

Evolving Viral Marketing Strategies

Forrest Stonedahl[†], William Rand[‡], Uri Wilensky[†]

[†]Center for Connected Learning and Computer-Based Modeling
Northwestern University, Evanston, Illinois, USA

[‡]Department of Marketing & Department of Computer Science
Center for Complexity in Business

University of Maryland, College Park, Maryland, USA

forrest@northwestern.edu, wrand@umd.edu, uri@northwestern.edu

ABSTRACT

One method of viral marketing involves seeding certain consumers within a population to encourage faster adoption of the product throughout the entire population. However, determining how many and which consumers within a particular social network should be seeded to maximize adoption is challenging. We define a strategy space for consumer seeding by weighting a combination of network characteristics such as average path length, clustering coefficient, and degree. We measure strategy effectiveness by simulating adoption on a Bass-like agent-based model, with five different social network structures: four classic theoretical models (random, lattice, small-world, and preferential attachment) and one empirical (extracted from Twitter friendship data). To discover good seeding strategies, we have developed a new tool, called BehaviorSearch, which uses genetic algorithms to search through the parameter-space of agent-based models. This evolutionary search also provides insight into the interaction between strategies and network structure. Our results show that one simple strategy (ranking by node degree) is near-optimal for the four theoretical networks, but that a more nuanced strategy performs significantly better on the empirical Twitter-based network. We also find a correlation between the optimal seeding budget for a network, and the inequality of the degree distribution.

Categories and Subject Descriptors: I.6.3 [Simulation and Modeling]: Applications, J.1 [Administrative Data Processing] – Marketing

General Terms: Management, Economics

Keywords: Viral Marketing, Agent-Based Modeling, Diffusion, Social Networks, Business, Genetic Algorithms

1. MOTIVATION

Viral marketing, or word-of-mouth marketing, is based on the idea that consumer discussions about a product are

more powerful than traditional advertising. One way to encourage positive word-of-mouth is by distributing reduced or free products to target consumers who will then discuss the product with their friends and encourage those friends to buy the product. However, whom to seed with these initial products in order to maximize the amount and rate of product adoption is not obvious. Given an arbitrary social network and a limited seeding budget, choosing the optimal seeding locations has been shown to be an NP-Hard problem [16]. Furthermore, it is not clear what the proper seeding budget should be for a particular network. Assuming that the product is beneficial and that seeded consumers are inclined to speak positively about it, seeding more consumers will increase the speed of product adoption. However, giving away more free products increases the overall expense of the promotional campaign. In addition, seeded consumers are removed from the pool of potential customers, which may decrease total revenue for the product. Thus, it is important to choose both the correct target consumers to seed and correct seeding budget to maximize adoption.

This problem has direct implications for real-world marketing managers. The growth of YouTube, Twitter, Facebook, and other digital social media capabilities, has given marketing managers a new platform by which to advertise and market their products to consumers. The compelling aspect of these platforms is that they encourage consumers to develop online social networks which provide a formalization of the social interactions of individuals. However, despite the power of this new media it has been difficult for marketing managers to use this platform successfully [2]. In many cases, due to privacy considerations, the full network described by these social media applications is not known, so advertisers are forced to rely on third party information about the consumers they are targeting.

To account for the challenges that marketers face, we propose a version of the general viral marketing problem, which we call the *local viral marketing problem* or *LVMP*. We will first overview related research, then formally define the LVMP, and discuss the agent-based model we use for simulating adoption and the five networks we will test it on. We propose a range of strategies to solve the LVMP, then discuss experimental results from exploring this strategic space using a new evolutionary tool (*BehaviorSearch*), and conclude with recommendations for future work.

2. RELATED WORK

Recently there has been work on viral marketing from two different disciplines, computer scientists, and marketing

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GECCO'10, July 7–11, 2010, Portland, Oregon, USA.

Copyright 2010 ACM 978-1-4503-0072-8/10/07 ...\$10.00.

researchers. Originally introduced to computer science by Domingos and Richardson [8], the problem was formalized by Kempe, Kleinberg, and Tardos [16] who described the problem as selecting the correct individuals to seed with a product in an arbitrary network given a fixed marketing budget. They showed that their formalization of this problem is in fact NP-hard, but presented some heuristic solutions to the problem, with some provable approximation guarantees. However, their best approximation algorithm requires global knowledge of the network; in other words, in order to be implemented the marketing manager would need to know every node in the network and how it is connected to every other node; unfortunately, this is an unrealistic requirement in many real-world cases.

Leskovec, Adamic, and Huberman [17], on the other hand, take a descriptive approach to viral marketing. Similarly within marketing research, Goldenberg, Libai and Muller [13] use a cellular automata model to describe adoption processes and characterize which individuals have the greatest effect on adoption. Goldenberg and others have also examined the role of hubs (individuals with a high number of friends) in the adoption process [11]. Other marketing researchers have explored how innovations diffuse across a variety of different topologies [23], and how word-of-mouth affects product adoption [6, 22]. In contrast to this previous work, our goal is to make prescriptive suggestions for seeding within viral marketing campaigns, but at a knowledge level that could be available to marketing managers.

To accomplish this task, we use a genetic algorithm (GA) [14] to search for optimal (or high-performing) strategies in the space of possible consumer seeding strategies. Our task is equivalent to the problem of optimizing the parameters of a multi-agent simulation, where the parameters control the seeding strategy. In a different context, one of the earliest uses of a GA was to characterize the parameters of a cell simulation [30]. Later, Miller proposed the use of nonlinear optimization techniques for a variety of model exploration and testing tasks, dubbed as “active nonlinear testing” or ANT [20]. Calvez and Hutzler used a genetic algorithm for several parameter search/calibration tasks in an agent-based model of ant food foraging [4]. Within the marketing domain, Midgley, Marks, and Kunchamwar [19] have used a genetic algorithm to examine agent-based models in a consumer retail environment. Building on this research, we have constructed a general tool, which we call *BehaviorSearch* [24], for using evolutionary computation to explore the parameters of agent-based models created using the NetLogo agent-based modeling toolkit [31].

3. LOCAL VIRAL MARKETING PROBLEM

The global viral marketing problem (GVMP) consists of selecting a group of individuals who will be seeded with a product in order to encourage their friends to adopt a product at a quicker rate than they normally would have. The problem assumes that there is a graph G , of vertices and edges, where each vertex is a consumer in the network and each edge represents a social connection between two vertices. In addition to the social network, there is also an adoption function, $f_i(t)$, which specifies the likelihood that a vertex, i , will adopt a product at time t , given the adoption state of its immediate neighbors. For the purposes of the results presented herein, the adoption function $f_i(t)$ is

assumed to be the same for all individuals, so we will use the notation $f(t)$.

In order to simultaneously consider both the amount and rate of adoption, we will use the notion of the net present value (NPV) of an adoption network [12]. Intuitively, the NPV measure accounts for the fact that it is worth more to a company if people buy its product now, rather than several months from now, especially since new competing products may enter the marketplace. The NPV , given an adoption function ($f(t)$), social network (G), and seeded vertices (S), is the sum of vertices that adopt the product multiplied by the profit from the product and a discount factor for time of adoption, specifically:

$$NPV(G, S, f(t)) = \sum_{t=0}^{\infty} a(t)p\lambda^t$$

where $a(t)$ is the number of adopters at time t , p is the profit for adoption of a product, and λ is the discount factor. In our experimental results, we chose a 10% discount rate ($\lambda = 0.9$), which has previously been used in related marketing literature [12, 18]. This discount rate represents the cumulative effect of several factors, including the opportunity cost of not having the money earlier and the potential necessity to lower prices over time to stay competitive.¹ The fully specified GVMP is to identify a set of vertices S that will maximize the network’s NPV , given that $|S| \times c \leq b$, where c is the cost of seeding one vertex, and b is a specified budgetary constraint.

The local viral marketing problem (LVMP) is similar to the GVMP, except that we remove knowledge of the structure of the global network (G), instead offering only characteristics of each vertex which provide summary statistics about the vertex and its role in the network. There are many different network measures that could be chosen [26]; in Section 4.3 we will describe the specific measures we used, but one example measure is the vertex degree (i.e., the number of neighbors). Specifically, the problem is to find a weighting function, $w(i)$, that determines where to place vertex i in a priority queue for seeding. Once the queue has been created, vertices to be seeded are chosen in rank order, until the budget (b) is exhausted. Also, in contrast to the GVMP, in our formulation of the LVMP we allow b to be varied as part of the strategy, which includes finding an optimal budget amount as part of the problem definition. Thus, we define a seeding strategy, S , to be a weighting function $w(i)$ together with a specified budget, b , as this is sufficient information to seed an arbitrary network.

Our examination of the LVMP is arguably more relevant to the real-world than the GVMP for a number of reasons. As discussed in Section 1, often the best budgetary value to use for viral marketing seeding is unknown. Moreover, in many real-world cases the global social network is also unknown. In face-to-face interactions, no one knows the full network of any reasonably sized market, and even in the case of social networking web sites, privacy constraints may prevent access to the whole network (e.g., Facebook), or data collection limitations may be prohibitive (e.g., Twitter). Even in cases where data is available, running simulations on the entire network to determine the optimal seeding

¹Preliminary comparison experiments suggest that using other reasonable discount values is unlikely to change our qualitative results.

strategies would be computationally difficult, if not impossible. Solving the LVMP for realistic networks of moderate size could provide marketing managers with a way to specify solutions that are not reliant on global network knowledge. Moreover, since the LVMP strategies are specified in a generalizable way that is not dependent on a particular network structure, they may facilitate learning of solutions that perform well across a variety of network architectures. Finally, solutions to the LVMP could be used to drive new business models. If the role of an individual in diffusion is known, then social media platforms, such as Facebook, or intermediaries who work with these platforms, such as a third-party advertising firm, could charge different premiums to brands for advertising to different types of consumers, based on the consumers network characteristics. For instance, they might charge more for an advertising campaign targeting well-connected users than for a campaign using random sampling. Solutions to the LVMP would provide a way to quantify the differential utility, and appropriately price these campaigns.

In this paper, we specifically address these questions: How do different social networks affect the optimal seeding budget and strategy? Does providing a complex strategy space yield better solutions than simple strategies? How robust are LVMP strategies to different adoption “virality” levels?

4. THE MODEL

In order to investigate the LVMP, we must specify a model for the diffusion of products throughout the network. Specifically we must describe an adoption function, $f(t)$, the network structure, G , and the strategy space, S .

4.1 Adoption Function

There are at least two classes of product adoption function that have been examined, Bass-like models [21, 13] (sometimes called “cascade” models), and “threshold” models [27, 28]. In the Bass-like model (so-called because of its resemblance to the aggregate-level Bass model [3]), the adoption decision consists of two factors, whether to adopt due to individual innovation, and whether to adopt due to peer imitation. In a “threshold” model, each individual adopts only if the fraction of their neighbors that has adopted is above a certain threshold. We will use a Bass-like adoption function that is the most immediate translation of the aggregate Bass model to an individual level and is an example of an independent interaction model that has been previously examined in similar forms [11, 28, 23, 25]. In our model, the heuristic for adoption of individual i can be written as, $f(t) = p + q(\frac{n_a(t)}{n})$ where p is the effect of external influences on adoption, q is the effect of social influences on adoption, n is the number of neighbors of i , and $n_a(t)$ is the number of neighboring vertices who have already adopted the product at time t . This adoption function has previously been shown to validate well against empirical data [21].

In the present work, we examine two different diffusion scenarios: a ‘medium virality’ scenario ($p = 0.01535$ and $q = 0.455$) and a ‘high virality’ scenario ($p = 0.0007$ and $q = 0.53$), which are at the middle and extremes (respectively) of empirically observed values [5]. We do not examine a ‘low virality’ scenario (high p and low q), since the dominance of individual adoption over peer-based word-of-mouth minimizes the network-effects that interest us, and viral marketing does not significantly affect adoption.

Different values of p and q may be seen to represent different types of products. A product which has a high p relative to other products is one that consumers will naturally adopt on their own; this could represent a product which is just clearly useful, such as a refrigerator. A product which has a high q relative to other products is one which consumers are more likely to adopt if many of their friends have adopted; this could represent a product with considerable network efforts, such as a fax machine, or a product which encourages social discussions, such as the Flixster Facebook app for sharing movie recommendations.

4.2 The Networks

In the experimental results below, we investigate four abstract networks created using network generation routines from the social network literature, along with one empirically derived network. In all five cases, the number of nodes² in the network is exactly 1000. For the generated networks, we also chose parameters that would yield a similar³ edge density to that of the empirically derived network. Specifically, the networks are:

1. **random** - an Erdős-Renyi random graph [9], with a uniform probability ρ of an edge being present between any two vertices ($\rho = 0.26712$ in the results below).
2. **lattice** - a regular network, where each node in the network is located on a circle and connected to a particular number of neighbors (26 in the results below) on either side of them.
3. **small-world (sw)** - this network is generated by starting with a lattice network, and randomly rewiring some of the edges as described in [29] (in the results below, we used a degree of 26 and a rewiring probability of 0.01).
4. **preferential attachment (pa)** - this network is generated with the preferential attachment mechanism described in [1]. Nodes are incrementally added to the network and connect in a way that is preferentially biased toward individuals who already have many connections (in the results below, 14 connections created per added node).
5. **twitter** - this network (shown in Figure 1) was extracted from data available via the public Twitter API. It represents a small connected subgraph of the complete Twitter social network. Starting with a random Twitter UID between 1 and 10 million, we used breadth-first search to add the 999 nodes closest to our starting node, and all friendship links (13,343 in this case) between these nodes. (Note: we define A and B as friends when A “follows” B and B “follows” A).

The *lattice* and *random* networks are not realistic social networks, but they are used for comparison purposes, as well-studied examples of extreme order and disorder (respectively). The small-world (*sw*) and preferential attachment

²The terms *graph*, *vertex*, and *edge* come from graph theory, whereas *network*, *node*, and *link* are often used in network science – we will use these terms interchangeably.

³Matching the exact number of edges is not possible with these network generation algorithms, but reasonably similar edge densities were obtained.

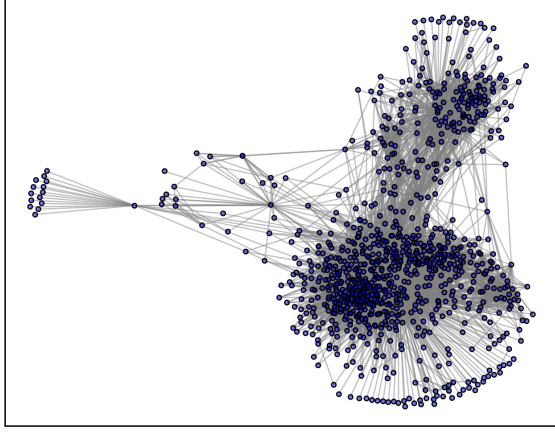


Figure 1: Visualization of the *twitter* network.

(*pa*) networks have been shown to model certain types of social and constructed networks fairly well [1, 29]. The *sw* network has a high level of clustering, while maintaining a short average path length. The *pa* network exhibits a power law (or scale-free) relationship between the degree of nodes and their frequency of occurrence. The *twitter* network provides an example from a real digital social network. It displays a more skewed degree distribution than even the *pa* network, indicating that a very small number of individuals have a disproportionately large number of social connections.⁴

4.3 Strategies

In order to evolve solutions for the *LVMP*, we need to define the search space for optimal strategies. In Section 3, we define a strategy to consist of two elements: the budget b , which we will operationalize as the fraction of the total network to be seeded, f_s , and a priority weighting function $w(i)$. For the experiments presented here, we will assume no additional cost c for seeding a node beyond the loss in potential profit p that would otherwise have been gained from a node if it had adopted, thus the budget cost b is reflected in the ineligibility to adopt the product of the initial $f_s \times n$ seeded nodes (where n is the size of the network). This is a generally optimistic view of seeding costs, but may be realistic for digital media products, where after the sunk development costs, the marginal production cost is near 0. A useful weighting function for determining seeding priority requires information about individuals. In this work, we will assume knowledge is available about several characteristics of nodes in the network, illustrated by the following five simple weighting functions:

1. *degree* - the number of neighbors of the target node normalized by the maximum possible value, i.e., $w_d(i) = \frac{\text{degree}(i)}{\max(\text{degree})}$. Higher degree nodes influence more neighbors, directly encouraging more adoption.
2. *twostep* - a measure corresponding to the number of nodes that are reachable within two steps of the given

⁴This may partially be an artifact of our subgraph extraction method, but the degree distribution of the complete Twitter network is likely to be similarly skewed.

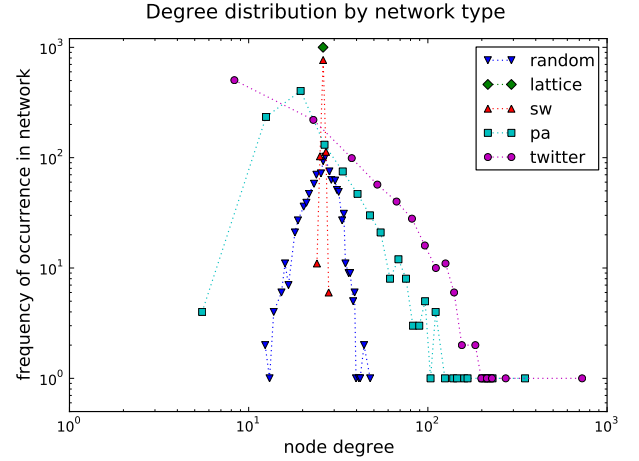


Figure 2: Degree distributions for each network, displayed on a log-log plot. As the precise shape is dependent on binning choices, this histogram is meant only to give a general sense of the degree distributions. The dotted lines serve only to guide the reader between data points.

node (by following edges in the network), $w_t(i) = \frac{\text{twostep}(i)}{\max(\text{twostep})}$. An extension to the degree measure.

3. *average path length (apl)* - the average number of steps from this node to any other node subtracted from the maximum average path length of any node in the network and normalized: $w_a(i) = \frac{\max(\text{apl}) - \text{apl}(i)}{\max(\text{apl})}$. Nodes with lower average path length are better connected to the entire graph, potentially encouraging adoption.
4. *clustering coefficient (cc)* - 1.0 minus the fraction of neighbors of the node whose neighbors are also neighbors of the target node, normalized by the highest clustering-coefficient in the network, i.e., $w_c(i) = 1.0 - \frac{\text{cc}(i)}{\max(\text{cc})}$. The lower the clustering coefficient of a node, the less overlap there is among its neighbors, encouraging wider adoption more quickly.
5. *random* - the priority of an individual is determined randomly, i.e., $w_r(i) = U[0, 1]$.

(Note that each weighting function is normalized so that values fall within the range of [0, 1], and higher values of $w(i)$ will correspond to a better ranking in the priority queue, and that ties will be broken randomly.) In past work on the *GVMP*, the degree and *apl* have been shown to be important factors, while random seeding performs poorly [16].

We hypothesize that better solutions (using the same available information) than these 5 simple strategies may be possible if the strategies are employed together. Thus we consider weighting functions that use a linear combination of the strategies above:

$$w_{\text{comb}}(i) = \alpha_d w_d(i) + \alpha_t w_t(i) + \alpha_a w_a(i) + \alpha_c w_c(i) + \alpha_r w_r(i) \quad (1)$$

where the α 's express the normalized weights assigned to each of these various characteristics of the node. Finally,

a linear combination might still not be expressive enough; what if it were better to alternate seeding between two different strategies? For instance, first seed the highest degree node, then the node with the lowest normalized path length, and back and forth until the budget is exhausted. Therefore we expand our space to include “mixed strategies”, consisting of two sub-strategies, along with an additional parameter for how often each substrategy should be used. This gives us our final $w(i)$ function, which is:

$$w(i) = \begin{cases} w_{comb,1}(i) & \text{if } x < p \\ w_{comb,2}(i) & \text{otherwise} \end{cases} \quad (2)$$

where $w_{comb,1}$ and $w_{comb,2}$ are both of the form described in Equation 1 with their own α 's, x is a random variable drawn⁵ from $U[0,1]$, and p is the parameter which specifies the probability with which $w_{combined,1}$ is to be used. Without loss of generality, we restrict $p > 0.5$, meaning that $w_{comb,1}$ will always be the primary sub-strategy and $w_{comb,2}$ is the secondary sub-strategy (chosen less often for seeding). Given this space, we can now describe an individual in the population of our genetic algorithm. Each individual will specify weights for all the α values described above (ten different values, five for each of the two strategies), a p which is the probability with which the first strategy is used, and f_s which is the fraction of the population to seed. This results in 12 real-valued genes for each individual, which is not especially many, yet the search space is too large for a brute-force approach. Also, given the complexity and stochasticity of the fitness function, we speculate that the space will be highly nonlinear, and there will be noise in the fitness determination (discussed below). These factors motivate our choice of genetic algorithms for exploring this problem.

5. IMPLEMENTATION

In order to explore the LVMP, we constructed an agent-based model of it using NetLogo [31]. In the model, we first create a number of agents (1000 for the experiments presented herein), and then we connect them according to one of the social network topologies described in Section 4.2. Then we take the strategy currently being investigated, and we sort the list of all agents using Equation 2. After this we select the fraction of agents at the top of this priority queue using the f_s specified by the strategy, and we seed each of these agents with the product (setting their adoption state to true). Then at each time step of the model, every agent who has not adopted the product runs the decision rule described in Section 4.1 to decide if they will adopt the product. Once all the agents have decided whether to adopt the product in a particular time step, we record the total number of consumers who have adopted and we begin the next time step. In our experiments, we make the simplifying assumption that the product has some appeal to every agent in the population, thus the simulation ends once all agents have adopted, and we calculate the NPV of the current run. Since the adoption heuristic is stochastic and the seeding strategy may be stochastic, we run the simulation multiple times to more accurately calculate the expected NPV for a given strategy. Specifically, an individual's fitness is the average NPV from 10 simulations (with different random seeds). While this Monte Carlo averaging cannot eliminate noise, in practice we found that using

⁵ x is only drawn once per seeding choice

10 replicates sufficiently reduced the noise so the GA could progress toward good solutions. Moreover, GAs are often successful despite the presence of noise or uncertainty [15].

To automate the process of exploration, we have created a tool called *BehaviorSearch* [24] that interfaces with NetLogo, and which can run a genetic algorithm over the parameters of any NetLogo simulation. In this case, the parameters of our model correspond to the seeding strategy to be evaluated. The genetic algorithm used is reasonably simple: we generate an initial population consisting of 50 random individual strategies, each containing 12 different genes as described in Section 4.3. The numeric values that make up a strategy are discretized at a resolution of 0.01, and encoded as a binary string, using a Gray code⁶.

The fitness of each individual is evaluated by decoding the binary string into the 12 strategy parameters, initializing the agent-based model with these parameters, and observing the mean NPV from 10 independent replications of the simulation. Using these fitness values, BehaviorSearch performs a standard generational GA [14] evolution step (70% one-point crossover, 1% mutation rate, tournament selection with tournament size 3) on the population. This process is repeated for 200 generations. For both the ‘medium’ and ‘high’ virality scenarios, we used BehaviorSearch to conduct multiple instances (30) of these searches on each of the five different networks (Section 4.2); this resulted in a total of $2 \text{ scenarios} \times 5 \text{ networks} \times 30 \text{ searches} \times 50 \text{ individuals} \times 200 \text{ generations} = 3 \text{ million fitness evaluations}$. As each fitness evaluation requires averaging 10 runs, the grand total is 30 million simulation runs, which took approximately 11,000 hours (or 462 days⁷) of compute time.

6. RESULTS AND DISCUSSION

The first result we will examine is the GA's performance across the different networks types. Figure 3 shows the best-of-run performance for the GA on each network topology, for the ‘medium virality’ scenario (performance trends for the ‘high virality’ scenario were very similar, and are omitted due to space constraints). The GA finds fairly good solutions for each topology early on and then the rate of improvement slows after that. The effect of noisy fitness evaluation is observable, in that the actual NPV values (dotted lines in the figure, approximated by the average NPV from 1000 simulation runs with the GA's best individual) are considerably lower than the best-of-run fitness values the GA reports (solid lines). This is because the GA only averages 10 simulation runs to determine fitness, and then it chooses the best from the population, so the noise causes an overly-optimistic estimate of the best individual's fitness. However, individuals with the highest noisy fitness are likely to also have highest actual NPV, and the correlation between the increase in fitness and the increase in the actual NPV confirms that the GA does make real progress despite the noisy environment. Figure 3 also demonstrates that there are different maximum NPV values achievable for each social network. In fact, there is substantial variability in the capacity of these different networks to transmit/diffuse information which directly affects NPV. In general higher NPV values

⁶Preliminary comparison experiments using real-coded genes did not provide a noticeable performance benefit.

⁷Less than a month in real-time because these searches were distributed across a computing cluster.

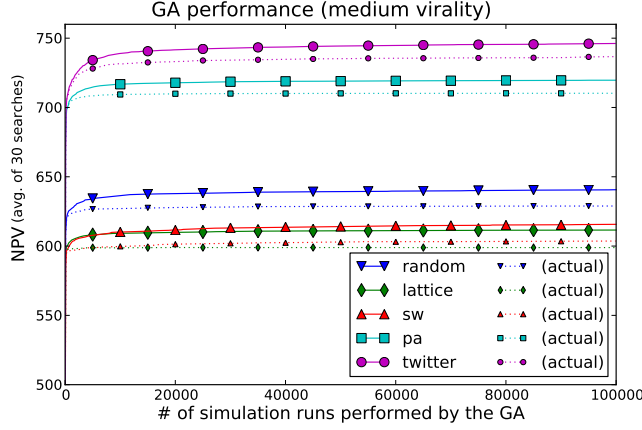


Figure 3: GA progress (averaged across 30 searches) by network topology, for the ‘medium virality’ scenario. GA’s reported best-of-run fitness (solid lines) are compared with the actual NPV values (dotted lines), estimated by 1000 simulation runs, showing the effect of noise. (Error bars too small to show.)

were possible on the networks with degree distributions that were more skewed, or inequitably distributed (in particular, the *pa* and *twitter* networks). The NPV values have a theoretical maximum of 1000 (unattainable), which would correspond to every person spontaneously deciding to adopt the product immediately, without any seeded individuals.

Before examining the evolved strategies, we will discuss results for the seeding budgets (seeding fraction, f_s) discovered by the GA. In all of the 30 search replications, the chosen f_s was always centered tightly around a specific value, which indicates a high degree of confidence that the seeding fraction values that were found are indeed optimal. However, the specific value of f_s varied substantially between network types, and also slightly based on the virality scenario (see Figure 4). In general, f_s was lower for those networks with degree distributions that were skewed such that a small number of nodes had a disproportionately large number of connections. Figure 4 displays this relationship quantitatively by plotting the optimal seeding budgets (as discovered by the GA) against the Gini coefficient [10] of the network’s degree distributions, which is a standard measure of distributional inequality ranging from 0.0 (flat equal distribution) to 1.0 (all connections concentrated in a single individual). This relationship also mirrors how the maximum achievable NPV varies by network type: essentially networks with uneven degree distributions have lower optimal seeding budgets, and a higher payoff in terms of adoption (NPV). This result is sensible, given that degree (w_d) turns out to be very important component of seeding strategies for all of the networks, as we will discuss below.

The next question we investigated was what the best strategies discovered by the GA looked like. For each of the 5 network types, in each of the 2 virality scenarios, BehaviorSearch provides us with the best strategies found in each of 30 GA searches. Due to space constraints, we present only the results for the *twitter* network (which proved to be the most interesting case) on the ‘medium virality’ scenario (see Figure 5). As shown, there is a fair amount of

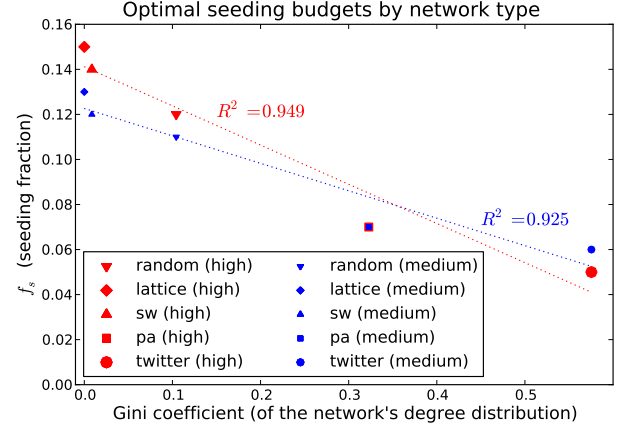


Figure 4: The best seeding budgets found by the GA for each network type. These are plotted against (on the x-axis) the Gini coefficient of the degree distributions. The regression lines are not intended to propose a linear relationship, but merely to illustrate the correlation.

variation among the GA’s best strategies. This is likely due to large plateau areas in the landscape resulting in neutral evolution among a variety of different strategies, though it could also indicate non-convexities in the space that make it difficult to search. It is worth mentioning that the weight given to the “random” strategy (α_r) was very low in all the best strategies found by the GA (except for in the *lattice* network, where no LVMP strategy outperforms random because all nodes have identical characteristics), which shows that choosing an informed strategy for seeding is important.

Our next inquiry was whether the GA’s best strategies gave better performance than using very simple strategies with the same available information. For each set of 30 strategies generated by the evolutionary search, we determined the “best strategy” by testing them with an additional independent 1000 simulation runs, and choosing the one with the highest average NPV. As a baseline for comparison, for each network, we also determined an NPV value by seeding using each of the basic component weighting functions individually: degree (w_d), two-step neighbors (w_t), average-path-length (w_a), clustering-coefficient (w_c), and random (w_r). On each of the five network types, w_d proved to be the best basic strategy of the five basic strategies. On four of the five networks, the best strategies found by the GA were either only very marginally better, or not significantly different than w_d , with the notable exception being the *twitter* network, where the GA found a strategy that outperformed w_d by more than 19 NPV units, or 2.5% ($p < 0.01$ significance), in both the ‘medium’ and ‘high’ virality scenarios. A performance comparison is shown in Figure 6, and the GA’s best strategies for the *twitter* network are displayed in Figure 7. Further testing showed that in the best strategy for the ‘high virality’ scenario, the 1% use of a secondary strategy had no impact, and in both ‘medium’ and ‘high’ scenarios, the small amount of α_a (apl) weighting included in these strategies was not significant in affecting performance. Thus, the key strategic ingredient turned out to be the combination of high degree (α_d) with low clustering-

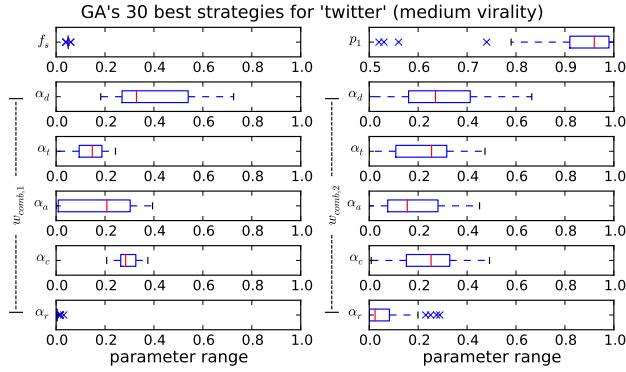


Figure 5: Box and whisker plots showing the variation among parameters for the best strategies that the GA found for the *twitter* network ('medium virality' scenario). These strategies' NPV performance varied slightly but was consistently high (from 733 to 741). (Boxes show middle quartiles with median marked red, and outliers as \times s.)

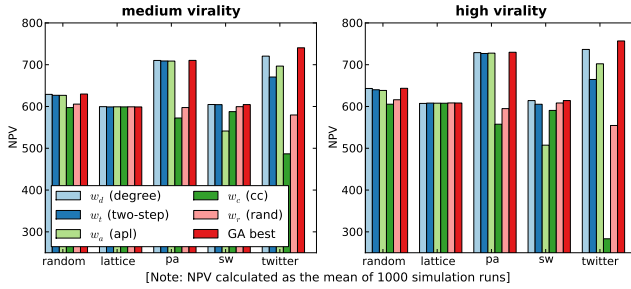


Figure 6: Best strategies found by the GA compared against the 5 basic component strategies.

coefficient (α_c). This is interesting, since using w_c (cc) alone as a weighting strategy performs worse than random seeding on the *twitter* network. The poor performance of w_c comes from a sizable number of degree one nodes (only a single friend) in the *twitter* network, which (trivially) have clustering coefficients of 0, but make poor choices for seeding. These findings beg the question: what is special/different about the *twitter* network, that was not captured in any of the 4 abstract generated social networks, which makes clustering coefficient information important for seeding? Our hypothesis is that many of the highest degree nodes (hubs) in this *twitter* network are closely linked with one another, but that there are some important individuals in the network that are further away from the central hubs, and serve as "brokers" to individuals or groups that are not directly connected to the hub. The visualization of seed choices within the *twitter* network (Figure 8) supports this explanation. Logically, it makes sense to seed individuals that are both reasonably high degree, and also play the role of brokers in the network – and yet, the $w_d + w_c$ combination does not outperform pure w_d on the other four networks. This might indicate that the four artificial networks fail to capture an important component of real social networks.

One of our research questions was whether mixed strategies (i.e. alternating between two sub-strategies) offered any

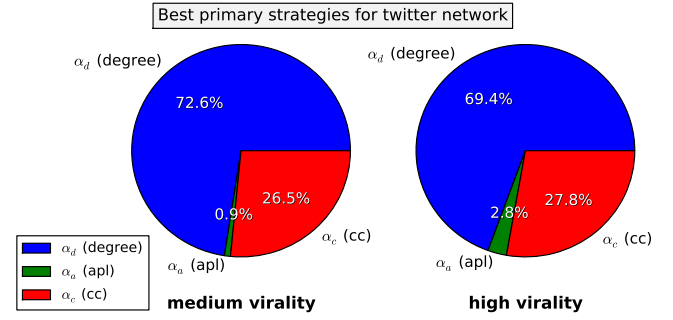


Figure 7: Components of the best primary sub-strategies the GA found for the *twitter* network. Secondary sub-strategies were basically unused: $p_1 = 1.00$ ('medium') and $p_1 = 0.99\%$ ('high').

advantages over pure (single) linear-combination strategies. In our current results we do not see any benefit, as the GA was only able to find a strategy that outperformed the simple w_d degree strategy in the *twitter* network, and that turned out to be a pure strategy as well, only requiring a combination of w_d and w_c to succeed. However, this does not rule out the possibility that a mixed strategy could be useful with a different network from the 5 investigated here, or with a different set of available network characteristics.

All evidence so far suggests that LVMP strategies are robust across different "virality" levels. In particular, the simple w_d strategy performed fairly well across the board, and the improved $w_d + w_c$ combination strategy for the *twitter* network was very consistent between the 'medium' and 'high' virality scenarios. This is a hopeful sign for marketing managers, in that results may be generalizable across different types of products, and (to a lesser extent) across different social network structures.

7. FUTURE WORK AND CONCLUSIONS

In this particular paper we have only explored one potential adoption heuristic, but a wide range of adoption heuristics exists within the space of contagion/adoption models [7]. The Bass-like model investigated here may be the best validated of extant viral marketing models, but it could be useful to look at others, especially since the applicability of adoption heuristics may vary according to product types, e.g., consumer durables vs. software.

In our seeding cost we have assumed that every node costs the same amount to be seeded, but it may be more expensive to seed an individual with more friends. Influentials are influential because people respect their advice, and thus they are not as easily swayed by promotions. The landscape of best possible strategies will alter when different cost functions for seeding are used, especially those that take into account the underlying network characteristics, which are the same features used by the seeding strategies.

While we attempted to choose a representative set of social networks that covered a range of network types, we found that our results were substantially different for our single empirically-based network (*twitter*) than for the theoretically-based networks. This reminds us that it is important to work with empirical network data in addition to abstract theoretical models. It would be worthwhile to explore alternative

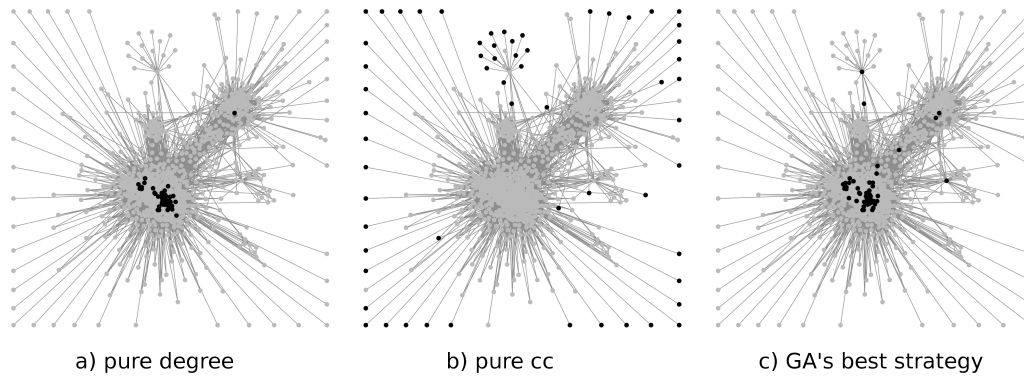


Figure 8: Visualization of three seeding strategies on the *twitter* network.

network structures, with different degree distributions and different topologies, and most importantly, other empirical networks should be gathered and examined. Also, it is unclear how well LVMP strategies generalize from a sampled sub-network to the whole network. Future work should include examining how well the strategy derived for our small *twitter* network might apply to the whole Twitter network, or to successively larger subgraphs, to see if the results scale.

In conclusion, we have presented a novel problem, called the local viral marketing problem, and have shown that evolutionary computation provides a useful method for exploring this space, and discovering unexpected features of the problem and the social networks being investigated.

Acknowledgments: We gratefully acknowledge support from NSF grant IIS-0713619, and we also wish to thank Luis Amaral, Northwestern's SSCC/Quest, and the UMD OIT HPC for supplying computational resources.

8. REFERENCES

- [1] BARABASI, A., AND ALBERT, R. Emergence of scaling in random networks. *Science* 286, 5439 (1999), 509.
- [2] BARUH, L. Social Media Marketing: Web X.0 of Opportunities. In *Handbook of Research on Social Interaction Technologies and Collaboration Software: Concepts and Trends*, T. Dumova and R. Fiordo, Eds. Idea Group, Inc., 2009, pp. 33–45.
- [3] BASS, F. A new product growth model for consumer durables. *Management Science* 15, 5 (1969), 215–227.
- [4] CALVEZ, B., AND HUTZLER, G. Automatic Tuning of Agent-Based Models Using Genetic Algorithms. In *MABS 2005: Proceedings of the 6th International Workshop on Multi-Agent-Based Simulation* (2005).
- [5] CHANDRASEKARAN, D., AND TELLIS, G. A critical review of marketing research on diffusion of new products. In *Review of marketing research*. ME Sharpe, 2007, pp. 39–80.
- [6] CHEVALIER, J., AND MAYZLIN, D. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research* 43, 3 (2006), 345–354.
- [7] DODDS, P. S., AND WATTS, D. J. Universal behavior in a generalized model of contagion. *Phys. Rev. Lett.* 92, 21 (May 2004), 218701.
- [8] DOMINGOS, P., AND RICHARDSON, M. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (2001), ACM New York, NY, USA, pp. 57–66.
- [9] ERDOS, P., AND RENYI, A. On the evolution of random graphs. *Publication* 5 (1951), 17–61.
- [10] GINI, C. Variabilità e mutabilità. *Studi Economico-Giuridici Fac. Giurisprudenza Univ. Cagliari, A III, Parte II* (1912), 3–159.
- [11] GOLDENBERG, J., HAN, S., LEHMANN, D., AND HONG, J. The role of hubs in the adoption process. *Journal of Marketing* 73, 2 (2009), 1–13.
- [12] GOLDENBERG, J., LIBAI, B., MOLDOVAN, S., AND MULLER, E. The NPV of bad news. *International Journal of Research in Marketing* 24, 3 (2007), 186–200.
- [13] GOLDENBERG, J., LIBAI, B., AND MULLER, E. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters* 12, 3 (2001), 211–223.
- [14] HOLLAND, J. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI, 1975.
- [15] JIN, Y., AND BRANKE, J. Evolutionary optimization in uncertain environments—a survey. *Evolutionary Computation, IEEE Transactions on* 9, 3 (2005), 303–317.
- [16] KEMPE, D., KLEINBERG, J., AND TARDOS, É. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (2003), ACM New York, NY, USA, pp. 137–146.
- [17] LESKOVEC, J., ADAMIC, L., AND HUBERMAN, B. The dynamics of viral marketing. *ACM Transactions on the Web (TWB)* 1, 1 (2007), 5.
- [18] LIBAI, B., MULLER, E., AND PERES, R. The Social Value of Word-of-Mouth Programs: Acceleration versus Acquisition. Working Paper, 2009.
- [19] MIDGLEY, D., MARKS, R., AND KUNCHAMWAR, D. Building and assurance of agent-based models: An example and challenge to the field. *Journal of Business Research* 60, 8 (2007), 884–893.
- [20] MILLER, J. Active nonlinear tests (ANTs) of complex simulation models. *Management Science* (1998), 820–830.
- [21] RAND, W., AND RUST, R. Agent-Based Modeling in Marketing. Working Paper, 2010.
- [22] RYAN, B., AND GROSS, N. The diffusion of hybrid seed corn in two Iowa communities. *Rural sociology* 8, 1 (1943), 15–24.
- [23] SHAIKH, N., RANGASWAMY, A., AND BALAKRISHNAN, A. Modeling the diffusion of innovations using small-world networks. Tech. rep., Working paper. Penn State University, 2006.
- [24] STONEDAHL, F. BehaviorSearch [computer software]. Available: <http://www.behaviorsearch.org/>, 2010.
- [25] TOUBIA, O., GOLDENBERG, J., AND GARCIA, R. A new approach to modeling the adoption of new products: aggregated diffusion models. Tech. rep., MSI Report, 2008.
- [26] WASSERMAN, S., AND FAUST, K. *Social network analysis: Methods and applications*. Cambridge Univ Pr, 1994.
- [27] WATTS, D. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences of the United States of America* 99, 9 (2002), 5766.
- [28] WATTS, D., AND DODDS, P. Influentials, networks, and public opinion formation. *Journal of Consumer Research* 34, 4 (2007), 441–458.
- [29] WATTS, D., AND STROGATZ, S. Collective dynamics of ‘small-world’ networks. *Nature* 393, 6684 (1998), 409–10.
- [30] WEINBERG, R. *Computer Simulation of a Living Cell*. PhD thesis, University of Michigan, 1970.
- [31] WILENSKY, U. NetLogo. <http://ccl.northwestern.edu/netlogo/>, 1999.