential B2B buyers. Based on this rather noisy browsing data, hidden buying stages can be inferred using hidden Markov models. HMM has been used in various marketing contexts such as in modeling the effect of dynamics of customer interactions, in estimating the relationship of customer value and marketing investment (Kumar, Sriram, Luo, & Chintagunta, 2011), in measuring the return of investment in B2B marketing (Luo & Kumar, 2013), in analyzing the impact of pricing on B2B relationship (Zhang, Netzer, & Ansari, 2014) and in assessing the effectiveness of relationship marketing strategies in B2B (Zhang, Watson Iv, Palmatier, & Dant, 2016).

In our study, we divide the B2B buying journey into three core stages related to buying (which are hidden for the B2B seller): Early-funnel, Middle-funnel, and Late-funnel (Toman, Adamson, & Gomez, 2017). As in each of these stages, B2B buyers study a wide range of information and explore numerous options. Thus, sharing meaningful and appropriate information that fits their needs in each stage becomes crucial from the B2B sellers' point of view (Toman et al., 2017). In our conceptualization, we also take into account that B2B buyers may not be in the mode of problem-solving/buying at all, and therefore, can be identified to be in a No-funnel stage.

3. Theory

In the B2B setting, the buyers are engaged in a lengthy digital information search where the objective is to collect the necessary information required at the various stages of the buying process. When users are in pursuit of the information, to meet their informational need, complete a task, or execute a transaction, they adopt a various set of search strategies with different keywords, browsing across various web pages. B2B buyers also endure a similar journey where they follow an information search trajectory led by cues and signals directed towards collecting relevant information. To understand this online information-seeking behavior of the B2B buyer, at the various buying stages, we draw on Information Foraging Theory (IFT) (Pirolli & Card, 1999). While searching information for online transactions, users exhibit a tendency to follow information scent (Spool, Perfetti, & Brittan, 2004). The information scent of a search keyword is its ability to return the results with the information that the users perceive valuable. For instance, if a google search with certain keywords lists out the websites with the words that a user perceives valuable, the user is more likely to go after information scent by clicking a specific link that redirects the user to a website. As stated earlier, IFT theory posits that the users "modify their strategies or the structure of the environment to maximize their rate of gaining valuable information" (Pirolli & Card, 1999, p 643). Similar information search behavior is prevalent in B2B buying where the individuals from the B2B buyer company follow the information cues and scent to meet their information need. In doing so, the browsing also provides valuable information traces of the browsing activities which B2B sellers can utilize as a proxy of buyers' behavior. The online browsing data can be used as a proxy to the buyers' interest (see, e.g., D'Agostino et al., 2016; Siriaraya et al., 2017) as the data capture digital traces generated when the B2B buyers browse for information following information scent. Thus, in this study, we rely on browsing data to infer the buyers' interest in the content provided by sellers. Further, we posit that the browsing data also encapsulates the changes in information-seeking behavior of the buyers at the various B2B buying stages and the evolving interest of buyers is a proxy to the transition of the latent buying stage. As the information is spread across multiple silos within the websites and across various websites, the B2B buyers have to navigate across a "patchy" structure and thus, generates a significant amount of digital traces that vary across different stages of buying. For example, a B2B buyer at an early stage of buying stages is more likely to browse a

wider range of information on various alternatives while the information search gets narrower and deeper as the buying progresses to the later stages. Based on the given theory, our hypothesis for the rest of this paper implies that stages of the buying process can be estimated from online browsing behavior.

4. Technical implementation

4.1. Proposed modeling framework

Fig. 1 illustrates four stages of the B2B buying process, namely 1: No-funnel, 2: Early-funnel, 3: Middle-funnel, and 4: Late-funnel, and the possible transition between them during the buying journey. As can be observed in this figure, the B2B buying process can always go forward, remain in the same stage, or transit back to the 1st stage. Transition to the "No funnel" -stage means that the buying process ends (i.e., B2B buyer is no longer in the mode of buying).

The potential buyers' online browsing behavior can be monitored by their visits to B2B seller's website, which are considered as observations in our method and shown by vector $\boldsymbol{f}$. In total, eight web page categories for a seller's web pages (f_1: products, f_2: detailed-products, f_3: news, f_4: recruitment, f_5: references, f_6: support, f_7: investors, and f_8: contacts) are assigned based on the expert interviews conducted in the same research project (Saaranto, 2019).

Fig. 2 illustrates a sequence of observations during a year with a time step of one month. We assume that each buying stage has a specific buying-related browsing behavior, which is targeted to be revealed here.

As shown in Fig. 3, the buying stage of a buyer at time step t, in addition to its observed online behavior at that time, also depends on buying stage of the buyer in the previous time step $t - 1$. A potential statistical model to analyze such dependencies is HMM. Note that, in the methodological part of this study, the term "state" can be interchangeably utilized to refer to the buying stage of the B2B buyer.

4.2. Hidden Markov model

In this model, a sequence of noisy data is observed, while the underlying hidden states, which emit such observations, are unknown. This model contains three types of parameters: the initial probability vector ($\boldsymbol{\pi}$), the transition probability matrix ($\boldsymbol{A}$), and the emission probability matrix ($\boldsymbol{B}$). The initial probability ($\boldsymbol{\pi}$) defines the probability of starting the journey at each specific state. The transition probability, $a_{i,j} = P(h_t = j | h_{t-1} = i)$, shows the dynamics of the state, i.e., moving from a specific state i at previous time step ($h_{t-1} = i$) to state j in the current time step ($h_t = j$). The emission probability, $\boldsymbol{b_i} = P(\boldsymbol{f_t} | h_t = i)$, shows the probability of emitting/producing a specific observation vector (browsing behavior), $\boldsymbol{f_t}$, in a given state (i). In general, if these probabilities are known, the probability of state ($\boldsymbol{h_t}$) for each of its possible values conditioned on all of the observations ($\boldsymbol{f}_{1:t}$) up to that time, $P(\boldsymbol{h_t}|\boldsymbol{f}_{1:t})$, can be computed from the joint probability $P(\boldsymbol{h_t},\boldsymbol{f}_{1:t})$ via the forward (filtering) algorithm (Bishop, 2006). The final equation is shown in Eq. (1)

$$P(\boldsymbol{h_t}|\boldsymbol{f}_{1:t}) \propto P(\boldsymbol{h_t},\boldsymbol{f}_{1:t}) = P(\boldsymbol{f_t}|\boldsymbol{h_t})\sum_{\boldsymbol{h}_{t-1}}P(\boldsymbol{h}_{t-1},\boldsymbol{f}_{1:t-1})P(\boldsymbol{h_t}|\boldsymbol{h}_{t-1}) \quad (1)$$

However, since the parameter set of our model φ, which consists of initial probability vector, transition probability matrix, and emission probability matrix ($\varphi = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}\}$) is also unknown, it should initially be estimated from the data (observed buying behavior) itself. To this end,
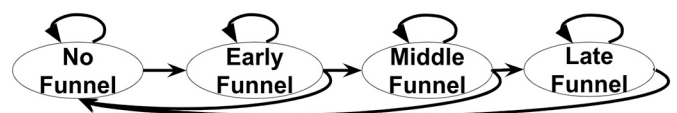

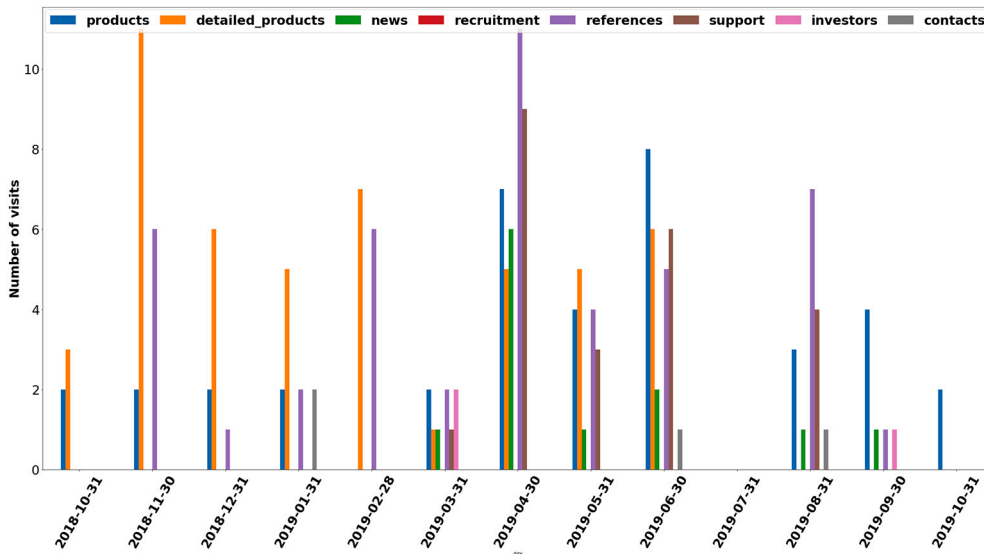**Fig. 1.** The stage transition diagram of the B2B buying process.

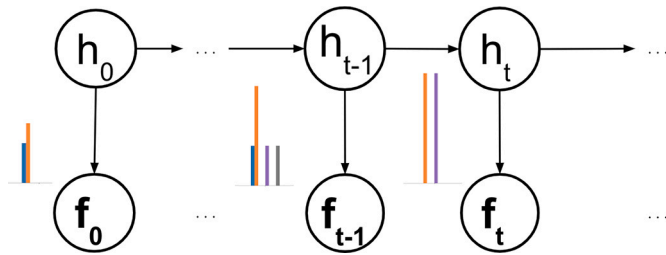**Fig. 2.** The visit frequencies of a buyer company to a seller website.

Fig. 3. Relation between B2B buying stages and their corresponding observations, where t stands for the time, and h_t and f_t represent hidden buying stage value and observed browsing behavior vector at time t, respectively.

the well-known Baum-Welch algorithm is utilized, which iteratively maximizes the log-likelihood of observations (Eq. (2)) and updates the parameters of the model to make it more similar to the optimal set, that explains the observed data the best (Baum, Petrie, Soules, & Weiss, 1970; Bishop, 2006). Each iteration is guaranteed to increase the log-likelihood of the observed data. However, as the solution may converge to a local maximum instead of the global optimal solution, the parameter estimation task is performed multiple times with different initialization to select the parameters, which best describe the observations.

$$log(P(\boldsymbol{f}_{1:T})) = log\left(P(\boldsymbol{f}_1)\prod_{t=2}^{T}P(\boldsymbol{f}_t|\boldsymbol{f}_{1:t-1})\right) \tag{2}$$

4.3. Model setup

Baum-Welch algorithm has two steps: the expectation step and the maximization step. Starting from an initial guess of the model parameters in the expectation step, the probability of each buying state given the observed data and initial parameters are computed ($P(h_t^{(s)}|\boldsymbol{f}_{1:T}^{(s)},\varphi)$), where s is the buyer company index in the dataset, and T shows the total number of time steps in the observed data. Then, in the maximization step, parameters of the model are updated ($\varphi \leftarrow \varphi^*$) by solving Eqs. (3)–(5), where i and k are the state and category indices, respectively.

$$\pi_i^* = \frac{\sum_{s=1}^{S}P\left(h_1^{(s)} = i|\boldsymbol{f}_{1:T}^{(s)},\varphi\right)}{S} \tag{3}$$

$$a_{i,j}^* = \frac{\sum_s \sum_{t=2}^{T} P\left(h_{t-1}^{(s)} = i, h_t^{(s)} = j|\boldsymbol{f}_{1:T}^{(s)},\varphi\right)}{\sum_s \sum_{t=2}^{T} P\left(h_{t-1}^{(s)} = i|\boldsymbol{f}_{1:T}^{(s)},\varphi\right)} \tag{4}$$

$$b_{k,i}^* = \frac{\sum_s \left[\sum_{t=2}^{T} f_{k_t} P\left(h_t^{(s)} = i|\boldsymbol{f}_{1:T}^{(s)},\varphi\right)\right]}{\sum_s \left[\sum_{t=2}^{T} \left(\sum_k f_{k_t}\right) P\left(h_t^{(s)} = i|\boldsymbol{f}_{1:T}^{(s)},\varphi\right)\right]} \tag{5}$$

5. Simulation study

5.1. Simulator

To assess the proposed algorithm performance, a simulator is designed with parameters matching with those of the actual B2B buying process to replicate the behavior of a buyer company visiting a seller 'company's website for purchasing complex B2B products. A set of artificial data generated by this simulator is employed to estimate the underlying buying state of each simulated company. Results are then compared to the original buying state in the artificial dataset.

The simulator (Fig. 4) is designed based on the extant conceptual work in B2B marketing (Hutt & Speh, 2007; Toman et al., 2017) and interviews available in Saaranto's thesis (Saaranto, 2019). Note that, although the general buying behavior of all buyer companies is considered to be similar, we individually allowed the possibility of having heterogeneous preferences and browsing behavior for each of them.
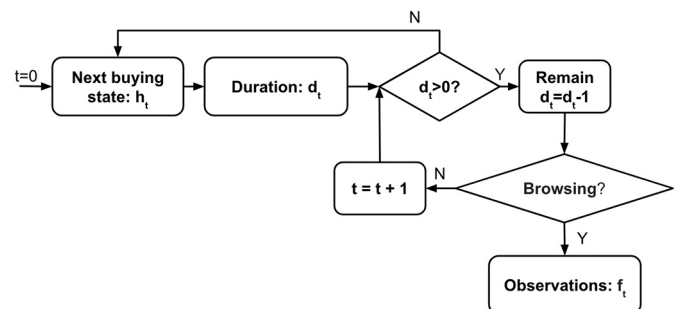

Fig. 4. Block diagram of the designed simulator for the B2B buying process.

As illustrated in Fig. 4, the process starts from time step zero, in which the initial buying stage of the buyer company is randomly selected from one of the buying stages by sampling from a categorical distribution $h_0 \sim categorical(\pi)$, where π is defined as a sample from Dirichlet distribution, i.e., $\pi \sim Dir(\lambda_{1:4})$ and $\lambda_{1:4}$ are Dirichlet parameters defining the prior probability on each buying stage.

At each time step, the simulator should decide the buying stage (h_t) of the simulated buyer and also the duration (d_t) of the buying stage. As shown in Eqs. (6) and (7), if the assigned duration of the previous buying state (d') is zero, new values for the next state and its corresponding period will be given, otherwise, the state remains unchanged while its duration decrements one time step.

$$P\left(h_t = j | h_{t-1} = i, d_{t-1} = d'\right) = \begin{cases} a_{i,j}, & if\ d' = 0 \\ \delta(i,j), & if\ d' > 0 \end{cases} \tag{6}$$

$$P(d_t = d | d_{t-1} = d', h_t = j) = \begin{cases} P(d_t = d | h_t = j); & if\ d' = 0 \\ \delta(d, d'-1); & if\ d' > 0 \end{cases} \tag{7}$$

The next state is selected via Eq. (6), which depends on both duration and state of the previous time step where $\delta(i,j)$ an indicator function is set to one if $i = j$ and zero otherwise and $a_{i,j}$ is the transition probability from state i to state j defined in Eq. (8) in which vector a_i is the transition probability from state i to each of four different states, including itself and $\alpha_{i,j}$ are the Dirichlet parameters.

$$\begin{cases} a_1 = Dir([0, 1, 0, 0]) \\ a_2 = Dir([\alpha_{2,1}, 0, \alpha_{2,3} + d, 0]) \\ a_3 = Dir([\alpha_{3,1}, 0, 0, \alpha_{3,4} + d]) \\ a_4 = Dir([1, 0, 0, 0]) \end{cases} \tag{8}$$

The duration that a buyer spends in each stage (d_t) is randomly generated with probability $P(d | h_t = i)$ according to Eq. (9), where γ_j is the average duration spent in state j and $P(h_{t+1} = j + 1 | h_t = i)$ is the stationary transition probability between different buying stages defined in Eq. (8). The second (Early) and third (Middle) stages of B2B buying may take months due to the complexity of the B2B buying process (Lilien, 2016). Thus, mixed distribution is defined for the duration of different buying states, and also the longer the duration, the more probable the process goes forward rather than quit.

$$P(d | h_t = j) = \begin{cases} Geometric(P(h_{t+1} = j + 1 | h_t = j)); & if\ j = 1\ or\ 2 \\ Gamma(\gamma_j); & if\ j = 3\ or\ 4 \end{cases} \tag{9}$$

In every time step, to check whether a browsing behavior on the 'seller's website is observed, a binary variable O_t is sampled from Bernoulli distribution $(O_t \sim Bernouli(\theta))$, where $\theta \sim Beta(\kappa, 30 - \kappa)$ control the sparsity of the 'users' online behavior. If the browsing behavior is observed $(O_t > 0)$, an observation vector containing the frequency of visits to different web page categories is randomly generated by sampling from the emission distribution conditioned on the current state as defined in Eq. (10), where $b_i \sim Dir(\beta_i)$ is a vector representing the probability of visiting each web page category conditioned on state i, and $N_i \sim Poisson(n_i)$ and n_i are respectively the total number of web page visits and the average of total daily visits in state i. The process generates a sequence of observations (web page visits) for different simulated companies.

$$f_t = Multinomial(N_i, b_i) \tag{10}$$

5.2. Experimental results

Three different sets of simulated data are generated using the simulator above to test the performance of the proposed algorithm. Each dataset contains 300 different simulated B2B buyer companies, for each of which, a sequence of observations (web page visits) are generated. The length of these observations is different due to the randomness of the duration d_t and observation sparsity o_t.

In the first dataset (Experiment 1), web page visits along with the underlying buying stages are generated on a daily basis $(o_t \sim \delta(t))$, i.e. browsing is observed daily. In the second and third datasets (Experiments 2 and 3), the average monthly observation rates are 66% and 33%, respectively, which stands for ($\kappa = 20\ and\ 10$) average visits in a month.

Parameters of the simulator obtained and validated through the survey and interviews with B2B experts are listed in Tables 1 and 2, where $\varepsilon = 10^{-5}$. The Dirichlet parameters related to the probability of starting the process at each state are defined as $\lambda = [60, 22, 16, 2]$. The average duration spent in states 3 and 4 are set to $\gamma_3 = 60\ and\ \gamma_4 = 60$ days, respectively, and the average number of visits to each category is $n_{1:4} = 20$. Note that the randomness of the parameters $(\pi, A, B, N, d\ and\ \theta)$ allows the possibility of having heterogeneous preferences and behavior for different B2B simulated buyers. Finally, to make simulated data more similar to real-world data, we added Poisson noise to the generated browsing data.

Assuming no knowledge on the underlying buying process of each observation sequence and considering similar browsing behavior for all buyers visiting a seller's web page, the HMM is employed to estimate the underlying buying processes in each dataset. The Baum-Welch algorithm has been applied to estimate the initial, transition, and emission probabilities of the model. Both parameter learning and buying state estimations are performed on the same dataset. To ensure a better inference for parameters, we run the algorithm multiple times (200), each with different random initialization. In our computer simulation, the "mhsmm R-package" (O'Connell & Højsgaard, 2011) has been utilized for the HMM-based inference implementations. Buying stage estimation performance is evaluated by overall accuracy obtained from the confusion matrix of estimated state sequence and actual state sequence. The accuracy is defined by:

$$ACC = \frac{TTP_{all}}{Total\ number\ of\ test\ entries} \tag{11}$$

where TTP_{all} is the total true positive calculated by summing all the diagonal elements in the confusion matrix (Freitas, De Carvalho, Oliveira, Aires, & Sabourin, 2007).

Tables 3–5 show the confusion matrices for three experiments, in which the number of correct and incorrect predictions are summarized with count values and broken down by each buying stage. The achieved accuracy rates for these three experiments are 83%, 79%, and 71%, respectively, which shows the significance of modeling the buying browsing behavior with hidden Markov models. Besides that, the fewer browsing activities are observed, the lower the performance is achieved. Moreover, after learning the parameters of the model, the running time for buying process estimation is less than 5 s, which is very suitable for real case scenarios. Additionally, the estimated parameters can be utilized to estimate the future buying stages of potential buyers based on their previous online behaviors.

The transition probability of some stages is dependent on the duration of the previous stage, and also the duration distribution is mixed, i.e., the No-funnel and Early-funnel stages have geometric distributions whereas the Middle-funnel and Late-funnel stages have gamma distribution. Despite this, the HMM algorithm can estimate a stationary distribution for the transition probability, ultimately leading to a better estimation of the buying stages.

Table 1
Dirichlet parameters for transition probability.

Next Current	$\alpha_{\cdot,1}$	$\alpha_{\cdot,2}$	$\alpha_{\cdot,3}$	$\alpha_{\cdot,4}$
$\alpha_{1,\cdot}$	80	20	ε	ε
$\alpha_{2,\cdot}$	60	30	10	ε
$\alpha_{3,\cdot}$	40	ε	40	20
$\alpha_{4,\cdot}$	10	ε	ε	90

Table 2
Dirichlet parameters for emission probability.

State	β_1	β_2	β_3	β_4
Category				
f_1: Product	6	16	20	14
f_2: Detailed-Product	6	24	31	21
f_3: News	30	4	3	2
f_4: Recruitment	30	ε	ε	ε
f_5: References	6	24	23	30
f_6: Support	6	7	5	9
f_7: Investors	6	19	15	20
f_8: Contacts	6	6	3	4

Table 3
Confusion matrix of the experiment 1.

Estimated	1	2	3	4
Actual				
1	21,784	837	53	6
2	832	40,839	3809	107
3	27	553	27,427	11,782
4	4	1	734	5172

6. Real-world dataset experiments

In this section, we apply the proposed method on the anonymized browsing dataset provided by a B2B seller company. In the scope of this research, the authors have collaborated with the Account-Based Marketing platform, namely N.Rich. N.Rich has confirmed that the data has been collected considering the framework outlined in the EU General Data Protection Regulation (GDPR). Two different datasets are provided for this experiment; the first one is the "browsing dataset" including four rows: Company-id, Page-URL, visit-Date, and Visit-duration. The second dataset is the "opportunity dataset", which provides the starting and closing dates of different sales opportunities of different buyers.

In the browsing dataset, there are 648 that companies visited this seller company website during 2018.09.28 and 2019.10.08. The minimum, maximum, and average days of visits are 1, 246, and 15, respectively. The average duration of each visit is 6 min. There are around 3800 unique web pages on this seller's website, which are semi-automatically categorized into one of the eight previously described categories. Note that, these data statistics are computed on the processed data, in which visits with duration less than 1 min have been eliminated, and also stays with duration larger than 20 min are capped to 20 min.

In the Opportunity dataset, there are 268 companies defined as potential buyers during the time period of 2018.01.02 and 2019.05.31.

Table 4
Confusion matrix of the experiment 2.

Estimated	1	2	3	4
Actual				
1	13,919	691	63	7
2	651	27,478	14,512	149
3	19	433	27,427	9791
4	3	0	174	3513

Table 5
Confusion matrix of the experiment 3.

Estimated	1	2	3	4
Actual				
1	7083	415	11	53
2	471	13,790	212	1552
3	23	495	6531	6283
4	8	5	1761	187

The minimum, maximum, and average numbers of opportunities per company are 1.0, 1256, and 48, respectively. As shown in Table 6, there are 15 different opportunity types defined in this dataset mapped to previously described, four buying states with the help of a sales specialist. We have employed the part of the dataset which matches the browsing dataset, i.e., the time period of 2019.01.20 and 2019.10.08.

The described Baum-Welch algorithm is applied to the defined HMM model. The observations of the model are derived from the browsing dataset, and the parameters of the model are learned from them. After that, the state of the buying process each day is estimated through the forward-backward (smoothing) algorithm. To validate the estimated states, true state values extracted from the opportunity dataset are utilized. In other words, the buying stage value in each day is calculated using the browsing behavior on that specific day and then evaluated by comparing it to the stage value extracted from the opportunity data. Note that we have considered an interval of 2 weeks before and after each opportunity date as a valid interval for each buying stage.

As described in the previous section, the accuracy of results is computed according to Eq. (11). Table 7 illustrates the resulted confusion matrix, and the corresponding accuracy is 62%.

As can be seen, the first row is all zero because of the fact that in dates that stage 1 (no funnel) is estimated by the proposed algorithm, there is no ground truth (opportunity data) available to show the actual state value. Also, the opportunity data does not have any information related to no funnel stage. Moreover, it is observed that the confusion matrix seems non-diagonal, which can be interpreted due to the fact that the utilized data (the only available dataset) includes parallel simultaneous buying processes. This issue is illustrated in Fig. 5, in which the performance of the proposed algorithm on one of the buyer companies' browsing behavior is discussed.

Circles in this figure are indicators of the estimated stats (extracted by observing browsing dataset) and squares are indicators of true state (extracted from opportunity data). The X-axis indicates the dates in which browsing behavior is provided, and opportunity dates are displayed as text on top of their indicators. Coloring of both estimated and true states are coded as {black: No funnel, red: Early funnel, green: Mid funnel and blue: Late funnel.

Fig. 5 shows an example of buying state estimation of a potential buyer in days that browsing state is available. As can be seen, this company has a browsing activity on the 23rd of January, which the proposed method identifies as Late funnel activity. On the same date, there is a Late funnel opportunity in the dataset. Moreover, in August 2019, there are three different opportunities created, showing there is more than one buying process going on simultaneously, out of which our algorithm identified the Early funnel activity. Note that, having simultaneous buying process from different buying stages going on at the same day may degrade the accumulated observations, hence can result

Table 6
Opportunity types defined in the "Opportunity dataset" and their corresponding defined buying stages.

Index	Opportunity-name	Corresponding-state
1	Canceled	Early funnel
2	On Hold	Early funnel
3	Customer No Go	Early funnel
4	Owner No Go	Early funnel
5	Discover Sales	Early funnel
6	Submitted for Review	Early funnel
7	Qualify	Early funnel
8	Plan	Early funnel
9	Negotiate	Middle Funnel
10	Customer Canceled	Middle Funnel
11	Customer Lost	Middle Funnel
12	Tailor	Middle Funnel
13	Closed Won	Late Funnel
14	Closed Lost	Late Funnel
15	Finalize	Late Funnel

Table 7
Confusion matrix of the predicted states using real data browsing behavior.

Actual	1	2	3	4
Predicted				
1	0	0	0	0
2	53	138	16	134
3	66	137	34	122
4	223	277	68	1660

in erroneous estimated buying state at that day. Another important observation on this data is that in each month, at most, four days include some browsing behavior, which further demonstrates the sparsity of the browsing data that may not a proper data for a machine learning method. Another fact to consider is that the provided ground truth may not be precise, and the seller may have missed existing opportunities on some dates. Altogether, although the proposed method is designed to estimate the buying state, where there is only one buying process at a time, the estimation performance on this sparse and highly complex dataset, according to experts, is of sufficient merit.

7. Discussion

The results reported in both Simulation Study and Real-data Experiments sections have experimentally confirmed the effectiveness of our proposed method in revealing the underlying buying state in the B2B buying journey by solely observing their browsing behavior. Simulation results show that in the case of having consecutive new buying processes, the more browsing behavior provided to the algorithm, the better the estimation performance becomes.

Although the proposed method was designed for consecutive B2B buying processes, the only available dataset for real-data experiments was the dataset containing multiple parallel buying processes (a mixture of new, modified, and re-buy). Using this data resulted in having accumulated observations of multiple buying processes, each being in different stages of purchasing that may lead the observations to be corrupted. In other words, the browsing data is available only at an aggregated level, where it is not possible to separate browsing associated with different products or decision-makers. Also, many of the companies in the browsing dataset have highly sparse visits each month to the seller's website, which decreases the algorithm performance, as illustrated in simulation experiments. Moreover, the provided ground truth may not be precise, meaning sales staff may miss some opportunities, or there may have been errors in submitting the opportunity dates. Despite these limitations, the achieved results illustrate the capability of the

proposed method in estimating the buying states somewhat with acceptable accuracy. The results prove that more detailed browsing data is needed to infer the stages of B2B buying better.

To improve the performance of this method, in addition to browsing behavior, other information such as previous contacts with the seller, previous purchases, or interests in seller's products, similar searches performed by the company, could be added to the HMM model. Furthermore, qualitative validation of the results can be pursued to assess the performance of the results, which means running the program on an ongoing project and validating the estimated state by contacting the buyers. In the proposed method, the accumulated browsing behavior of all people inside the buyer company is utilized to reveal the underlying buying stages, which makes it useful mostly for consecutive (or few parallel) buying processes. Whereas a more robust method, as our next target, would be to have the product and individual-based browsing data to be able to differentiate different buying processes as well as identify the person's role in the seller company. We will model the actions of each individual in the company separately, along with their interactions. The buying process is considered as an emerging behavior of all people taking part in that specific process.

8. Conclusion

As digital information technologies have shaped the B2B buying journey centrally buyer-driven (Grewal et al., 2015; Schwartz & Kim, 2012), understanding the buying stage of potential B2B buyers allows the B2B sellers to properly adjust their marketing and sales, which has been found to increase the efficiency and hit rate of the activities (Kiang & Chi, 2001). Our proposed method provides valuable tools for practicing seller-side marketing and helps B2B sellers with the means to gain more control over the buying process. Note that, to the best of our knowledge, this is the first attempt to model the B2B buying process statistically.

We have proposed hidden Markov statistical models to uncover the underlying stages of the B2B buying process by observing the online browsing behavior of potential B2B buyers. The B2B buying stages can be (from B2B sellers' point of view) either No-funnel, Early-funnel, Middle-funnel, or Late-funnel (Toman et al., 2017). Results reveal the effectiveness of modeling different stages of buying and their connection to browsing behavior using HMM. These results also partially confirm our hypothesis, which was the stages of the buying process can be estimated from the browsing behavior.

Two different sets of experiments have been performed: simulation study and real-data experiments. Due to the lack of an exact dataset of a consecutive complex buying process, we used the simulation studies to
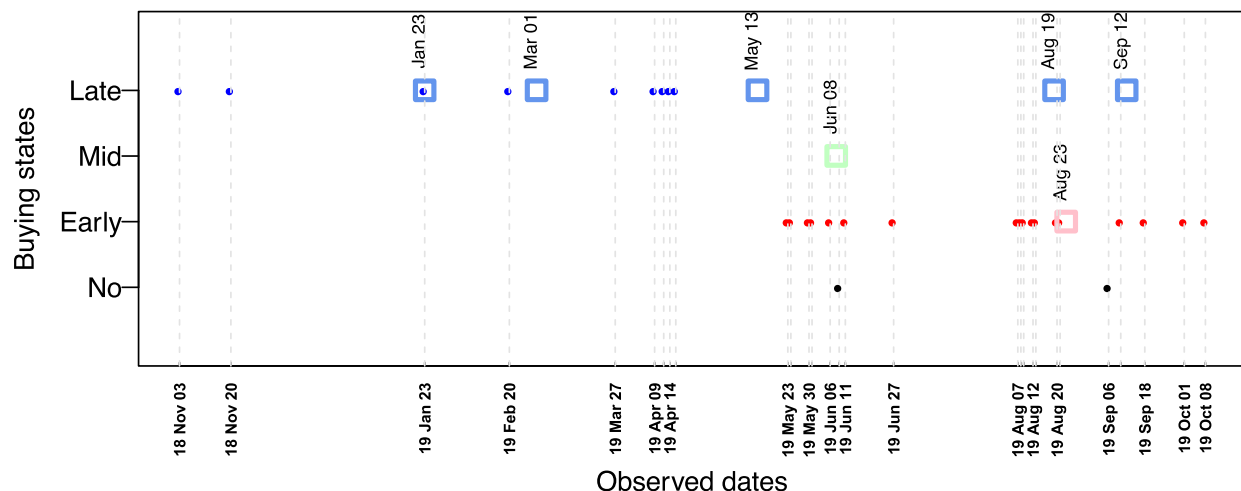

Fig. 5. Buying state estimation by only observing the online browsing behavior.

assess the performance of the proposed modeling approach. The simulator is designed with parameters that match with those of the actual B2B buying process to replicate the behavior of a buyer company visiting a seller web page purchasing complex B2B products. In this model, although the buying behavior of all companies visiting a seller company's website is considered to be similar, we allowed the possibility of having heterogeneous preferences and behavior for different B2B buyers. The real-world dataset experiments have been performed on a noisy dataset containing accumulated browsing behavior related to different parallel buying processes, each connected to various product lines. Although the proposed modeling framework is not suitable for this dataset, the achieved results illustrate the effectiveness of it to reveal the underlying stages to acceptable performance. The results prove that it is possible to infer the stages of the B2B buying process from browsing behavior, but more detailed browsing data is required to improve the results.

Although this model is yet restricted in several ways, it is already an improvement, as reflected by the observed performance. Our plan for future work is 1) to add more detailed observations from user behavior to both model and dataset, 2) increase the accuracy of the system by modeling the individuals inside the buying center who are involved in the buying process, and 3) model parallel buying processes.

Financial disclosure

Acknowledgement

References

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics, 41*(1), 164–171.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Brennan, R., & Croft, R. (2012). The use of social media in B2B marketing and branding: An exploratory study. *Journal of Customer Behaviour, 11*(2), 101–115.

Chavan, G. D., Chaudhuri, R., & Johnston, W. J. (2019). Industrial-buying research 1965-2015: Review and analysis. *The Journal of Business and Industrial Marketing, 34*(1), 205–229.

Cyert, R. M., & March, J. G. (1963). *A behavioral theory of the firm.*

D'Agostino, D., Gasparetti, F., Micarelli, A., & Sansonetti, G. (2016, July). A social context-aware recommender of itineraries between relevant points of interest. In *International conference on Human-Computer Interaction* (pp. 354–359). Cham: Springer.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., … Williams, M. D. (2019). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57*, 101994.

Freitas, C. O., De Carvalho, J. M., Oliveira, J., Aires, S. B., & Sabourin, R. (2007, November). Confusion matrix disagreement for multiple classifiers. In *Iberoamerican congress on pattern recognition* (pp. 387–396). Berlin, Heidelberg: Springer.

Grewal, R., Lilien, G. L., Bharadwaj, S., Jindal, P., Kayande, U., Lusch, R. F., … Spekman, R. (2015). Business-to-business buying: Challenges and opportunities. *Customer Needs and Solutions, 2*(3), 193–208.

Hutt, M. D., & Speh, T. W. (2007). *Business marketing management: B2B.* Cengage Learning.

Johnston, W. J., & Lewin, J. E. (1996). Organizational buying behavior: Toward an integrative framework. *Journal of Business Research, 35*(1), 1–15.

Kiang, M. Y., & Chi, R. T. (2001). A framework for analyzing the potential benefits of internet commerce. *Journal of Electronic Commerce Research, 2*(4), 157–163.

Kumar, V., Sriram, S., Luo, A., & Chintagunta, P. K. (2011). Assessing the effect of marketing investments in a business marketing context. *Marketing Science, 30*(5), 924–940.

Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management, 34*(3), 387–394.

Lawrance, J., Bellamy, R., & Burnett, M. (2007, September). Scents in programs: Does information foraging theory apply to program maintenance?. In *IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC 2007)* (pp. 15–22). IEEE.

Lilien, G. L. (2016). The B2B knowledge gap. *International Journal of Research in Marketing, 33*(3), 543–556.

Luo, A., & Kumar, V. (2013). Recovering hidden buyer–seller relationship states to measure the return on marketing investment in business-to-business markets. *Journal of Marketing Research, 50*(1), 143–160.

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management, 58*(3), 103434.

O'Connell, J., & Højsgaard, S. (2011). Hidden semi markov models for multiple observation sequences: The mhsmm package for R. *Journal of Statistical Software, 39*(4), 1–22.

Paschen, U., Pitt, C., & Kietzmann, J. (2020). Artificial intelligence: Building blocks and an innovation typology. *Business Horizons, 63*(2), 147–155.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review, 106*(4), 643.

Saaranto, A. (2019). *A B2B Buyer's purchasing behavior: An agent-based modeling approach.* Master's thesis. Espoo, Finland: Aalto University.

Schwartz, J., & Kim, D. (2012). *ITSMA how buyers consume information survey. (December) www. itsma. com* (assessed November 2014).

Siriaraya, P., Yamaguchi, Y., Morishita, M., Inagaki, Y., Nakamoto, R., Zhang, J., … Nakajima, S. (2017, December). Using categorized web browsing history to estimate the "user" latent interests for web advertisement recommendation. In *2017 IEEE International Conference on big data (big data)* (pp. 4429–4434). IEEE.

Spool, J. M., Perfetti, C., & Brittan, D. (2004). *Designing for the scent of information: The essentials every designer needs to know about how users navigate through large web sites* (User Interface Engineering).

Steward, M. D., Narus, J. A., & Roehm, M. L. (2018). An exploratory study of business-to-business online customer reviews: External online professional communities and internal vendor scorecards. *Journal of the Academy of Marketing Science, 46*(2), 173–189.

Steward, M. D., Narus, J. A., Roehm, M. L., & Ritz, W. (2019). From transactions to journeys and beyond: The evolution of B2B buying process modeling. *Industrial Marketing Management, 83*, 288–300.

Toman, N., Adamson, B., & Gomez, C. (2017). The new sales imperative. *Harvard Business Review, 95*(2), 118–125.

Woodside, A. G. (1996). Theory of rejecting superior, new technologies. *The Journal of Business and Industrial Marketing, 11*(3–1), 25–43.

Wouters, J. P. (2004). Customer service strategy options: A multiple case study in a B2B setting. *Industrial Marketing Management, 33*(7), 583–592.

Zhang, J. Z., Netzer, O., & Ansari, A. (2014). Dynamic targeted pricing in B2B relationships. *Marketing Science, 33*(3), 317–337.

Zhang, J. Z., Watson Iv, G. F., Palmatier, R. W., & Dant, R. P. (2016). Dynamic relationship marketing. *Journal of Marketing, 80*(5), 53–75.