Viral Marketing: Identifying likely adopters via consumer networks

Shawndra Hill
New York University
44 W 4th St. 8th Floor
New York, NY 10012
shill@stern.nyu.edu

Foster Provost
New York University
44 W 4th St., 8th Floor
New York, NY 10012
fprovost@stern.nyu.edu

Chris Volinsky
AT&T Labs Research
180 Park Avenue
Florham Park, NJ 10012
volinsky@research.att.com

ABSTRACT

We investigate the hypothesis: those consumers who have communicated with a customer of a particular service have increased likelihood of adopting the service. We survey the diverse literature on such "viral marketing," providing a categorization of the specific research questions asked, the data analyzed, and the statistical methods used. We highlight a striking gap in the literature: no prior study has had both of the two key types of data necessary to provide direct support for the hypothesis: data on communications between consumers, and data on product adoption. We suggest a type of service for which both types of data are available---telecommunications services. Then, for a particular telecommunication service, we show support for the hypothesis. Specifically, we show three main results. 1) there is such a "viral" effect and it is statistically significant, resulting in take rates 3-5 times greater than a baseline group; 2) attributes constructed from the consumer network can improve models for ranking of targeted customers by likelihood of adoption, and 3) observing the network allows the firm to target new customers that would have fallen through the cracks, because they would not have been identified based solely on the traditional set of attributes used for marketing by the firm. We close with a discussion of challenges and opportunities for research in this area. For example, can one determine whether the reason for the viral effect is customer advocacy (e.g., via "word of mouth") versus network-identified homophily?

General Terms

Measurement, Economics, Experimentation

Keywords

Viral marketing, word-of-mouth marketing, target marketing

1. INTRODUCTION

Viral marketing, seeks to increase awareness or adoption of a product by taking advantage of the relationship network among consumers—awareness or adoption spreads from consumer to consumer. For example, friends or acquaintances may tell each other about a product or service, increasing awareness and possibly exercising explicit advocacy. Consumer networks may also provide leverage to the advertising or marketing strategy of the firm, as we will discuss below in the case of Hotmail. Firms also may use their websites to facilitate consumer-to-consumer advocacy via product recommendations (Kautz, Selman et al. 1997) or from reputation via online customer feedback mechanisms (Dellarocas 2004).

Instances of viral marketing have been called *word-of-mouth marketing, diffusion of innovation, buzz marketing*, and *network marketing*. Viral marketing campaigns typically consider a consumers' social networks and seek to exploit social behavior to increase brand recognition and profit. There are three, possibly complementary, modes of achieving viral marketing.

Explicit advocacy: Individuals become vocal advocates for the product or service, recommending them to their friends or acquaintances. Particular individuals such as Oprah, with her monthly book club reading list, or Francis McInerney with the Da Vinci Code may represent "hubs" of advocacy in the consumer relationship network. The success of "The Da Vinci Code," by Dan Brown, may be due to its initial marketing (Paumgarten 2005). The best seller was marketed in the following way: ten thousand free books were initially delivered to readers whose opinions supposedly matter (e.g., readers, booksellers) enough to stimulate the traffic in editions that are not free.

Implicit advocacy: Even if individuals do not speak about a product, they may advocate implicitly through their actions—especially through their own adoption of the product. Designer labeling has a long tradition of using consumers as implicit advocates. Firms commonly capitalize on influential individuals (such as athletes) to advocate products simply by conspicuous adoption. More recently, firms have tried to induce the same effect by convincing particularly "cool" members of smaller social groups to adopt products (Gladwell 1997; Baker 2005).

Network targeting: The third mode of achieving viral marketing is for the firm to market to prior purchasers' social-network neighbors, possibly without any advocacy at all by customers. For network targeting, the firm must have some means of identifying these social neighbors.

These three modes may be used in combination. The most well-known example of viral marketing combines network targeting and implicit advocacy. The Hotmail free email service appended to the bottom of every outgoing e-mail message, the hyperlinked advertisement, "Get your free e-mail at Hotmail," thereby targeting the social neighbors of every current user (Montgomery 2001). This strategy simultaneously took advantage of the user's implicit advocacy. Hotmail saw an exponentially increasing customer base. Started in July 1996, in the first month alone, Hotmail acquired 20,000 customers. By September 1996 the firm had acquired over 100,000 accounts, and by early 1997 it had over one-million subscribers. Today Hotmail has found that hotmail sales followed the well-known model of diffusion proposed by Bass (Bass 1969).

Firms believe it is possible that viral marketing is more profitable than traditional marketing, not only because targeting costs can be low, but also because adoption rates are suspected to be higher (Rosen 2000). In addition, traditional marketing methods do not appeal to some segments of consumers. For various reasons, some consumers value the appearance of being on the cutting edge or "in the know," and therefore derive satisfaction from promoting new, exciting products. In fact, the firm BzzAgents (Walker 2004) has managed to entice voluntary (unpaid) marketing of new products. Furthermore, with the Internet, more and more information is available on consumer products. However, parsing such information is costly to the consumer. Explicit advocacy, such as word-of-mouth advocacy can be a useful way to filter out noise.

A key assumption of viral marketing through explicit advocacy is that consumers propagate "positive" information about products after they have either been made aware of the product by traditional marketing vehicles or have experienced the product themselves. Under this assumption, a subset of consumers may have greater value to firms because they have a higher propensity to propagate product information to their friends and acquaintances (Gladwell 2002), based on a combination of their being particularly influential and their having more friends (Domingos 2005). Firms should want to find these influencers and to promote their behavior.

This paper makes two contributions. First we survey the burgeoning research literature on viral marketing, in particular on statistical analysis for viral marketing. We review the research questions posed, and the data and analysis techniques used. The literature review highlights a clear deficiency in the current research. To answer the question, "does viral marketing improve over traditional marketing techniques?" it is necessary to know about communication between customers and about product adoption. As far as we have found no prior study has analyzed data

containing both, and so this question remains open. The second contribution is to provide empirical support that viral marketing indeed does improve over traditional marketing techniques. We point out that telecommunications networks present a natural testbed for viral marketing models, because communication linkages, as well as product adoption rates can be observed. Then, for a particular telecommunication service, we show three main results: (1) viral-marketed consumers, those who have previously communicated with a person who subscribes to the service, respond to direct mailers at a much higher rate (3-5 times greater) than non-viral marketed consumers; (2) modeling attributes constructed from the consumer network can improve the ranking of customers by likelihood of adoption, leading to more precise target marketing; and (3) observing the network allows the firm to target new customers that would have fallen through the cracks, because they would not have been identified based solely on the traditional set of attributes used for marketing by the firm.

We close the paper with a discussion of challenges and opportunities for research in this area. For example, can one determine whether the reason for the viral effect is customer advocacy versus network-identified homophily (Touhey 1974) Most prior research has assumed, but not shown, that the effect is due to advocacy (e.g., via "word-of-mouth").

2. LITERATURE REVIEW

Viral marketing, takes advantage of the relationship network among consumers—awareness or adoption spreads from consumer to consumer. Therefore, interdependency among consumer preferences is a necessary condition for viral marketing to exist. In this section, we will review the wide range of research topics that have had impact on viral marketing research. The research topics fall into three categories: 1) consumer preference, where the focus is on the consumer attributes that influence one's choice of product; 2) consumer value, where the focus is on consumer attributes that influence one's lifetime value to the firm; and 3) product diffusion, where the focus is on describing the process that describes product adoption at the aggregate level. Not only is the focus of the questions that fall within each of the three categories of research different but so is the data and analysis needed to answer the associated research questions.

Although the related work discussed in this section spans the fields of sociology, computer science, statistics, marketing and economics, the related work all meet at the intersection of consumer behavior models that consider interdependent preferences. Before we introduce interdependent preferences, we will first place our work in the broader field of consumer behavior.

In general, the purpose of the study of consumer behavior is to enable firms to design and improve their marketing strategies by understanding individuals, groups, or organizations and the processes they use to select, secure, use, and dispose of products, services, experiences, or ideas to satisfy needs and the impacts that these processes have on the consumer and society (Hawkins, Best et al. 2004). Within the broader discipline of consumer behavior our work fits with consumer decision making. Consumer decision making comes about as an attempt to solve consumer problems (Hawkins, Best et al. 2004). Consumers often note problems when by comparing their current situation to some desired situation discrepancies arise. Once a discrepancy is found a search for solutions may be initiated.

The choice of solution (eg., product) has been found to be influenced by a number of factors. At some level, consumer choice research investigates the effects of various consumer attributes, intrinsic and extrinsic, on the consumer's propensity to consume a particular product or group of products. Among the factors that are studied are: 1) how consumers think, feel, reason (e.g., habits, loyalty, bias); 2) limitations in consumer knowledge or information processing abilities (e.g., awareness, experience, education); 3) consumer motivation (e.g., importance of product, switching costs, search costs); 4) the consumer's environment (e.g., signs, media, geographic location); and 5) the consumer's interdependent preferences (e.g., culture, family, friends). Consumer preference models seek to find and explain the consumer attributes or dimensions along which consumers are similar in order to explain their choices. The same factors may influence a consumer's propensity to spread the word about a product. In this study, we add to the body of literature that explores the impact of interdependent preferences on consumer preferences. However, based on prior studies, we acknowledge we must control (to the extent possible) for additional factors known to influence consumer choice.

The term interdependent preference, preferences which depend on other people's preferences, has been used to describe a number of interactions, both inferred and assumed, between consumers. To our knowledge, the first work on interdependent preferences describes several examples of interdependence in consumer consumption (Duesenberry 1949). Duesenberry uses data on consumer purchases made in 1935 and 1936. He finds that the percentage of income spent on consumption is highly correlated with the person's rank order in the local income distribution. The finding suggests that individuals are arrayed in a hierarchy in which each individual's preferences are influenced by the consumption behavior to those directly above him in the hierarchy. Using

this finding as an assumption, the role of expenditure distribution as a determinant of per capita consumption patterns was formalized to consider consumer patterns over time (Pollak 1976), introducing the concept of lagged interdependent consumer preferences. Pollak presents a theoretical model of demand analysis that incorporates interdependent preference and its influence on the distributions of overall consumption. In both of these ground-breaking studies, and some that follow (Darrough, Pollak et al. 1983; Alessie and Kapteyn 1991), the link between consumers is defined by their relative income and the dependent variable is the consumer's overall expenditure allocation.

Since the pioneering studies that relied on a rather crude dataset, firms and researchers have gained access to a wide range of consumer data including not only income and overall expenditures but transaction, demographic, habits, and more recently data on consumer-to-consumer (C2C) networks. This has made the study of different types of interdependencies, or links between customers, possible. All of the work below studies the impact of interdependent preferences on product adoption at three different levels of analysis: 1) overall product diffusion; 2) individual customer value; and 3) customer adoption (both individual and at the aggregate level) .

The following review is organized by the types of targets (e.g., preference, value, model) studied and to some degree the types of interdependencies that are assumed or inferred between the individual consumers. The goal of the review is to provide a useful guide for researchers studying interdependent preferences within explicit consumer networks. The review is organized by types of targets in reverse order of relevance: 3) consumer preference; 2) consumer value; and 1) product diffusion,

2.1 Diffusion

The speed of product diffusion depends to some extent on the speed of product adoption. Like individual preferences and group preferences, product sales/adoption/diffusion,/etc, are impacted by many factors. These factors are both endogenous and exogenous to the firm. Despite the fact that some of these factors are uncontrollable, it is very important that firms have an idea of the sales trajectory prior to launching a new product. Therefore, understanding process adoption in general is fundamental to the firm.

There are many theories of product adoption, the most notable and most influential has been the Bass (Bass 1969; Dodds 1973; Ueda 1990; Bass, Krishnan et al. 1994; Evans 1995; Satoh 2001;

Niu 2002) model of product adoption. This model is so popular there is software (Lilien 2000) available for use in real world applications.

The Bass model of product diffusion predicts the number of users that will adopt an innovation. It assumes that the total population has m individuals that will eventually adopt the product and defines the number of individuals who have adopted at time t as $n(t)$. In addition to word-of-mouth instigated product diffusion, the model assumes a constant proportion adopt because of advertising. The model has three parameters m: market potential, α: the effect of advertising and β: the effect of word of mouth. There have been many extensions that fit the parameters to various data sets. In fact, there have been over 250 papers including applications, refinements and extensions. A review of can be found here (Mahajan, Muller et al. 1990).

Another popular model familiar to many practitioners can be found in Geoffry Moore's "Crossing the Chasm" (Moore 1999). Moore discusses the market for new products in terms of a product life cycle. The area of each segment corresponds roughly to the number of people who fit its profile. His thesis for product adoption, specifically technology product adoption, is more of a schematic that includes the following five types of consumers, 1) technology enthusiasts; 2) visionaries; 3) pragmatists; 4) conservatives; and 5) skeptics. His thesis relates to viral marketing in that he suggests the technology adoption happens in the following two ways, technology enthusiasts tell visionaries who in turn tell the pragmatists. In other words, he proposes that viral marketing is necessary for wide spread new technology adoption.

The most recent work on product diffusion explore the extent to which the Internet (Fildes 2003) as well as globalization (Kumar and Krishnan 2002) have played a role in product diffusion. In general, the empirical studies which test and extend accepted theories of product diffusion rely on aggregate level data for both the customer attributes and overall consumption of the product. The empirical studies show the extent to which existing models fit new data by fitting parameters to predict future sales. In some research, parameters are added to accommodate the appropriate business landscape. For example, to account for competition (Ueda 1990; Evans 1995), attitudes and imitation (Ueda 1990; Evans 1995).

Recall our research question "do viral-marketed consumers, those who have previously communicated with a person who subscribes to the service, respond to direct mailers at a much higher rate than non-viral marketed consumers?" Diffusion models do not focus on the same target. Instead, the focus is on aggregate level product adoption as opposed to individual adoption

as is the case on our research or the mechanisms that generate individual level explicit advocacy as is the case in the next section describing work on customer value. However, models of product diffusion assume (and firms hope) that viral marketing is effective. However, the understanding of when it occurs and to what extent it is effective, is important for the firm. In the next section, we will discuss studies that consider when viral marketing occurs.

2.2 Customer value

Models that describe the process by which overall product adoption propagates through a customer network have been studied for the diffusion of technological innovations as described above. However, these models rely on the assumption that viral marketing exists and is effective. If the assumption that viral marketing is effective holds, one way for firms to influence the diffusion rate of a product is to convince a subset of influential individuals to adopt the new product or innovation, so they may in turn influence additional customers.

Therefore, in addition to figuring out the right customers to target based on whether or not they will adopt a product, firms also care about the lifetime value of the customer. If a firm knows the lifetime customer value for individual customers, they can decide how much effort they should spend on targeted individuals. In fact, precise cost/benefit analysis is necessary for large scale projects, for which the cost to acquire customers is very large (and the opportunity costs of not acquiring customers is large) both in dollars and resources. In the past, this number has been estimated solely based in terms of spending over time. Recently, the idea of targeting individual customers based on network value was proposed (Domingos 2005). With this proposal, the value is not only based on how much income the customer themselves will generate but also the amount of income the network of customers they influence will generate.

2.2.1 Finding the Best Subset

In (Domingos and Richarson 2001), the social network of customers is modeled as a Markov random field. They authors test their model on a collaborative filtering database of movie reviews from the EachMovie database. The customers are assumed to be linked when a customer reads a review of a movie by another customer and the n sees the movie as evidenced by their own review in turn. The authors show that they can offer value to a firm by selecting a good subset of customers to target. The authors find their proposed methodology outperforms naïve methods for estimating customer value.

New strategies for selecting the best subset of customers based on their network value has been proposed since the pioneering work by Domingos and Richardson in 2001. For example, Kempe et al. (Kempe, Kleinberg et al. 2003), treat the same problem of selecting the best subset of customers as an optimization problem and provide the first provable approximation guarantees for efficient algorithms. Using an analysis framework based on submodular functions, the authors show that a natural greedy strategy for selecting customers based on network attributes obtains a solution that is provably within 63% of optimal for several classes of models; The authors provide computational experiments on large collaboration networks, showing that in addition to their provable guarantees, their approximation algorithms significantly out-perform node selection heuristics based on the well-studied notions of degree centrality and distance centrality from the field of social networks.

In both of these best subset studies, the authors make an assumption about the links. Their methods assume that customers that communicate with each other will adopt at a higher rate and therefore the customers that are central are most influential. Although these studies provide very interesting empirical results, they don't have data on customer response and so therefore, they must use reasonable proxies to assess customer response.

Even though the focus here is on customer value and not customer adoption, the answer to their question depends on the result of the investigation of our hypothesis: those consumers who have communicated with a customer of a particular service have increased likelihood of adopting the service. However, due to the proxies, they cannot sufficiently answer our research question. Therefore, the answer to their research questions would benefit from knowing the actual explicit links between consumers. Their models depend on the assumption that we provide evidence for, that customer choice is interdependent.

2.2.2 Who Will Spread Information by Word of Mouth?

Word-of-mouth studies are probably the most prevalent in viral marketing studies. There are two different sides to word-of-mouth research. In (Godes and Mayzlin 2004), the authors identify word-of-mouth as a driver of consumer behavior and word-of-mouth as an outcome of consumer behavior. In this section, we discuss research on word-of-mouth as an outcome. In many of these studies, the authors ask the question who are the people that are most likely to spread the word and under what conditions.

It is important for firms to identify those influential people because it has been shown that consumers sometimes place more weight on their friends' and acquaintances' preferences than their own, potentially leading to irrational outcomes (Straffin 1977). The models of diffusion discussed above assume that consumer-to-consumer interaction occur both as an outcome and precursor of sales.

As an outcome, authors have studied when people pass along product information, both good and bad. For example, (Richins 1983) studies word-of-mouth communication by the dissatisfied customers. This work investigates the moderating factors that determined whether and when one passes along the information associated with their negative experience. The result is that people spread information at rate correlated with the severity of dissatisfaction. Furthermore, if the customer feels the firm is guilty (as opposed to having the result be partially their own fault), they are even more likely to spread negative information about the product or service.

Similarly, (Anderson 1998)looks at both positive and negative word-of-mouth communication. He finds that very dissatisfied customers and very satisfied customers are most likely to engage in word of mouth. Those "in the middle" do not do so.

Most word-of-mouth studies rely on survey data. This is the strength of these studies because they actually ask people who and how many people they told about the product. For example, Bowman and (Bowman and Narayandas 2001)specify a model that identify product attributes that influence word-of-mouth. The authors survey customers of 60 brands of consumer products manufactured in the US. By using survey data, they were able to capture both whether the customer told someone else of their experience and how many people they told. Furthermore, the authors find that self-reported loyal customers were more likely to participate in word-of-mouth when they are dissatisfied but interestingly not more likely when they are satisfied.

Although these studies have information on the surveyed consumer's word-of-mouth behavior, they do not know which of the customers purchased the product and therefore do not address our research question to know if word-of-mouth actually impacts individual sales.

2.3 Consumer preferences

Consumer choice depends on two things, consumer preferences and the set of feasible alternatives. Consumer preferences are influenced by a number of factors. In this section, we explore the models that include some variant of interdependent preferences as a factor in a consumer choice model. We divide the work into three categories organized by the types of links used to connect

consumers: 1) spatial models which use geographic links; 2) collabarative filtering models which use product, transaction links, and demographic; and 3) word-of-mouth use some form of communication about a product between customers as a link.

2.3.1 Preference based on spatial characteristics

Additional empirical work on interdependent preferences include the impact of consumer interdependence on rice consumption (Case 1991) , automobile purchases (Yang and Allenby 2003), and elections (Smith and LeSage 2004). Each of these approaches measures the impact of geographic region on aggregate consumption. Yang and Allenby, however, develop a spatial autoregressive mixture model that incorporates additional explanatory variables for each consumer such as age, annual income, and education level. These variables allow for the alternative explanation that their result is capturing explicit similarity. The authors show that the geographically defined network (customers that are close geographically), is more important than the demographic network in explaining consumer behavior as it relates to purchasing Japanese cars. The authors only have data on 857 customers living in 122 different zip codes and don't report about other characteristics that may be influencing the result such as location of dealerships.

Due to the lack of real-world data to test the impact of actual interdependent preferences (both implicit and explicit) on consumption (individual and aggregate), simulation based estimations of parameters have been proposed for the identification of geographic target markets (Ter Hofstede, Wedel et al. 2002) and to forecast brand sales (Bronnenberg and Mahajan 2001) . In the former, the authors suggest that segments of consumers are likely to demonstrate spatial patterns. And to forecast brand sales (Bronnenberg and Mahajan 2001), the authors use simulated data to extrapolate the market response effect from the reverse retailer effect by computing responses to price and promotion net of any spatial-and therefore retailer-influence. The proposed model allows to test for endogeneity of prices and promotion variables in the cross-sectional dimension of the data. In both cases, the authors are able to test theories of complex data, fit parameters and apply them to empirical data. However, the authors still only take as the response aggregate level data.

The related work presented in this section study the impact of interdependent consumer preferences on consumption. In most cases, the authors wanted to get at whether or not interdependence of the explicit interaction type influenced product consumption. But often they did not have the appropriate data to do so. These studies do not have feedback for a specific

product, such as from a direct marketing campaign, so they cannot directly answer our research question. Recall our research question "do viral-marketed consumers, those who have previously communicated with a person who subscribes to the service, respond to direct mailers at a much higher rate than non-viral marketed consumers?" In some prior work the authors ask similar questions for example, does a particular type of interdependence (like relative income (Duesenberry 1949; Darrough, Pollak et al. 1983; Alessie and Kapteyn 1991) or geography(Case 1991; Yang and Allenby 2003; Smith and LeSage 2004)) influence overall consumption (demand distributions (Duesenberry 1949; Pollak 1976) and aggregate level consumption (Alessie and Kapteyn 1991; Case 1991; Yang and Allenby 2003)). Yang and Allenby are the only authors in this group that look at individual level consumption on real data.

In general, data used in the aforementioned research on the impact of interdependent preferences on consumption does not have information on explicit links. Therefore, they are not able to compare non-viral and viral targets (e.g., the impact of the interdependent attribute on consumption). Because the actual links were not available, studies rely on proxies for those links. The data sets are relatively small and attributes derived from the explicit consumer-to-consumer network are not used. Therefore methods to address networked data are not necessary for the aforementioned studies. The first work we know of the accounts for links between customers (implicit or explicit) is in work on collaborative filtering.

2.3.2 Collaborative Filtering

The purpose of collaborative filtering systems is to automate the process of "word-of-mouth" by which people recommend products or services to one another. Collaborative filtering systems make personal recommendations to individual consumers via the internet based on the purchase preferences of the most similar consumers. Collaborative filtering marketing techniques involve associating customers primarily based on descriptive data, for example demographics-based data mining or content-based recommendation, and/or transaction data. To our knowledge current collaborative filtering systems do not add to these a third type of data: direct interaction between consumers, for example explicit communication.

Unlike most of the literature in consumer choice that relies on traditional statistical methods, work in collaborative filtering has proposed techniques that enable the exploitation of information in the links between customers (Huang, Chung et al. 2004). Establish the connection between the recommendation problem and the relational learning framework through the application of a

recently developed statistical relational learning method called Probabilistic Relational Models (PRMs). The authors use purchase data from a large internet company, which has individual level response data but no explicit interaction between customers. Likewise, (Newton and Greiner 2004) applied PRMs to the Collaborative Filtering task. The authors use recommendations for links between customers and whether or not the customer sees the recommended movie as the target. Another example of links used to assess similarity for better product recommendations is location as identified by a mobile device. For example, (Huang, Terry et al. 2002)) uses physical proximity between mobile devices to help users filter incoming information and determine its relevance.

Although collaborative filtering is very related to viral marketing, the goal is to automate recommendations as opposed to studying the effect of viral behavior for the purpose of better targeting for a single product. We suspect firms that use recommendation systems could benefit from the additional link of explicit consumer interaction. This customer link would allow the firm to combine one additional, if not the most important, aspect of customer similarity.

2.3.3 Who is influenced by word-of-mouth?

The theoretical models of diffusion all incorporate to some degree the impact of viral marketing on product sales. In the Bass model, the attribute is imitation. Another study by Banerjee (1992, 1993) suggests that in some cases people may place significant weight on the opinions of others. In these models, social phenomena such as herding where actors take similar actions even if they each have information that favors a different outcome take place.

These theoretical studies rely on the fact that people tell or influence other people in some way. However, these studies do not have individual response. Recent studies have had the ability to get explicit links between customers via web data. For example, (Godes and Mayzlin 2004) use discussion boards to establish links between people. They compare the frequency of mention of new television shows within and across different television discussion groups. They find that both the frequency and dispersion of mention are related to take rates. In another study they look at customer reviews of books. The compare consumer choice of two books across two different book sellers and they show that the number of reviews and quality of reviews influences the overall relative book sales. In both of their studies, they use aggregate level data for the response variable.

A similar study considers the actual business impact of proxied word-of-mouth generated from customer review sites. The movie industry is used as its testbed (Dellarocas, Awad et al. 2004). The authors develop a revenue-forecasting model based on the Bass model discussed above (Bass

1969) that incorporates the impact of both publicity and word of mouth on a movie's revenue trajectory. The authors successfully predict a movie's total revenues during the first week of a new movie's release.

The web has made it easier to establish links between customers. In addition, the web has enabled both researchers and firms to take advantage of publicly available data. For example, by using information found in discussion groups and product review sites above. In addition, reputation response sites (Dellarocas 2004) have been suggested as a way to study proxied word-of-mouth. In the Dellarocas's survey of reputation sites and their potential for helping researchers understand word of mouth. He identifies a number of studies that provide evidence that customer reputation on the internet greatly influences consumer trust which greatly influences customer response (or price). Online auction sites, such as eBay, rely on reputation mechanisms for quality signaling and quality control. Although, reputation site data contain explicit links between individuals, the explicit links that can be observed are between consumers and buyers, the links between consumers (one consumer communicates to another via a customer feedback site) are implicit .

2.3.4 The Gap in the Literature

A number of firms increasingly utilize advanced information tracking and surveillance technologies that incessantly log consumer behavior. In addition, e-commerce firms are also able to monitor the results of data on refer a friend programs. The data gathered using such technologies can used to derive explicit networks of consumer interactions. Firms whose business involves providing telecommunication and Internet infrastructure have had access to such data sets for many years. More recently, a variety of other firms are able to gather such networked data, by offering their own email and IM services (recent examples being Gmail and GoogleTalk), or by mining the "blogosphere", which represents an immense network of interconnected consumers. However, because these data is often proprietary and unavailable for research studies rely on proxies for direct interaction. The different types of proxies discussed above make it impossible for prior research to answer our research question.

As should be evident from above, consumer-to-consumer (C2C) data are widely inferred in many target viral marketing studies. Sometimes, proxies are used when links cannot be explicitly observed. C2C models for product adoption often do not rely that people have personal contact however sometimes it is required for people to have experience with the product.

Often viral marketing studies take as input different types of data such as demographics, geography, product, and interactions between consumers and products (such as purchases and

ratings). In addition, these data vary in their reliability and availability. The links between the consumers and attributes such as geography and product attributes have been used as proxies for explicit links. It is a challenge to determine to what extent and which links are important for predicting individual take rates. We argue that firms should be able to use these dynamic networked data to improve their real-time targeted marketing and personalization.

In addition, individual purchasing behavior is not available in most studies therefore studies rely on aggregate level behavior in most cases. Therefore, the strength of relationship between customers and the impact that relationship has on sales cannot be modeled directly. In order to provide evidence that viral marketing exists, we need both explicit customer to customer *interaction and individual response.*

In this section we have provided a summary of prior work showing that no one has had data on both explicit consumer interaction and individual sales rate. We see that in the case where explicit interaction can be measured as is the case in online discussion groups, the individual sales rates and/or adoption have not been observed. Furthermore in the cases where individual sales can be observed as is the case in some recommendation studies, the explicit links between consumers was not observed. Our data set is unique in that it has both product adoption and explicit links between consumers. In the next section we describe this data set.

3. VIRAL MARKETING TESTBED

In late 2004, a large telecommunications firm undertook a large direct-mail marketing campaign to potential customers of a new communications service. In keeping with standard practice, the marketing team used profitability models and behavioral models to identify prospective customers to be *targets* for the mailing. Separately, we created a list of *viral* marketing prospects based on the following definition: a viral prospect was a consumer who had communicated with a user of the service some time prior to the campaign. Our list overlapped with the marketing list, but also contained many prospects they had not identified. The list was scrubbed of prospects that were not able to purchase the service for various reasons (the same criteria apply to all lists of prospects). We handed this list off to the marketers, who applied their models to this new set of prospects. By amending the thresholds on these models, they defined a set of the best prospects from our viral list to augment their existing list. Overall, the firm sent a marketing solicitation to about 15% of the viral prospects. We call these consumers *viral targets*. The rest of the targets are *non-viral targets*.

Table 1: Data Categories. *The data for our study are broken down into whether or not they were targeted for the marketing campaign, and whether or not they were on the viral list. The "relative size" value shows the number of prospects that show up in each group, relative to the Non-viral Target group.*

	Target =Y	**Target = N**
Viral=Y	*Viral Targets* *Segments 1-22* ***Relative Size=0.015*** Prospects who were defined by marketing models and also are viral. Those in Segment 22 are only marketed to because of their viral status.	*Viral Non-targets* ***Relative size = 0.10*** Prospects who were viral but were not marketed to because they were not considered to be good prospects.
Viral=N	*Non-viral Targets* *Segments 1-21* ***Relative Size = 1*** Prospects who were defined by marketing models and did not satisfy the viral criterion.	*Non-viral Non-targets* ***Relative Size > 8*** Those prospects that were not viral and were not considered prospects by the marketing model.

The remaining viral prospects comprise potential consumers that the firm had no prior relationship with, or were believed to be very poor prospects. In the absence of marketing to this group, we are still able to observe the take rates of these prospects. We call these consumers *viral non-targets*.

In summary, we have four categories of potential customers at the time of the mailer, viral targets, non-viral targets, viral non-targets and the rest of the prospect universe, the non-viral non-targets. Table 1 summarizes the four categories and shows their relative sizes, using the viral non-targets as the reference group.

In the sections that follow, we describe in detail the data we had available to investigate the impact of viral marketing on targeted sales. Unfortunately, we cannot disclose the specific attributes of the entire data dictionary. Therefore, next we present a more general discussion of attributes in the following 5 categories: 1) segment data; 2) loyalty data, 3) demographic data, 4) geographic data, and 5) social-network data, plus a discussion of missing data.

3.1.1 Segment data

The prospects initially defined by the marketing team were binned into 21 different marketing segments in order to stratify across traditional marketing attributes that were known from experience to be important. These attributes are included in a profitability model which included the prospects' previous relationships with the firm, any currently subscribed services,

demographics, and characteristics of their communication behavior. In addition, a small number of segments were identical to others, except for the marketing channel chosen.

When the marketing prospects and the viral prospects were combined, each of the 21 marketing segments contained a number of viral targets along with the non-viral targets, allowing us to assess the "viral effect" in each of those segments. In addition, we selected a set from the viral targets who did not appear on the marketing list to make up a 22nd viral-only segment. We had a budget of a fixed number of prospects in this segment, and so chose those that were believed to be the best prospects, based on the same profitability models the marketing team used to select the original segments. Notably, these targets are potential customers the firm would have otherwise ignored. The remainder of the viral list was not marketed to, and make up the viral non-target group.

3.1.2 Loyalty Data

The firm records information on previous relationships it has had with its customers, including previous orders of this and other services. These data include past spending, types of service, how often the customer responded to prior mailers, a loyalty score generated by a proprietary model, and information about length of tenure. In addition, we know if the customer is an employee (who often gets incentives to sign up for novel services).

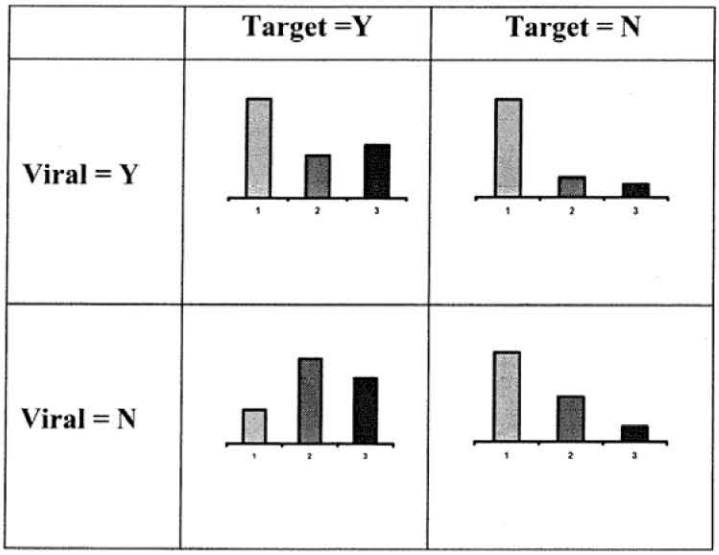
We use the loyalty data to identify three levels of customer loyalty based on the target's prior relationship with the firm. Loyalty level 3 comprises the customers with moderate-to-long tenure and/or those who have subscribed to a number of services in the past. Loyalty level 2 comprises to those customers with which the firm has had limited prior experiences. Loyalty level 1 comprises the consumers who did not have service with the firm at the time of mailer; therefore, little (if any) information is available on level-1 customers.

A look at the distribution of the loyalty groups across the four categories (Figure 1) of prospects shows that the firm targeted customers in Segments 2 and 3 quite heavily, trying to appeal to customers who had shown some loyalty in the past. The viral target group appears to skew toward the less loyal prospects; this is due to the fact that Segment 22, which makes up a large part of the viral population, is predominantly a low-loyalty group.

3.1.3 Demographic Data

We purchased demographic data from an external vendor. These data include information such as gender, education level, credit score, head of household, number of children in the household, age of members in the household, occupation, and home ownership information. Although these

Figure 1: Loyalty Distribution by Customer Category. *The three bars show relative size of the three loyalty groups for our four data categories. Loyalty group 1 is the low loyalty group that have no previous relationship with the firm. Group 2 had a limited relationship and Group 3 a strong relationship.*



demographic data are important for evaluating the results of marketing campaigns, they do have limitations. First, they are costly. Second, for those consumers who did not subscribe previously to a service from the firm, often either relatively little information was available or the information was outdated.

3.1.4 Geographic Data

Since the primary marketing channel was a direct mail piece, geographic data were available for all targeted customers. These data include information such as city, state, zip code (and the corresponding census data associated with it), area code, and metropolitan city code.

3.1.5 Network Data

The contributions of this paper result from the fact that we can observe the communications of the current subscribers to the service with other consumers. The firm has data on all subscribers of the service, as well as some limited usage information on those in loyalty levels 1 and 2. In addition, for the viral targets, we have information about their communications with current customers. Therefore, for evaluation purposes we have network information on many targets, prior to their order. For each target, we generate a list of their transactions. Each entry in the list includes the

transactors, a time stamp, and the transaction duration. For the purposes of this research, all analyses are anonymized so that individual identities are protected.

We define the "viral attribute" as simply a flag indicating whether or not the targeted consumer had current users of the service in her communication neighborhood just prior to the marketing campaign. In addition to the viral attribute, we construct other network attributes for each viral target, which are described later.

3.1.6 Data Limitations

We encounter missing values for customers across all loyalty levels. However, the amount of missing information is directly related to the level of experience we have had with the customer just prior to the direct mailer. For example, geography data are the most available data—data are available for all targets across all three loyalty levels. On the other hand, we only see communications that are carried on this firm's network. As the number of services and tenure decline, so does the amount of information available on each user. Given the difference in information as loyalty varies, we generally group customers by loyalty level, and treat the levels separately when we present our analysis.

Although, the data used in this study will enable us to investigate many questions about social network influence on product adoption, the data presents some challenges. For example, the data set is large, and as with most direct-marketing campaigns, the overall response rate is very low. This creates challenges for the analysis inherent with having a heavily skewed response variable. An analysis that stratifies over many different variables may have several strata with no sales in them whatsoever, rendering them mostly useless. Therefore, we have restricted ourselves to fairly simple statistical analyses focused on estimating the viral effect while controlling for the appropriate factors.

4. ANALYSIS

We now turn to our fundamental question of interest: can we show direct, statistical evidence that consumers who have communicated with prior customers are more likely to become customers? We can observe a large-scale consumer-to-consumer network augmented with the responses resulting from a large direct marketing campaign. These data enable us to compare the response of viral targets to non-viral targets and to further investigate methods to utilize the consumer network for target marketing.

In order to show that the consumer-to-consumer network provides (additional) value, we first will assess the effect of viral marketing at an aggregate level, showing that viral-targeted consumers

have much higher response rates than other customers, and that the viral attribute contributes significantly over all other information in a statistical targeting model. Then we report on a fine-grained analysis showing that this network information improves the actual marketing scores given to individual consumers. Finally, we will demonstrate that measuring more complex network attributes allows us to rank the viral targets and identify the best prospects.

4.1 Viral-targeting improves response

The segment-based construction provided an ideal setting to test for the significance and magnitude of the viral effect, while stratifying by many of the attributes which are known to be important to the marketers, including history with the company, and marketing message.

The response variable is the sales rate (*take rate*) for the targets in the two months following the direct mail drop. Due to proprietary restrictions in reporting the data, we are not able to present the actual take rates, but instead present the odds ratio of adoption, i.e. the odds of adoption for the viral group divided by the odds of the non-viral group. For each segment, we construct a 2x2 contingency table for the independent viral attribute versus the dependent sales response. From each of these tables we define an odds ratio, which we present on the log scale due to the skewness of its distribution. Figure 2 shows these log-odds ratios for 20 of the 21 segments (one of the segments had only a small number of viral prospects and zero viral sales, and therefore an infinite log-odds). By running a logistic regression on each segment, we use the standard error of the regression coefficient to calculate confidence bounds on the log-odds scale (Hosmer and Lemeshow 1989), which are also plotted in Figure 2. Looking at Figure 2, we see that in all 20 of the measurable segments the viral effect is positive (log odds greater than zero), demonstrating an increased sales rate for the viral group within the segment. For 17 of these segments, the log-odds ratio is significantly different from the null hypothesis value of 0 ($p < 0.05$), indicating that the viral effect significantly impacted sales in those segments.

While odds ratios allow for tests of significance of an independent variable, they are not as directly interpretable as simple ratios of take rates between the viral and non-viral group in a given segment. These simple *take ratios* give the factor by which the take rates are increased in the viral group over the non-viral group. The take ratios are shown graphically in Figure 3, where they are plotted as a function of the size of the segment. The lower horizontal line in the plot is the "no-effect" value of 1. All of our segments are above this value, since all of the segments had increased sales in the viral part of the segment. The median ratio of our 20 values is 5.0, but that is a bit misleading to use as an overall measure, since some of the smaller segments have very high

Figure 2: ***Viral Effect by Segment***: *Comparison of sales rates for viral and non-viral customers from a direct mailing. Due to proprietary restrictions, the figure presents only the log odds ratio.*

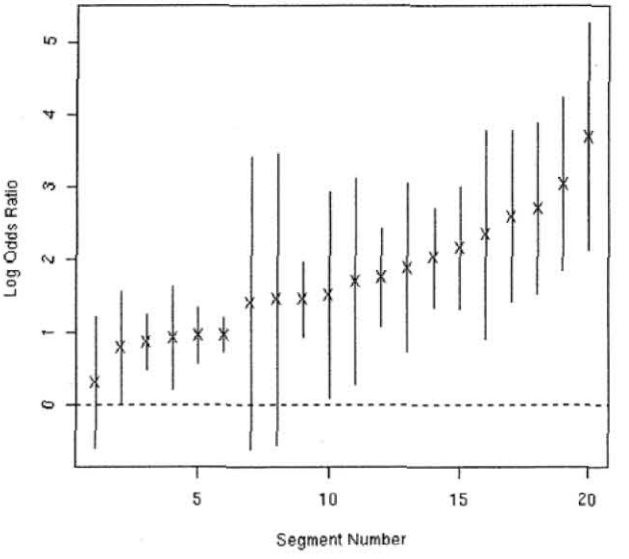
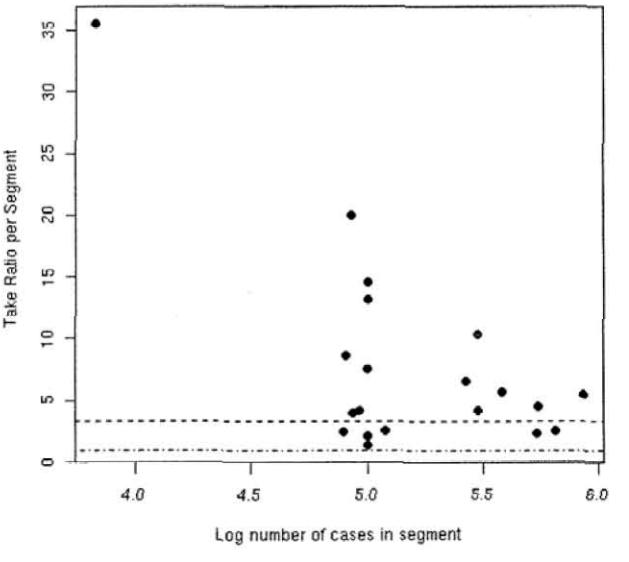


Figure 3: ***Take Ratios for Marketing Segments***. *Take ratios are plotted as a function of log of segment size. Horizontal lines are drawn at the null ratio of 1, and the mean ratio for the entire data of 3.4.*



take ratios. To get a more representative number, we look at the take ratio for all virals vs. all non-virals, regardless of segment, and get an overall viral take ratio of 3.4. This number has a simple interpretation: the viral group had a take rate 3.4 times greater than the non-viral rate. This value is plotted as the higher of the two horizontal lines in Figure 3.

Some of the segments had much higher take rates than others, due to the attributes used to create them. In order to get a better sense of the statistical significance of the viral effect after accounting for this segment effect, we ran a logistic regression across all segments, including main effects for the viral attribute, dummy variables for each segment, and for the interaction terms between the two. Two of the interaction terms had to be deleted, one from Segment 22, which only had viral cases, and one from another segment where there were no sales in the viral portion of the segment (and hence an inestimable interaction effect). We ran a full logistic regression, and used stepwise regression for subset selection.

The results of the logistic regression reiterate the significance of the viral attribute. The final model can be found in Table 2. The coefficient of 2.0 for the viral attribute in the final model is an estimate of the log-odds, which we exponentiate to get an odds ratio of 7.49, with a 95% confidence interval of (5.64, 9.94). This means that the odds of take for a viral target is estimated to be 7.49 times greater than the odds for a similar non-viral case (with respect to segment). Also, more than half of the segment effects and many of the interactions are significant.

In Table 3 we present an analysis of deviance table, an analogy of analysis of variance used for nested logistic regressions (McCullagh and Nelder 1983) The table confirms the significance of the main effects and of the interactions. Each level of the nested model is significant when using a chi-squared approximation for the differences of the deviances. The fact that so many interactions are significant demonstrates that the viral effect varies for different segments of the prospect population.

4.2 Segment 22

The segment data enables us to compare sales rates of viral and non-viral targets for the segments that contained both types of targets. However, many of the viral targets fell into the viral-only segment 22. Segment 22 is made up of prospects that were not identified by the original marketing models to be good prospects. As we can see in the loyalty distribution in Figure 1, the segment is mostly made up of those that did not have a prior relationship with the firm.

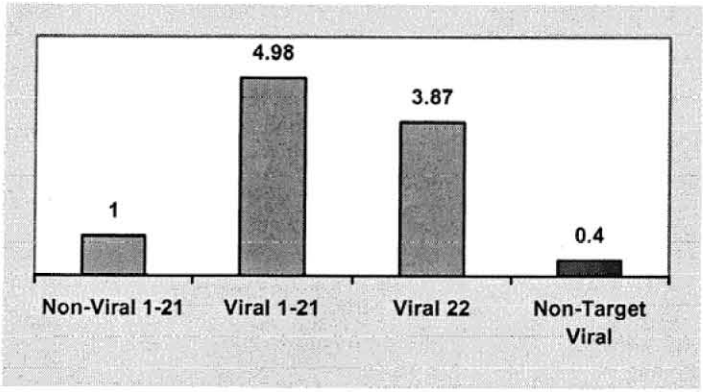
Table 2: Coefficients for Final Segment Model. *Significance of the variable in the logistic regression model is shown at the .05 (*) and .01 (**) level.*

Attribute	Coef (CI)	significance
Viral	2.0 (1.7,2.3)	**
Segment = 4	1.9 (1.2,2.5)	**
Segment = 5	1.5 (0.7,2.2)	**
Segment = 8	2.2 (1.6,2.9)	**
Segment = 11	1.7 (0.9,2.5)	**
Segment = 12	1.8 (1.2,2.4)	**
Segment = 13	1.4 (1.0,1.9)	**
Segment = 14	1.3 (0.9,1.7)	**
Segment = 16	2.1 (1.3,3.0)	**
Segment = 19	1.9 (0.4,3.3)	**
Viral X Segment = 4	-1.5 (-2.6,-0.6)	**
Viral X Segment = 5	-1.6 (-2.8,-0.5)	**
Viral X Segment = 8	-1.1 (-1.9,-0.3)	**
Viral X Segment = 11	-1.1 (-2.1,0.0)	*
Viral X Segment = 12	-0.9 (-1.7, -0.2)	**
Viral X Segment = 13	-1.2 (-1.7, -0.6)	**
Viral X Segment = 14	-0.8 (-1.3, -0.4)	**
Viral X Segment = 16	-1.8(-4.0, 0.4)	**

Table 3:Analysis of Deviance table for the viral study. *Significance of the group of variables at each step is shown at the .05 (*) and .01 (**) levels.*

Variable	Deviance	DF	Change in Deviance	Significance
Intercept	11200			
Segment	10869	9	63	**
Segment + Cell	10733	1	370	**
Segment + Cell + Interactions	10687	8	41	**

Figure 4: Relative Take Rates for Marketing Segments. *Take rates are shown for the Viral and Non-Virals in Segments 1-21 compared with the all-viral Segment 22, and with the non-target virals. All take rates are relative to the Non-Viral Segments 1-21 group.*



Since Segment 22 does not have any non-viral prospects, we cannot compare odds ratios from a model with the viral attribute. However, we can compare the take rates for this group against the take rates for the combined group Segment 1-21. Again, for proprietary reasons we cannot share the actual take rates, but do present relative take rates. These are shown in the 3 leftmost bars in Figure 4. We can see that the viral Segment 22 is not as successful as the Viral groups in Segment 1-21, thus validating some of the marketing targeting models. However, we also see that the Segment 22 virals outperform the non-virals in Segments 1-21, indicating that these viral prospects initially ignored by the marketing group are at least as valuable as the non-virals that the marketing models selected.

Remember that those in the viral segment are those that the marketing models either were not able to identify, or were deemed to be unworthy prospects. In some sense, this group represent customers that would have "fallen through the cracks" in a traditional marketing event. This demonstrates an additional revenue-enhancing benefit of viral marketing, that it may complement traditional marketing by uncovering additional good prospects.

4.3 Improving a multivariate targeting model

As described above, the segments used in the last section were defined by the marketing team based on previously determined important characteristics such as loyalty, communications behavior, and marketing message. The previous sections demonstrated that viral-targeted consumers are more likely to respond than are non-viral consumers. Now we will assess whether this "viral attribute" can improve a multivariate targeting model.

We tried to address the concern that there may be some confounding variable that makes viral appear significant. For instance, perhaps viral prospects simply communicate more, or had higher income, and it is this other variable that is affecting sales rate, not the fact that they spoke to a current customer. We collected as many descriptive attributes as we could and included them in the analysis.

The attributes that we included comprised all we could find out about the customer from internal and external sources, including demographic, geographic, and loyalty data. They include:

- **Marketing channel**: callblast, postcard, letter
- **Geography**: MSA, DMA, State, Zip, Area Code
- **Demographics**: Aggregate data on Census tracts on income, gender, household information on gender, household size
- **Loyalty**: firm specific variables, previous services, tenure, churn model, employee

In all, we considered over 150 different attributes for their effect on sales rate. Note that some of these variables are only available for the higher loyalty categories. Because of this, and also because we believe the effect of viral to vary across loyalty categories we fit three different models. For each of the three loyalty categories, we fit a full logistic regression and selected a final model using stepwise variable selection. All variables were checked for collinearity and any significant correlations were accounted for.

Each model found the viral attribute to be significant along with a host of other variables – the final models had from 7 – 22 variables in them, too many to show in a table here. Interestingly, none of the attributes were significant in all three models, but there were some consistencies, for instance, when tenure (as a customer of the firm) was available (in loyalty groups 1 and 2), it was a significant variable. The results are in Table 4. For each of the three loyalty groups we see that confidence interval for the log odds does not include zero, i.e. the viral attribute is significant. The table also shows the relative take rates in the three loyalty groups. As we expect, the take rates increase in the groups with higher loyalty to the firm. However, the effect of the viral attribute (as measured by the log odds ratio) increases with the lower loyalty groups. So, the impact of viral is stronger for those market segments with low loyalty, where actual take rates are weakest. This conclusion could have important implications for future marketing campaigns.

Table 4: Results of multivariate model. *Coefficient of Viral attribute from logistic regression across loyalty levels.*

Loyalty	Log Odds (CI)	Relative Take Rate	Significant Variables
1	1.828 (1.28.2.38)	1	7
2	1.257 (0.64 1.88)	1.1	12
3	0.988 (0.73,1.25)	3.2	22

4.4 The Others

As discussed above, a list was constructed with viral targets based on prior customer-to-customer interaction. Only the top prospects from this group were selected as targets in the marketing campaign. The remainder of the list, the **non-target virals,** made up the majority of the list. They were left out for various reasons, in some cases, the customers were on a do-not-contact list. In others, address information was unknown, or not reliable. In even more cases, they were believed to be poor prospects for various reasons.

Since we knew the prospects in this group, we were able to identify whether they purchased the product in the follow-up time period. The take rate relative to the target groups is presented as the right-most bar in Figure 4. We find that the take rate for this group is about half of the non-viral group hand-picked by the marketing team, a*nd they were not even part of the marketing campaign.* The marketing team was surprised to realize that a group that they had not even considered could have such relatively strong take rates.

Finally, we will briefly discuss the last category, the non-viral non-target group. Unfortunately, it is very difficult to estimate a take rate in this category, because we would need to estimate the size of the space of all prospects. This includes all of the prospects the firm knows about, as well as those customers of the firm's competitors, and consumers who might purchase this product that do not have current telecommunications service with any provider. It has been shown that the size of the communications market is difficult to estimate (Poole 2004). Nonetheless, given our best estimates of the of this category, we believe the take rates in this group to be at least an order of magnitude less than even the non-target virals.

These numbers perhaps allow us to estimate the factor of increased sales attributable to the marketing campaign. The difference between the targeted virals and the non-targeted virals is

about 10 to 1. This difference cannot all be attributed to the marketing effect, since the targeted group was specifically chosen to be better prospects, and it is likely that more of them would have signed up for the service even in the total absence of marketing. But is does seem reasonable to call this factor of 10 an upper bound for the effect of marketing.

The results here present two potentially conflicting arguments: 1) viral prospects are more likely to purchase the product, even in the absence of marketing, so there is no need to spend marketing dollars on them, or 2) viral prospects are our best prospects among all market segments, and marketing to them allows us to get the best take rates possible. We suggest that both of these are somewhat true. A potential way to capitalize on both phenomena is to market differently to the viral group, in a way that will capitalize on the possibility that they are already familiar with the product. Or, perhaps the firm could adopt strategies that would encourage their existing customers to talk more about the product, or refer a friend, through incentive programs. This would increase the pool of viral prospects, who don't necessarily need to be directly marketed to.

4.5 Improving targeting scores

We have shown that consumers who have communicated with a customer of this service do indeed have increased likelihood of adopting the service. The results of Section 4 suggest that it may be possible not only to make such an aggregate claim, but in fact to give fine-grained predictions as to which customers are more or less likely to respond to an offer. Such fine-grained predictions can be quite valuable: the consumer pool is immense, and a campaign will have a limited budget. Therefore, being able to pick a better list of "top-k" prospects will lead directly to increased profit (assuming targeting costs are not much higher for higher-ranked prospects).

In this section, we build a model which scores customers by their likelihood of responding to an offer. We show that combining the "viral attribute" with the traditional attributes improves the models in all three loyalty categories, in terms of their ability to rank customers accurately—and therefore by the profit that the models would produce

It is noteworthy that in different business scenarios, different types and amounts of data are available. For example, for low-loyalty customers very few descriptive attributes are known. We report results here using all attributes; the findings are qualitatively similar for every different subset of attributes we have tried (viz., segment, loyalty, geography, demographic).

For each targeted customer, we create a record comprising all of the loyalty, demographic and geographic attributes, and whether they were viral or not. Recall from Section 4 that this attribute indicates whether or not the consumer communicated with a service subscriber. The response variable for this prediction task is binary: whether or not the consumer adopted the product in the specified response period. We used the same logistic regression models chosen in the last section for the purpose of prediction. We measure the predictive impact by an increase in the Wilcoxon-Mann-Whitney statistic, equivalent to the area under the ROC curve (AUC). The AUC measures the quality of a ranking based on an underlying binary variable. Specifically, it is the probability that a randomly chosen taker will be ranked higher than a randomly non-taker; AUC=1.0 means the classes are perfectly separated (all the takers are at the top of the list), and AUC=0.5 means the list is randomly shuffled.

All reported AUC values are the average result using 10-fold cross-validation. Table 5 shows the AUC values for the models for models from the three loyalty groups. As it turns out, the increase in predictive value for the loyalty groups 1 (0.81 to 0.82) and 2 (0.82 to 0.84) appears minimal as compared to the improvement in the high-loyalty group 3 (0.74 to 0.83). We present the corresponding ROC curves for loyalty level 3, where we see the most significant improvement in Figure 5. Although initially this may appear in contrast to the results in the last section (where it was shown that viral has the most impact for loyalty group 1), this is due to the fact that the ROC curves represent our ability to separate sales from non-sales. Since the majority of sales were in Loyalty group 3 (see Table 4), this is the group that shows the most improvement in the ability to rank customers. Although the viral attribute is significant in the other two levels, unfortunately there are simply not enough sales in those groups to show the effect in the entire ROC curve. One way to show the predictive benefit when there are low take rates would be to do a focused analysis on the top k predicted probabilities, for some value of k, and show that the model does better at identifying the very best prospects. We hope to report on this type of analysis in future work.

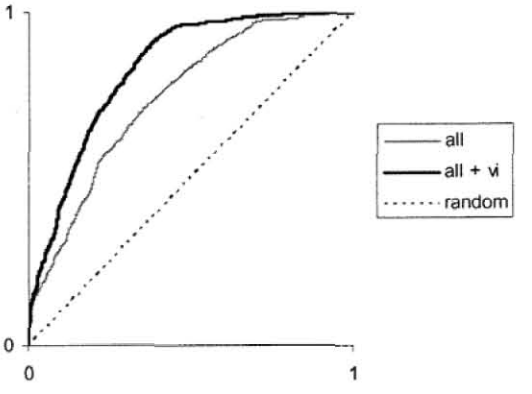
4.5.1 Network

We have already seen that the viral variable can have a significant impact in a direct marketing campaign. In this section, we describe some network-defined variables that may allow firms to further fine-tune their targeting by ranking the prospects in the viral group.

Table 5. AUC values *resulting from the application of logistic regression models built using all available attributes with and without the viral attribute. We use the two subset of attributes to predict customer response for each level of customer loyalty.*

Loyalty	all	all + viral
1	0.80	0.81
2	0.82	0.84
3	0.74	0.83

Figure 5: ROC Curve for Loyalty Group 3 *resulting from the application of logistic regression models built using all available attributes with (all +vi) and without (all) the viral attribute. The associated AUC values are 0.83 and 0.74 respectively.*



The attributes are described in Table 6. **Degree**, **Transactions**, and **Seconds of Communication** are simple attributes derived from the communication records of the prospects.

We define *influencers* as those subscribers who signed up for the service, and subsequently we see one of their network neighbors sign up for the service (We appreciate that we do not actually know if there was influence). **Connected to Influencer** is an indicator of whether the prospect is connected to one of these influencers. **Connected Component Size** is the size of the largest

Table 6 Network Attributes and their description.

Attribute	Description
Degree	Number of unique customers communicated with before the mailer
Transactions	Number of transactions to/from customers before the mailer
Seconds of communication	Number of seconds communicated with customers before mailer
Connected to influencer	Is an influencer in your local neighborhood?
Connected component Size	Size of the connected component target belongs to.
Similarity (structural equivalence)	Max overlap in local neighborhood with existing customer

network connected component the prospect is connected to. *Similarity* is defined as the size of the overlap in the neighborhoods of the prospect and the customer: **Max Similarity** is the maximum of this value across all neighbors of the prospect.

We built predictive models on the viral customers only (because they are the only customers that will have positive values for the network attributes) using all of the aforementioned attributes, and show AUC values for these predictive models in Table 8. We find that the network attributes do have predictive power individually, and have even more value when combined with all other customer attributes.

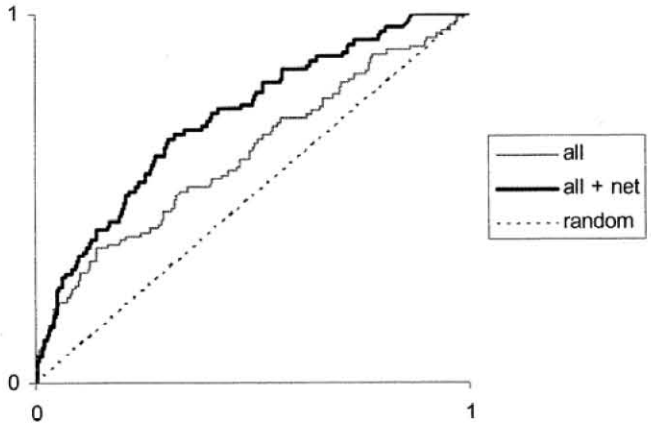
Our result is confirmed by the ROC curve in Figure 5. The main conclusions are that the more one talks on the network and the more friends they have on the network, the more likely they are to take the product. This is indicated by AUC of .68 for both Transactions and Seconds of Communication. The low values of Connected Component Size, Similarity and Connected to Influencer indicate that these variables might not have as much predictive power

Interestingly, when we add all of the other attributes to the network attributes, there is no gain in AUC. Although many of these variables were shown to be significant in the larger analysis before, remember that this analysis is done only on the viral customers. For these viral customers, we know that the network is important, and these results suggest that the network information might be all that we need to rank them for targeting, and that traditional demographics and other marketing variables do not add value over and above the network.

Table 7. AUC Values resulting from logistic regression models built on constructed network attributes. *Results are presented for loyalty level 3 customers.*

Attribute	AUC
Transactions	.68
Seconds of Communication	.68
Degree	.59
Connected to Influencer	.53
Connected Component Size	.55
Similarity	.55
All network	.70
All other	.61
All other + all network	.70

Figure 7: ROC Curve for Network attributes for Loyalty Group 3 *resulting from the application of logistic regression models built using all available attributes with (all + net) and without (all) the viral attribute. The associated AUC values are 0.61 and 0.70 respectively.*



5. LIMITATIONS

We believe our study is the first to combine direct customer communication with product adoption in order to show the effect of viral marketing. However, there are several limitations in our study that are worth mentioning. There are several types of missing, incomplete, or unreliable data which could influence our outcomes. First, due to the nature of our data, we can see all of the transactions to and from the people who have already purchased the service, but that is not true of the viral candidates, who have not yet purchased the service. As such, we do not have complete information about the viral targets (as well as the non-viral targets) for the modeling task.

Some of the variables we used were collected by purchasing customer data from external sources. These data are known to be at least partially erroneous and outdated, although it is not well known quite how much so. An additional problem is joining data on customers from the external sources to internal communication data, leading to missing data or perhaps just blatantly incorrect data. Telecommunications companies are not legally able to collect information regarding the actual content of the call, so we were not actually able to determine if the consumers in question were actually discussing the product. In this regard, our data is inferior to some other domains where content is visible, such as Internet bulletin boards, or product discussion forums.

We expect the viral effect to manifest itself differently for different types of products. Most of the studies done to date on viral behavior have focused on the types of products that people are likely to talk about, such as a new, high-tech gadget or a recently released movie. We expect there to be less buzz for less "sexy" products, like a new deodorant, or a sale on grapes at the supermarket. The study presented in this paper involves a new telecommunications service, which involves a new technology and features that consumers have perhaps never been exposed to before. The firm hopes the new technology and features are such that they would encourage word of mouth. But what can we say about other products that might not be quite so buzz-worthy?

In order to study this, we compared the new service studied here to a rollout of another product by the same firm. This other "product" was simply a new pricing plan for an older telecommunications service. Customers who signed up for this new plan could stand to save a significant amount of money, depending on their current usage patterns. However, the range and variety of telecommunications pricing plans out on the marketplace is so extensive, and so confusing to the typical consumer, that we do not believe that this is the type of product that would generate a lot of word-of-mouth discussion between communicators. As such, this is a good product to compare to the new communications service analyzed earlier.

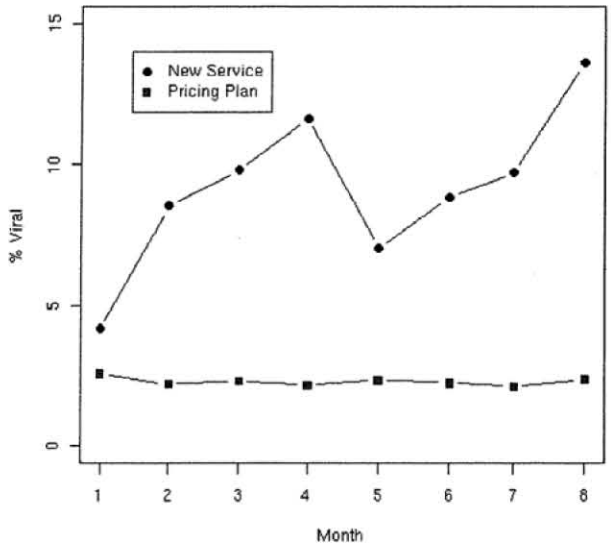
We refer to the two products as the "pricing plan" and the "new technology". For the pricing plan, we have the same knowledge of the network as we do for the new technology. For those consumers that belong to the pricing plan, we know who they communicate with, and then we can follow these viral candidates to see if they ultimately sign up for the plan. We construct a measure of "viral-ness" as follows. For a series of consecutive months, gather data for all of the customers who ordered the product in that month. Calculate the percentage of these new customers that were viral, i.e. those that had previously communicated with a user of the product. The higher this percentage, the higher the amount of new customers that purchases the product. Perhaps this result is due to some viral effect. By comparing these percentages, we can see the relative impact of the viral effect over other effects which may cause someone to purchase a product.

We now look at this value for our two products over an eight month period. The time period for the two products was chosen so that it would be within the first year after the product was broadly available. The results are shown in Figure 6. The two main points to take away are that the New Service has a higher percent of purchasers that are viral, and also an increasing one (except for the dip in month 5). In contrast the Pricing Plan has a flat viral percentage, never increasing above 3%.

Interestingly, the dip in the plot for the New Service corresponds exactly to the month that we did the direct marketing campaign discussed earlier. Before the campaign, we can see that the viral effect was increasing, that more and more of the purchasers in a given month were viral. During the mass marketing campaign, we exposed many non-virals to the service, and many of them ended up purchasing it, temporarily dropping the viral percent. After the campaign, we see the viral percent starting to increase again.

This viral-ness measure should not be confused with the success of the product, as the Pricing Plan was actually quite successful from a sales perspective. But it does suggest that the Pricing Plan is a product that generates less word of mouth than the high-tech New Service. This type of analysis can be done to give a preliminary indication of which products are worth trying to generate word of mouth among customers.

Figure 6: Viralness Plot for New Service vs. Pricing Plan.



6. DISCUSSION

One of the main concerns for any firm is when, how and to whom they should market their products. Based on how much a firm knows about their target customer and potential customers, they may choose to mass market when they don't know much or to target market based on some desirable observed characteristics of current or potential customers or, more recently based on the network that they may have influence on. We take a network marketing approach to this problem and provide evidence on real world data that there is indeed information in communication links.

We have chosen to use data from a telecommunications firm because such firms collect network data generated from the observed communication between their customers and between their customers and other consumers. We chose a large diversified provider of telecommunication services, because it provides a variety of business contexts for taking advantage of the data. This combination facilitates the investigation of how we obtain value from consumer networks. However, this combination also provides considerable challenges. The entire network is extremely large, with data on hundreds of millions of consumers, with hundreds of millions of interactions daily, and representing dynamic consumer behavior that is not well understood.

Current data-driven marketing techniques involve associating customers based on descriptive data, for example demographics-based data mining or content-based recommendation, and/or transaction data, for example RFM modeling or collaborative filtering. We add to these a third type of data: direct communication between consumers as measured by a connection in our network.

The results of the study show that adding information from the network improves targeting consistently and considerably, when compared to alternative techniques that do not use the network, but that use a wide variety of other data (demographics; prior purchase data; hand-crafted problem-specific segments, etc.)

Our preliminary results indicate, we can benefit from the use of social networks to predict purchases. It is tempting to argue that we have shown that customers discuss the product, and that discussion helps to improve take rates. However, viral marketing is not the only possible explanation for our result. Some other theories include homophily and collective choice. The theory of homophily suggests that most communication will occur between individuals who are alike across attributes, such as demographic variables, beliefs, values, and consumer choice. Therefore, homophily would suggest that the viral attribute allows us to find customers who are similar across some or all of our attribute subsets. In other words, if homophily is true, then even if there is no direct discussion about the product (word of mouth), then the effect that we are seeing is simply the network showing us who is similar. In that case it is the similarity, not the communication that is the cause of the so-called viral effect. Since our model includes many of the variables which indicate similarity, we might argue that our analysis is a strong indicator that homophily is not the driver of the viral effect. However, we never can be sure that there is some exogenous variable which is really the cause of the viral effect that we have not accounted for.

Whether or not this is evidence of word-of-mouth or homophily is interesting from a research perspective, but does not necessarily bear on the importance to the firm -- for example, if the reason is purely homophily based on some hidden variable, the firm can still use the network to improve marketing.

Another point of discussion is that when interested in the effect of a treatment of a marketing campaign on sales rates, true randomization is difficult or impossible. One way to address this problem is to use propensity score matching. Propensity score matching matches members of different groups based on a range of characteristics. Under certain assumptions, comparisons of the

matched groups reveal the true impact of the treatment of interest. In addition, in the case presented here, where we have relatively low take rates, a good way to compensate is to take a matched sample for both the viral and non viral customers. Propensity scores are well suited for this, and we plan to use propensity scores in future work.

7. ACKNOWLEDGEMENTS

We would like to thank DeDe Paul and Deepak Agarwal of AT&T, as well as Chris Dellarocas of the University of Maryland for useful discussions and helpful comments and suggestions.

REFERENCES

Alessie, R. and A. Kapteyn (1991). "Habit Formation, Interdependent Preferences and Demographic Effects in the Almost Ideal Demand System." Economic Journal **101**(406): 404-419.

Anderson, E. W. (1998). "Customer Satisfaction and Word of Mouth " Journal of Service Research **1**(1): 5-17.

Baker, P. (2005). Word of mouth advocacy: Right people, right message. Alternative Advertising and Marketing Conference Melbourne

Bass, F. (1969). "A New Product Growth Model for Product Diffusion." Management Science **15**: 215-227.

Bass, F. M., T. V. Krishnan, et al. (1994). "Why the Bass Model Fits without Decision Variables." Marketing Science **13**(3): 203-223.

Bowman, D. and D. Narayandas (2001). "Managing customer-initiated contacts with manufacturers: The impact on share of category requirements and word-of-mouth behavior." Journal of Marketing Research **38**(3): 281-297.

Bronnenberg, B. J. and V. Mahajan (2001). "Unobserved retailer behavior in multimarket data: Joint spatial dependence in market shares and promotion variables." Marketing Science **20**(3): 284-299.

Case, A. C. (1991). "Spatial Patterns in Household Demand." Econometrica **59**(4): 953-965.

Darrough, M. N., R. A. Pollak, et al. (1983). "Dynamic and Stochastic Structure - an Analysis of 3 Time-Series of Household Budget Studies." Review of Economics and Statistics **65**(2): 274-281.

Dellarocas, C. (2004). The Digitization of Word-of-Mouth: Promise and Challenges of Online Feedback. Massachusetts Institute of Technology (MIT), Sloan School of Management.

Dellarocas, C., N. F. Awad, et al. (2004). Using Online Reviews as a Proxy of Word-of-Mouth for Motion Picture Revenue Forecasting.

Dodds, W. (1973). "Application of Bass Model in Long-Term New Product Forecasting." Journal of Marketing Research **10**(3): 308-311.

Domingos, P. (2005). "Mining social networks for viral marketing." Ieee Intelligent Systems **20**(1): 80-82.

Domingos, P. and M. Richarson (2001). Mining the Network Value of Customers. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining

Duesenberry, J. S. (1949). Income, saving, and the theory of consumer behavior. Cambridge,, Harvard University Press.

Evans, R. H. (1995). "Incorporating Attitudes and Imitation in the Bass Model of Diffusion." Psychological Reports **77**(3): 1043-1048.

Fildes, R. (2003). "New-product diffusion models." International Journal of Forecasting **19**(2): 327-328.

Gladwell, M. (1997). The Coolhunt New Yorker. The New Yorker**:** 78-88.

Gladwell, M. (2002). The tipping point : how little things can make a big difference. Boston, Back Bay Books.

Godes, D. and D. Mayzlin (2004). "Using online conversations to study word-of-mouth communication." Marketing Science **23**(4): 545-560.

Hawkins, D. I., R. J. Best, et al. (2004). Consumer behavior : building marketing strategy. Boston, McGraw-Hill Irwin.

Hosmer, D. W. and S. Lemeshow (1989). Applied logistic regression. New York, Wiley.

Huang, E. M., M. Terry, et al. (2002). "Distributing event information by simulating word-of-mouth exchanges." Human Computer Interaction with Mobile Devices **2411**: 60-68.

Huang, Z., W. Y. Chung, et al. (2004). "A graph model for e-commerce recommender systems." Journal of the American Society for Information Science and Technology **55**(3): 259-274.

Kautz, H., B. Selman, et al. (1997). "Referral web: Combining social networks and collaborative filtering." Communications of the Acm **40**(3): 63-65.

Kempe, D., J. Kleinberg, et al. (2003). Maximizing the Spread of Influence through a Social Network. SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining.

Kumar, V. and T. V. Krishnan (2002). "Multinational diffusion models: An alternative framework." Marketing Science **21**(3): 318-330.

Lilien, G. L., Arvind Rangaswamy, and Christophe Van den Bulte (2000). Diffusion Models: Managerial Applications and Software. New-Product Diffusion Models. E. M. Vijay Mahajan, and Jerry Wind. Boston Kluwer**:** 295-336.

Mahajan, V., E. Muller, et al. (1990). "New Product Diffusion-Models in Marketing - a Review and Directions for Research." Journal of Marketing **54**(1): 1-26.

McCullagh, P. and J. A. Nelder (1983). Generalized linear models. London ; New York, Chapman and Hall.

Montgomery, A. L. (2001). "Applying Quantitative Marketing Techniques to the Internet." Interfaces **3**(2): 90-108.

Moore, G. A. (1999). Crossing the chasm : marketing and selling high-tech products to mainstream customers. New York, HarperBusiness.

Newton, J. and R. Greiner (2004). Hierarchical Probabilistic Relational Models for Collaborative Filtering Statistical Relational Learning and its Connections to other Fields.

Niu, S. C. (2002). "A stochastic formulation of the Bass Model of new-product diffusion." Mathematical Problems in Engineering **8**(3): 249-263.

Paumgarten, N. (2005). NO. 1 FAN DEPT ACKNOWLEDGED. The New Yorker.

Pollak, R. A. (1976). "Interdependent Preferences." American Economic Review **66**(3): 309-320.

Poole, D. J. (2004). Estimating the size of the telephone universe. *Proceedings of the Tenth ACM* SIGKDD International Conference
on Knowledge Discovery and Data Mining (KDD-04), Seattle, WA, ACM.

Richins, M. L. (1983). "Negative Word-of-Mouth by Dissatisfied Consumers - a Pilot-Study." Journal of Marketing **47**(1): 68-78.

Rosen, E. (2000). The Anatomy of Buzz. New York, Doubleday.

Satoh, D. (2001). "A discrete bass model and its parameter estimation." Journal of the Operations Research Society of Japan **44**(1): 1-18.

Smith, T. E. and J. P. LeSage (2004). "A Bayesian probit model with spatial dependencies." Spatial and Spatiotemporal Econometrics **18**: 127-160.

Straffin, P. D. (1977). "Bandwagon Curve." American Journal of Political Science **21**(4): 695-709.

Ter Hofstede, F., M. Wedel, et al. (2002). "Identifying spatial segments in international markets." Marketing Science **21**(2): 160-177.

Touhey, J. C. (1974). "Situated Identities, Attitude Similarity, and Interpersonal Attraction." Sociometry **37**: 363-374.

Ueda, T. (1990). "A Study of a Competitive Bass Model Which Takes into Account Competition among Firms." Journal of the Operations Research Society of Japan **33**(4): 319-334.

Walker, R. (2004). The Hidden (In Plain Sight) Persuaders. The New York Times Magazine.

Yang, S. and G. M. Allenby (2003). "Modeling interdependent consumer preferences." Journal of Marketing Research **40**(3): 282-294.