Mining Knowledge-Sharing Sites for Viral Marketing

Matthew Richardson and Pedro Domingos
Department of Computer Science and Engineering
University of Washington
Box 352350
Seattle, WA 98195-2350
{mattr,pedrod}@cs.washington.edu

ABSTRACT

Viral marketing takes advantage of networks of influence among customers to inexpensively achieve large changes in behavior. Our research seeks to put it on a firmer footing by mining these networks from data, building probabilistic models of them, and using these models to choose the best viral marketing plan. Knowledge-sharing sites, where customers review products and advise each other, are a fertile source for this type of data mining. In this paper we extend our previous techniques, achieving a large reduction in computational cost, and apply them to data from a knowledge-sharing site. We optimize the amount of marketing funds spent on each customer, rather than just making a binary decision on whether to market to him. We take into account the fact that knowledge of the network is partial, and that gathering that knowledge can itself have a cost. Our results show the robustness and utility of our approach.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *data* mining; I.2.6 [**Artificial Intelligence**]: Learning – *induction*; I.5.1 [**Pattern Recognition**]: Models – *statistical*; J.4 [**Computer Applications**]: Social and Behavioral Sciences

Keywords

Probabilistic models, linear models, direct marketing, viral marketing, social networks, knowledge sharing

1. INTRODUCTION

Marketing has been one of the major applications of data mining since the field emerged. Typically, the decision of whether or not to market to a particular person is based solely on their characteristics (*direct marketing*), or those of the population segment to which they belong (*mass marketing*). This often leads to sub-optimal marketing decisions by not taking into account the effect that members of a market have on each other's purchasing decisions. In many markets, customers are strongly influenced by the opinions of their peers. *Viral marketing* takes advantage of this to inexpensively promote a product by marketing primarily to those with the strongest influence in the market. The use of relationships between people makes viral marketing potentially more profitable than direct marketing.

Data mining techniques have been successfully employed for direct marketing [9]. By building models that predict future purchasing behavior from past behavior, marketing can be more targeted and lead to increases in profit [18][22]. In previous work [5], we showed that the same could be done for viral marketing. By explicitly modeling the market as a social network [24], we were able to use the influence between customers to our advantage to significantly increase profits.

Viral marketing uses the customers in a market to promote a product. This "word-of-mouth" advertising can be much more cost effective than traditional methods since it leverages the customers themselves to carry out most of the promotional effort. Further, people typically trust and act on recommendations from friends more than from the company selling the product.

Examples of viral marketing are becoming increasingly common. A classic example of this is the Hotmail free email service, which grew from zero to 12 million users in 18 months on a miniscule advertising budget, thanks to the inclusion of a promotional message with the service's URL in every email sent using it [13]. Competitors using conventional marketing fared far less well. Many markets, notably those associated with information goods (e.g., software, media, telecommunications, etc.) contain strong network effects (known in the economics literature as network externalities). In these, ignoring the relationships between customers can lead to a severely sub-optimal marketing plan.

In the presence of strong network effects, it is crucial to consider not only a customer's *intrinsic value* (his value as a customer based on the products he is likely to purchase), but also his *network value*. The network value of a customer is high when he is expected to have a very positive influence on others' probabilities of purchasing the product. A customer whose intrinsic value is less than the cost of marketing may in fact be worth marketing to when his network value is considered. The immediate effect of marketing to him may be negative, but the overall effect may be positive once his influence on his friends, their influences on their friends, and so on is taken into account. Further, a customer who looks valuable based on intrinsic value alone may in fact not be worth marketing to if he is expected to have an overall negative effect on others in the market (e.g., a person who tends to give very low product ratings). Ignoring the network value can result in incorrect marketing decisions, especially in a market with strong network effects.

To estimate the network value of its customers, a company needs to know the relationships between them. One source of such information is the Internet, with its plethora of chat rooms, discus-

SIGKDD 02, Edmonton, Alberta, Canada.

sion forums, and knowledge-sharing web sites. In these is found a wealth of social interaction, often product-related, which a company could use to gather information on the relationships between its customers. Knowledge-sharing sites in particular are often product-oriented. On these sites, information about product likes and dislikes, ratings of quality, benchmarks, and comparisons are exchanged, making them an ideal source for data about customer preferences and interactions.

In this paper, we extend ideas from our earlier work [5] and apply them to the domain of knowledge-sharing sites. We show how to find optimal viral marketing plans, use continuously valued marketing actions, and reduce computational costs (Sections 2 and 3). In Sections 4 and 5, we apply the model to Epinions, a popular knowledge-sharing site. In practice, the relationships between customers is often unknown, but may be obtained at some cost. We introduce a technique for marketing in such a situation and show that it performs well even with very limited marketing research funds. We conclude with a discussion of related work and future directions.

2. THE MODEL

Consider a set of n potential customers, and let X_i be a Boolean variable that takes the value 1 if customer i buys the product being marketed, and 0 otherwise. Let the *neighbors* of X_i be the customers who directly influence X_i: $\mathbf{N_i}=\{X_{i,1},\ldots,X_{i,n_i}\} \subseteq \mathbf{X}-\{X_i\}$, where $\mathbf{X}=\{X_1,\ldots,X_n\}$. The product is described by a set of attributes $\mathbf{Y}=\{Y_1,\ldots,Y_m\}$. Let M_i be the marketing action that is taken for customer i. For example, M_i could be a Boolean variable, with M_i=1 if the customer is (say) offered a discount, and M_i=0 otherwise. Alternatively, M_i could be a continuous variable indicating the size of the discount offered, or a nominal variable indicating which of several possible actions is taken. Let $\mathbf{M}=\{M_1,\ldots,M_n\}$ be the *marketing plan*. Then, for all X_i, we will assume that

$$\begin{aligned} &P(X_i \mid \mathbf{X}-\{X_i\}, \mathbf{Y}, \mathbf{M}) \\ &\quad = P(X_i \mid \mathbf{N_i}, \mathbf{Y}, \mathbf{M}) \\ &\quad = \beta_i P_0(X_i \mid \mathbf{Y}, M_i) + (1-\beta_i) P_N(X_i \mid \mathbf{N_i}, \mathbf{Y}, \mathbf{M}) \end{aligned} \quad (1)$$

$P_0(X_i \mid \mathbf{Y}, M_i)$ is X_i's *internal* probability of purchasing the product. $P_N(X_i \mid \mathbf{N_i}, \mathbf{Y}, \mathbf{M})$ is the effect that X_i's neighbors have on him. β_i is a scalar with $0 \le \beta_i \le 1$ that measures how self-reliant X_i is. For many products, such as cellular telephones, multi-player computer games, and Internet chat programs, a customer's probability of purchasing depends strongly on whether his friends have also purchased the product. In previous work [5] we modeled this interaction with a non-linear function. In this paper, we employ a simple linear model to approximate this effect:

$$P_N(X_i = 1 \mid \mathbf{N_i}, \mathbf{Y}, \mathbf{M}) = \sum_{X_j \in \mathbf{N_i}} w_{ij} X_j \quad (2)$$

where w_{ij} represents how much customer i is influenced by his neighbor j, with $w_{ij} \ge 0$ and $\sum_{X_j \in \mathbf{N_i}} w_{ij} = 1$ (Note, $w_{ij} = 0$ if $j \notin \mathbf{N_i}$). While not exact, we believe it is a reasonable approximation when the probabilities are all small, as is typically the case for marketing domains. Linear models often perform well, especially when data is sparse [4], and provide significant advantages for computation. Note that we are modeling only positive interactions between customers, which we found in our previous work to be the most common type.

Combining Equations 1 and 2, we get

$$\begin{aligned} &P(X_i = 1 \mid \mathbf{N_i}, \mathbf{Y}, \mathbf{M}) = \\ &\quad \beta_i P_0(X_i = 1 \mid \mathbf{Y}, M_i) + (1-\beta_i) \sum_{X_j \in \mathbf{N_i}} w_{ij} X_j \end{aligned} \quad (3)$$

For the purposes of this paper, we will be calculating the optimal marketing plan for a product that has not yet been introduced to the market. In this situation, the state of the neighbors will not be known, so we derive a formula for computing P(X_i = 1 | **Y**, **M**). We first sum over all possible neighbor states:

$$P(X_i = 1 \mid \mathbf{Y}, \mathbf{M}) = \sum_{\tilde{\mathbf{N}} \in C(\mathbf{N_i})} P(X_i = 1 \mid \tilde{\mathbf{N}}, \mathbf{Y}, \mathbf{M}) P(\tilde{\mathbf{N}} \mid \mathbf{Y}, \mathbf{M})$$

where C($\mathbf{N_i}$) is the set of all possible configurations of the neighbors of X_i, and hence $\tilde{\mathbf{N}}$ is an set of neighbor state assignments. Substituting equation 3, we get:

$$\begin{aligned} &P(X_i = 1 \mid \mathbf{Y}, \mathbf{M}) = \\ &\quad \sum_{\tilde{\mathbf{N}} \in C(\mathbf{N_i})} \beta_i P_0(X_i = 1 \mid \mathbf{Y}, M_i) P(\tilde{\mathbf{N}} \mid \mathbf{Y}, \mathbf{M}) \\ &\quad + \sum_{\tilde{\mathbf{N}} \in C(\mathbf{N_i})} (1-\beta_i) \sum_{X_j \in \mathbf{N_i}} w_{ij} \tilde{N}_j P(\tilde{\mathbf{N}} \mid \mathbf{Y}, \mathbf{M}) \end{aligned}$$

where $\tilde{N}_j$ is the value of X_j specified by $\tilde{\mathbf{N}}$. $P_0(X_i \mid \mathbf{Y}, M_i)$ is independent of $\tilde{\mathbf{N}}$, so the first term simplifies to it. We swap the summation order in the second term, and note that it is zero whenever $\tilde{N}_j$ is zero. This leads to:

$$\begin{aligned} &P(X_i = 1 \mid \mathbf{Y}, \mathbf{M}) \\ &= \beta_i P_0(X_i = 1 \mid \mathbf{Y}, M_i) + (1-\beta_i) \sum_{X_j \in \mathbf{N_i}} \sum_{\substack{\tilde{\mathbf{N}} \in C(\mathbf{N_i}) \\ \text{with } \tilde{N}_j = 1}} w_{ij} P(\tilde{\mathbf{N}} \mid \mathbf{Y}, \mathbf{M}) \end{aligned}$$

Since the inner summation is over all possible values of $\tilde{\mathbf{N}}$ whenever $\tilde{N}_j$=*1*, it is equivalent to w_{ij}P(X_j = 1 | **Y**, **M**), hence:

$$\begin{aligned} &P(X_i = 1 \mid \mathbf{Y}, \mathbf{M}) \\ &= \beta_i P_0(X_i = 1 \mid \mathbf{Y}, M_i) + (1-\beta_i) \sum_{X_j \in \mathbf{N_i}} w_{ij} P(X_j = 1 \mid \mathbf{Y}, \mathbf{M}) \end{aligned} \quad (4)$$

Because Equation 4 expresses the probabilities $P(X_i = 1 \mid \mathbf{Y}, \mathbf{M})$ as a function of themselves, it can be applied iteratively to find them, starting from a suitable initial assignment. A natural choice for initialization is to use the internal probabilities $P_0(X_i = 1 \mid \mathbf{Y}, M_i)$.

The marketer's goal is to find the marketing plan that maximizes profit. For simplicity, assume that M is a Boolean vector (i.e., only one type of marketing action is being considered, such as offering the customer a given discount). Let c be the cost of marketing to a customer (assumed constant), r_0 be the revenue from selling the product to the customer if no marketing action is performed, and r_1 be the revenue if marketing is performed. r_0 and r_1 will be the same unless the marketing action includes offering a discount. Let $f_i^1(\mathbf{M})$ be the result of setting M_i to 1 and leaving the rest of **M** unchanged, and similarly for $f_i^0(\mathbf{M})$. The *expected lift in profit* from marketing to customer i in isolation (i.e., ignoring his effect on other customers) is then [3]

$$\begin{aligned} ELP_i^1(\mathbf{Y}, \mathbf{M}) = &\; r_1 P(X_i = 1 \mid \mathbf{Y}, f_i^1(\mathbf{M})) \\ &- r_0 P(X_i = 1 \mid \mathbf{Y}, f_i^0(\mathbf{M})) - c \end{aligned}$$

We also refer to this as the customer's *intrinsic value*. Let $\mathbf{M_0}$ be the null vector (all zeros). The global lift in profit that results from a particular marketing plan $\mathbf{M}$ is then

$$ELP(\mathbf{Y},\mathbf{M}) = \sum_{i=1}^{n}\left[r_i P(X_i=1|\mathbf{Y},\mathbf{M}) - r_0 P(X_i=1|\mathbf{Y},\mathbf{M}_0) - c_i\right]$$

where $r_i=r_1$ and $c_i=c$ if $M_i=1$, and $r_i=r_0$ and $c_i=0$ if $M_i=0$.

A customer's *total value* is the global lift in profit from marketing to him: $ELP(\mathbf{Y}, f_i^1(\mathbf{M})) - ELP(\mathbf{Y}, f_i^0(\mathbf{M}))$. A customer's *network value* is the difference between his total and intrinsic values. A customer with a high network value is one who, when marketed to, directly or indirectly influences many others to purchase.

Our previous work was based on this Boolean marketing case, but in this paper we explore continuous valued marketing actions as well. The expected lift in profit in the continuous case is a straightforward extension of the Boolean one. Let z be a marketing action, with $0 \le z \le 1$, and $z = 0$ when no marketing is performed. Let $c(z)$ be the cost of performing the action (with $c(0)=0$), and $r(z)$ be the revenue obtained if the product is purchased. Let $f_i^z(\mathbf{M})$ be the result of setting M_i to z and leaving the rest of $\mathbf{M}$ unchanged. The expected lift in profit from performing marketing action z on customer i in isolation is then

$$\begin{aligned} ELP_i^z(\mathbf{Y},\mathbf{M}) = {} & r(z)P(X_i=1|\mathbf{Y}, f_i^z(\mathbf{M})) \\ & - r(0)P(X_i=1|\mathbf{Y}, f_i^0(\mathbf{M})) - c(z) \end{aligned} \quad (5)$$

The global lift in profit is

$$ELP(\mathbf{Y},\mathbf{M}) = \sum_{i=1}^{n}\left[r(M_i)P(X_i=1|\mathbf{Y},\mathbf{M}) - r(0)P(X_i=1|\mathbf{Y},\mathbf{M}_0) - c(M_i)\right]$$

3. INFERENCE AND SEARCH

Our goal is to find the $\mathbf{M}$ that maximizes $ELP(\mathbf{Y}, \mathbf{M})$. In our previous work, we assumed marketing actions were Boolean, and heuristically searched through the vast space of possible marketing plans. Because of the linearity of the model presented here (see Equation 3), the effect that marketing to a person has on the rest of the network (their *network effect*) is independent of the marketing actions to other customers. From a customer's network effect, we can directly compute whether he is worth marketing to. Let the $\Delta_i(\mathbf{Y})$ be the *network effect* of customer i for a product with attributes $\mathbf{Y}$. It is defined as the total increase in probability of purchasing in the network (including X_i) that results from a unit change in $P_0(X_i)$:

$$\Delta_i(\mathbf{Y}) = \sum_{j=1}^{n}\frac{\partial P(X_j=1|\mathbf{Y},\mathbf{M}_0)}{\partial P_0(X_i=1|\mathbf{Y},M_i)} \quad (6)$$

Since $\Delta_i(\mathbf{Y})$ is the same for any $\mathbf{M}$, we define it for $\mathbf{M} = \mathbf{M}_0$. We can calculate $\Delta_i(\mathbf{Y})$ using the following recursive formula (see the Appendix for a proof)

$$\Delta_i(\mathbf{Y}) = \sum_{j=1}^{n} w_{ji}\Delta_j(\mathbf{Y}) \quad (7)$$

Intuitively, customer i's network effect is simply the effect that he has on people he influences, times their effect on the network. $\Delta_i(\mathbf{Y})$ is initially set to 1 for all i, then recursively re-calculated using equation 7 until convergence (note this takes approximately linear time in the number of non-zero w_{ij}'s). Empirically, we found it converged quickly (10-20 iterations).

Note that while the *network value* of a customer depends on the marketing scenario, the *network effect* does not. The *network effect* simply describes how much influence a customer has on the network. The *network value* depends on the network effect, the customer's responsiveness to marketing, and the costs and revenues associated with the marketing scenario.

With the network effects in hand, we can calculate the expected lift in profit of marketing to each customer. For convenience, we define $\Delta P_i(z,\mathbf{Y})$ to be the immediate change in customer i's probability of purchasing when he is marketed to with marketing action z:

$$\Delta P_i(z,\mathbf{Y}) = \beta_i\left[P_0(X_i=1|\mathbf{Y},M_i=z) - P_0(X_i=1|\mathbf{Y},M_i=0)\right]$$

From Equation 6, and given that $P(X_j=1|\mathbf{Y},\mathbf{M}_0)$ varies linearly with $P_0(X_i=1|\mathbf{Y},M_i)$, the change in the probability of purchasing across the entire network is then

$$\begin{aligned} \sum_{j=1}^{n}\Delta P(X_j=1|\mathbf{Y},\mathbf{M}_0) &= \Delta_i(\mathbf{Y})\cdot\Delta P_0(X_i=1|\mathbf{Y},M_i) \\ &= \Delta_i(\mathbf{Y})\cdot\Delta P_i(z,\mathbf{Y}) \end{aligned}$$

Typically, only a small portion of the network will be marketed to. Therefore, it is relatively safe to approximate the increase in revenue from the network due to marketing to customer i as his influence on the network multiplied by $r(0)$. The total lift in profit is this increase in revenue on the network, plus the change in revenue from customer i, minus the cost of the marketing action:

$$\begin{aligned} ELP_{i,total}^z(\mathbf{Y},\mathbf{M}) = {} & r(0)\left[(\Delta_i(\mathbf{Y})-1)\cdot\Delta P_i(z,\mathbf{Y})\right] \\ & + \left[r(z)P(X_i=1|\mathbf{Y}, f_i^z(\mathbf{M})) - r(0)P(X_i=1|\mathbf{Y},\mathbf{M})\right] \\ & - c(z) \end{aligned}$$

Notice that this approximation is exact when $r(z)$ is constant, which is the case in any marketing scenario that is advertising-based (i.e., if it does not offer discounts). When this is the case, the equation simplifies to:

$$\begin{aligned} ELP_{i,total}^z(\mathbf{Y},\mathbf{M}) &= r\left[(\Delta_i(\mathbf{Y})-1)\cdot\Delta P_i(z,\mathbf{Y})\right] + r\left[\Delta P_i(z,\mathbf{Y})\right] - c(z) \\ &= r\Delta_i(\mathbf{Y})\cdot\Delta P_i(z,\mathbf{Y}) - c(z) \end{aligned} \quad (8)$$

With Equation 8, we can directly estimate customer i's lift in profit for any marketing action z. Typically, we will want to find the z that maximizes the lift in profit. To do this, we take the derivative with respect to z and set it equal to zero, resulting in:

$$r\Delta_i(\mathbf{Y})\frac{d\Delta P_i(z,\mathbf{Y})}{dz} = \frac{dc(z)}{dz} \quad (9)$$

Assuming $\Delta P_i(z,\mathbf{Y})$ is differentiable, this allows us to directly calculate the z which maximizes $ELP_{i,total}^z(\mathbf{Y}, \mathbf{M})$ which, because our model is linear, is the optimal value for M_i in the $\mathbf{M}$ that

maximizes $ELP(\mathbf{Y}, \mathbf{M})$. Hence, from the customers' network effects, $\Delta_i(\mathbf{Y})$, we can directly calculate the optimal marketing plan. We now show how this model can be applied to knowledge-sharing sites.

4. MINING KNOWLEDGE-SHARING SITES

Internet use has exploded over the past decade. Millions of people interact with each other online, and, in many instances, those social interactions are recorded in archives that reach back twenty years or more[1]. As a result, there are many online opportunities to mine social networks for the purposes of viral marketing. UseNet newsgroups, IRC, instant messaging, online forums, and email mailing lists are examples of possible sources.

In this paper, we concentrate on knowledge-sharing sites. On such sites, volunteers offer advice, product ratings, or help to other users, typically for free. Social interaction on knowledge-sharing sites comes in a variety of forms. One feature that is often found is some form of explicit trust between users. For example, at many sites, users rate reviews according to how helpful or accurate they are. On others, users directly rate other users. Without a filtering feature such as this, knowledge-sharing sites can quickly become mired in inaccurate or inappropriate reviews.

We have chosen to mine Epinions[2], possibly the best known knowledge-sharing site. On Epinions, members submit product reviews, including a rating (from 0 to 5 stars) for any of over one hundred thousand products. As added incentive, reviewers are paid each time one of their reviews is read. Epinions users interact with each other in both of the ways outlined above, by rating reviews, and also by listing reviewers that they trust. The network of trust relationships between users is called the "web of trust", and is used by Epinions to re-order the product reviews such that a user first sees reviews by users that they trust. The trust relationships between users, and thus the entire web of trust, can be obtained by crawling through the pages of the individual users[3]. With over 75k users and 500k edges in its web of trust, and 586k reviews over 104k products, Epinions is an ideal source for experiments on social networks and viral marketing. Interestingly, we found that the distribution of trust relationships in the web of trust is Zipfian [25], as has been found in many social networks [24]. This is evidence that the web of trust is a representative example of a social network, and thus is a good basis for our study. A Zipfian distribution of trust is also indicative of a skewed distribution of network values, and therefore of the potential utility of viral marketing.

To apply our model to Epinions, we needed to estimate some parameters, such as the effect that marketing has on a customer's probability of purchasing, the self-reliance factor β_i, and the amount of influence between customers w_{ij}. In practice, the marketing research department of a company, or the maintainers of the knowledge-sharing site itself, would typically have the resources and access to customers necessary to experimentally determine these parameters. For instance, the effect that marketing has on a customer could be measured by selecting users at random and recording their responses (both when being marketed to and not). The parameters could be estimated individually for each user, or (requiring far less data) as the same for all users, as was done in Chickering and Heckerman [3]. If this is not feasible, they could be set using a combination of prior knowledge and any demographic information available.

For Epinions, we made the simplifying assumption that a user is more likely to purchase a product if it was reviewed by a person he trusts. Though not required by the model, we considered all trusted people to have equal influence, as there is no data in Epinions to inform otherwise. Thus, $\mathbf{N_i}=\{X_j$ such that i trusts $j\}$ and $w_{ij}=1/|\mathbf{N_i}|$ for $X_j \in \mathbf{N_i}$. For the product attribute vector $\mathbf{Y}$, we used a single attribute: the product category (from one of 25 top-level categories defined by Epinions). The model supports more complex attribute vectors. For example, one could imagine using the text description of products, possibly augmented by the product category and sub-category. We plan to explore their effect in future work. All that remained to define was $P_0(X_i \mid \mathbf{Y}, M_i)$, which we estimated using a naïve Bayes model[4] for X_i as a function of $\mathbf{Y}$ and M_i.

$$P_0(X_i \mid \mathbf{Y}, M_i) = \frac{P_0(\mathbf{Y} \mid X_i) P_0(M_i \mid X_i) P_0(X_i)}{\sum_{X_i} P_0(\mathbf{Y} \mid X_i) P_0(M_i \mid X_i) P_0(X_i)} = \frac{P_0(X_i \mid \mathbf{Y}) P_0(M_i \mid X_i)}{\sum_{X_i} P_0(X_i \mid \mathbf{Y}) P_0(M_i \mid X_i)}$$

We used a naïve Bayes model for $P_0(X_i \mid \mathbf{Y})$. We equated reviewing a product with purchasing it[4], so training the model was simply a matter of counting. In the case of Epinions, measuring the effectiveness of marketing on the users was not possible for us. We expected marketing to have a larger effect on a customer who was already inclined to purchase the product, so we followed our previous work and set $P_0(M_i \mid X_i)$ so as to obtain (for the Boolean marketing scenario):

$$P_0(X_i = 1 \mid M_i = 1) = \min\{\alpha P_0(X_i = 1 \mid M_i = 0), 1\} \quad (10)$$

where $\alpha > 1$ is a parameter that specifies the magnitude of the marketing effect[5].

5. EXPERIMENTS

We built the model based on Epinions data, as discussed above, and used it to gather empirical results. For all of the experiments, we used just one of the 25 product categories, "Kids & Family", as it had the most reviews per product (10.2, on average) and

[1] See http://groups.google.com/ and http://www.archive.org/.

[2] http://www.epinions.com

[3] Epinions does not provide a list of all of its users, so we seeded the crawl with the top reviewers in each product category and followed both "trusts" and "trusted-by" links to find other users.

[4] We expect that more users purchase the product than review it. However, purchasers who do not review have no additional effect on the network, so knowing the ratio of purchasers to reviewers would simply scale the results. The results would be affected if we knew, *per user*, the probability of purchasing vs. reviewing, but this information is not available to us.

[5] To fully specify $P(M_i \mid X_i)$ we used the additional constraint that $P(\mathbf{Y}, M_i=1) = P(\mathbf{Y}, M_i=0)$. With the values of α we used it was always possible to satisfy Equation 10 and this constraint simultaneously.

reviews per person who submitted at least one review in the category (5.8, on average). We first tested the Boolean marketing case. We hypothesized a simple advertising situation with α=2, r_0=1, r_1=1, which meant revenues were in units of the number of products sold, and a person's internal probability of purchasing a product doubled after being advertised to[6]. In earlier work, we varied α and found that, while it affected the scale of the results, it had little effect on the qualitative nature of them. Thus, for this paper, we fixed α and instead varied other characteristics of the model. We had no data to estimate users' self-reliance, so we simply chose to set β_i=0.5 for all customers. To combat data sparseness, $P_0(X_i \mid \mathbf{Y})$ was smoothed using an m-estimate with m=2 and the population average as the prior. These parameters were all chosen before running any experiments.

Table 1: Profit results for Boolean marketing scenario for various costs of marketing.

	α=2, r_0=1, r_1=1		
	c = 0.1	c = 0.01	c = 0.001
No Marketing	37.78	37.78	37.78
Direct Marketing	37.78	42.71	66.08
Viral Marketing	47.25	60.54	70.23

5.1 Profits and Network Values

Viral marketing resulted in a considerable increase in profit over direct marketing (see Table 1). Notice that when the cost of marketing is a significant fraction of the revenue, the direct marketer will choose to market to no one because the cost of marketing exceeds the expected revenue from the customer (since the customers' influences on each other are being ignored). As this scenario illustrates, assuming the model is accurate, viral marketing will always perform at least as well as direct marketing, often outperforming it by a substantial margin.

We measured the network value of all of the customers. Figure 1 shows the 500 highest network values (out of 75888) in decreasing order. The unit of value in this graph is the average revenue that would be obtained by marketing to a customer in isolation, without costs or discounts. Thus, a network value of 200 for a given customer implies that by marketing to him we essentially get free marketing to an additional 200 customers. The scale of the graph depends on the marketing scenario (e.g., network values increase with α), but the shape generally remains the same. The figure shows that a few users have very high network value. This is the ideal situation for the type of targeted viral marketing we propose, since we can effectively market to many people while incurring only the expense of marketing to those few.

A customer with high network value is one who: (1) Is likely to purchase the product, and thus is more affected by the marketing, and (2) is trusted by many other people in the network, who tend

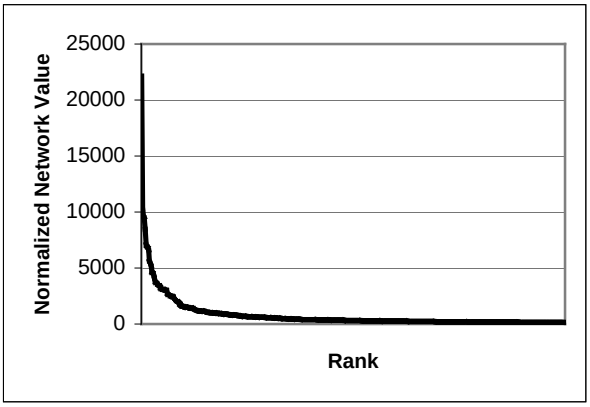

Figure 1: Typical distribution of network value.

to have low β_i, and who also have characteristic 2, and so on recursively. For instance, the customer with the highest network value (22,000) influences 784 people, and has a probability of purchasing of 0.03, which is 23 times that of the average person.

5.2 Speed

The linear model introduced in this paper has tremendous speed advantages over a non-linear model such as that introduced in our previous work. Because of the independence that linearity provides, we are able to simultaneously calculate the network value for all customers. The network value is independent of the marketing actions being performed on others, which allows us to find the optimal marketing plan[7] without performing a heuristic search over plans. It would take approximately 100 hours to perform the single-pass search (the fastest of the heuristic search methods introduced in our previous work) with this model, or about 10-15 minutes if we make approximations in the inference. In contrast, the linear model takes 1.05 seconds to find the optimal marketing plan. At these speeds, our model could be used to find optimal marketing plans for markets involving hundreds of millions of customers in just hours.

5.3 Continuous Marketing Actions

Continuous-valued marketing actions ($M_i \in [0,1]$) allow the marketer to better optimize the marketing plan – tailoring the action for each person specifically to his characteristics. Our framework allows for any function to be used to model $P_0(X_i \mid \mathbf{Y}, M_i)$, as long as it is differentiable in M_i. As in the Boolean case, we have chosen to model the effect of marketing as a multiplicative factor on the internal probability of purchasing:

$$P_0(X_i \mid M_i = z) = \alpha(z) \cdot P_0(X_i \mid M_i = 0)$$

$\alpha(z)$ could be any differentiable function, and we assume $\alpha(0)$=1. $c(z)$ also could be any differentiable function. We have chosen $c(z)=c_1 z$ such that the cost of marketing is directly proportional to the amount of marketing being performed.

[6] In previous work we varied the value of α and found that, while it affected the scale of results, they remained qualitatively similar.

[7] The plan is optimal if $r_0=r_1$ (or if $r(z)$ is constant in the continuous marketing scenarios). If $r_1<r_0$ then the plan overestimates the revenues from influence on the network, potentially resulting in a sub-optimal marketing plan. In our experience, this overestimation ranged from 1% to 10% of the profits. We thus believe the resulting plan was still nearly optimal.

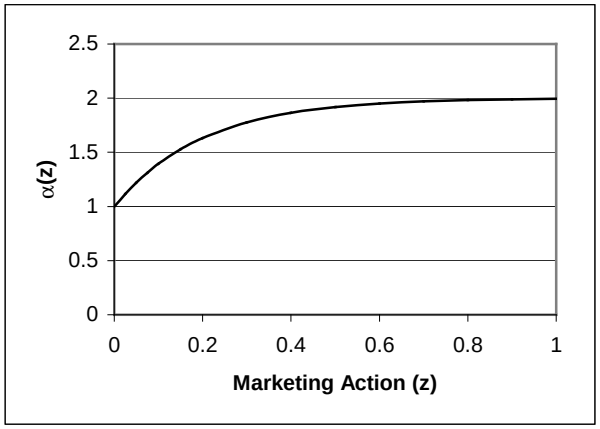

Figure 2: Marketing effect vs. marketing action.

We believe an exponentially asymptotic function for $\alpha(z)$ is reasonable; it models the phenomenon of diminishing returns (i.e., the more money that is spent on marketing, the less improvement is derived from it). We also experimented with logarithmic and inverse polynomial functions, which gave similar results. The function we used was:

$$\alpha(z) = \alpha_\infty + (1-\alpha_\infty)e^{-\lambda z}$$

Note that $\alpha(0)=1$, and $\alpha(z)\to\alpha_\infty$ as $z\to\infty$. The parameter λ affects the curvature of the function; $\alpha(z)$ converges to α_∞ more quickly with a large λ. In the experiments below, we used $\lambda=5$, which is large enough that $\alpha(1) \approx \alpha_\infty$, yet low enough that $\alpha(z)$ does not converge to α_∞ too quickly. The resulting curve, for $\alpha_\infty=2$, is shown in Figure 2.

From equation 9, we can find the optimal marketing action for each customer

$$\frac{dc(z)}{dz} = r\Delta_i(\mathbf{Y})\beta_i P_0(X_i=1\,|\,\mathbf{Y}, M_i=0)\frac{d(\alpha(z)-1)}{dz}$$

$$\Rightarrow \quad c = -r\Delta_i(\mathbf{Y})\beta_i\lambda(1-\alpha_\infty)e^{-\lambda z}$$

$$\Rightarrow \quad z = -\frac{\ln\left(\dfrac{-c}{r\Delta_i(\mathbf{Y})\beta_i\lambda(1-\alpha_\infty)P_0(X_i=1\,|\,\mathbf{Y}, M_i=0)}\right)}{\lambda}$$

The second derivative is negative, implying the point is a maximum.

We ran the same experiments as in the Boolean case, with $\alpha_\infty=2$ so that marketing fully to a customer will double their internal probability of purchasing the product, as before. The results are presented in Table 2. In all three scenarios, and for both direct and viral marketing, continuous marketing actions resulted in a higher lift in profit than Boolean actions, sometimes by a very significant amount. Viral marketing also continued to consistently outperform direct marketing.

The increased lift in profit is due to two factors: (1) At low z, the $\alpha(z)$ curve provides a more favorable ratio of marketing effect to cost, and (2) tailoring the marketing action for each customer allows us to optimize the tradeoff between the cost and benefit of marketing on a per customer basis.

Table 2: Profit results for continuous marketing scenario for various costs of marketing.

	$\alpha_\infty=2$, $r(z)=1$, $\lambda=5$		
	$c_1 = 0.1$	$c_1 = 0.01$	$c_1 = 0.001$
No Marketing	37.78	37.78	37.78
Direct Marketing	37.84	51.71	68.38
Viral Marketing	51.14	63.23	71.28
Lift over Boolean Viral Marketing	3.89 (41.08%)	2.69 (11.82%)	1.05 (3.24%)

To verify that factor (1) is not the sole cause of the increase in profit, we ran Boolean marketing experiments with $\alpha=\alpha(z)$ and $c=c(z)$ for z ranging from 0 to 1. Doing so simulates a company which globally optimizes its choice of marketing action, but still performs that same (or no) action on each customer. The maximum realizable profits in this case were 49.90, 61.60, and 70.23 for a c_1 of 0.1, 0.01, and 0.001. These results show that tailoring the marketing action for each customer is indeed a significant cause of the increase in profits derived from the continuous marketing case.

One interesting question is what happens if the marketing effect function $\alpha(z)$ is linear, $\alpha(z)= \alpha z$. In this case, continuous-valued marketing reduces to Boolean marketing. If it would be profitable to market to a customer some ($z>0$), then the benefit of marketing to him must be higher than the cost for any z (since both the cost and the benefit are linear), and it would thus advantageous to market to him the maximum possible ($z=1$).

5.4 Incomplete Network Knowledge

So far, we have considered only markets where the entire social network between customers is known. This is often not the case. In fact, most companies today have little or no knowledge of the actual relationships between their customers. In such a situation, companies may simply choose to use direct marketing, but if they do, they will likely lose profit opportunities, as demonstrated in earlier sections. In the following sections, we will demonstrate that even with little network knowledge, our viral marketing methods still outperform direct marketing. In all of the experiments that follow, we used continuous-valued marketing actions, with the same parameters as those used for Section 5.3 (Table 2) and $c_1 = 0.1$.

5.4.1 Viral marketing is robust

We simulated partial knowledge by randomly removing members from the neighbor sets, which corresponds to randomly removing edges from the social network. This is the situation a company would be in if they had only a random sample of the neighbor relations between customers. We devised the optimal marketing plan on the incomplete network, and then tested this plan on the complete network, which simulates the "real-world". Naturally, when no edges are known, viral marketing is equivalent to direct marketing.

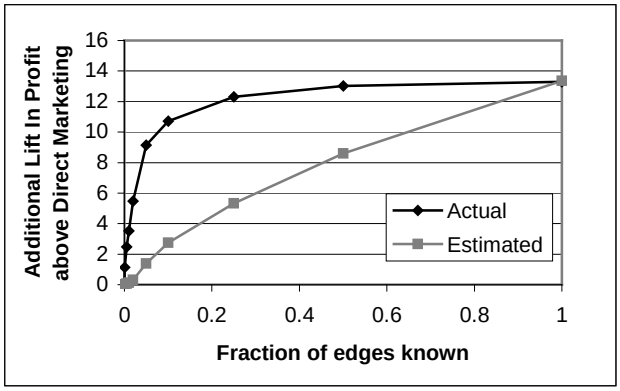

Figure 3: Actual and estimated difference between viral marketing and direct marketing profits with only partial network knowledge.

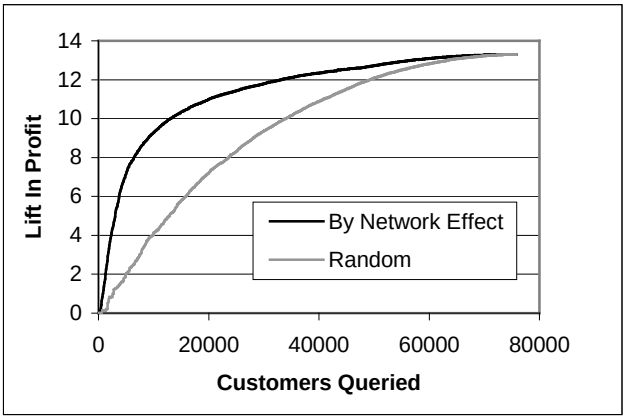

Figure 4: Lift in profit (on the full network) for the viral marketing plan in which the given number of customers has been queried for their neighbor information.

In Figure 3 ("Actual"), we show the difference in profit between direct and viral marketing for partially known networks. Surprisingly, the company can achieve 69% of the lift in profit knowing only 5% of the edges in the network. Further, the algorithm considerably underestimates the lift in profit that will result (Figure 3, "Estimated"), meaning that for a company with only partial network knowledge, not only are viral marketing plans robust but the actual results of viral marketing will be significantly better than the algorithm estimates.

We hypothesize that this robustness will occur whenever the edges are missing at random (or approximately so), resulting in a correlation between the number of people who trust a given person in the partial network and the number who trust him in the true network. A customer who appears to have a high network value in the partial network is likely to have a high network value in the full network, and would thus be chosen to be marketed to. We also believe that the algorithm could use an estimate of the fraction of edges that are missing to construct an even better viral marketing plan; we plan to investigate this in future work.

5.4.2 Acquiring new network knowledge

In many instances, a company will have little or no knowledge about the relationships between its customers, but may be willing to spend marketing research funds to acquire it. More knowledge about the influences between customers will allow the company to form a marketing plan with a higher lift in profit. If the company could compute the *value of information* [8] of knowing the neighbors of each customer, it could then make a decision-theoretic choice of which, and how many, customers to query.

The acquisition of neighbor relations could be done in many ways. For the purposes of this paper, we assume that it is done by selecting a user to query, spending money to persuade him to provide a list of the people he trusts, selecting another user to query, and so on. We assume the company has a fixed amount of money it is willing to spend for this, and that the cost of querying a user is constant. The interesting problem is thus not how many users to query, but how to select the subset of users to query that leads to the most profit.

A customer with a high network effect has a large influence on the network, and is thus one that we wish to influence to purchase the product. Apart from directly marketing to the customer, we can indirectly influence him by marketing to those that he trusts, which we can discover by querying him. One estimate for a customer's network effect on the full network is his network effect on the partial network. We thus query the customer with the highest network effect on the partial network, recalculate network effects with the new information, query the next customer with highest network effect, and so on until the marketing funds have been spent.

We performed this experiment, starting with a network containing no neighbor information[8]. Figure 4 shows the resulting lift in profit, compared to randomly selecting customers to query. Our method performs well, lifting profits an order of magnitude more than random choice would when 1000 customers are queried, and by almost 3 times the lift achieved by random choice when 10% of the customers are queried.

We must re-calculate the customers' network effects each time we query a user. We can drastically speed this up by querying the 100 customers with highest network effect at each iteration, with a potential loss of accuracy. Interestingly, the lift in profit when selecting 100 customers at a time is only (on average) 0.008 less than when selecting one at a time, a negligible amount compared to the lift in profit itself. Since it takes $1/100^{th}$ the time to run, this approximation could be used to make knowledge acquisition tractable for non-linear models, or for markets of tens of millions of customers.

In future work, we would like to find a measure that estimates the increase in ELP of querying one more user, thus informing the company when to stop acquiring network knowledge. This would allow us to optimize the overall profit (lift in profit minus funds spent to acquire network knowledge). We believe such a measure could be formed from the ELP, an estimate of the number of missing edges, and other statistics on the partial network.

[8] The first customers to query are therefore chosen at random.

6. RELATED WORK

In our previous work [5], we mined a collaborative filtering system to demonstrate the advantages of our viral marketing approach over direct or mass marketing. There, we used a more complicated, piecewise linear function over product ratings to determine the influences of customers on each other. In this paper, we used a model with stronger linearity assumptions to achieve greater scalability. A disadvantage of our previous work is that it required full knowledge of network structure, and restricted the marketer to selecting Boolean marketing actions. Both of these limitations were addressed and overcome in this paper.

Interestingly, the computation of network effect (see Equation 7) is very similar to the PageRank[21] algorithm, used by Google[2] for determining important web pages. In PageRank, a web page is valued highly if many highly valued pages point to it. Similarly, in viral marketing a customer is valued highly if he influences many highly valued customers. The computation is equivalent to finding the primary eigenvector of the matrix W, where $W_{ij}=w_{ij}$ (w_{ji} for PageRank). The network effect of a customer is also proportional to the probability that a random walker, who randomly traverses the links of influence in the network backwards, is at that customer. Also related is the HITS[15] algorithm, which would find bipartite "trusts/trusted-by" sub-graphs in the web of trust. Interestingly, social networks, the World-Wide Web, and many naturally occurring networks all exhibit Zipfian, or "scale free" characteristics, and have been the topic of much recent research [17] [1].

Social networks have been the object of much research. One classic paper is that by Milgram [20], which estimated that every person in the world is only six acquaintances away from every other. Some recent social network research uses the Internet as a source of data. For instance, Schwartz and Wood [23] mined social relationships from email logs, the ReferralWeb project mined a social network from a wide variety of publicly-available online information [14], and the COBOT project gathered social statistics from participant interactions in the LambdaMoo MUD [11]. Our network was mined from a knowledge-sharing site. A good overview of Epinions and other sites like it can be found in Frauenfelder [6].

Several researchers have studied the problem of estimating a customer's lifetime value from data [12], generally focusing on variables like an individual's expected tenure as a customer [19] and future frequency of purchases [7]. Networks of customers have received some attention in the marketing literature [10] but most of these studies are purely qualitative, or involve very small data sets and overly simplified models. Krackhardt [16] proposes a model for optimizing which customers to offer a free sample of a product to, but the model only considers the impact on the customer's immediate friends, assumes the relevant probabilities are the same for all customers, and is only applied to a made-up network with seven nodes.

7. FUTURE WORK

We have developed models for viral marketing on social networks mined from real-world data. There are many directions in which these models, or their use, could be extended. In this section, we describe some of the main ones.

In this paper, we mined a network from a single source. In general, multiple sources of relevant information will be available; the ReferralWeb [14] project exemplified their use. Methods for combining diverse information into a sound representation of the underlying influence patterns are thus an important area for research.

Here, we considered only constant $r(z)$. In preliminary experiments, a decreasing $r(z)$ caused the algorithm to somewhat overestimate the lift in profit that would result from a particular marketing plan, therefore likely leading to a sub-optimal marketing plan (though it still outperformed direct marketing). In future work, we would like to investigate methods for handling variable $r(z)$, which may involve, for instance, a correction factor based on the expected number of customers that will be marketed to.

We have introduced methods for developing a marketing plan when the structure of the network is unknown or only partially known, but there are still many directions in which the methods could be extended. In particular, we would like to explore the effect of having a biased network sample on the resulting viral marketing plan. Knowing how the sample is biased should lead to better marketing plans. Also, with more information it may be possible to make more intelligent selections about which users to query. All information known about a user (e.g., demographic characteristics, past purchasing behavior, and partial knowledge about "trusts/trusted-by" relations) could be used to estimate the value of querying him. We would like to further develop the application of the theory of value of information [8] to optimizing the tradeoff between the cost and expected benefits of acquiring knowledge about the network.

This paper considered making marketing decisions at a specific point in time. A more sophisticated alternative would be to plan a marketing strategy by explicitly simulating the sequential adoption of a product by customers given different interventions at different times, and adapting the strategy as new data on customer response arrives. A further time-dependent aspect of the problem is that social networks are not static; they evolve, and particularly on the Internet can do so quite rapidly. Some of the largest opportunities may lie in modeling and taking advantage of this evolution. If the network evolution is understood, it may be possible to affect the structure itself, driving the network toward one which has a higher profit potential.

We would also like to investigate further the algorithmic similarities between viral marketing and web page ranking algorithms such as PageRank[21] and HITS[15]. Applying the techniques and lessons learned in viral marketing to the web domain, or vice versa, could result in new insights into the problems found in each. For instance, recent work on mining significant Web subgraphs such as bipartite cores, cliques and web rings (e.g., [17]) may be applicable to viral marketing. Their techniques could possibly be used to study network sub-structures and identify those with the highest profit potential.

8. CONCLUSION

This paper uses data mining to improve viral marketing. We apply our techniques to data mined from a real-world knowledge-sharing site, and show that they scale efficiently to networks of hundreds of millions of customers. We extend our techniques to handle continuously variable marketing actions and partial network knowledge. Our results show the promise of our approach.

9. ACKNOWLEDGEMENTS

This research was partly funded by NSF CAREER and IBM Faculty awards to the second author.

10. REFERENCES

[1] A. L. Barabási, R. Albert, and H. Jong. Scale-free characteristics of random networks: The topology of the World Wide Web. *Physica A*, 281:69-77, 2000.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, 1998. Elsevier.

[3] D. M. Chickering and D. Heckerman. A decision theoretic approach to targeted advertising. In *Proceedings of the Sixteenth Annual Conference on Uncertainty in Artificial Intelligence*, Stanford, CA, 2000. Morgan Kaufmann.

[4] P. Domingos and M. Pazanni. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103-130, 1997.

[5] P. Domingos and M. Richardson. Mining the Network Value of Customers. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*, pages 57-66, San Francisco, CA, 2001. ACM Press.

[6] M. Frauenfelder. Revenge of the know-it-alls: Inside the Web's free-advice revolution. *Wired* 8(7):144-158, 2000.

[7] K. Gelbrich and R. Nakhaeizadeh. Value Miner: A data mining environment for the calculation of the customer lifetime value with application to the automotive industry. In *Proceedings of the Eleventh European Conference on Machine Learning*, pages 154-161, Barcelona, Spain, 2000. Springer.

[8] R. A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, SSC-2:22-26. 1966

[9] A. M. Hughes. *The Complete Database Marketer: Second-Generation Strategies and Techniques for Tapping the Power of you Customer Database*. Irwin, Chicago, IL, 1996.

[10] D. Iacobucci, editor. *Networks in Marketing*. Sage, Thousand Oaks, CA, 1996.

[11] C. L. Isbell, Jr., M. Kearns, D. Korman, S. Singh, and P. Stone. Cobot in LambdaMOO: A social statistics agent. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 36-41, Austin, TX, 2000. AAAI Press.

[12] D. R. Jackson. Strategic application of customer lifetime value in direct marketing. *Journal of Targeting, Measurement and Analysis for Marketing*, 1:9-17, 1994.

[13] S. Jurvetson. What exactly is viral marketing? *Red Herring*, 78:110-112, 2000.

[14] H. Kautz, B. Selman, and M. Shah. ReferralWeb: Combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63-66, 1997.

[15] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668-677, Baltimore, MD, 1998. ACM Press.

[16] D. Krackhardt. Structural leverage in marketing. In D. Iacobucci, editor, *Networks in Marketing*, pages 50-59. Sage, Thousand Oaks, CA, 1996.

[17] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the Web. In *Proceedings of the Twenty-Fifth International Conference on Very Large Databases*, pages 639-650, Edinburgh, Scotland, 1999. Morgan Kaufmann.

[18] C. X. Ling and C. Li. Data mining for direct marketing: Problems and solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 73-79, New York, NY, 1998. AAAI Press.

[19] D. R. Mani, J. Drew, A. Betz, and P. Datta. Statistics and data mining techniques for lifetime value modeling. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 94-103, New York, NY, 1999. ACM Press.

[20] S. Milgram. The small world problem. *Psychology Today*, 2:60-67, 1967.

[21] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report, Stanford University, Stanford, CA. 1998.

[22] G. Piatetsky-Shapiro and B. Masand. Estimating campaign benefits and modeling lift. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 185-193, San Diego, CA, 1999. ACM Press.

[23] M. F. Schwartz and D. C. M. Wood. Discovering shared interests using graph analysis. *Communications of the ACM*, 36(8):78-80, 1993.

[24] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, UK, 1994.

[25] G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Boston, MA, 1949.

11. APPENDIX

In this appendix, we give a proof for Equation 7:

$$\Delta_i(\mathbf{Y}) = \sum_{j=1}^{n} w_{ji}\Delta_j(\mathbf{Y})$$

As this is an iterative equation, we identify which iteration we are on by a super-script. Let

$$\Delta_i^n = \Delta_i(\mathbf{Y})^n \text{ and } P_i^n = P(X_i \mid \mathbf{Y}, \mathbf{M})^n$$

be the n^{th} estimate of customer *i*'s network effect and probability of purchasing, respectively, and let

$$P_i = P_i^0 = P_0(X_i \mid \mathbf{Y}, M_i)$$

since on the 0^{th} iteration no network effect is taken into account. For notational convenience, we also define

$$w'_{km} = (1-\beta_k)w_{km}$$

The iterative update from Equation 4 is:

$$P_k^n = \beta_k P_0(X_k \mid \mathbf{Y}, M_k) + \sum_m w'_{km} P_m^{n-1}$$

Thus,

$$\frac{\partial P_k^n}{\partial P_i} = \sum_m w'_{km} \frac{\partial P_m^{n-1}}{\partial P_i}$$

Note that

$$\frac{\partial P_k^0}{\partial P_i} = \frac{\partial P_k}{\partial P_i} = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{if } k \neq i \end{cases}$$

Also note, from Equation 6, we have

$$\Delta_i^n = \sum_k \frac{\partial P_k^n}{\partial P_i}$$

and also

$$\Delta_i^0 = \sum_k \frac{\partial P_k^0}{\partial P_i} = 1$$

We first will prove by induction that

$$\frac{\partial P_k^n}{\partial P_i} = \sum_{a_1} \sum_{a_2} \cdots \sum_{a_{n-1}} w'_{a_1 i} w'_{a_2 a_1} \cdots w'_{k a_{n-1}} \text{ for } n \geq 2 \qquad (11)$$

We first show this is true for the case where $n = 2$:

$$\begin{aligned} \frac{\partial P_k^2}{\partial P_i} &= \sum_{a_1} w'_{k a_1} \frac{\partial P_{a_1}^1}{\partial P_i} \\ &= \sum_{a_1} w'_{k a_1} \sum_m w'_{a_1 m} \frac{\partial P_m^0}{\partial P_i} \\ &= \sum_{a_1} w'_{k a_1} w'_{a_1 i} \end{aligned}$$

We now prove Equation 11 is true for n if we assume it is true for n-1:

$$\begin{aligned} \frac{\partial P_k^n}{\partial P_i} &= \sum_m w'_{km} \frac{\partial P_m^{n-1}}{\partial P_i} \\ &= \sum_m w'_{km} \sum_{a_1} \sum_{a_2} \cdots \sum_{a_{n-2}} w'_{a_1 i} w'_{a_2 a_1} \cdots w'_{m a_{n-2}} \\ &= \sum_{a_{n-1}} w'_{k a_{n-1}} \sum_{a_1} \sum_{a_2} \cdots \sum_{a_{n-2}} w'_{a_1 i} w'_{a_2 a_1} \cdots w'_{a_{n-1} a_{n-2}} \\ &= \sum_{a_1} \sum_{a_2} \cdots \sum_{a_{n-1}} w'_{a_1 i} w'_{a_2 a_1} \cdots w'_{k a_{n-1}} \end{aligned}$$

This completes the proof of Equation 11.

We will now prove by induction that

$$\Delta_i^n = \sum_k w'_{ki} \Delta_k^{n-1} \quad \text{for } n \geq 1 \qquad (12)$$

We first prove that the induction hypothesis is true for the case where $n = 1$:

$$\begin{aligned} \Delta_i^1 &= \sum_k \frac{\partial P_k^1}{\partial P_i} \\ &= \sum_k \sum_m w'_{km} \frac{\partial P_m^{n-1}}{\partial P_i} \\ &= \sum_k w'_{ki} \\ &= \sum_k w'_{ki} \Delta_k^0 \end{aligned}$$

We now prove Equation 12 is true for n if we assume it is true for n-1.

$$\Delta_i^{n-1} = \sum_k w'_{ki} \Delta_k^{n-2}$$

By "unrolling" the recursion, we obtain

$$\Delta_i^{n-1} = \sum_{a_1} \sum_{a_2} \cdots \sum_{a_{n-1}} w'_{a_1 i} w'_{a_2 a_1} \cdots w'_{a_{n-1} a_{n-2}}$$

From the definition of Δ_i^n, and from Equation 11:

$$\Delta_i^n = \sum_k \frac{\partial P_k^n}{\partial P_i} = \sum_k \sum_{a_1} \sum_{a_2} \cdots \sum_{a_{n-1}} w'_{a_1 i} w'_{a_2 a_1} \cdots w'_{k a_{n-1}}$$

renaming a_1 as k, a_j as a_{j-1} for $2 \leq j \leq n\text{-}1$, and k as a_{n-1}, we obtain:

$$\begin{aligned} \Delta_i^n &= \sum_{a_{n-1}} \sum_k \sum_{a_1} \cdots \sum_{a_{n-2}} w'_{ki} w'_{a_1 k} \cdots w'_{a_{n-1} a_{n-2}} \\ &= \sum_k w'_{ki} \left(\sum_{a_1} \cdots \sum_{a_{n-1}} w'_{a_1 k} \cdots w'_{a_{n-1} a_{n-2}} \right) \\ &= \sum_k w'_{ki} \Delta_k^{n-1} \end{aligned}$$

We have shown that $\Delta_i^n = \sum_k w'_{ki} \Delta_k^{n-1}$. If this recursion is iterated until it reaches a fixed point, the resulting values for $\Delta^n{}_i$ satisfy

$$\Delta_i(\mathbf{Y}) = \sum_{j=1}^{n} w_{ji} \Delta_j(\mathbf{Y})$$

This completes the proof of Equation 7.