# Mining Social Networks for Viral Marketing

**Pedro Domingos**
Department of Computer Science and Engineering
University of Washington

Traditionally, social network models have been descriptive, rather than predictive: they are built at a very coarse level, typically with only a few global parameters, and are not useful for making actual predictions of the future behavior of the network. In the past, this was largely due to lack of data: the networks available for experimental study were small and few, and contained only minimal information about each node. Fortunately, the rise of the Internet has changed this dramatically. Massive quantities of data on very large social networks are now available from blogs, knowledge-sharing sites, collaborative filtering systems, online gaming, social networking sites, newsgroups, chat rooms, etc. These networks typically number in the tens of thousands to millions of nodes, and often contain substantial quantities of information at the level of individual nodes, sufficient to build models of those individuals. Assembling these models into models of the larger network they are part of gives us an unprecedented level of detail in social network analysis, with the corresponding potential for new understanding, useful predictions, and their productive use in decision-making.

We have begun to build social network models at this scale, using data from the Epinions knowledge-sharing site, the EachMovie collaborative filtering system, and others [1, 6]. These models allow us to design "viral marketing" plans that maximize positive word-of-mouth among customers. In our experiments, this makes it possible to achieve much higher profits than if we ignore interactions among customers and the corresponding network effects, as traditional marketing does.

## The Network Value of Customers

Customer value is usually defined as the expected profit from sales to that customer, over the lifetime of the relationship between the customer and the company. Customer value is of critical interest to companies, because it determines how much it is worth spending to acquire a particular customer. However, traditional measures of customer value ignore the fact that, in addition to buying products himself, a customer may influence others to buy them. For example, if, in addition to seeing a particular movie myself, I persuade three friends to see it with me, my customer value with respect to that movie has effectively quadrupled, and the movie studio is thus justified in spending more on marketing the movie to me than it otherwise would. Conversely, if I tend to make decisions on what movies to see purely based on what my friends tell me, marketing to me may be a waste of resources, which would be better spent marketing to my friends. We call the *network value* of a customer the expected increase in sales to *others* that results from marketing to that customer.

Clearly, ignoring the network value of customers, as is done in traditional direct marketing, may lead to very suboptimal marketing decisions. But, while the existence of network effects has been acknowledged in the marketing literature, they have generally been considered to be unquantifiable, particularly at the level of individual customers. This is what is changed by the data sources now available. Our models enable us to measure the network value of a customer. For each customer, we model how probable that customer is to buy some product, as a function of both the intrinsic properties of the customer and the product, and of the influence of the customer's neighbors in the network. By performing probabilistic inference over the joint model of all the customers, we can answer questions like "If we market to this particular set of customers, what is the expected profit from the whole network, after the influence of those customers has propagated throughout?" Using this capability, we can now search for the optimal set of customers to market to, in the sense that marketing to this set will yield the highest return on investment. Intuitively, we can look for the customers with highest network value, market to them, and reap the benefits of the ensuing wave of word of mouth.

## Factors that Influence Network Value

What makes for a customer with high network value? Clearly, high connectivity in the network should help, but there are other factors, which our model identifies. First of all, it is important that the customer like the product, preferably a lot. Customers who have high connectivity but dislike a product can have negative network value, and marketing to them should be avoided. Indeed, in our experiments with the EachMovie collaborative filtering system, the fact that our model took this into account was one of the reasons it outperformed a standard direct marketing approach. The latter assumed that the most it had to lose by marketing to a customer who did not like the product was the cost of the marketing, which is typically small per customer, and thus marketed even to customers whose chances of liking the product were relatively low.

Another key aspect is that, to have high network value, a customer should influence her acquaintances more (ideally much more) than they influence her. If influence is symmetric, there is no advantage in searching for the most influential customers. Fortunately, asymmetric influence is widespread in practice, and our approach takes advantage of it. While in various fields there are well-known opinion leaders (e.g., celebrities), our approach makes it possible to identify them at the local level.

The third (and perhaps most important) aspect is that a customer's network value does not end with her immediate acquaintances. Those acquaintances in turn influence other people, and so on recursively until potentially the entire network is reached. These acquaintances should in turn like the product and have many other people they influence. A customer who is not widely connected may in fact have high network value if one of her acquaintances is highly connected (for example, an advisor to an opinion leader). In our experiments with the Epinions knowledge-sharing Web site, the most valuable customer had a network value of over 20,000, meaning that marketing to that customer was as effective as marketing to over 20,000 others in the absence of network effects, but the customer's number of direct links to others in the network (i.e., people who read his reviews) was much smaller.

One consequence of our model is that word-of-mouth marketing may not be effective in some markets, because the requisite networks of influence are not present. While this is known at a high level for some market types, many startup companies have failed by investing heavily to unleash network effects that never materialized. Conversely, trials for some products, like cash cards and interactive television, have resulted in "failure" because giving the product to a small sample of isolated customers does not allow network effects to take hold, and this was not appreciated. When the data is available, our models make it possible to measure these effects precisely and make correspondingly better decisions.

Another interesting consequence of our model is that it may pay to lose money on some customers, if they are influential enough. In traditional direct marketing, customers only receive an offer if the expected profits from them exceed the cost of the offer. In viral marketing, giving a product for free to a well-chosen customer could pay off many times in sales to other customers.

## Maximizing Word of Mouth

Given a model of a social network, we have a well-defined optimization problem: choose the set of customers to market to so as to maximize net profits (profits from sales minus the cost of marketing). Kempe, Kleinberg and Tardos have shown this problem to be NP-hard, but approximable within 63% of the optimal using a simple hill-climbing search procedure [4]. In our experiments, similar results were obtained with an even faster approach where each customer is added to the current "marketing set" as long as this improves overall profit. With careful implementation, the potentially prohibitive cost of performing probabilistic inference over the whole network at each search step, necessary to measure the effect of adding a customer to the "marketing set," also turns out to not be a problem. This is because the vast majority of customers has very small network value; their influence in the network does not propagate very far, and thus the computation for them converges quickly. For the few customers that have high network value the computation can indeed take substantial time, but amortized over all search steps it becomes quite manageable. In our experiments, the optimal marketing set for a network with tens of thousands of nodes was found in minutes.

No matter how much data we have, completely capturing the network of social interactions among people in the real world will never be feasible. Thus the important question arises of whether our approach to maximizing word of mouth still works when our knowledge of the network is incomplete. We have tested this by randomly removing a variable number of edges from the network before passing it to the data mining system, and found the system to be quite robust, with 70% of the lift in profit obtained when only 5% of the edges were known. Our model can also be used to determine the most cost-effective way to gather additional knowledge. We have found that the simple heuristic of iteratively asking the customers with the highest network value in the current network who their acquaintances are is quite effective.

## Prospects

Traditional marketing is in crisis, because customers are increasingly inured to television commercials, direct mailings, etc. At the same time, companies like Amazon, Google and Hotmail succeed with virtually no marketing, based solely on word of mouth [2]. A recent study found that positive word of mouth among customers is by far the best predictor of a company's growth [5]. Word-of-mouth marketing has the key advantage that a recommendation from a friend or other trusted source has the credibility that advertisements lack [3]. Because it leverages customers themselves to do the marketing, it can also produce unparalleled returns on investment. However, until now it has been somewhat of a black art. The goal of our work is to put it on a firmer foundation, and the results so far are very promising.

Beyond marketing, word-of-mouth optimization is potentially applicable in any setting where we desire to produce a large social outcome with only limited resources. Examples include reducing the spread of HIV, combatting teenage smoking, and grass-roots political initiatives. Until recently, sociology lagged behind other sciences in developing a computational branch; the wealth of social data provided by the Internet has the potential to change this, and our work can be seen as a step in this direction.

Needless to say, we have only begun to scratch the surface of the very rich set of possibilities opened up by building predictive, as opposed to descriptive, models of social networks. Real social networks evolve in time, have multiple types of arcs and nodes, are affected by the actions of multiple players, and can be mined from a combination of sources. Because data points are not independent and identically distributed, subtle statistical issues arise. We are currently designing a rich language for modeling these and other aspects of social networks, and developing learning and inference algorithms for it. This language, called Markov logic networks, combines the probabilistic modeling of Markov random fields and the expressiveness of first-order logic [7]. In preliminary experiments, it greatly speeded development of a complex social network model, and yielded more accurate predictions than standard methods.

We are all familiar with the notion that a butterfly flapping its wings in Beijing can cause a storm in New York. At the same time, the chances that a given butterfly flapping its wings will indeed cause a storm in New York are very small. Our approach, in a nutshell, is to ask: "If we wanted to cause a storm in New York, and could make a few butterflies flap their wings, which ones would we choose?" Our experiments so far show that, at least in the world of marketing, this is an effective way to unleash storms on demand.

## References

[1] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 57–66, San Francisco, CA, 2001. ACM Press.

[2] R. Dye. The buzz on buzz. *Harvard Business Review*, 78(6):139–146, 2000.

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

[4] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence in a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, Washington, DC, 2003. ACM Press.

[5] F. Reichheld. The one number you need to grow. *Harvard Business Review*, 81(12):47–54, 2003.

[6] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 61–70, Edmonton, Canada, 2002. ACM Press.

[7] M. Richardson and P. Domingos. Markov logic networks. Technical report, Department of Computer Science and Engineering, University of Washington, Seattle, WA, 2004. http://-www.cs.washington.edu/homes/pedrod/mln.pdf.