Mining the Network Value of Customers

Pedro Domingos
Dept. of Computer Science & Engineering
University of Washington
Box 352350
Seattle, WA 98195-2350, U.S.A.
pedrod@cs.washington.edu

Matt Richardson
Dept. of Computer Science & Engineering
University of Washington
Box 352350
Seattle, WA 98195-2350, U.S.A.
mattr@cs.washington.edu

ABSTRACT

One of the major applications of data mining is in helping companies determine which potential customers to market to. If the expected profit from a customer is greater than the cost of marketing to her, the marketing action for that customer is executed. So far, work in this area has considered only the intrinsic value of the customer (i.e, the expected profit from sales to her). We propose to model also the customer's *network value*: the expected profit from sales to other customers she may influence to buy, the customers those may influence, and so on recursively. Instead of viewing a market as a set of independent entities, we view it as a social network and model it as a Markov random field. We show the advantages of this approach using a social network mined from a collaborative filtering database. Marketing that exploits the network value of customers—also known as viral marketing—can be extremely effective, but is still a black art. Our work can be viewed as a step towards providing a more solid foundation for it, taking advantage of the availability of large relevant databases.

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

General Terms

Markov random fields, dependency networks, direct marketing, viral marketing, social networks, collaborative filtering

1. INTRODUCTION

Direct marketing is one of the major applications of KDD. In contrast to mass marketing, where a product is promoted indiscriminately to all potential customers, direct marketing attempts to first select the customers likely to be profitable, and market only to those [19]. Data mining plays a key role in this process, by allowing the construction of models that predict a customer's response given her past buying behavior and any available demographic information [29]. When successful, this approach can significantly increase profits [34]. One basic limitation of it is that it treats each customer as making a buying decision independently of all other customers. In reality, a person's decision to buy a product is often strongly influenced by her friends, acquaintances, business partners, etc. Marketing based on such word-of-mouth networks can be much more cost-effective than the more conventional variety, because it leverages the customers themselves to carry out most of the promotional effort. A classic example of this is the Hotmail free email service, which grew from zero to 12 million users in 18 months on a minuscule advertising budget, thanks to the inclusion of a promotional message with the service's URL in every email sent using it [23]. Competitors using conventional marketing fared far less well. This type of marketing, dubbed *viral marketing* because of its similarity to the spread of an epidemic, is now used by a growing number of companies, particularly in the Internet sector. More generally, network effects (known in the economics literature as network externalities) are of critical importance in many industries, including notably those associated with information goods (e.g., software, media, telecommunications, etc.) [38]. A technically inferior product can often prevail in the marketplace if it better leverages the network of users (for example, VHS prevailed over Beta in the VCR market).

Ignoring network effects when deciding which customers to market to can lead to severely suboptimal decisions. In addition to the intrinsic value that derives from the purchases she will make, a customer effectively has a *network value* that derives from her influence on other customers. A customer whose intrinsic value is lower than the cost of marketing may in fact be worth marketing to when her network value is considered. Conversely, marketing to a profitable customer may be redundant if network effects already make her very likely to buy. However, quantifying the network value of a customer is at first sight an extremely difficult undertaking, and to our knowledge has never been attempted. A customer's network value depends not only on herself, but potentially on the configuration and state of the entire network. As a result, marketing in the presence of strong network effects is often a hit-and-miss affair. Many startup companies invest heavily in customer acquisition, on the basis that this is necessary to "seed" the network, only to face bankruptcy when the desired network effects fail to materialize. On the other hand, some companies (like Hotmail and the ICQ instant messenger service) are much more successful than expected. A sounder basis for action in network-driven markets would thus have the potential to greatly reduce the risk of companies operating in them.

We believe that, for many of these markets, the growth of the Internet has led to the availability of a wealth of data from which the necessary network information can be mined. In this paper we propose a general framework for doing this, and for using the results to optimize the choice of which customers to market to, as well as estimating what customer acquisition cost is justified for each. Our solution is based on modeling social networks as Markov random fields, where each customer's probability of buying is a function of both the intrinsic desirability of the product for the customer and the influence of other customers. We then focus on collaborative filtering databases as an instance of a data source for mining networks of influence from. We apply our framework to the domain of marketing motion pictures, using the publicly-available EachMovie database of 2.8 million movie ratings, and demonstrate its advantages relative to traditional direct marketing. The paper concludes with a discussion of related work and a summary of contributions and future research directions.

2. MODELING MARKETS AS SOCIAL NETWORKS

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. In what follows we will often slightly abuse language by taking X_i to "be" the ith customer. Let the *neighbors* of X_i be the customers which directly influence X_i: $\mathbf{N}_i = \{X_{i,1}, \ldots, X_{i,n_i}\} \subseteq \mathbf{X} - \{X_i\}$, where $\mathbf{X} = \{X_1, \ldots, X_n\}$. In other words, X_i is independent of $\mathbf{X} - \mathbf{N}_i - \{X_i\}$ given $\mathbf{N}_i$. Let $\mathbf{X}^k$ ($\mathbf{X}^u$) be the customers whose value (i.e., whether they have bought the product) is known (unknown), and let $\mathbf{N}_i^u = \mathbf{N}_i \cap \mathbf{X}^u$. Assume the product is described by a set of attributes $\mathbf{Y} = \{Y_1, \ldots, Y_m\}$. Let M_i be a variable representing 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 offered a given discount, and $M_i = 0$ otherwise. Alternately, 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 $\mathbf{M} = \{M_1, \ldots, M_n\}$. Then, for all $X_i \notin \mathbf{X}^k$,

$$\begin{aligned} & P(X_i|\mathbf{X}^k, \mathbf{Y}, \mathbf{M}) \\ &= \sum_{C(\mathbf{N}_i^u)} P(X_i, \mathbf{N}_i^u|\mathbf{X}^k, \mathbf{Y}, \mathbf{M}) \\ &= \sum_{C(\mathbf{N}_i^u)} P(X_i|\mathbf{N}_i^u, \mathbf{X}^k, \mathbf{Y}, \mathbf{M}) P(\mathbf{N}_i^u|\mathbf{X}^k, \mathbf{Y}, \mathbf{M}) \\ &= \sum_{C(\mathbf{N}_i^u)} P(X_i|\mathbf{N}_i, \mathbf{Y}, \mathbf{M}) P(\mathbf{N}_i^u|\mathbf{X}^k, \mathbf{Y}, \mathbf{M}) \qquad (1) \end{aligned}$$

where $C(\mathbf{N}_i^u)$ is the set of all possible configurations of the unknown neighbors of X_i (i.e., the set of all possible $2^{|\mathbf{N}_i^u|}$ assignments of 0 and 1 to them). Following Pelkowitz [33], we approximate $P(\mathbf{N}_i^u|\mathbf{X}^k, \mathbf{Y}, \mathbf{M})$ by its maximum entropy estimate given the marginals $P(X_j|\mathbf{X}^k, \mathbf{Y}, \mathbf{M})$, for $X_j \in \mathbf{N}_i^u$. This yields[1]

$$\begin{aligned} & P(X_i|\mathbf{X}^k, \mathbf{Y}, \mathbf{M}) \\ &= \sum_{C(\mathbf{N}_i^u)} P(X_i|\mathbf{N}_i, \mathbf{Y}, \mathbf{M}) \prod_{X_j \in \mathbf{N}_i^u} P(X_j|\mathbf{X}^k, \mathbf{Y}, \mathbf{M}) \end{aligned} \qquad (2)$$

The set of variables $\mathbf{X}^u$, with joint probability conditioned on $\mathbf{X}^k$, $\mathbf{Y}$ and $\mathbf{M}$ described by Equation 2, is an instance of a *Markov random field* [2, 25, 7]. Because Equation 2 expresses the probabilities $P(X_i|\mathbf{X}^k, \mathbf{Y}, \mathbf{M})$ as a function of themselves, it can be applied iteratively to find them, starting from a suitable initial assignment. This procedure is known as relaxation labeling, and is guaranteed to converge to locally consistent values as long as the initial assignment is sufficiently close to them [33]. A natural choice for initialization is to use the network-less probabilities $P(X_i|\mathbf{Y}, \mathbf{M})$. Notice that the number of terms in Equation 2 is exponential in the number of unknown neighbors of X_i. If this number is small (e.g., 5), this should not be a problem; otherwise, an approximate solution is necessary. A standard method for this purpose is Gibbs sampling [16]. An alternative based on an efficient k-shortest-path algorithm is proposed in Chakrabarti et al. [6].

Given $\mathbf{N}_i$ and $\mathbf{Y}$, X_i should be independent of the marketing actions for other customers. Assuming a naive Bayes model for X_i as a function of $\mathbf{N}_i$, $Y_1, \ldots, Y_m$ and M_i [11],

$$\begin{aligned} & P(X_i|\mathbf{N}_i, \mathbf{Y}, \mathbf{M}) \\ &= P(X_i|\mathbf{N}_i, \mathbf{Y}, M_i) \\ &= \frac{P(X_i)P(\mathbf{N}_i, \mathbf{Y}, M_i|X_i)}{P(\mathbf{N}_i, \mathbf{Y}, M_i)} \\ &= \frac{P(X_i)P(\mathbf{N}_i|X_i)P(M_i|X_i)}{P(\mathbf{N}_i, \mathbf{Y}, \mathbf{M})} \prod_{k=1}^{m} P(Y_k|X_i) \\ &= \frac{P(X_i|\mathbf{N}_i)P(M_i|X_i)}{P(\mathbf{Y}, M_i|\mathbf{N}_i)} \prod_{k=1}^{m} P(Y_k|X_i) \qquad (3) \end{aligned}$$

where $P(\mathbf{Y}, M_i|\mathbf{N}_i) = P(\mathbf{Y}, M_i|X_i = 1)P(X_i = 1|\mathbf{N}_i) + P(\mathbf{Y}, M_i|X_i = 0)P(X_i = 0|\mathbf{N}_i)$. The corresponding network-less probabilities are $P(X_i|\mathbf{Y}, \mathbf{M}) = P(X_i)P(M_i|X_i) \prod_{k=1}^{m} P(Y_k|X_i)/P(\mathbf{Y}, M_i)$. Given Equation 3, in order to compute Equation 2 we need to know only the following probabilities, since all terms reduce to them: $P(X_i|\mathbf{N}_i)$, $P(X_i)$, $P(M_i|X_i)$, and $P(Y_k|X_i)$ for all k. With the exception of $P(X_i|\mathbf{N}_i)$, all of these are easily obtained in one pass through the data by counting (assuming the Y_k are discrete or have been pre-discretized; otherwise a univariate model can be fit for each numeric Y_k). The form of $P(X_i|\mathbf{N}_i)$ depends on the mechanism by which customers influence each other, and will vary from application to application. In the next section we focus on the particular case where $\mathbf{X}$ is the set of users of a collaborative filtering system.

For simplicity, assume that $\mathbf{M}$ 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

[1] The same result can be obtained by assuming that the X_j are independent given $\mathbf{X}^k$, $\mathbf{Y}$ and $\mathbf{M}$.

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 $\mathbf{M}$ unchanged, and similarly for $f_i^0(\mathbf{M})$. The *expected lift in profit* from marketing to customer i in isolation (i.e., ignoring her effect on other customers) is then [8]

$$\begin{aligned} &ELP_i(\mathbf{X}^k, \mathbf{Y}, \mathbf{M}) = \\ &\quad r_1 P(X_i = 1 | \mathbf{X}^k, \mathbf{Y}, f_i^1(\mathbf{M})) \\ &\quad - r_0 P(X_i = 1 | \mathbf{X}^k, \mathbf{Y}, f_i^0(\mathbf{M})) - c \end{aligned} \tag{4}$$

Let $\mathbf{M}_0$ be the null vector (all zeros). The global lift in profit that results from a particular choice $\mathbf{M}$ of customers to market to is then

$$\begin{aligned} &ELP(\mathbf{X}^k, \mathbf{Y}, \mathbf{M}) = \\ &\quad \sum_{i=1}^{n} r_i P(X_i = 1 | \mathbf{X}^k, \mathbf{Y}, \mathbf{M}) \\ &\quad - r_0 \sum_{i=1}^{n} P(X_i = 1 | \mathbf{X}^k, \mathbf{Y}, \mathbf{M}_0) - |\mathbf{M}| c \end{aligned} \tag{5}$$

where $r_i = r_1$ if $M_i = 1$, $r_i = r_0$ if $M_i = 0$, and $|\mathbf{M}|$ is the number of 1's in $\mathbf{M}$. Our goal is to find the assignment of values to $\mathbf{M}$ that maximizes ELP. In general, finding the optimal $\mathbf{M}$ requires trying all possible combinations of assignments to its components. Because this is intractable, we propose using one of the following approximate procedures instead:

Single pass For each i, set $M_i = 1$ if $ELP(\mathbf{X}^k, \mathbf{Y}, f_i^1(\mathbf{M}_0)) > 0$, and set $M_i = 0$ otherwise.

Greedy search Set $\mathbf{M} = \mathbf{M}_0$. Loop through the M_i's, setting each M_i to 1 if $ELP(\mathbf{X}^k, \mathbf{Y}, f_i^1(\mathbf{M})) > ELP(\mathbf{X}^k, \mathbf{Y}, \mathbf{M})$. Continue looping until there are no changes in a complete scan of the M_i's. The key difference between this method and the previous one is that here later changes to the M_i's are evaluated with earlier changes to the M_i's already in place, while in the previous method all changes are evaluated with respect to $\mathbf{M}_0$.

Hill-climbing search Set $\mathbf{M} = \mathbf{M}_0$. Set $M_{i_1} = 1$, where $i_1 = \text{argmax}_i\{ELP(\mathbf{X}^k, \mathbf{Y}, f_i^1(\mathbf{M}))\}$. Now set $M_{i_2} = 1$, where $i_2 = \text{argmax}_i\{ELP(\mathbf{X}^k, \mathbf{Y}, f_i^1(f_{i_1}^1(\mathbf{M})))\}$. Repeat until there is no i for which setting $M_i = 1$ increases ELP.

Each method is computationally more expensive than the previous one, but potentially leads to a better solution for $\mathbf{M}$ (i.e., produces a higher ELP).

The intrinsic value of a customer is given by Equation 4. The total value of a customer (intrinsic plus network) is the ELP obtained by marketing to her: $ELP(\mathbf{X}^k, \mathbf{Y}, f_i^1(\mathbf{M})) - ELP(\mathbf{X}^k, \mathbf{Y}, f_i^0(\mathbf{M}))$. The customer's network value is the difference between her total and intrinsic values. Notice that, in general, this value will depend on which other customers are being marketed to, and which others have already bought the product.

Suppose now that M_i is a continuous variable, that we can choose to incur different marketing costs for different customers, and that there is a known relationship between c_i and $P(X_i|M_i)$. In other words, suppose that we can increase a customer's probability of buying by increasing the amount spent in marketing to her, and that we can estimate how much needs to be spent to produce a given increase in buying probability. The optimal customer acquisition cost for customer i is then the value of c_i that maximizes her total value $ELP(\mathbf{X}^k, \mathbf{Y}, f_i^1(\mathbf{M})) - ELP(\mathbf{X}^k, \mathbf{Y}, f_i^0(\mathbf{M}))$, with $|\mathbf{M}|c$ replaced by $\sum_{i=1}^{n} c_i$ in Equation 5.

3. MINING SOCIAL NETWORKS FROM COLLABORATIVE FILTERING DATABASES

Arguably, a decade ago it would have been difficult to make practical use of a model like Equation 2, because of the lack of data to estimate the influence probabilities $P(X_i|\mathbf{N_i})$. Fortunately, the explosion of the Internet has drastically changed this. People influence each other online (and leave a record of it) through postings and responses to newsgroups, review and knowledge-sharing sites like epinions.com, chat rooms and IRC, online game playing and MUDs, peer-to-peer networks, email, interlinking of Web pages, etc. In general, any form of online community is a potentially rich source of data for mining social networks from. (Of course, mining these sources is subject to the usual privacy concerns; but many sources are public information.) In this paper we will concentrate on a particularly simple and potentially very effective data source: the collaborative filtering systems widely used by e-commerce sites (e.g., amazon.com) to recommend products to consumers.

In a collaborative filtering system, users rate a set of items (e.g., movies, books, newsgroup postings, Web pages), and these ratings are then used to recommend other items the user might be interested in. The ratings may be implicit (e.g., the user did or did not buy the book) or explicit (e.g., the user gives a rating of zero to five stars to the book, depending on how much she liked it). Many algorithms have been proposed for choosing which items to recommend given the incomplete matrix of ratings (see, for example, Breese et al. [3]). The most widely used method, and the one that we will assume here, is the one proposed in GroupLens, the project that originally introduced quantitative collaborative filtering [35]. The basic idea in this method is to predict a user's rating of an item as a weighted average of the ratings given by similar users, and then recommend items with high predicted ratings. The similarity of a pair of users (i, j) is measured using the Pearson correlation coefficient:

$$W_{ij} = \frac{\sum_k (R_{ik} - \overline{R}_i)(R_{jk} - \overline{R}_j)}{\sqrt{\sum_k (R_{ik} - \overline{R}_i)^2 \sum_k (R_{jk} - \overline{R}_j)^2}} \tag{6}$$

where R_{ik} is user i's rating of item k, $\overline{R}_i$ is the mean of user i's ratings, likewise for j, and the summations and means are computed over the items k that both i and j have rated. Given an item k that user i has not rated, her rating of it is then predicted as

$$\hat{R}_{ik} = \overline{R}_i + \rho \sum_{X_j \in \mathbf{N}_i} W_{ji}(R_{jk} - \overline{R}_j) \tag{7}$$

where $\rho = 1/\sum_{X_j \in \mathbf{N}_i} |W_{ij}|$ is a normalization factor, and $\mathbf{N}_i$ is the set of n_i users most similar to i according to Equation 6 (her neighbors). In the limit, $\mathbf{N}_i$ can be the entire database of users, but for reasons of noise robustness and computational efficiency it is usually much smaller (e.g.,

$n_i = 5$). For neighbors that did not rate the item, R_{jk} is set to $\overline{R}_j$.

The key advantage of a collaborative filtering database as a source for mining a social network for viral marketing is that the mechanism by which individuals influence each other is known and well understood: it is the collaborative filtering algorithm itself. User i influences user j when j sees a recommendation that is partly the result of i's rating. Assuming i and j do not know each other in real life (which, given that they can be anywhere in the world, is likely to be true), there is no other way they can substantially influence each other. Obviously, a user is subject to many influences besides that of the collaborative filtering system (including the influence of people not on the system), but the uncertainty caused by those influences is encapsulated to a first degree of approximation in $P(X_i|\hat{R}_{ik})$, the probability that a user will purchase an item given the rating the system predicts for her. It is also reasonable to assume that an individual would not continue to use a collaborative filtering system if she did not find its recommendations useful, and therefore that there is a causal connection (rather than simply a correlation) between the recommendations received and the purchases made.

To extract a social network model from a collaborative filtering database, we view an item as a random sample from the space $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{Y}$ is a set of properties of the item (assumed available), and X_i represents whether or not user i rated the item. For simplicity, we assume that if a user rates an item then she bought it, and vice-versa; removing this assumption would be straightforward, given the relevant data. The prior $P(X_i)$ can then be estimated simply as the fraction of items rated by user i. The conditional probabilities $P(Y_k|X_i)$ can be obtained by counting the number of occurrences of each value of Y_k (assumed discrete or pre-discretized) with each value of X_i. Estimating $P(M_i|X_i)$ requires a data collection phase in which users to market to are selected at random and their responses are recorded (both when being marketed to and not). $P(M_i|X_i)$ can be estimated individually for each user, or (requiring far less data) as the same for all users, as done in Chickering and Heckerman [8]. If the necessary data is not available, we propose setting $P(M_i|X_i)$ using prior knowledge about the effectiveness of the type of marketing being considered, given any demographic information available about the users. (It is also advisable to test the sensitivity of the outcome to $P(M_i|X_i)$ by trying a range of values.)

The set of neighbors $\mathbf{N}_i$ for each i is the set of neighbors of the corresponding user in the collaborative filtering system. If the ratings are implicit (i.e., yes/no), a model for $P(X_i|\mathbf{N}_i)$ (e.g., a naive Bayes model, as we have assumed for $P(Y_k|X_i)$) can be fit directly to the observed $\mathbf{X}$ vectors. If explicit ratings are given (e.g., zero to five stars), then we know that X_i depends on $\mathbf{N}_i$ solely through $\hat{R}_i$, X_i's predicted rating according to Equation 7 (for readability, we will omit the item indexes k). In other words, X_i is conditionally independent of $\mathbf{N}_i$ given $\hat{R}_i$. If the neighbors' ratings are known, $\hat{R}_i$ is a deterministic function of $\mathbf{N}_i$ given by Equation 7, with $X_j \in \mathbf{N}_i$ determining whether the contribution of the jth neighbor is $R_j - \overline{R}_j$ or 0 (see discussion following Equation 7). If the ratings of some or all neighbors are unknown (i.e., the ratings that they would give if they were to rate the item), we can estimate them as their expected values given the item's attributes. In other words, the contribution of a neighbor with unknown rating will be $E[R_j|\mathbf{Y}] - \overline{R}_j$. $P(R_j|\mathbf{Y})$ can be estimated using a naive Bayes model (assuming R_j only takes on a small number of different values, which is usually the case). Let $\hat{R}_i(\mathbf{N}_i)$ be the value of $\hat{R}_i$ obtained in this way. Then, treating this as a deterministic value,

$$\begin{aligned} P(X_i|\mathbf{N}_i) &= \int_{R_{min}}^{R_{max}} P(X_i|\hat{R}_i, \mathbf{N}_i)\, dP(\hat{R}_i|\mathbf{N}_i) \\ &= P(X_i|\hat{R}_i(\mathbf{N}_i), \mathbf{N}_i) = P(X_i|\hat{R}_i(\mathbf{N}_i)) \qquad (8) \end{aligned}$$

All that remains is to estimate $P(X_i|\hat{R}_i)$. This can be viewed as a univariate regression problem, with $\hat{R}_i$ as the input and $P(X_i|\hat{R}_i)$ as the output. The most appropriate functional form for this regression will depend on the observed data. In the experiments described below, we used a piecewise-linear model for $P(X_i|\hat{R}_i)$, obtained by dividing $\hat{R}_i$'s range into bins, computing the mean $\hat{R}_i$ and $P(X_i|\hat{R}_i)$ for each bin, and then estimating $P(X_i|\hat{R}_i)$ for an arbitrary $\hat{R}_i$ by interpolating linearly between the two nearest means. Given a small number of bins, this approach can fit a wide variety of observations relatively well, with little danger of overfitting.

Notice that the technical definition of a Markov random field requires that the neighborhood relation be symmetric (i.e., if i is a neighbor of j, then j is also a neighbor of i), but in a collaborative filtering system this may not be the case. The probabilistic model obtained from it in the way described will then be an instance of a *dependency network*, a generalization of Markov random fields recently proposed by Heckerman et al. [17]. Heckerman et al. show that Gibbs sampling applied to such a network defines a joint distribution from which all probabilities of interest can be computed. While in our experimental studies Gibbs sampling and relaxation labeling produced very similar results, the formal derivation of the properties of dependency networks under relaxation labeling is a matter for future research.

4. EMPIRICAL STUDY

We have applied the methodology described in the previous sections to the problem of marketing motion pictures, using the EachMovie collaborative filtering database (*www.research.compaq.com/src/eachmovie/*). EachMovie contains 2.8 million ratings of 1628 movies by 72916 users, gathered between January 29, 1996 and September 15, 1997 by the eponymous recommendation site, which was run by the DEC (now Compaq) Systems Research Center. EachMovie is publicly available, and has become a standard database for evaluating collaborative filtering systems (e.g., Breese at al. [3]). Motion picture marketing is an interesting application for the techniques we propose because the success of a movie is known to be strongly driven by word of mouth [12].

EachMovie is composed of three databases: one containing the ratings, one containing demographic information about the users (which we did not use), and one containing information about the movies. The latter includes the movie's title, studio, theater and video status (old or current), theater and video release dates, and ten Boolean attributes describing the movie's genre (action, animation, art/foreign, classic, comedy, drama, family, horror, romance, and thriller; a movie can have more than one genre). The movie's URL in the Internet Movie Database (*www.imdb.-*

com) is also included. This could be used to augment the movie description with attributes extracted from the IMDB; we plan to do so in the future. The ratings database contains an entry for each movie that each user rated, on a scale of zero to five stars, and the time and date on which the rating was generated.

The collaborative filtering algorithm used in EachMovie has not been published, but we will assume that the algorithm described in the previous section is a reasonable approximation to it. This assumption is supported by the observation that, despite their variety in form, all the many collaborative filtering algorithms proposed attempt to capture essentially the same information (namely, correlations between users).

The meaning of the variables in the EachMovie domain is as follows: X_i is whether person i saw the movie being considered. $\mathbf{Y}$ contains the movie attributes. R_i is the rating (zero to five stars) given to the movie by person i. For simplicity, throughout this section we assume the $\hat{R}_i$'s are centered at zero (i.e., $\overline{R}_i$ has been subtracted from $\hat{R}_i$; see Equation 7).

4.1 The Model

We used $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_{10}\}$, the ten Boolean movie genre attributes. Thus $P(\mathbf{Y}|X_i)$ was in essence a model of a user's genre preferences, and during inference two movies with the same genre attributes were indistinguishable. The network consisted of all people who had rated at least ten movies, and whose ratings had non-zero standard deviation (otherwise they contained no useful information). Neighbor weights W_{ij} were determined using a modified Pearson correlation coefficient, which penalized the correlation by 0.05 for each movie less than ten that both X_i and X_j had rated. This correction is commonly used in collaborative filtering systems to avoid concluding that two users are very highly correlated simply because they rated very few movies in common, and by chance rated them similarly [18]. The neighbors of X_i were the X_j's for which W_{ji} was highest. With n_i=5, a number we believe provides a reasonable tradeoff between model accuracy and speed, the average W_{ji} of neighbors was 0.91. Repeating the experiments described below with $n_i = 10$ and $n_i = 20$ produced no significant change in model accuracy, and small improvements in profit. Interestingly, the network obtained in each case was completely connected (i.e., it contained no isolated subgraphs).

As discussed above, the calculation of $P(X_i|\mathbf{X}^k, \mathbf{Y}, \mathbf{M})$ requires estimating $P(X_i|\hat{R}_i)$, $P(X_i)$, $P(M_i|X_i)$, $P(Y_k|X_i)$, and $P(R_i|\mathbf{Y})$. $P(X_i)$ is simply the fraction of movies X_i rated. We used a naive Bayes model for $P(R_j|\mathbf{Y})$. $P(Y_k|X_i)$, $P(R_j|\mathbf{Y})$, and $P(X_i)$ were all smoothed using an m-estimate [5] with m=1 and the population average as the prior. We did not know the true values of $P(M_i|X_i)$. We expected marketing to have a larger effect on a customer who was already inclined to see the movie, and thus we set the probabilities $P(M_i|X_i)$ so as to obtain

$$P(X_i = 1|M_i = 1) = \min\{\alpha P(X_i = 1|M_i = 0), 1\} \qquad (9)$$

where $\alpha > 1$ is a parameter that we varied in the experiments described below.[2] As described in the previous section, $P(X_i|\hat{R}_i)$ was modeled using a piecewise linear function. We measured $P(X_i|\hat{R}_i)$ for each of nine bins, whose boundaries were -5.0, -2.0, -1.0, -0.5, -0.1, 0.1, 0.5, 1.0, 2.0, and 5.0. Note that while R_i must be between 0 and 5, $\hat{R}_i$ is a weighted sum of the neighbors' difference from their average, and thus may range from -5 to 5. We also had a zero-width bin located at $\hat{R}_i = 0$. Movies were seen with low probability (1–5%), and thus there was a high probability that a movie had not been rated by any of X_i's neighbors. In the absence of a rating, a neighbor's contribution to $\hat{R}_i$ was zero. 84% of the samples fell into this zero bin. Bin boundaries were chosen by examination of the distribution of data in the training set, shown in Figure 1. $\hat{R}_i$ was unlikely to deviate far from 0, for the reasons given above. We used narrow bins near $\hat{R}_i = 0$ to obtain higher accuracy in this area, which contained a majority of the data (96.4% of the data fell between -0.5 and 0.5). To combat data sparseness, both $P(X_i|\hat{R}_i)$ and the per-bin mean $\hat{R}_i$ were smoothed for each bin using an m-estimate with m=1 and the population average as the prior.

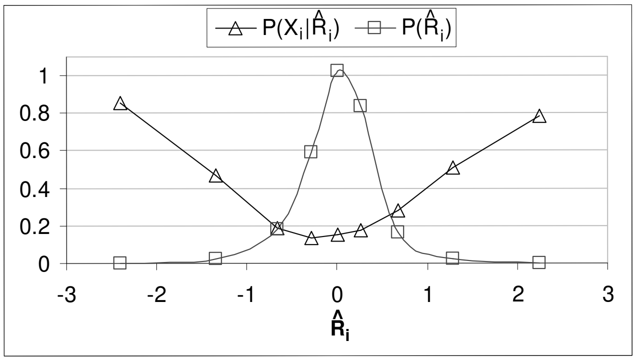

Figure 1: Empirical distribution of $\hat{R}_i$ and X_i given $\hat{R}_i$.

Initially, we expected $P(X_i|\hat{R}_i)$ to increase monotonically with $\hat{R}_i$. The actual shape, shown in Figure 1, shows increasing $P(X_i|\hat{R}_i)$ as $\hat{R}_i$ moves significantly away from 0 in either direction. This shape is due to a correlation between $|\hat{R}_i|$ and the popularity of a movie: for a popular movie, $\hat{R}_i$ is more likely to deviate further from zero and X_i is more likely to be 1. Note, however, that $P(X_i|\hat{R}_i)$ is indeed monotonically increasing in the $[-0.1, 0.1]$ interval, where the highest density of ratings is. Furthermore, $E[P(X_i|\hat{R}_i > 0)] = 0.203 > 0.176 = E[P(Xi|\hat{R}_i < 0)]$.

4.2 The Data

While the EachMovie database is large, it has problems which had to be overcome. The movies in the database which were in theaters before January 1996 were drawn from a long time period, and so tended to be very well known movies. Over 75% (2.2 million) of the ratings were on these movies. In general, the later a movie was released, the fewer ratings and thus the less information we had for it. We divided the database into a training set consisting of all ratings received through September 1, 1996, and a test set consisting of all movies released between September 1, 1996 and December 31, 1996, with the ratings received

[2] To fully specify $P(M_i|X_i)$ 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 9 and this constraint simultaneously.

for those movies any time between September 1, 1996 and the end of the database. Because there was such a large difference in average movie popularity between the early movies and the later ones, we further divided the training set into two subsets: S_{old}, containing movies released before January 1996 (1.06 million votes), and S_{recent}, containing movies released between January and September 1996 (90k votes). The average movie viewership of S_{old} was 5.6%, versus 1.4% for S_{recent}. Since 92% of the training data was in S_{old}, we could not afford to ignore it. However, in terms of the probability that someone rates a movie, the test period could be expected to be much more similar to S_{recent}. Thus, we trained using all training data, then rescaled $P(X_i)$ and $P(X_i|\hat{R}_i)$ using S_{recent}, and smoothed these values using an m-estimate with m=1 and the distribution on the full training set as the prior.

Many movies in the test set had very low probability (36% were viewed by 10 people or less, and 48% were viewed by 20 people or less, out of over 20748 people[3]). Since it is not possible to model such low probability events with any reliability, we removed all movies which were viewed by fewer than 1% of the people. This left 737,579 votes over 462 movies for training, and 3912 votes over 12 movies for testing. $P(Y|X_i)$, $P(R_i|\mathbf{Y})$, $P(X_i)$, and $P(X_i|\hat{R}_i)$ were learned using only these movies. However, because the EachMovie collaborative filtering system presumably used all movies, we used all movies when simulating it (i.e., when computing similarities (Equation 6), selecting neighbors, and predicting ratings (Equation 7)).

A majority of the people in the EachMovie database provided ratings once, and never returned. These people affected the predicted ratings $\hat{R}_i$ seen by users of EachMovie, but because they never returned to the system for queries, their movie viewing choices were not affected by their neighbors. We call these people *inactive*. A person was marked as inactive if there were more than τ days between her last rating and the end of the training period. In our tests, we used a τ of 60, which resulted in 11163 inactive people. Inactive people could be marketed to, since they were presumably still watching movies; they were just not reporting ratings to EachMovie. If an inactive person was marketed to, she was assumed to have no effect on the rest of the network.

4.3 Inference and Search

Inference was performed by relaxation labeling, as described in Section 2. This involved iteratively re-estimating probabilities until they all converged to within a threshold γ. (We used $\gamma = 10^{-5}$.) We maintained a queue of nodes whose probabilities needed to be re-estimated, which initially contained all nodes in the network. Each X_i was removed from the queue in turn, and its probability was re-estimated using Equation 2. If $P(X_i|\mathbf{X}^k, \mathbf{Y}, \mathbf{M})$ had changed by more than γ, all nodes that X_i was a neighbor of that were not already in the queue were added to it. Note that the probabilities of nodes corresponding to inactive people only needed to be computed once, since they are independent of the rest of the network.

The computation of Equation 2 can be sped up by noting that, after factoring, all terms involving the Y_k's are constant throughout a run, and so these terms and their combinations only need to be computed once. Further, since in a single search step only one M_i changes, most of the results of one step can be reused in the next, greatly speeding up the search process. With these optimizations, we were able to measure the effect of over 10,000 single changes in M per second, on a 1 GHz Pentium III machine. In preliminary experiments, we found relaxation labeling carried out this way to be several orders of magnitude faster than Gibbs sampling; we expect that it would also be much faster than the more efficient version of Gibbs sampling proposed in Heckerman et al. [17].[4] The relaxation labeling process typically converged quite quickly; few nodes ever required more than a few updates.

4.4 Model Accuracy

To test the accuracy of our model, we computed the estimated probability $P(X_i|\mathbf{X}^k, \mathbf{Y}, \mathbf{M})$ for each person X_i with $M = M_0$ and $\mathbf{X}^k = \emptyset$. We measured the correlation between this and the actual value of X_i in the test set, over all movies, over all people.[5] (Note that, since the comparison is with test set values, we did not expect to receive ratings from inactive people, and therefore $P(X_i|\mathbf{Y}) = 0$ for them.) The resulting correlation was 0.18. Although smaller than desirable, this correlation is remarkably high considering that the only input to the model was the movie's genre. We expect the correlation would increase if a more informative set of movie attributes $\mathbf{Y}$ were used.

4.5 Network Values

For the first movie in the test set ("Space Jam"), we measured the network value for all 9585 active people[6] in the following scenario (see Equations 4 and 9): $r_0 = 1$, $r_1 = 0.5$, $c = 0.1$, $\alpha = 1.5$, and $M = M_0$. Figure 2 shows the 500 highest network values (out of 9585) 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 20 for a given customer implies that by marketing to her we essentially get free marketing to an additional 20 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 good customer to market to is one who: (1) is likely to give the product a high rating, (2) has a strong weight in determining the rating prediction for many of her neighbors, (3) has many neighbors who are easily influenced by the rating prediction they receive, (4) will have a high probability of purchasing the product, and thus will be likely to actually submit a rating that will affect her neighbors, and finally (5) has many neighbors with the same four characteristics outlined above,

[3] This is the number of people left after we removed anyone who rated fewer than ten movies, rated movies only after September 1996, or gave the same rating to all movies.

[4] In our experiments, one Gibbs cycle of sampling all the nodes in the network took on the order of a fiftieth of a second. The total runtime would be this value multiplied by the number of sampling iterations desired and by the number of search steps.

[5] Simply measuring the predictive error rate would not be very useful, because a very low error rate could be obtained simply by predicting that no one sees the movie.

[6] Inactive people always have a network value of zero.

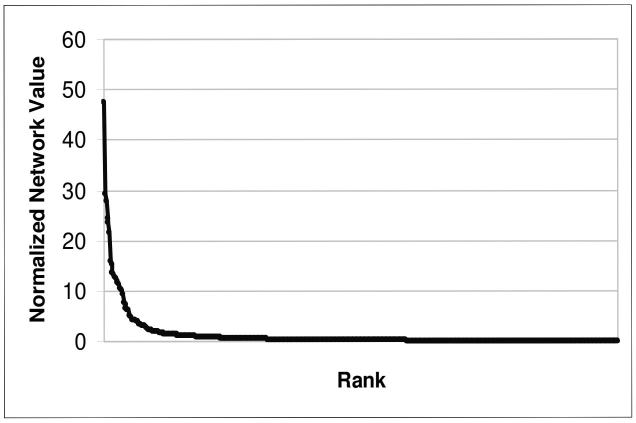

Figure 2: Typical distribution of network values.

and so on recursively. In the movie domain, these correspond to finding a person who (1) will enjoy the movie, (2) has many close friends, who are (3) easily swayed, (4) will very likely see the movie if marketed to, and (5) has friends whose friends also have these properties.

4.6 Marketing Experiments

We compared three marketing strategies: mass marketing, traditional direct marketing, and the network-based marketing method we proposed in Section 2. In mass marketing, all customers were marketed to ($M_i = 1$ for all i). In direct marketing, a customer X_i was marketed to ($M_i = 1$) if and only if $ELP_i(\mathbf{X}^k, \mathbf{Y}, \mathbf{M}_0) > 0$ (see Equation 4) ignoring network effects (i.e., using the network-less probabilities $P(X_i|\mathbf{Y}, \mathbf{M})$). For our approach, we compared the three approximation methods proposed in Section 2: single pass, greedy search and hill-climbing. Figure 3 compares these three search types and direct marketing on three different marketing scenarios. For all scenarios, $r_0 = 1$, which means profit numbers are in units of number of movies seen. In the *free movie* scenario $r_1 = 0$, and in the *discounted movie* scenario $r_1 = 0.5$. In both of these scenarios we assumed a cost of marketing of 10% of the revenue from a single sale: $c = 0.1$. In the *advertising* scenario no discount was offered ($r_1 = 1$), and a lower cost of marketing was assumed (corresponding, for example, to online marketing instead of physical mailings): $c = 0.02$. Notice that all the marketing actions considered were effectively in addition to the (presumably mass) marketing that was actually carried out for the movie. The average number of people who saw a movie given only this marketing (i.e., with $M = M_0$) was 311. The baseline profit would be obtained by subtracting from this the (unknown) original costs. The correct α for each marketing scenario was unknown, so we present the results for a range of values. We believe we have chosen plausible ranges, with a free movie providing more incentive than a discount, which in turn provides more incentive than simply advertising. $\mathbf{X}^k = \emptyset$ in all experiments.

In all scenarios, mass marketing resulted in negative profits. Not surprisingly, it fared particularly poorly in the free and discounted movie scenarios, producing profits which ranged from -2057 to -2712. In the advertising scenario, mass marketing resulted in profits ranging from -143 to -381 (depending on the choice of α). In the case of a free movie offer, the profit from direct marketing could not be positive, since without network effects we were guaranteed to lose money on anyone who saw a movie for free. Figure 3 shows that our method was able to find profitable marketing opportunities that were missed by direct marketing. For the discounted movie, direct marketing actually resulted in a loss of profit. A customer that looked profitable on her own may actually have had a negative overall value. This situation demonstrates that not only can ignoring network effects cause missed marketing opportunities, but it can also make an unprofitable marketing action look profitable. In the advertising scenario, for small α our method increased profits only slightly, while direct marketing again reduced them. Both methods improved with increasing α, but our method consistently outperformed direct marketing.

As can be seen in Figure 3, greedy search produced results that were quite close to those of hill climbing. The average difference between greedy and hill-climbing profits (as a percentage of the latter) in the three marketing scenarios was 9.6%, 4.0%, and 0.0% respectively. However, as seen in Figure 3, the runtimes differed significantly, with hill-climbing time ranging from 4.6 minutes to 42.1 minutes while greedy-search time ranged from 3.8 to 5.5 minutes. The contrast was even more pronounced in the advertising scenario, where the profits found by the two methods were nearly identical, but hill climbing took 14 hours to complete, compared to greedy search's 6.7 minutes. Single-pass was the fastest method and was comparable in speed to direct marketing, but led to significantly lower profits in the free and discounted movie scenarios.

The lift in profit was considerably higher if all users were assumed to be active. In the free movie scenario, the lift in profit using greedy search was 4.7 times greater than when the network had inactive nodes. In the discount and advertising scenarios the ratio was 4.1 and 1.8, respectively. This was attributable to the fact that the more inactive neighbors a node had, the less responsive it could be to the network. From the point of view of an e-merchant applying our approach, this suggests modifying the collaborative filtering system to only assign active users as neighbors.

5. RELATED WORK

Social networks have been an object of study for some time, but previous work within sociology and statistics has suffered from a lack of data and focused almost exclusively on very small networks, typically in the low tens of individuals [41]. Interestingly, the Google search engine [4] and Kleinberg's (1998) HITS algorithm for finding hubs and authorities on the Web are based on social network ideas. The success of these approaches, and the discovery of widespread network topologies with nontrivial properties [42], has led to a flurry of research on modeling the Web as a semi-random graph (e.g., Kumar et al. [28], Barabási et al. [1]). Some of this work might be applicable in our context.

In retrospect, the earliest sign of the potential of viral marketing was perhaps the classic paper by Milgram [31] estimating that every person in the world is only six edges away from every other, if an edge between i and j means "i knows j." Schwartz and Wood [37] mined social relationships from email logs. The ReferralWeb project mined a social network from a wide variety of publicly-available online information [24], and used it to help individuals find experts who could answer their questions. The COBOT project

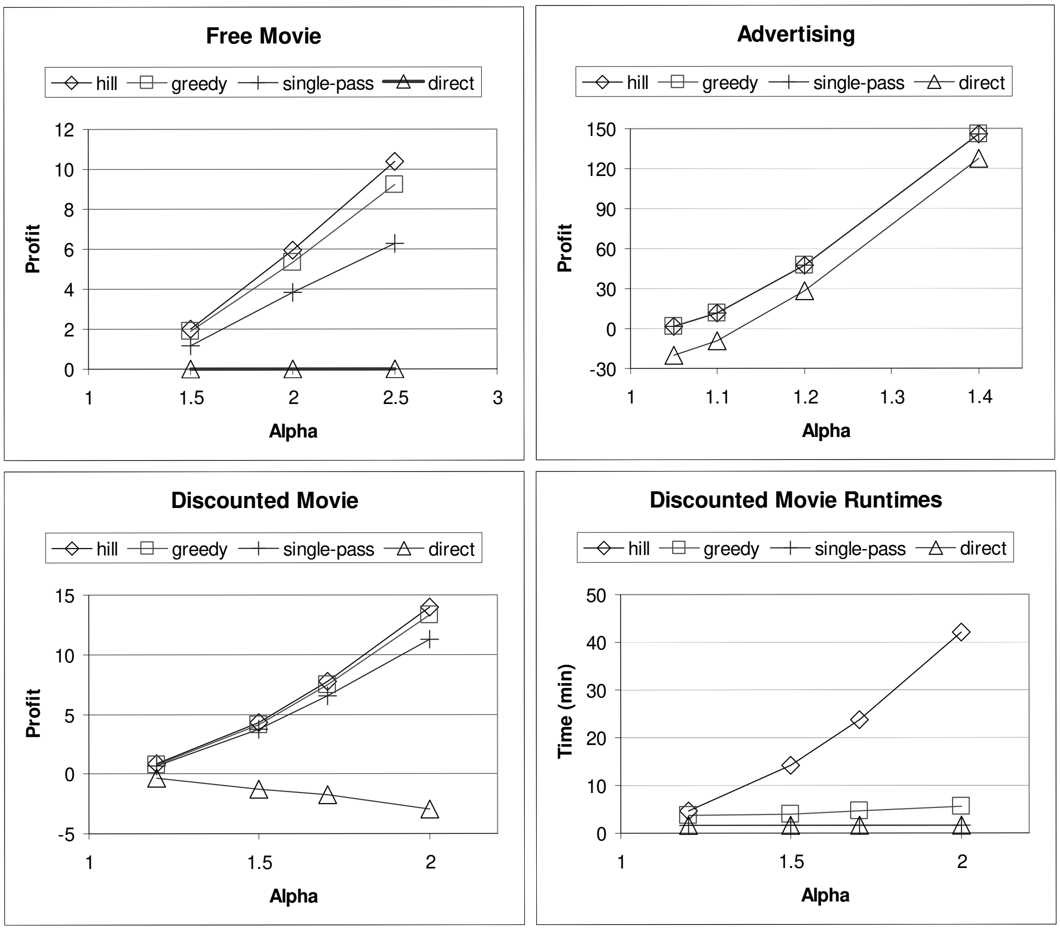

Figure 3: Profits and runtimes obtained using different marketing strategies.

gathered social statistics from participant interactions in the LambdaMoo MUD, but did not explicitly construct a social network from them [21]. A Markov random field formulation similar to Equation 2 was used by Chakrabarti et al. [6] for classification of Web pages, with pages corresponding to customers, hyperlinks between pages corresponding to influence between customers, and the bag of words in the page corresponding to properties of the product. Neville and Jensen [32] proposed a simple iterative algorithm for labeling nodes in social networks, based on the naive Bayes classifier. Cook and Holder [9] developed a system for mining graph-based data. Flake et al. [13] used graph algorithms to mine communities from the Web (defined as sets of sites that have more links to each other than to non-members).

Several researchers have studied the problem of estimating a customer's lifetime value from data [22]. This line of research generally focuses on variables like an individual's expected tenure as a customer [30] and future frequency of purchases [15]. Customer networks have received some attention in the marketing literature [20]. Most of these studies are purely qualitative; where data sets appear, they are very small, and used only for descriptive purposes. Krackhardt [27] proposes a very simple model for optimizing which customers to offer a free sample of a product to. The model only considers the impact on the customer's immediate friends, ignores the effect of product characteristics, assumes the relevant probabilities are the same for all customers, and is only applied to a made-up network with seven nodes.

Collaborative filtering systems proposed in the literature include GroupLens [35], PHOAKS [40], Siteseer [36], and others. A list of collaborative filtering systems, projects and related resources can be found at *www.sims.berkeley.-edu/resources/collab/*.

6. FUTURE WORK

The type of data mining proposed here opens up a rich field of directions for future research. In this section we briefly mention some of the main ones.

Although the network we have mined is large by the standards of previous research, much larger ones can be envisioned. Scaling up may be helped by developing search methods specific to the problem, to replace the generic ones we used here. Segmenting a network into more tractable parts with minimal loss of profit may also be important. Flake et al. [13] provide a potential way of doing this. A related approach would be to mine subnetworks with high profit potential embedded in larger ones. Recent work on mining significant Web subgraphs such as bipartite cores, cliques and webrings (e.g., [28]) provides a starting point. More generally, we would like to develop a characterization of network types with respect to the profit that can be obtained in them using an optimal marketing strategy. This would, for example, help a company to better gauge the profit potential of a market before entering (or attempting to create) it.

In this paper we mined a network from a single source (a collaborative filtering database). In general, multiple sources of relevant information will be available; the ReferralWeb project [24] exemplified their use. Methods for combining diverse information into a sound representation of the underlying influence patterns are thus an important area for research. In particular, detecting the presence of causal relations between individuals (as opposed to purely correlational ones) is key. While mining causal knowledge from observational databases is difficult, there has been much recent progress [10, 39].

We have also assumed so far that the relevant social network is completely known. In many (or most) applications this will not be the case. For example, a long-distance telephone company may know the pattern of telephone calls among its customers, but not among its non-customers. However, it may be able to make good use of connections between customers and non-customers, or to take advantage of information about former customers. A relevant question is thus: what can be inferred from a (possibly biased) sample of nodes and their neighbors in a network? At the extreme where no detailed information about individual interactions is available, our method could be extended to apply to networks where nodes are groups of similar or related customers, and edges correspond to influence among groups.

Another promising research direction is towards more detailed node models and multiple types of relations between nodes. A theoretical framework for this could be provided by the probabilistic relational models of Friedman et al. [14]. We would also like to extend our approach to consider multiple types of marketing actions and product-design decisions, and to multi-player markets (i.e., markets where the actions of competitors must also be taken into account, leading to a game-like search process).

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 objects; 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.

Once markets are viewed as social networks, the inadequacy of random sampling for pilot tests of products subject to strong network effects (e.g., smart cards, video on demand) becomes clear. Developing a better methodology for studies of this type could help avoid some expensive failures.

Many e-commerce sites already routinely use collaborative filtering. Given that the infrastructure for data gathering and for inexpensive execution of marketing actions (e.g., making specific offers to specific customers when they visit the site) is already in place, these would appear to be good candidates for a real-world test of our method. The greatest potential, however, may lie in knowledge-sharing and customer review sites like epinions.com, because the interaction between users is richer and stronger there. For example, it may be profitable for a company to offer its products at a loss to influential contributors to such sites. Our method is also potentially applicable beyond marketing, to promoting any type of social change for which the relevant network of influence can be mined from available data. The spread of online interaction creates unprecedented opportunities for the study of social information processing; our work is a step towards better exploiting this new wealth of information.

7. CONCLUSION

This paper proposed the application of data mining to viral marketing. Viewing customers as nodes in a social network, we modeled their influence on each other as a Markov random field. We developed methods for mining social network models from collaborative filtering databases, and for using these models to optimize marketing decisions. An empirical study using the EachMovie collaborative filtering database confirmed the promise of this approach.

8. 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] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.

[3] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, Madison, WI, 1998. Morgan Kaufmann.

[4] 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.

[5] B. Cestnik. Estimating probabilities: A crucial task in machine learning. In *Proceedings of the Ninth European Conference on Artificial Intelligence*, pages 147–149, Stockholm, Sweden, 1990. Pitman.

[6] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pages 307–318, Seattle, WA, 1998. ACM Press.

[7] R. Chellappa and A. K. Jain, editors. *Markov Random Fields: Theory and Application*. Academic Press, Boston, MA, 1993.

[8] 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.

[9] D. J. Cook and L. B. Holder. Graph-based data mining. *IEEE Intelligent Systems*, 15:32–41, 2000.

[10] G. F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1:203–224, 1997.

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

[12] R. Dye. The buzz on buzz. *Harvard Business Review*, 78(6):139–146, 2000.
[13] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of Web communities. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, 2000. ACM Press.
[14] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1300–1307, Stockholm, Sweden, 1999. Morgan Kaufmann.
[15] 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.
[16] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
[17] D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1:49–75, 2000.
[18] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 1999 Conference on Research and Development in Information Retrieval*, Berkeley, CA, 1999.
[19] A. M. Hughes. *The Complete Database Marketer: Second-Generation Strategies and Techniques for Tapping the Power of your Customer Database*. Irwin, Chicago, IL, 1996.
[20] D. Iacobucci, editor. *Networks in Marketing*. Sage, Thousand Oaks, CA, 1996.
[21] 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.
[22] D. R. Jackson. Strategic application of customer lifetime value in direct marketing. *Journal of Targeting, Measurement and Analysis for Marketing*, 1:9–17, 1994.
[23] S. Jurvetson. What exactly is viral marketing? *Red Herring*, 78:110–112, 2000.
[24] H. Kautz, B. Selman, and M. Shah. ReferralWeb: Combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–66, 1997.
[25] R. Kindermann and J. L. Snell. *Markov Random Fields and Their Applications*. American Mathematical Society, Providence, RI, 1980.
[26] 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.
[27] D. Krackhardt. Structural leverage in marketing. In D. Iacobucci, editor, *Networks in Marketing*, pages 50–59. Sage, Thousand Oaks, CA, 1996.
[28] 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.
[29] 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.
[30] 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.
[31] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.
[32] J. Neville and D. Jensen. Iterative classification in relational data. In *Proceedings of the AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, pages 42–49, Austin, TX, 2000. AAAI Press.
[33] L. Pelkowitz. A continuous relaxation labeling algorithm for Markov random fields. *IEEE Transactions on Systems, Man and Cybernetics*, 20:709–715, 1990.
[34] 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.
[35] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the ACM 1994 Conference on Computer Supported Cooperative Work*, pages 175–186, New York, NY, 1994. ACM Press.
[36] J. Rucker and M. J. Polanco. Siteseer: Personalized navigation for the web. *Communications of the ACM*, 40(3):73–76, 1997.
[37] M. F. Schwartz and D. C. M. Wood. Discovering shared interests using graph analysis. *Communications of the ACM*, 36(8):78–89, 1993.
[38] C. Shapiro and H. R. Varian. *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business School Press, Boston, MA, 1999.
[39] C. Silverstein, S. Brin, R. Motwani, and J. Ullman. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery*, 4:163–192, 2000.
[40] L. Terveen, W. Hill, B. Amento, D. McDonald, and J. Creter. PHOAKS: A system for sharing recommendations. *Communications of the ACM*, 40(3):59–62, 1997.
[41] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, UK, 1994.
[42] D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440–442, 1998.