Marketing Science
Vol. 29, No. 2, March–April 2010, pp. 348–365
issn 0732-2399 | eissn 1526-548X | 10 | 2902 | 0348

doi 10.1287/mksc.1090.0520
© 2010 INFORMS

A Viral Branching Model for Predicting the Spread of Electronic Word of Mouth

Ralf van der Lans, Gerrit van Bruggen
Rotterdam School of Management, Erasmus University, 3000 DR Rotterdam, The Netherlands
{rlans@rsm.nl, gbruggen@rsm.nl}

Jehoshua Eliashberg
The Wharton School of the University of Pennsylvania, Philadelphia, Pennsylvania 19104,
eliashberg@wharton.upenn.edu

Berend Wierenga
Rotterdam School of Management, Erasmus University, 3000 DR Rotterdam, The Netherlands,
bwierenga@rsm.nl

In a viral marketing campaign, an organization develops a marketing message and encourages customers to forward this message to their contacts. Despite its increasing popularity, there are no models yet that help marketers to predict how many customers a viral marketing campaign will reach and how marketers can influence this process through marketing activities. This paper develops such a model using the theory of branching processes. The proposed viral branching model allows customers to participate in a viral marketing campaign by (1) opening a seeding e-mail from the organization, (2) opening a viral e-mail from a friend, and (3) responding to other marketing activities such as banners and offline advertising. The model parameters are estimated using individual-level data that become available in large quantities in the early stages of viral marketing campaigns. The viral branching model is applied to an actual viral marketing campaign in which over 200,000 customers participated during a six-week period. The results show that the model quickly predicts the actual reach of the campaign. In addition, the model proves to be a valuable tool to evaluate alternative what-if scenarios.

Key words: branching processes; forecasting; Markov processes; online marketing; viral marketing; word of mouth
History: Received: March 4, 2008; accepted: May 27, 2009; processed by Carl Mela. Published online in *Articles in Advance* August 26, 2009.

1. Introduction

In October 2006, Unilever launched a 75-second viral video film, *Dove Evolution*. This campaign generated over 2.3 million views in its first 10 days and three times more traffic to its website than the 30-second commercial that aired during the Super Bowl (van Wyck 2007). More recently, Comic Relief, a British charity organization, achieved 1.16 million participants in the first week after launching their viral game "Let It Flow" that promoted Red Nose Day, their main money-raising event (*New Media Age* 2007). These two examples illustrate a new way of marketing communication in which organizations encourage customers to send e-mails to friends containing a marketing message or a link to a commercial website. Because information spreads rapidly on the Internet, viral marketing campaigns have the potential to reach large numbers of customers in a short period of time. Not surprisingly, many companies such as Microsoft, Philips, Sony, Ford, BMW, and Procter & Gamble have gone viral. However, not all viral marketing campaigns are successful, and because of competitive clutter, they need to become increasingly sophisticated in order to be effective and successful. It is also important that marketers be able to predict the returns on their expenditures and thus how many customers they will reach. As one marketing agency executive stated: "The move to bring a measure of predictability to the still-unpredictable world of viral marketing is being driven by clients trying to balance the risks inherent in a new marketing medium with the need to prove return on investment" (Morrissey 2007, p. 12). Despite their importance, no forecasting tools for these purposes are available yet. The aim of this research is to develop a model that predicts how many customers a viral marketing campaign reaches, how this reach evolves, and how it depends on marketing activities.

The structure of this paper is as follows. Section 2 defines viral marketing campaigns and describes how marketers can influence the viral process. Section 3 shows how the flow of communication among customers in viral marketing campaigns follows a branching process, and we introduce our viral branching model (VBM). Section 4 describes the data of a real-life viral marketing campaign that reached over 200,000 customers after only six weeks. The predictive performance of our model, analyzed using data from this campaign, is presented in §5. The final section discusses implications of our research and suggestions for further research.

2. Viral Marketing Campaigns

In a viral marketing campaign, an organization develops an online marketing message and stimulates customers to forward this message to members of their social network. These contacts are subsequently motivated to forward the message to their contacts, and so on. Because messages from friends are likely to have more impact than advertising and because information spreads rapidly over the Internet, viral marketing is a powerful marketing communication tool that may reach many customers in a short period of time (De Bruyn and Lilien 2008). Furthermore, the nature of the Internet allows marketers to use many different forms of communication such as videos, games, and interactive websites in their viral campaigns. The term *viral* marketing may (incorrectly) suggest that information spreads automatically (Watts and Peretti 2007). However, marketers need to actively manage the viral process to facilitate the spread of information (Kalyanam et al. 2007).

2.1. Marketing Activities for Managing Viral Marketing Campaigns

In viral marketing campaigns, marketers may use two types of strategies to influence the spread of information. The first focuses on motivating customers to forward marketing messages to their contacts (Chiu et al. 2007, Godes et al. 2005, Phelps et al. 2004). As suggested by Godes et al. (2005) motivations to forward messages are either intrinsic or extrinsic. The former can be triggered by the content of the marketing message. Important components of the marketing message are the subject line of the e-mail and the text in the e-mail itself (Bonfrer and Drèze 2009). Furthermore, marketers nowadays develop websites containing videos and games that attract customer attention and interests. These websites usually facilitate the viral process by providing tools to easily forward e-mails to friends, such as “Tell a Friend” or “Share Video” buttons. Examples of extrinsic motivations to forward marketing messages are prizes and other monetary incentives (Biyalogorsky et al. 2001).

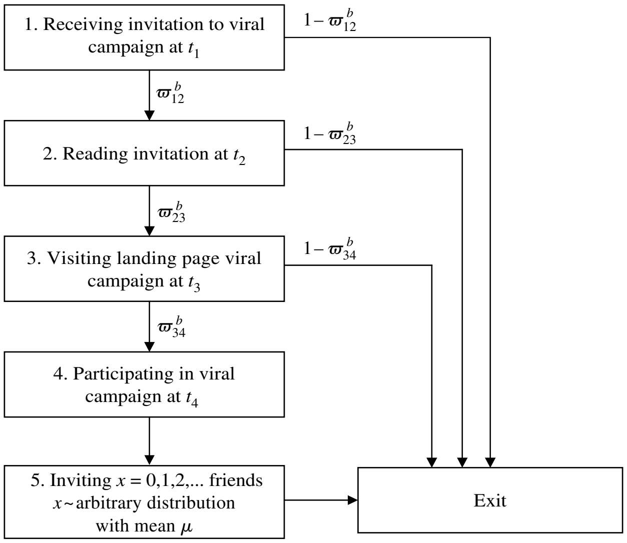
Although increasing customers’ motivation to forward messages to friends has a strong impact on the reach of the viral campaign, this is usually a difficult and expensive task. In contrast, controlling the number of initial or seeded customers is much more cost effective. In general, marketers can choose from three distinct categories to seed their viral marketing campaign: (1) seeding e-mails, (2) online advertising, and (3) offline advertising. Seeding e-mails are usually sent by the company itself or by a specialized marketing agency to customers who have given permission to receive promotional e-mails (Bonfrer and Drèze 2009). Using this seeding tool, a marketer can target a specific group of customers that are potentially interested in the campaign. The design and content of the e-mails are crucial because customers easily categorize such e-mails as spam and quickly delete them. For this reason, seeding e-mails are expected to be less effective than viral e-mails that are sent by friends or acquaintances of the recipient.

Online advertising is another important seeding tool that marketers can use to influence the viral process. The effectiveness of online advertising may differ depending on the customers as well as the websites on which the ads are placed. It is worth noting that marketers can directly observe when a specific online ad generates a visitor to the viral campaign. Hence, the effectiveness of online advertising can be monitored accurately, and based on its performance, marketers can decide to adapt their online advertising strategy. Furthermore, online advertising agencies offer contracts that guarantee a predetermined number of clicks to the campaign website within a certain time window. In such cases organizations usually pay for each click. Because online ads may be perceived as less obtrusive than promotional e-mails, this seeding tool may be very attractive.

Finally, besides online seeding tools, marketers may still use “traditional” offline advertising to seed their campaigns. Examples are magazine or TV ads that refer to the website of the viral marketing campaign, and package labels or coupons that try to attract visitors to the campaign website. However, offline seeding is less popular and expected to be less effective because customers cannot directly visit the campaign website by clicking a link. Another disadvantage of offline seeding is that it is more difficult to measure its effectiveness, because marketers cannot directly observe when offline advertising generates a customer to the viral campaign. Possible solutions for this problem are asking customers on the website how they were informed or to ask for the barcode of the product or coupon that was used to enter the website.

As described above, the appropriate strategic decision of the marketing activities at the right moment strongly depends on the spread of the process and the effectiveness of each marketing communication tool. Therefore, marketers need to closely monitor the spread of information in viral marketing campaigns.

Figure 1 **Decision Tree to Participate in Viral Marketing Campaign**


1. Receiving invitation to viral campaign at t_1
$1-\varpi_{12}^b$
ϖ_{12}^b
2. Reading invitation at t_2
$1-\varpi_{23}^b$
ϖ_{23}^b
3. Visiting landing page viral campaign at t_3
$1-\varpi_{34}^b$
ϖ_{34}^b
4. Participating in viral campaign at t_4
5. Inviting $x = 0,1,2,...$ friends $x \sim$ arbitrary distribution with mean μ
Exit

2.2. Monitoring Viral Marketing Campaigns

An important feature of viral marketing campaigns is that marketers are able to accurately measure the actions of customers, such as when they open an e-mail (Bonfrer and Drèze 2009) and which pages they visit (Moe 2003). Hence, marketers may obtain large databases containing detailed customer behavior. Monitoring such behavior is not straightforward, and it is therefore important to retain only those variables that are relevant to the viral process.

Figure 1 summarizes the five-stage process that a customer may go through during a viral marketing campaign. In the first stage, a customer receives an invitation at time t_1 from source b, i.e., through a viral e-mail from a friend or through one of the seeding tools of a company. At the end of this stage, the customer decides with probability ϖ_{12}^b to go to the second stage and read the invitation at time t_2, or with probability $1-\varpi_{12}^b$ to exit the campaign by deleting or ignoring the invitation. This probability ϖ_{12}^b is likely to depend on the source of invitation b, because customers are less likely to open and read a seeding e-mail from a company than a viral e-mail from a friend. After reading the invitation to the viral campaign, a customer decides to accept the invitation with probability ϖ_{23}^b by clicking a link to the landing page of the campaign website. After arriving on the landing page at time t_3 (stage 3), a customer decides to participate in the viral campaign (stage 4) with probability ϖ_{34}^b at time t_4. Participation may consist of watching a video, playing a game, and/or subscribing to a service. Finally, a customer decides to forward the message to x friends.

Figure 1 indicates that the number of customers receiving an e-mail is not necessarily the same as the number of customers who ultimately participate in the viral campaign, as this depends on the probabilities ϖ_{12}^b, ϖ_{23}^b, and ϖ_{34}^b. As described in §2.1, these probabilities depend on marketing activities such as the attractiveness of the subject line (ϖ_{12}^b), the content of the invitation (ϖ_{23}^b), and the design and content of the website (ϖ_{34}^b). Although the sequence of stages is quite generic for most viral marketing campaigns (De Bruyn and Lilien 2008), we recognize that it does not necessarily hold for all viral marketing campaigns. For instance, participation may consist of several stages (activities) such as watching a video, subscribing to a newsletter, and/or playing a game. In addition, it is possible that customers forward the message before participation, i.e., in cases where customers can only participate when they invite a certain number of friends. Therefore, marketers should adapt Figure 1 depending on the specific structure of their campaign. For the campaign of interest in our empirical application, Figure 1 accurately matches its structure. However, the agency executing our campaign did not store data for stages 2 and 3. Hence, for each participant we observed the transition from stages 1 to 4, which occurred with probability $\varpi_{12}^b\varpi_{23}^b\varpi_{34}^b$. Adapting our model (§3) to an alternative structure of a viral marketing campaign is straightforward.

To manage viral marketing campaigns, marketers need to monitor the stages represented in Figure 1 for each individual customer. Specifically, they should register the following variables: (1) the source of the invitation, (2) if and when a customer arrives at each stage, and (3) how many friends a customer invites. This leads to a dynamic database in which each row represents a customer and in which corresponding variables are updated when a customer switches to the next stage. New rows are added when new customers are invited. Such a database can be automatically generated in real time during the process of a viral marketing campaign.

In summary, viral marketing is an effective online marketing communication tool that may reach many customers in a short period of time. The reach of a viral marketing campaign is a function of seeding activities and the number of forwarded viral e-mails. Although the seeding activities are under the direct control of marketers, they can only influence the number of forwarded e-mails through incentives. To reach the campaign's goals, it is important for marketers to be able to forecast the reach of a viral marketing campaign as early as possible and to determine how this reach depends on marketing activities. Because tools for supporting these forecasts do not yet exist, we have developed such a forecasting model in the next section.

3. Modeling the Viral Marketing Process

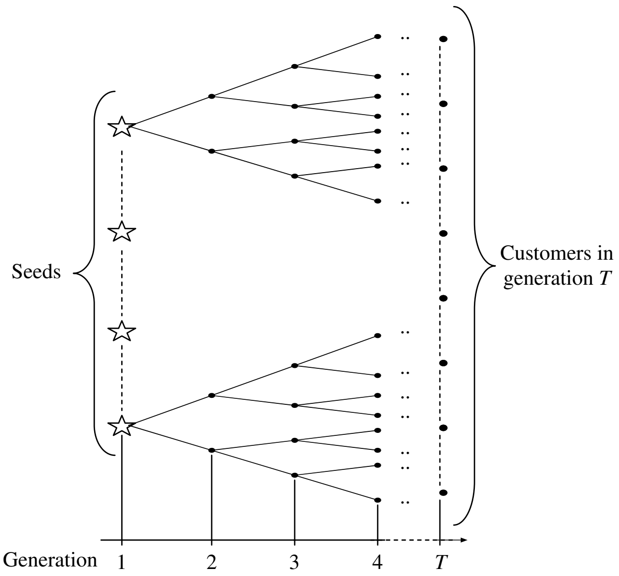
Insights from epidemics about the spread of viruses are useful to understand and model the spread of marketing messages in viral marketing campaigns. In epidemics, both aggregate- and disaggregate-level models have been developed to describe the spread of viruses (Bartlett 1960). Aggregate-level or diffusion models assume an underlying infection process, and the corresponding model parameters are inferred from the total number of infected individuals over time. Based on these insights, Bass (1969) developed his famous diffusion model and assumed adoption to depend on two forces: one that is independent of previous adoptions and one that depends positively on previous adoptions. As the number of customers in viral marketing campaigns (i.e., adopters) is also influenced by these two forces, the Bass model should be able to describe the spread of information during viral marketing campaigns. However, there are two important reasons why the Bass model does not optimally describe the viral marketing process. First, it assumes a specific process but does not include actual information on this process at the individual level. Such information becomes readily available in viral marketing campaigns and can be used to describe the process accurately at the aggregate level. Second, the Bass model assumes that every customer who has adopted the product increases the probability of others adopting in each time period after adoption. However, in viral marketing campaigns, customers only influence each other right after participation when they invite their friends.

Disaggregate-level or branching process models (Athreya and Ney 1972, Dorman et al. 2004, Harris 1963) may alleviate these two limitations as parameters are estimated based on individual-level information, and they assume that customers only influence each other right after participation by infecting a fixed number of others. Although branching process models have proven to be very useful in describing the spread of viruses theoretically, to our knowledge, they have not been applied to real empirical process data so far. The reason for this is that, similar to the diffusion of products, the process of the actual spread of viruses is typically not observed. Interestingly, in viral marketing campaigns, marketers can observe the actual spread of information across customers, and branching processes might therefore be a promising tool to describe and predict the reach of these campaigns. Furthermore, because standard branching models and their extensions are not capable of describing viral marketing campaigns, another contribution of our research is to extend the standard branching model. To do so, we now first explain the standard branching process.

3.1. Viral Marketing as a Branching Process

Branching processes, or Galton-Watson processes, were originally developed at the end of the 19th century to derive the probability of extinction of families (Athreya and Ney 1972, Dorman et al. 2004, Harris 1963). Generalizations of these processes, of which the birth-and-death process is an example, have been applied to model phenomena in physics, biology, and epidemiology to describe the spread of viruses in populations. Figure 2 graphically demonstrates the spread of information according to a standard branching process. The process represents T generations of customers that all invite $x = 2$ other customers. In the branching literature, x is crucial and has an arbitrary probability distribution with mean μ, which is called the infection or reproduction rate of the process. In Figure 2, the first generation (represented by stars) consists of an initial seed of n "infected" customers that forward the message to a second generation of customers that subsequently forward the message to a third generation, etc. Therefore, the total number of customers $V(t)$ in generation t equals nx^{t-1}, and the total reach of the campaign at generation T equals $n\sum_{t=1}^{T} x^{t-1}$. In situations where the infection rate is greater than one, it is sufficient for marketers to seed only a few initial customers to start the viral process, after which the whole population will ultimately be infected. However, unlike in an epidemic, the infection rate in viral marketing campaigns is generally smaller than one (Watts and Peretti 2007), which means that the spread of information dies out quickly as each customer generates on average less than one new customer. In such situations, marketers should influence the viral process by (1) increasing

Figure 2 Spread of a Message in a Viral Marketing Campaign as a Branching Process



the campaign's infection rate μ or (2) increasing the number of seeded customers n.

Although the standard branching model is useful to understand the underlying process in viral marketing campaigns, a more detailed model is needed to accurately describe and predict the actual spread of information. Therefore, we have extended this standard model as follows. First, whereas the standard branching model is a Markov process with fixed transition times, we allow customers to participate at any moment in time leading to a Markov process with stochastic transition times. Second, we incorporate two different types of marketing seeding activities; the first type allows seeding via sources Q such as banners and traditional advertising, and the second type allows seeding through e-mails. To incorporate this second type, we add the dimension $M(t)$ to the standard branching process, which represents the number of unopened seeding e-mails at time t. Third, whereas branching models typically count the number of "infected" customers $V(t)$ (i.e., customers who received e-mails and did not participate or delete the e-mail yet), we also count the cumulative number of customers who actually participate by introducing a third dimension $N(t)$ to the branching model. Fourth, standard branching processes assume parameters to be constant over time. However, it is likely that new invitations become less effective during the course of the campaign because invitations may be sent to customers who already received one or already participated in the campaign. Interestingly, invitations by seeding activities are less likely to be affected by this, because companies observe participants and invitations in real time during viral marketing campaigns. Hence, if a company carefully selects e-mail addresses, seeding e-mails should be sent to customers that did not receive an invitation yet. Furthermore, as discussed in §2.1, online marketing agencies frequently offer banner contracts generating a prespecified number of clicks. Also, these clicks are likely to come from new customers that did not participate yet. However, the probability that a participant invites a friend who already received an invitation or already participated increases as a function of the number of participants and sent invitations. In this research, we explicitly model this dynamic phenomenon by allowing μ to decrease as a function of $N(t)$ and already-invited customers. Next, we describe how the three processes $M(t)$, $V(t)$, and $N(t)$ interact in our viral branching model.

3.2. The Viral Branching Model

In this study, we decided, without loss of generality, to count those customers who participated in the viral campaign as the reach metric (stage 4 in Figure 1). Before introducing our model, we present its notations. Let

$t \in [0, \dots, T]$: denote continuous time, with $t = 0$ the start and $t = T$ the end of the campaign;

$N(t)$: denote the cumulative number of participants in the viral campaign at time t;

$V(t)$: denote the number of customers who received a viral e-mail from a friend and who did not participate or delete this e-mail yet;

$M(t)$: denote the number of customers who received a seeding e-mail from an organization and who did not participate or delete this e-mail yet;

$Z(t)$: denote the vector $\{M(t), V(t), N(t)\}$;

$q \in Q$: denote the set of seeding sources excluding seeding e-mails (i.e., banners, advertising);

b: denote the index over all invitation sources, i.e., $b \in \{\text{viral mail}, \text{seeding mail}, Q\}$;

μ^*: denote the average number of invited contacts, given participation;

θ: denote the average proportion of invited contacts that have already been invited or already participated in the campaign;

μ: denote the average number of invited contacts who have not been invited or participated in the campaign, given participation; hence, $\mu = \mu^* \cdot (1-\theta)$;[1]

π_b: denote the probability of participation upon receiving an invitation by source b (i.e., $\pi_b = \varpi_{12}^b \varpi_{23}^b \varpi_{34}^b$ [2]);

$1/\lambda_v$: denote the average time between receiving a viral e-mail and participating;

$1/\lambda_m$: denote the average time between receiving a seeding e-mail and participating; and

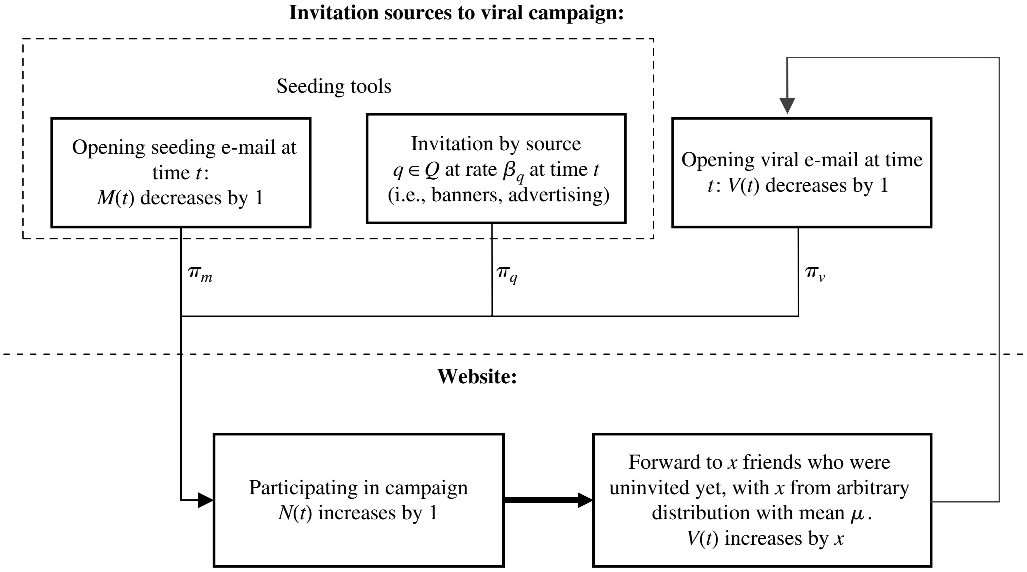
β_q: denote the rate by which customers are invited by seeding tool q.

Figure 3 summarizes our VBM and shows how $Z(t)$ changes over time. It shows how customers are invited to participate in the viral campaign by (1) receiving a seeding e-mail from a company, (2) another seeding source $q \in Q$ such as a banner or traditional advertising, or (3) receiving a viral e-mail from a friend. When a customer participates in the viral campaign at time t, the number of participants $N(t)$ increases by one and $M(t)$ or $V(t)$ decreases by one if this participant was invited by a seeding or viral e-mail, respectively. Furthermore, customers may invite y friends, of which w friends are have already invited to or have already participated in the viral campaign. Hence, the number of customers that has an invitation by viral e-mail increases by

[1] Without loss of generality, in the derivations of the viral branching model in §§3.2 and 3.3, we express the processes $Z(t)$ as a function of μ. In §3.4 we show how μ^* and θ are incorporated.

[2] To count the number of customers in another stage of Figure 1, it is sufficient to change the definition of π_b and μ. For instance, to count the number of participants in stage 2, π_b becomes equal to ϖ_{12}^b, and μ needs to be multiplied by $\varpi_{23}^b \varpi_{34}^b$.

Figure 3 Flow Diagram of a Viral Marketing Campaign


Invitation sources to viral campaign:
Seeding tools
Opening seeding e-mail at time t: $M(t)$ decreases by 1
Invitation by source $q \in Q$ at rate β_q at time t (i.e., banners, advertising)
Opening viral e-mail at time t: $V(t)$ decreases by 1
π_m
π_q
π_v
Website:
Participating in campaign $N(t)$ increases by 1
Forward to x friends who were uninvited yet, with x from arbitrary distribution with mean μ. $V(t)$ increases by x

Notes. Customers are invited to participate in the viral campaign by either (1) receiving a seeding e-mail from the company, (2) via another seeding source q such as a banner or advertising, or (3) receiving a viral e-mail from a friend. A customer participates with probability π_b, depending on the source b of the invitation. If the customer decides to participate in the viral campaign, $N(t)$ increases by one. After participation, the customer invites x friends who did not receive an invitation or participate yet, where x is generated from an arbitrary chosen distribution with mean μ. These x invited customers become members of $V(t)$; hence, $V(t)$ increases by x.

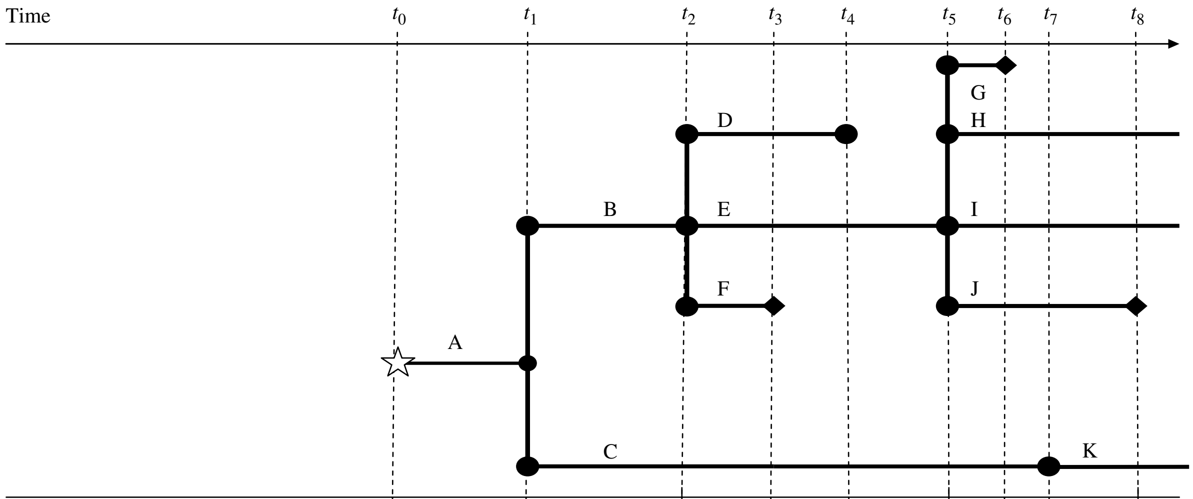
$x = y - w$. Because each participant may decide to invite a different number of friends, we assume that $y \in \{0, 1, 2, 3, \ldots\}$ comes from an arbitrary distribution with mean μ^*. Furthermore, we assume that $w \in \{0, 1, 2, \ldots, y\}$ is an arbitrarily distributed proportion θ of y. Hence, x comes from an arbitrary distribution with mean $\mu = \mu^*(1-\theta)$. As shown in Figure 3, every time t a customer decides to participate, the process variables $M(t)$, $V(t)$, and $N(t)$ change to new values. These process variables only depend on the parameters β_q, π_b, $\mu = \mu^*(1-\theta)$. Finally, to incorporate the speed at which people open viral and seeding e-mails, we assume the time between receiving an invitation and participation to be exponentially distributed with means $1/\lambda_v$ and $1/\lambda_m$ for viral and seeding e-mails, respectively. Although other distributions may fit better, the exponential distribution for the time between receiving an e-mail and participation is a reasonable approximation (Bonfrer and Drèze 2009). In addition, the exponential distribution is the only distribution that leads to mathematically tractable solutions (Dorman et al. 2004).

Based on the flow diagram in Figure 3, Figure 4 illustrates one possible realization of the stochastic process that is generated by our VBM. In Figure 3, we assume for simplicity that only a single customer is seeded by an e-mail from a company to customer A at time t_0. Therefore, $M(t_0)$, indicating the number of unopened seeding e-mails at t_0 sent by a company, equals one. After $t_1 - t_0$ time units, which are assumed to have an exponential distribution with mean $1/\lambda_m$, customer A opens the e-mail message and participates in the viral campaign, for example, by clicking a link directed to the campaign website. Consequently, $M(t_1) = M(t_0) - 1 = 0$, and $N(t_1)$, indicating the reach of the viral marketing campaign up to time t_1, equals $N(t_1) = N(t_0) + 1 = 1$. After participation, customer A sends two e-mails to friends B and C via the Invite a Friend button. For that reason $V(t_1)$, representing the number of customers with an unopened viral e-mail in their mailbox, equals $V(t_1) = V(t_0) + 2 = 2$. The time that customers B and C need to open this message is assumed to be independent and identically exponentially distributed with mean $1/\lambda_v$, which may be different from the time assumed for customer A. In the example in Figure 4, customer B opens the e-mail from friend A after $t_2 - t_1$ time units, and customer C takes $t_7 - t_1$ time units. Finally, in this example, at time t_8, we observe that $M(t_8) = 0$, $V(t_8) = 3$, and $N(t_8) = 5$. In the next subsection we derive the equations of our VBM for $M(t)$, $V(t)$, and $N(t)$.

3.3. Derivation of the Viral Branching Process Equations

Branching processes are an important class of Markov processes (Ross 1997). The memoryless property of the exponential distribution of the time between state transitions leads to a continuous-time Markov process. Hence, the vector $Z(t) = (M(t), V(t), N(t))'$ follows a three-dimensional continuous-time Markov process since $P(Z(t+t') = \mathbf{j} \mid Z(t') = \mathbf{i}, Z(r) = \mathbf{k}, 0 \leq r < t')$ equals $P(Z(t+t') = \mathbf{j} \mid Z(t') = \mathbf{i})$, where

Figure 4 Realization of the Stochastic Viral Branching Process When a Company Initially Seeds One Customer (t Is a Continuous Clock Time)



Number of customers with unopened seeding e-mail from company $M(t)$	1	0	0	0	0	0	0	0	0
Number of customers with unopened viral e-mail from a friend $V(t)$	0	2	4	3	2	5	4	4	3
Total cumulative number of participants $N(t)$	0	1	2	2	3	4	4	5	5

Notes. At t_0 customer A is invited to the viral marketing campaign, in this case through receiving a seeding e-mail sent by the company (☆), but this could also be because of a banner or advertising. Hence, $M(t_0)=1$. At t_1 customer A participates in the viral campaign (indicated by ●) ($N(t_1)=1$) after opening the e-mail ($M(t_1)=0$) and decides to forward the message to two friends, B and C ($V(t_1)=2$). At t_2, customer B participates in the campaign ($N(t_2)=2$) after opening the e-mail from friend A and forwards it to three new friends: D, E, and F ($V(t_2)=V(t_1)-1+3=4$). Subsequently, at t_3, customer F opens the e-mail and is not interested in the campaign (indicated by ◆), i.e., $V(t_3)=V(t_2)-1=3$, after which customer D opens the e-mail ($V(t_4)=V(t_3)-1=2$) and participates in the campaign ($N(t_4)=N(t_3)+1=3$) but does not forward the message to friends. At t_5, customer E opens the e-mail, starts participating in the campaign, and forwards the message to four friends: G, H, I, and J; i.e., $N(t_5)=4$ and $V(t_5)=V(t_4)-1+4=5$. At t_6, customer G opens the e-mail from friend E but is not interested in the campaign, ($V(t_6)=4$). Then at t_7, customer C opens the e-mail and participates in the campaign ($N(t_7)=5$) and forwards a message to friend K ($V(t_7)=V(t_7)-1+1=4$). Finally, at t_8, customer J opens the e-mail but does not participate; hence, $V(t_8)=3$ and $M(t_8)$ and $X(t_8)$ do not change.

$\mathbf{i}=(i_m,i_v,i_n)'$, $\mathbf{j}=(j_m,j_v,j_n)'$, and $\mathbf{k}=(k_m,k_v,k_n)'$ are nonnegative integers counting, respectively, the number of unopened seeding e-mails (indicated by subscript m), unopened viral e-mails (indicated by subscript v), and number of participants (indicated by subscript n) for different time periods: t', $t+t'$, and r, respectively. In the viral marketing process without a company's interfering, the variable $M(t)$ strictly decreases and switches to state i_m-1 every time a customer opens a seeding e-mail, given that $M(t)$ was in state i_m. An important tool for a marketer to increase the value of $M(t)$ with a value K is by sending K seeding e-mails to a list of customers. The transitions of $V(t)$ in the viral process are more complex, because these depend on the process $M(t)$ and may both decrease as well as increase over time. When a customer opens a viral e-mail, $V(t)$ may decrease by one if the customer does not forward the message to friends. However, $V(t)$ increases if (1) a customer opens a *seeding* e-mail and forwards it to one or more friends, (2) a customer opens a *viral* e-mail and forwards it to two or more friends, and (3) a customer participates via another source ($q\in Q$) in the campaign and forwards it to one or more friends. The third possibility—i.e., that customers randomly enter the viral marketing campaign from "outside"—is an important extension of traditional branching processes and is called immigration (Kendall 1949, Sevast'yanov 1957). We assume that the immigration rate equals $\pi_q\beta_q$ for source $q\in\{1,2,\ldots,Q\}$; hence, the average time between two customers that participate in the viral campaign because of immigration is exponentially distributed with rate $1/\sum_{q=1}^{Q}\pi_q\beta_q$. Finally, the variable $N(t)$, which depends on both processes $M(t)$ and $V(t)$, strictly increases and does so every time a customer participates in the viral campaign. This may be due to opening an e-mail from a friend or due to seeding by a company.

Differential equations play a crucial role in determining the values of the interrelated state variables $Z(t)$ over time in a continuous-time Markov process. Kolmogorov's backward and forward equations are convenient to derive the differential equations that the state transition probabilities should satisfy (Ross 1997). This research uses the forward equations to derive these differential equations, because these are more convenient to solve compared to the backward equations and also lead to unique solutions for all generalizations of branching processes (Harris 1963). Because the VBM is new to the literature, we derive and solve these differential equations in Technical Appendix A of the electronic companion to this paper, available as part of the online version that can be found at http://mktsci.pubs.informs.org. Next, we provide the solutions of the expectations of $M(t)$, $V(t)$, and $N(t)$.

3.3.1. The Conditional Expected Number of Unopened Seeding E-mails $M(t)$**.** As derived in Technical Appendix A of the electronic companion, the conditional expected number of unopened seeding e-mails at time t, given that at time t', with $0 \le t' \le t$, there are i_m unopened seeding e-mails, equals

$$E(M(t) \mid M(t') = i_m) = i_m e^{-\lambda_m (t-t')}. \qquad (1)$$

Clearly, as λ_m is always positive, $M(t)$ decreases exponentially over time and reaches zero as time passes. A marketer, however, may increase $M(t)$ by sending an additional set of seeding e-mails to a list of customers; i.e., marketers control the value i_m directly.

3.3.2. The Conditional Expected Number of Unopened Viral E-mails $V(t)$**.** The conditional expected number of unopened viral e-mails at time t, given i_v unopened viral e-mails at time t', equals (see Technical Appendix A of electronic companion):

$$\begin{aligned} E(V(t) \mid V(t') = i_v) &= i_v e^{\lambda_v(\pi_v\mu-1)(t-t')} \\ &\quad + K_1(e^{\lambda_v(\pi_v\mu-1)(t-t')} - e^{-\lambda_m(t-t')}) \\ &\quad + K_2(e^{\lambda_v(\pi_v\mu-1)(t-t')} - 1), \qquad (2) \end{aligned}$$

with $K_1 = \lambda_m \pi_m \mu i_m / (\lambda_v(\pi_v\mu - 1) + \lambda_m)$, and $K_2 = \sum_{q=1}^{Q} \pi_q \beta_q \mu / (\lambda_v(\pi_v\mu - 1))$. In (2), $\pi_v\mu$ represents the infection rate of the viral marketing campaign, which is smaller than μ because not every customer who receives an e-mail decides to participate. Note that if $\pi_v\mu > 1$, $V(t)$ grows exponentially and reaches infinity when t becomes very large.

3.3.3. The Conditional Expected Number of Participants in the Viral Campaign $N(t)$**.** Technical Appendix A of the electronic companion shows that the conditional expected number of participants $N(t)$, given i_n participants at time t', equals

$$\begin{aligned} E(N(t) \mid N(t') = i_n) &= i_n + K_3(e^{\lambda_v(\pi_v\mu-1)(t-t')} - 1) \\ &\quad + K_4(e^{-\lambda_m(t-t')} - 1) + K_5(t-t'), \qquad (3) \end{aligned}$$

with

$$\begin{aligned} K_3 &= (\pi_v/(\pi_v\mu - 1))(K_1 + K_2 + i_v), \\ K_4 &= i_m \pi_m (\lambda_v - \lambda_m)/(\lambda_m + \lambda_v(\pi_v\mu - 1)), \end{aligned}$$

and

$$K_5 = -\Big(\sum_{q=1}^{Q} \pi_q \beta_q / (\pi_v\mu - 1)\Big).$$

Equation (3) represents highly nonlinear effects of the model parameters on the reach of the campaign $N(t)$. Fortunately, the model parameters are estimated on the disaggregate level, and hence Equation (3) is not used in the estimation procedure. In fact, it is relatively straightforward to code this equation in a spreadsheet program, which calculates the expected reach of the campaign based on the individual-level parameter estimates μ, π_b, λ_m, λ_v, and β_q.

3.4. Estimating the Model Parameters

The strength of the VBM is that its parameters can be estimated using the individual-level data obtained from viral marketing campaigns as described in §2.2. Hence, in contrast to most models in marketing, we do not estimate the model parameters using the functional form as represented by Equations (1)–(3) and data on the actual process variables $Z(t)$. Instead, we use the dynamically generated database (see §2.2) containing the individual-level data of the process from which we infer the model parameters. The estimates based on these individual-level data are subsequently inserted into the model to predict the number of participants over time. This approach is similar to pretest market models (Hauser and Wisniewski 1982, Shocker and Hall 1986), including Sprinter (Urban 1970), Perceptor (Urban 1975), ASSESSOR (Silk and Urban 1978), Tracker (Blattberg and Golanty 1978), and MOVIEMOD (Eliashberg et al. 2000) that predict market shares or diffusion curves based on customers' trial and adoption processes. For these models, the process parameters are estimated before the start of the diffusion process using data from surveys and experiments. For our VBM, we estimate the parameter values directly from the individual-level data that become available from the viral process of interest and that are stored in a dynamic database. The model parameters can be quickly estimated reliably because this database contains many customers already in the campaign's early stages.

We now describe how the basic parameters of the VBM can be estimated for a given time period. To do so, we first discretize the time period $[0, \ldots, T]$ into $d = 1, \ldots, D$ time periods, with period $d = [t_{d-1}, \ldots, t_d]$, $t_0 = 0$, and $t_D = T$. Note that we still account for a continuous time viral branching process but allow the model parameters to vary across time periods d. Hence, we estimate μ_d, π_{bd}, β_{qd}, λ_{md}, and λ_{vd} for each time period d. In the empirical application, each time period d corresponds to one day that the viral campaign is online. For each period d, we observe $c = 1, \ldots, n_d$ customers that participate in the viral campaign.

3.4.1. Estimating the Average Number of Forwarded E-mails ($\mu = \mu^*(1-\theta)$). Each customer c in period d forwards y_{cd} e-mails to friends. We introduce variable u_{cdj}, which equals one if e-mail $j \in \{1, \ldots, y_{cd}\}$ forwarded by customer c in period d reaches a customer who already participated or already received an invitation, and zero otherwise. Hence, the effective number of forwarded e-mails equals $x_{cd} = y_{cd} - \sum_{j=1}^{y_{cd}} u_{cdj}$. These x_{cd} e-mails are automatically stored in the dynamically updated database by adding x_{cd} rows, i.e., rows $R_{c-1,\,d} + 1$ to $R_{c-1,\,d} + x_{cd}$ (see §2.2). $R_{c-1,\,d}$ represents the number of rows in the database up to customer $c-1$ in period d, which corresponds to the cumulative number of customers who already participated or were already invited up to customer $c-1$ in period $d-1$. Given variables y_{cd} and u_{cdj}, it is relatively easy to estimate both parameters, μ^* and θ_d, as follows:

$$\mu^* = \frac{1}{n_d} \sum_{d=1}^{D} \sum_{c=1}^{n_d} y_{cd}, \quad \text{and} \tag{4}$$

$$\theta_d = \frac{1}{n_d} \sum_{c=1}^{n_d} \frac{\sum_{j=1}^{y_{cd}} u_{cdj}}{y_{cd}}. \tag{5}$$

As described above, for prediction we expect the probability that an e-mail is ineffective, i.e., $P(u_{cdj} = 1)$, to increase as a function of $R_{n_{d-1},\,d-1}$. We use a binary logit specification to estimate this increase:

$$P(u_{cdj} = 1) = \frac{\exp(\alpha_1 + \alpha_2 R_{c-1,\,d})}{1 + \exp(\alpha_1 + \alpha_2 R_{c-1,\,d})}. \tag{6}$$

For prediction of $R_{n_{d'},\,d'}$ in period $d' > D$ after the observation period $[1, \ldots, D]$, we use the following equation:

$$R_{n_{d'},\,d'} = n_{d'}\mu_{d'} + \sum_{q=1}^{Q} \pi_{qd'}\beta_{qd'} \cdot (t_{d'} - t_{d'-1}) + K_{d'}, \tag{7}$$

where $n_{d'}\mu_{d'} = (N(t_{d'}) - N(t_{d'-1}))\mu_{d'}$ represents the expected number of forwarded e-mails in period d', $\sum_{q=1}^{Q} \pi_{qd'}\beta_{qd'} \cdot (t_{d'} - t_{d'-1})$ represents the expected number of customers who join the campaign because of seeding activities $q \in Q$, and $K_{d'}$ represents the number of seeding e-mails that a company sends in period d'. Given the predicted value of $R_{n_{d'},d'}$, we use (6) to predict $\theta_{d'+1}$ as $P(u_{n_{d'},\,d'j} = 1)$, which in combination with (4) leads to the predicted value of $\mu_{d'+1} = \mu^*(1 - \theta_{d'+1})$. We use this procedure iteratively to forecast the viral process for all future periods of interest.

3.4.2. Estimating the Probabilities (π_m, π_v) and the Distribution Parameters (λ_m, λ_v) of the Time to Participate. In general, we do not observe when an invited customer opens an e-mail and decides to delete it and hence to exit the campaign (see Figure 2). Therefore, we need to infer π_{md} and λ_{md}, and π_{vd} and λ_{vd}[3] simultaneously from the observed number of participants in the viral marketing campaign for each period d. Because the time between receiving a seeding e-mail and participation is assumed to be exponentially distributed, the probability that customers open an e-mail in period d, given they receive a seeding e-mail before this period, equals $\int_{t_{d-1}}^{t_d} \lambda_{md} e^{-\lambda_m t}\, dt = e^{-\lambda_{md} t_{d-1}} - e^{-\lambda_{md} t_d}$. Hence, the probability of participating in period d, after receiving a seeding e-mail, equals $\psi_d = \pi_{md}(e^{-\lambda_{md} t_{d-1}} - e^{-\lambda_{md} t_d})$. Given that K_d customers receive a seeding e-mail in period d, we observe in each time period $d, d+1, \ldots, D$, how many of these customers h_d participate, which has a multinomial distribution[4] $[h_d, h_{d+1}, \ldots, h_D] \sim MN(K_d; \psi_d, \psi_{d+1}, \ldots, \psi_D)$. Because of the many observations available after only short time periods, the parameters π_{md} and λ_{md} can be estimated using maximum likelihood. π_{vd} and λ_{vd} are estimated in a similar fashion.

3.4.3. Estimating the Immigration Rate $\pi_q\beta_q$ Because of Seeding Tool q. Parameters β_{qd} and π_{qd}, representing the number of customers who visit the campaign website because of seeding tool q in time period d, and π_{qd} representing the fraction of these customers who also start participating, are directly observed and stored in the dynamically updated database. For specific seeding tools such as banners, a marketer frequently has the opportunity to buy a specific amount of clicks on the banner to the website. In this case, β_{qd} does not need to be estimated and can be directly determined (i.e., set) by the marketing manager.

4. Empirical Study: A Real-Life Viral Campaign

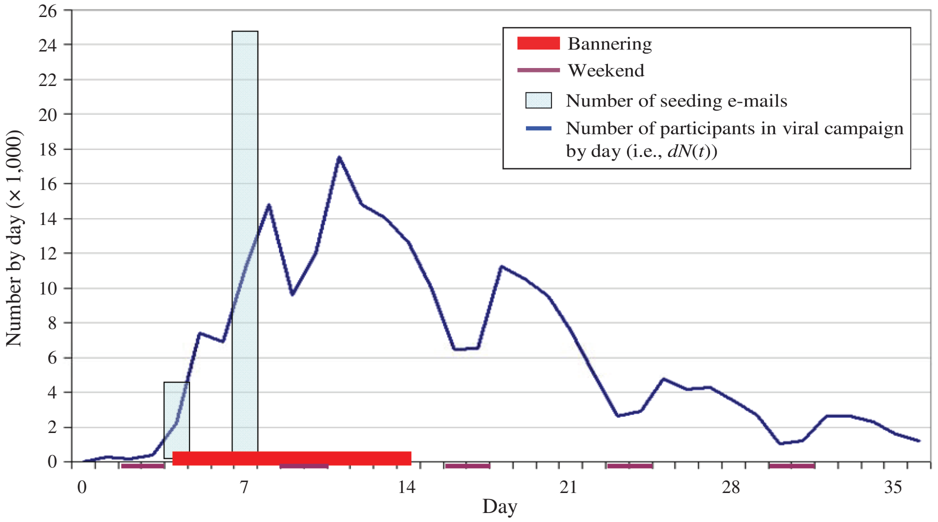
4.1. Description of the Campaign

From Friday, April 1, 2005 to Friday, May 6, 2005, a large financial services provider ran a viral marketing campaign. The goal of this campaign was to promote financial services to highly educated potential customers aged between 20 and 29. The structure of the campaign is as shown in Figure 1. Customers participated in the campaign while playing a game during which they answered questions, which then

[3] In the empirical application, we assume both λ_{md} and λ_{vd} to be equal across days during the week and across days during weekends. However, both λ_{md} and λ_{vd} are allowed to be different during weekends and weekdays.

[4] In the empirical application, we assume that the number of e-mails sent in period d is uniform over time; hence, the expected probability that a customer opens a seeding e-mail in period d, given that it was received at time τ in period d, equals $\int_{t_{d-1}}^{t_d} \int_0^{t_d-\tau} \lambda_{md} e^{-\lambda_{md} t}\, dt\, d\tau = 1 - (1/\lambda_{md})(1 - e^{-\lambda_{md}(t_d - t_{d-1})})$.

Figure 5 **Events and Number of Participants by Day During the Viral Campaign**



Notes. The viral campaign started on a Friday and was online for 36 days. On Day 4, the number of participants grew rapidly because of marketing activities. On this day, the company sent 4,500 seeding e-mails and placed banners on websites that generated 200 participants by day for 11 consecutive days. On Day 7, the company sent an additional set of 24,258 seeding e-mails to further promote the viral campaign.

led to a career profile. Then, in return for a guaranteed prize, participants could fill out an online form requesting personal information. After filling out this information, participants were informed that they could win bigger prizes if they invited one or more of their friends to the campaign by sending e-mails via the Send to a Friend button. Software connected to the campaign website checked in real time whether the e-mail addresses of these friends were valid (i.e., each e-mail address was filled out only once, e-mails were not sent to the participants themselves, and the viral e-mail did not bounce within a prespecified time period).

The viral campaign was online on April 1, but the organization started seeding on April 4. However, because of the novelty of the campaign, employees of the organization already started participating and inviting their contacts before the campaign was formally seeded. This resulted in 846 participants at the end of Day 3. To seed the campaign, the organization bought 6,400 banner clicks to the campaign website between April 4 and April 14 by placing a banner on a popular website. Of the 6,400 visitors, 2,200 people decided to participate in the viral campaign. Furthermore, on April 4 and April 7, the marketing agency sent 4,500 and 24,258 seeding e-mails, respectively, to customers who agreed to receive promotional e-mails. These marketing activities and the resulting viral process resulted in a total of 228,351 participants by Day 36 since the viral campaign was online. Figure 5 summarizes the marketing activities around the viral campaign and the resulting number of participants by day over time. This figure shows that the daily number of participants grew rapidly during the first 11 days, after which it slowly decreased over time. Note that during weekends the number of participants is lower, which is because during these days customers read their e-mail less frequently compared to weekdays, as is also shown in the following section.

4.2. Data Description

All 228,351 participants in the viral campaign registered on the campaign website by providing their e-mail addresses. Hence, we know the e-mail address of each participant and the time they participated in the viral campaign. Furthermore, we also obtained the e-mail addresses of over one million friends who were invited (some of which are also among the 228,351 because they actually participated) and the 28,758 seeding e-mail addresses that the marketing agency used to seed the campaign. Given these data, we coded, for each participant, how many viral e-mails were sent by counting the number of viral e-mails that were sent to new customers who had not participated yet or had not received an invitation at the moment the e-mails were sent.

Next to the number of e-mails a participant sent, we also coded how and when a participant was invited. Unfortunately, the marketing agency did not retain the source by which a participant was invited in their database. Therefore, we were only able to identify the source through which participants were invited by matching sent seeding and viral e-mail addresses with the registered e-mail addresses of participants. Using this procedure we were able to determine the source of invitation to the campaign website for 73% of the participants. Most of the remaining

27% of the customers registered under a different e-mail address through which they were invited, most likely because of privacy concerns. This percentage closely corresponds to findings of a recent survey that showed that 42% of Internet users have more than one e-mail account and that 33% of them provide e-mail addresses that would not identify them personally (*Wireless News* 2006). From this 27%, we know that between April 4 and April 14, 2,200 participated because of bannering. Hence, we randomly assigned 2,200 of these participants, equally distributed over the 11 days, to the banner as the source of invitation. We subsequently computed for each day the proportions of participants for which we knew whether they were invited by a viral or seeding e-mail. For example, on Sunday, April 10, 9,245 participants (98.5%) participated because of a viral e-mail and 145 participants (1.5%) participated after being invited by a seeding e-mail. On this day, after excluding 200 participants because of banners, there were 2,406 participants for which we did not observe the source of invitation. Hence, we randomly selected 98.5% of these 2,406 participants, and we assumed that they started participating because of a viral e-mail. For the remaining 1.5% of the participants, we assumed they were invited by a seeding e-mail. Sensitivity analyses showed that our results are not sensitive to different choices of proportions to allocate these customers to seeding e-mail or viral e-mail invitation sources.[5] We repeated this procedure for all days during the campaign so that all participants were assigned a source through which they were invited.

In summary, after these computations, our data set consists of 228,351 lines corresponding to participants. Each line contains the identity of the participant, the date of participation, the source of invitation, the date that the participant received the invitation, the number of e-mails that are sent to friends, and how many of these friends already participated or were already invited.

5. Results

5.1. Performance of the Viral Branching Model

Using the procedures as described in §§3.4.1–3.4.3, we were able to estimate the model parameters, which were subsequently plugged into Equations (1)–(3) to predict the number of participants by day. To capture the effect that customers read their e-mail less frequently during weekends, we estimated different distribution parameters of the time to participate for the weekdays and for the weekends. Using our parameter estimates, we assessed the VBM fit and its predictive performance. In addition to using all data during the 36 days that the campaign was online, we also estimated the parameters using only the first part of our data set and then developed forecasts for the remaining days of the 36-day period. Because we were interested in how early in the process we would be able to accurately predict the spread of the campaign, we estimated the parameters using the data obtained in four different time periods and then developed forecasts for the remaining days of the 36-day period (i.e., holdout periods). Because marketing activities only started on Day 4, we chose the first calibration period to be Days 1–7, just after the company seeded the campaign. This led to the following five scenarios:

1. Calibration period: Days 1–7
   Forecasting (holdout) period: Days 8–36
2. Calibration period: Days 1–14
   Forecasting (holdout) period: Days 15–36
3. Calibration period: Days 1–21
   Forecasting (holdout) period: Days 22–36
4. Calibration period: Days 1–28
   Forecasting (holdout) period: Days 29–36
5. Calibration period: Days 1–36.

Furthermore, we examined whether it is worthwhile to treat viral e-mails separately from seeding e-mails in our model. To test this, we also estimated a restricted version of our model by setting $\pi_m = \pi_v$ and $\lambda_m = \lambda_v$, which we call the nested VBM. Finally, we also compared the predictive accuracy of the nested and the nonnested VBM with the simplest form of the Bass (1969) model, and with an extended version of the Bass model that served as benchmarks. For the extended Bass model, we followed Kamakura and Balasubramanian (1988) and Parker (1992), and allowed the market potential $\bar{N}_d$[6] to be a function of marketing activities and the innovation parameter a_d to be different for weekdays and days of the weekend, leading to the following extended Bass model:

$$N(d)-N(d-1)=\left(a_d+b\frac{N(d-1)}{\bar{N}_d}\right)(\bar{N}_d-N(d-1)). \quad (8)$$

In (8), b represents the imitation parameter; $a_d = \gamma_{a0}+\gamma_{a1}\cdot \text{weekend}(d)$, where weekend$(d)$ represents a dummy that equals one if Day d is during the weekend, zero otherwise; and $\bar{N}_d = \gamma_{\bar{N}0} + \gamma_{\bar{N}1}\cdot \sum_{i=1}^{d} K_i + \gamma_{\bar{N}2}\cdot \sum_{i=1}^{d}\beta_i$, with K_i the number of seeding e-mails sent on Day i, and β_i the number of customers who start participating because of bannering on Day i. The parameters of the Bass model are estimated so that they optimally fit the process $N(t)$, whereas the VBM approach estimates parameters at the disaggregate

[5] In the sensitivity analyses, we varied the proportions to allocate customers to seeding e-mails from zero to twice as many customers as expected from the observed proportions.

[6] To avoid confusion with the parameters of the VBM, we slightly deviated from conventional notation of the Bass model.

Table 1 **Model Performance—Cumulative Number of Participants in a Time Period**

Estimation		In-sample fit		Out-of-sample forecast (MAPE) for days			
period	Model	RMSE[a]	MAPE	8–14	15–21	22–28	29–36
Days 1–7	VBM	1.79	0.07	0.09	0.03	0.07	0.14
	Nested VBM	4.02	0.23	0.39	0.60	0.37	0.25
	Standard Bass model	8.73	2.58	0.51	0.77	0.82	0.84
	Extended Bass model	0.48	0.24	0.08	0.19	0.33	0.39
Days 1–14	VBM	4.47	0.05	—	0.02	0.03	0.03
	Nested VBM	44.41	0.48	—	0.21	0.38	0.46
	Standard Bass model	15.85	2.66	—	0.09	0.25	0.32
	Extended Bass model	1.12	0.40	—	0.15	0.33	0.39
Days 1–21	VBM	6.06	0.06	—	—	0.01	0.02
	Nested VBM	83.60	0.58	—	—	0.06	0.14
	Standard Bass model	14.79	2.51	—	—	0.03	0.10
	Extended Bass model	2.35	0.43	—	—	0.02	0.02
Days 1–28	VBM	3.48	0.04	—	—	—	0.01
	Nested VBM	116.54	0.66	—	—	—	0.01
	Standard Bass model	12.85	2.07	—	—	—	0.04
	Extended Bass model	2.04	0.28	—	—	—	0.00
Days 1–36	VBM	6.98	0.05	—	—	—	—
	Nested VBM	119.70	0.61	—	—	—	—
	Standard Bass model	9.90	1.65	—	—	—	—
	Extended Bass model	1.83	0.22	—	—	—	—

[a] Root mean squared errors are multiplied by 1,000.

level and does not choose parameter values to optimize the fit of $N(t)$. The Bass model and its extended version, therefore, serve as a strong benchmark for our VBM. This is particularly true when we compare the in-sample fit over the calibration period.[7]

In Tables 1 and 2 we present the results of the five scenarios for the different models. Table 1 shows the in-sample fit statistics (root mean square error (RMSE) and mean absolute percentage error (MAPE)) and the forecasting accuracy (MAPE) for the *cumulative* number of participants (i.e., the reach $N(t)$) of the viral marketing campaign. Table 2 presents these statistics for the fit and prediction of the models for the *increase* (i.e., $dN(t)$) in the number of participants by day.

Overall, when analyzing the fit of the models, the results in Tables 1 and 2 (see also Figure 6) indicate that our VBM does very well in fitting the spread of the viral marketing campaign. The fit of the nested VBM, where the effectiveness of seeding e-mails is assumed to be equal to that of viral e-mails, is extremely low. This confirms the importance of incorporating different parameters for viral and seeding e-mails. Furthermore, although the standard Bass model does not seem to fit the process well, the extended Bass model fits the process $N(t)$ better than our viral branching model based on RMSE (1.83 versus 6.98 for the total estimation period). Interestingly, however, compared to the extended Bass model, the viral branching model fits the cumulative process better based on MAPE (0.05 versus 0.22) and the differenced process, $dN(t)$, based on both measures (RMSE: 1.23 versus 1.30; MAPE: 0.18 versus 0.31).

These results are due to the fact that the extended Bass model optimizes the RMSE of the cumulative number of participants, and suggests that the VBM better captures the actual process, which becomes apparent in the other fit statistics and the forecasting performance. As indicated by the results in Tables 1 and 2, and in contrast to all three competing models, the VBM was able to accurately predict the spread of the campaign already on Day 7, when the campaign was still not fully seeded. The nested version of the model is not able to predict the number of participants accurately in the early stages of the campaign and only starts doing better at the end of the campaign when the viral process has almost died out and does not attract many new customers. A similar phenomenon is true for the standard Bass model. Although the extended Bass model does slightly better, it is not able to predict the number of customers in the campaign after Day 7 or Day 14. As a matter of fact, after Day 14, the extended Bass model hugely underpredicts at 134,682, whereas the prediction of the VBM is at 221,429, which is very close to the true ultimate level of 228,351. The extended Bass model

[7] We tried several alternative specifications to incorporate marketing activities and weekend effects by incorporating these in functions for the innovation parameter a, imitation parameter b, and the market potential $\bar{N}$. We selected the best-performing model as the extended Bass model.

Table 2 Model Performance—Participants by Day

		In-sample fit		Out-of-sample forecast (MAPE) for days			
Estimation period	Model	RMSE[a]	MAPE	8–14	15–21	22–28	29–36
Days 1–7	VBM	1.12	0.11	0.15	0.50	0.61	0.93
	Nested VBM	1.91	0.25	0.79	0.88	0.83	0.99
	Standard Bass model	3.73	3.26	1.00	1.00	1.00	1.00
	Extended Bass model	0.84	0.32	0.30	0.59	0.93	0.99
Days 1–14	VBM	1.16	0.08	—	0.22	0.24	0.35
	Nested VBM	8.40	0.57	—	0.92	1.51	1.43
	Standard Bass model	3.18	2.80	—	0.75	0.98	1.00
	Extended Bass model	0.84	0.36	—	0.82	1.00	1.00
Days 1–21	VBM	0.96	0.07	—	—	0.15	0.31
	Nested VBM	7.65	0.68	—	—	0.80	1.35
	Standard Bass model	3.18	2.62	—	—	0.63	0.91
	Extended Bass model	1.62	0.46	—	—	0.18	0.29
Days 1–28	VBM	1.01	0.11	—	—	—	0.33
	Nested VBM	8.49	0.64	—	—	—	0.34
	Standard Bass model	2.85	2.23	—	—	—	0.75
	Extended Bass model	1.45	0.35	—	—	—	0.24
Days 1–36	VBM	1.23	0.18	—	—	—	—
	Nested VBM	6.57	0.62	—	—	—	—
	Standard Bass model	2.57	1.88	—	—	—	—
	Extended Bass model	1.30	0.31	—	—	—	—

[a] Root mean squared errors are multiplied by 1,000.

starts to predict the process relatively well only after Day 21, whereas the nested model and standard Bass model only start to predict well after Day 28. The fact that the extended Bass model is not able to predict the process at Day 7 or Day 14 confirms previous research findings that forecasts can only be made after the inflection point (Lenk and Rao 1990), which seems to occur after Day 14 (see Figure 6).

5.2. Parameter Estimates of the Viral Branching Model

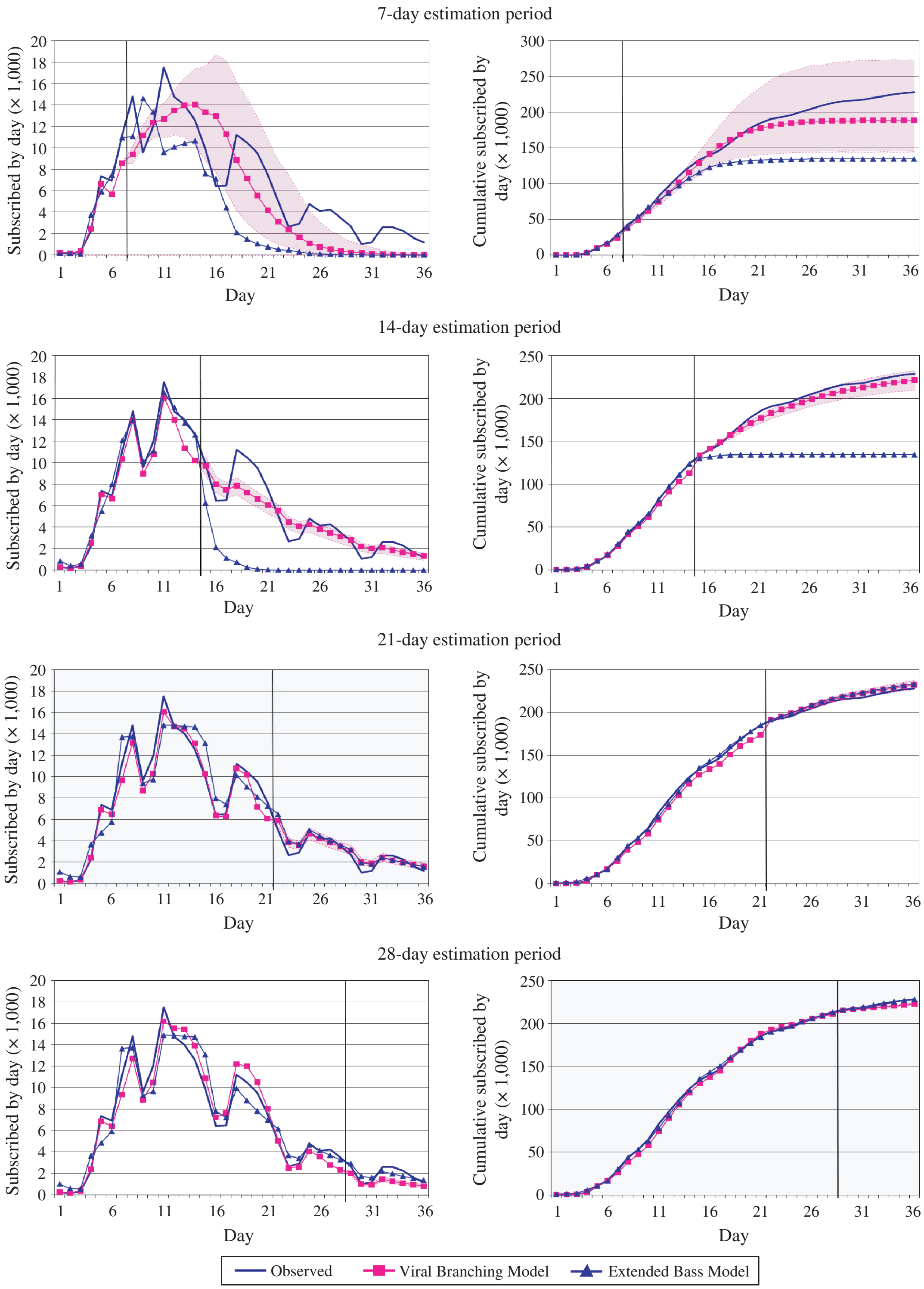
In addition to using the VBM for forecasting the spread of the viral marketing campaign, we also used its parameter estimates to gain insight into the spread of information in the viral campaign. Table 3 presents the parameter estimates for our VBM.[8] When we examine the parameter estimates, a number of observations can be made. First, on average, participants sent out over four ($\mu^* = 4.15$) viral e-mails to friends. Second, the probability that these friends start participating after receiving such an e-mail is, on average, 0.26.

Interestingly, this leads to an average infection rate of 1.08 (i.e., $\pi_v\mu^*$) at the start of the campaign, which shows that this particular viral campaign is extremely successful as the infection rate is larger than one. Hence, the number of participants grows exponentially. Note that as expected, the proportion of e-mails sent to customers who already received an invitation or already participated, θ, gradually increases over time as a function of the number of participants and people who already received an invitation, R. As explained in §3.4.1, Equation (6), this increase is captured by a binary logit regression. The results of this analysis confirms our expectations with $\alpha_1 = 2.99$ ($p < 0.01$), and $\alpha_2 = 7.24 \cdot 10^{-7}$ ($p < 0.01$). Consequently, at the end of the campaign the average infection rate is smaller than one and equals 0.87, which means that the number of additional participants does decrease over time as shown in Figure 5. This infection rate is still substantially larger than those reported by Watts and Peretti (2007), who find infection rates between 0.041 and 0.769. This emphasizes the success of the specific campaign we studied.

As expected, the probability of participation after receiving an e-mail from a friend ($\pi_v = 0.26$) is substantially higher than the probability of participation after receiving a seeding e-mail sent by a company ($\pi_m = 0.12$). The source of the e-mail strongly influences its effectiveness, which is also apparent in the forecasts of the nested VBM. Interestingly, the probability of participation after a banner click is relatively high (i.e. $\pi_q = 0.34$) and even higher than that of customers who received a viral e-mail from a friend. This is probably because of the fact that customers who click on a banner are already interested in the campaign. Still, 66% of these customers decide not to participate and quickly leave the campaign's landing page. The source of the e-mail also affects the average

[8] We did not estimate β_q for the banners because the company bought a fixed amount of 6,400 clicks.

Figure 6 Model Performance for Different Estimation Periods



Notes. Left (right) graphs reflect the (cumulative) number of participants by day for the four different calibration periods for the viral branching model (-■-), and the Bass model (-▲-). The actual values are indicated by the line (—). The shaded areas represent 95% prediction intervals of the viral branching model (see Technical Appendix B of the electronic companion for its derivation).

time between receiving an e-mail and participating in the viral campaign (1/λ.) This is more than two times shorter when the e-mail is received from a friend than from a company (1.64 days versus 3.88 days during weekdays). Note that we allowed for different estimates for λ_m and λ_v for e-mails sent during weekdays and those sent during the weekend. On weekends, people probably read their e-mails less often, leading

Table 3 Parameter Estimates

						$1/\lambda_m$		$1/\lambda_v$	
	μ^*	θ (%)	π_m	π_v	π_q	Week	Weekend	Week	Weekend
Days 1–7	4.59	4.06	0.06	0.25	0.34	0.69	—[a]	1.12	1.06
Days 1–14	4.29	6.47	0.10	0.26	0.34	1.75	2.77	1.33	1.51
Days 1–21	4.23	7.13	0.11	0.26	0.34	2.80	3.85	1.53	2.15
Days 1–28	4.19	7.38	0.12	0.26	0.34	3.31	4.39	1.59	2.80
Days 1–36	4.15	7.64	0.12	0.26	0.34	3.88	5.03	1.64	3.24

[a] The response time to the seeding e-mails at the weekend could not be estimated because there were no responses, because the first seeding e-mails were sent just after the first weekend the campaign was online.

to longer times to participate, which results in fewer participants at weekends as shown in Figure 3.

In the next section, we explore further implications of the parameter estimates of our VBM by examining the effects of two alternative what-if scenarios.

5.3. What-If Analyses

The VBM not only allows us to predict the spread of the viral marketing campaign over time, but it also enables us to forecast the spread if different marketing activities are pursued. This possibility to perform what-if analyses allows marketers to use the model to support decisions about modifying the campaign to reach their objectives. To illustrate this possibility, we explore the effects of two alternative marketing activities. Using the model parameters of the VBM based on the estimation period of 14 days, we predict how the spread of the viral marketing campaign is different if (1) an additional 10,000 seeding e-mails are sent on Day 15, and (2) an additional 10,000 clicks are bought through banners that are set online for one week from Day 15 to Day 22.

Table 4 summarizes the effects of these two alternative marketing campaigns. The additional 10,000 seeding e-mails results in an additional reach of 6,211 participants at the end of the campaign on Day 36. This means that on average 0.62 additional participants will be reached for every seeding e-mail. This is the number of people that directly participate by responding to the seeding e-mail and indirectly through receiving a viral e-mail with an invitation from a friend. It is remarkable that the effect of buying 10,000 additional banner clicks is substantially higher. This leads to an additional reach of 21,166 participants at the end of the campaign and means that the additional reach for every click is 2.17. Again, this is the sum of people who start participating directly after they have clicked the banner and the subsequently invited contacts through viral e-mails. Apparently, the bannering approach benefits from a self-selection mechanism. People who click on a banner may have an interest in the campaign and are then also more likely to participate and send viral e-mails to their friends. These effects are reflected in the model by the different probabilities of participating after receiving a seeding e-mail ($\pi_m = 0.10$ for Days 1–14; see Table 3), and after clicking on a banner ($\pi_q = 0.34$; see §5.2). Of course, the difference between the effectiveness of these approaches will also depend on the quality of the mailing database, the characteristics of the website where the banners are placed, and the costs of these seeding tools. Figure 7 graphically shows the difference in the spread of the campaign if the two alternative scenarios are executed. It is interesting to see that effects of the additional marketing expenditures on Day 15 or shortly after have not only an immediate effect but also have a more long-term effect. This is due to the indirect or viral effect following the direct effect of these marketing activities. Hogan et al. (2004) label this the "ripple" effect, and they find that ignoring this effect may underestimate the effectiveness of advertising campaigns. The same is true for viral marketing campaigns, and the ripple effect is likely to be even stronger for these types of campaigns because participants are actively encouraged to further spread the campaign among their friends. Once the rates of banner clicks and seeding e-mails are known, a company can determine which seeding method is most cost effective. Once the company can also put a dollar value on a customer that participates (e.g., customer lifetime value), it is possible to determine if it is profitable to carry out a particular additional seeding.

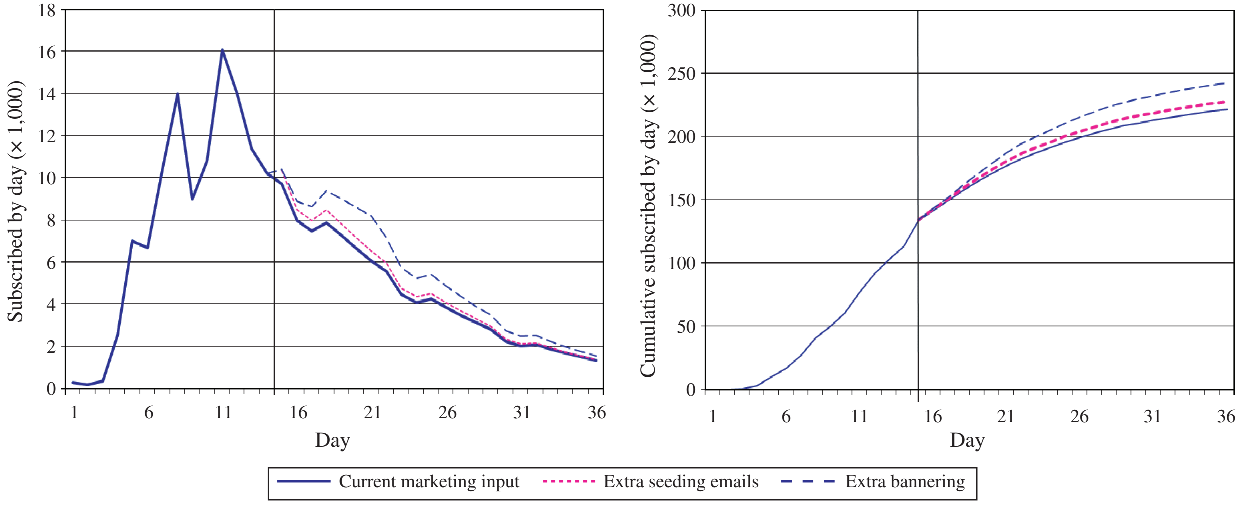
Table 4 Predicted Effects of What-If Scenarios

Marketing activity on Day 15	Predicted cumulative number of participants on Day 36	Predicted number of additional participants	Predicted number of additional participants per click/seed
Actual marketing strategy	221,429	—	—
Extra bannering for one week: 10,000 clicks	242,595	21,166	2.17 participants/click
Extra seeding: 10,000 e-mails	227,640	6,211	0.62 participants/seed

6. Discussion

Viral marketing is a relatively new way of approaching markets and communicating with customers and can potentially achieve a large reach and a fast spread among target audiences. Often these campaigns are relatively inexpensive because customer networks take care of spreading the messages and no expensive media exposure needs to be purchased. The dependency on these networks requires new modeling techniques to predict how a campaign will evolve over time and how many customers will receive the

Figure 7 Results of What-If Analyses



Notes. Left (right) panel reflects the predictions on Day 14 for the number (cumulative number) of participants by day for the current marketing activities (solid line) and for two different scenarios. In the first scenario, an additional set of 10,000 seeding e-mails is sent (dotted line); in the second scenario, an additional 10,000 clicks to the campaign website are generated via bannering (dashed line).

message and participate. Using insights from epidemiology to describe the spread of viruses as a branching process, we have derived and applied a new model to predict the reach of a viral marketing campaign. In addition to predicting the spread of information, our viral branching model also incorporates the effects of marketing activities such as seeding e-mails, bannering, and traditional advertising on this process, which standard branching models do not allow for. This enables marketers to accurately forecast the effects of their marketing activities and to analyze a variety of what-if scenarios. The application of our model on a real-life viral marketing campaign shows that it is able to accurately forecast the reach of a viral marketing campaign after only a few days when the campaign is online and the company has just started seeding the campaign.

Deriving the functional form of the viral branching model requires solving complex differential equations. This results in closed-form solutions for the expected reach of viral marketing campaigns. It is worth noting that this complex functional form of the reach is not needed to estimate the model parameters. Instead, they can be estimated relatively easily using the individual-level data that become available in large numbers early in the campaign. In fact, the functional form of the viral branching model can be implemented in a spreadsheet program such as Excel, and the values of the parameter estimates can be plugged into the model to derive the reach of the viral marketing campaign over time. This makes our viral branching model useful and implementable as a marketing decision-support system (Lilien and Rangaswamy 2004). In addition, the model parameters provide valuable insights for managers to improve their viral marketing campaigns because they are easily interpretable. For instance, it is insightful to monitor the switching probabilities as presented in Figure 1. A low probability means a bottleneck in the viral process, and marketers can then be advised to take appropriate measures to increase these probabilities. De Bruyn and Lilien (2008) show how these switching probabilities depend on characteristics of the sender and the receiver of the viral e-mail and their relationships. It would also be interesting to investigate how marketers could influence this process by changing, for example, the subject line of an e-mail, which in turn influences the probability of opening an e-mail (Bonfrer and Drèze 2009). The number of e-mails sent by a participant is another important parameter that positively influences the reach of the campaign. Marketers can influence this parameter by changing the incentives to forward viral e-mails. Finally, in our empirical example, customers seem to read their e-mails less frequently during weekends compared to weekdays. This implies that it is more effective to send seeding e-mails on a weekday. Next, to accurately forecasting and investigating alternative scenarios, managers can also use our model to compute the additional number of customers that a participant will generate in the viral marketing campaign. As shown by Hogan et al. (2004), the effectiveness of advertising is underestimated if word of mouth or the ripple effect is not taken into account. Our model incorporates this ripple effect directly.

In our research we only focused on the number of participants in a viral marketing campaign. However, an interesting feature of online marketing is the possibility to track the behavior of visitors on websites (Manchanda et al. 2006). This allows marketers not only to investigate the number of customers who visited the campaign website but also to inspect the quality of these visits. An interesting opportunity for future research would be to study the impact of viral marketing campaigns by integrating the reach of the campaign with behavioral data, such as the time customers spend on the website, which pages they visit, and whether they subscribe for a service or buy specific products.

We applied the viral branching model to one specific viral marketing campaign. Future research should investigate the performance of our model on other viral marketing campaigns. More interestingly, using a large set of viral marketing campaigns, it would be useful to determine the relationships between viral marketing campaign characteristics and the value of the model parameter estimates. This will provide interesting insights into what makes a campaign successful and under which circumstances. Furthermore, such insights could be useful to predict the reach of viral marketing campaigns even before their launch. In addition to relating model parameters to campaign characteristics, it would also be valuable to investigate how model parameters evolve over time during the course of a viral marketing campaign. For instance, in our research we found that response times are slower during weekends and that the number of effectively forwarded e-mails decreases as more customers are invited. It is possible that in other campaigns other parameters evolve as well. For instance, the effectiveness of seeding activities may change if more customers joined the campaign. How to design these seeding tools effectively is another fruitful area for future research. For example, in a field experiment one could study the effect of timing and different formats of seeding e-mails and banners on traffic to the campaign website. Moreover, the effect of other media, such as blogs, and search engines would be valuable to study.

In conclusion, this paper is the first to describe and predict the spread of electronic word of mouth in viral marketing campaigns. Our approach captures the interactions between customers as they are directly observed in viral marketing campaigns. Furthermore, it shows how offline and online marketing activities affect these interactions. We believe that our viral branching model is a valuable tool to develop and optimize viral marketing campaigns.

7. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://mktsci.pubs.informs.org/.

Acknowledgments

The authors thank Klaas Weima, Patrick Filius, and Ayse Geertsma of Energize for providing the data set and for their helpful suggestions during this project. The authors also gratefully acknowledge the valuable suggestions of the editor, area editor, and two anonymous reviewers.

References

Athreya, K. B., P. E. Ney. 1972. *Branching Processes*. Springer-Verlag, Berlin.

Bartlett, M. S. 1960. *Stochastic Population Models in Ecology and Epidemiology*. Methuen, London.

Bass, F. M. 1969. A new product growth for model consumer durables. *Management Sci.* **15**(5) 215–227.

Biyalogorsky, E., E. Gerstner, B. Libai. 2001. Customer referral management: Optimal reward programs. *Marketing Sci.* **20**(1) 82–95.

Blattberg, R., J. Golanty. 1978. Tracker: An early test market forecasting and diagnostic model for new product planning. *J. Marketing Res.* **15**(2) 192–202.

Bonfrer, A., X. Drèze. 2009. Real-time evaluation of e-mail campaign performance. *Marketing Sci.* **28**(2) 251–263.

Chiu, H.-C., Y.-C. Hsieh, Y.-H. Kao, M. Lee. 2007. The determinants of e-mail receivers' disseminating behaviors on the Internet. *J. Advertising Res.* **47**(4) 524–534.

De Bruyn, A., G. L. Lilien. 2008. A multi-stage model of word-of-mouth influence through viral marketing. *Internat. J. Res. Marketing* **25**(3) 151–163.

Dorman, K. S., J. S. Sinsheimer, K. Lange. 2004. In the garden of branching processes. *SIAM Rev.* **46**(2) 202–229.

Eliashberg, J., J.-J. Jonker, M. S. Sawhney, B. Wierenga. 2000. MOVIEMOD: An implementable decision support system for pre-release market evaluation of motion pictures. *Marketing Sci.* **19**(3) 226–243.

Godes, D., D. Mayzlin, Y. Chen, S. Das, C. Dellarocas, B. Pfeiffer, B. Libai, S. Sen, M. Shi, P. Verlegh. 2005. The firm's management of social interactions. *Marketing Lett.* **16**(3-4) 415–428.

Harris, T. E. 1963. *The Theory of Branching Processes*. Springer-Verlag, Berlin.

Hauser, J. R., K. J. Wisniewski. 1982. Application, predictive test, and strategy implications for a dynamic model of consumer response. *Marketing Sci.* **1**(2) 143–179.

Hogan, J. E., K. N. Lemon, B. Libai. 2004. Quantifying the ripple: Word-of-mouth and advertising effectiveness. *J. Advertising Res.* **44**(3) 271–280.

Kalyanam, K., S. McIntyre, J. T. Masonis. 2007. Adaptive experimentation in interactive marketing: The case of viral marketing at Plaxo. *J. Interactive Marketing* **21**(3) 72–85.

Kamakura, W. A., S. K. Balasubramanian. 1988. Long-term view of the diffusion of durables: A study of the role of price and adoption influence processes via tests of nested models. *Internat. J. Res. Marketing* **5**(1) 1–13.

Kendall, D. G. 1949. Stochastic processes and population growth. *J. Roy. Statist. Soc. Ser. B* **11**(2) 230–264.

Lenk, P. J., A. G. Rao. 1990. New models from old: Forecasting product adoption by hierarchical Bayes procedures. *Marketing Sci.* **9**(1) 42–53.

Lilien, G. L., A. Rangaswamy. 2004. *Marketing Engineering: Computer-Assisted Marketing Analysis and Planning*, Revised 2nd ed. Trafford Publishing, Victoria, BC, Canada.

Manchanda, P., J.-P. Dubé, K. Y. Goh, P. K. Chintagunta. 2006. The effect of banner advertising on Internet purchasing. *J. Marketing Res.* **43**(1) 98–108.

Moe, W. W. 2003. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *J. Consumer Psych.* **13**(1-2) 29–39.

Morrissey, B. 2007. Clients try to manipulate "unpredictable" viral buzz. *Adweek* **48**(March 19) 12.

New Media Age. 2007. Red Nose Day viral game played 1.16m times. (March 29) 3.

Parker, P. M. 1992. Price elasticity dynamics over the adoption life cycle. *J. Marketing Res.* **29**(3) 358–367.

Phelps, J. E., R. Lewis, L. Mobilo, D. Perry, N. Raman. 2004. Viral marketing or electronic word-of-mouth advertising: Examining consumer responses and motivations to pass along e-mail. *J. Advertising Res.* **44**(4) 333–348.

Ross, S. M. 1997. *Introduction to Probability Models*. Academic Press, San Diego.

Sevast'yanov, B. A. 1957. Limit theorems for branching stochastic processes of special form. *Theory Probab. Appl.* **2**(3) 321–331.

Shocker, A. D., W. G. Hall. 1986. Pretest market models: A critical evaluation. *J. Product Innovation Management* **3**(2) 86–107.

Silk, A. J., G. L. Urban. 1978. Pretest market evaluation of new packaged goods: A model and measurement methodology. *J. Marketing Res.* **15**(2) 171–191.

Urban, G. L. 1970. Sprinter Mod III: A model for the analysis of new frequently purchased consumer products. *Oper. Res.* **18**(5) 805–854.

Urban, G. L. 1975. Perceptor: A model for product postioning. *Management Sci.* **21**(8) 858–871.

Watts, D. J., J. Peretti. 2007. Viral marketing for the real world. *Harvard Bus. Rev.* **85**(May) 22–23.

Wireless News. 2006. TRUSTE/TNS survey: Most Internet users are not taking action to protect online privacy. (December 8) 1.

van Wyck, S. 2007. Viral is worth the investment. *B&T Weekly* **57**(February 23) 14.