Search for the optimal strategy to spread a viral video: An agent-based model optimized with genetic algorithms

Michal Kvasnička[1]

Abstract. Agent-based computational papers on viral marketing have been so far focused on the study of the word-of-mouth knowledge diffusion that merges the decisions to adopt a product and to share information about it. This approach is not suitable for the analysis of the viral video sharing because it is shared with no regard whether the sender has adopted the advertised product or not. This paper presents a more realistic model of viral video diffusion in which every agent that viewed the video shares it once with a random subset of her neighbors. The optimal seeding strategy is then searched with genetic algorithms. The seeding strategy found by the genetic algorithm includes into the initial seed the agents with most connections and lowest clustering ratios; some agents are also selected randomly. However, this complex seeding strategy does not perform significantly better than a simple strategy of selecting agents with many connections.

Keywords: viral video, viral marketing, social network, agent-based model, genetic algorithm.

JEL classification: C61, C63, D85, M31
AMS classification: 68U20, 37N40, 05C82, 90B15

1 Introduction

All previous agent-based models of viral marketing merge the individuals' decisions to adopt a new product and to share information about it. This approach is suitable to model the traditional word-of-mouth diffusion of information but it does not seem to capture well the diffusion of viral videos over a social network. It is because the decisions to adopt a product and to share a video advertising it are separated: an individual can share the video because she regards it as entertaining, interesting, or alarming without adopting the product. This paper provides an alternative model of viral content diffusion that is better suited to capture this feature of viral video sharing. The goal of this paper is to explore by means of an agent-based computational simulation what constitutes an optimal seeding strategy in viral video marketing, i.e. which and how many agents the marketer should initially "infect" with the video when she has to pay for accessing them. The optimal seeding strategy is searched by genetic algorithms in the way proposed by Stonedahl, Rand, and Wilensky [12] for the ordinary "word-of-mouth" viral marketing.

2 Review of the existing literature on agent-based viral marketing

There is a wide agent-based computational literature on viral marketing. (For an introduction into the agent-based computational modeling, ACE, see e.g. [13]; for its typical application, see e.g. [8].) The ACE models of viral marketing consist of three parts: an explicit description of the used social network, the activation mechanism (i.e how an agent gets ready to share the video, i.e. get "infected"), and the seeding strategy (i.e. which and how many agents get initially infected by the marketer).

The researchers use both various artificial and empirical networks. Among the artificial networks, the most used ones are random network [12, 14], lattice and its modifications [1, 6, 12], ring [9], small-world [12], power network and its modifications [3, 9, 12, 14], mix of ring and power network [9], and a network consisting of small complete sub-networks connected by scarce inter-cluster connections [5, 14]. Several empirical networks have been used too: a network consisting of a segment of Twitter users [12], a network of users of the Korean social network Cyworld [4], and a network of coauthors of scientific papers [2, 7].

[1] Masaryk University, Faculty of Economics, Department of Economics, Lipová 41a, Brno, 602 00, michal.kvasnicka@econ.muni.cz

Great majority of researchers uses one of two following activation mechanisms: either a variant of the threshold model or a variant of the epidemiological SIR model. In the threshold model, an agent gets infected when at least x her neighbors are already infected, where x can be either an absolute number or a relative share of her neighbors. In the SIR model, an agent gets infected with some probability every time she interacts with any of her infected neighbors; thus the probability she gets infected rises to unity over time. The pure threshold activation is used e.g. in [1, 2, 7, 12, 14], the pure SI action is used e.g. in [2, 7, 14]; the combination of the threshold and personal preference of the agents is used in [3, 6], a weighted mix between SI and threshold is used in [9], and a combination of an external contamination (an advertisement) and SI (the word-of-mouth) is used in [5]. In all these approaches, the knowledge is shared only by the adopters of the product, i.e. being infected means that an agent has bought and consumes the product. The intuition of the diffusion is such that an agent buys a product because she can see some of her neighbors consuming it.

The usual seeding strategy is such that one or more agents are randomly infected at the beginning of the simulation. Other randomly chosen agents can get infected later in the simulation run to mimic the impact of an advertisement, as in [5]. Few papers try to find the optimal seeding strategy and the optimal seed size. The optimization requires perfect knowledge of the network and an immense computational power. The formalization of the problem and an approximate algorithm to solve it for a given seed size is provided by [7]. However, it is more realistic to assume only local information about the agents such as number of their connections (degree), their clustering ratio, etc. Stonedahl, Rand, and Wilensky [12] use genetic algorithms to solve this "local viral marketing problem". Their strategies consist of the seed size and weights placed on agents' desirable properties. They infect the selected number of agents with the highest weighted sum of measures of these properties. There is a great variability in their optimal strategies but the most important factors to select an agent into the seed are her degree (i.e. the number of her connections) and her clustering ratio. The optimal seed size is the lower the higher is the Gini coefficient of agents' degree (i.e. the more unequally the degree is distributed). However, their optimal strategies do not perform significantly better than a simple seeding strategy that selects the agents with the highest degree.

3 Model

3.1 Activation mechanism

I call every agent that has viewed the video *infected* with no regard whether she has adapted the advertised product or not. An infected agent's decision whether and with whom to share a video depends on three determinants: how much the agent is used to share videos in general, how much she finds the video appealing, and how much she believes her neighbors would like to see it. The decision to view a video shared by another person depends on one's personality and the relationship she has with the sender. In this paper, these two complex decisions are modeled as a simple probabilistic act: an infected person shares the video with each of her neighbors with some probability. If she shares it with a person, the person gets infected. Each infected person shares the video only once. The precise mechanism of the activation is following: At the initialization, every agent i draws a probability p_i that she shares the video; p_i is drawn from the continuous uniform distribution $U(0, v)$ where v is the maximal *virality* of the video. Then she creates a list l_i of agents to share the video with: she adds each of her neighbors to the list l_i with probability p_i. When agent i is first infected at time t, she shares the video with all agents in her list l_i at time $t + 1$, and each of these agents get instantly infected. Agent i shares the video with no one after time $t + 1$.

3.2 Network structure

An agent's neighborhood is defined by a social network in which the agent is located. Actual internet social networks seem to consist of two kinds of relations between the agents: friendship and following. Friendship is a symmetric relationship between two agents that can share stuff, e.g. videos, with each other. The stylized facts are that friends are highly clustered (i.e. one's two friends are likely to be friends together), the mean length of path between any two agents is quite short (about seven), and the number of one's friends has a power distribution, i.e. many people have few friends while few people have many friends. Following is an asymmetric relationship between a followed person and a follower, e.g. between a celebrity and her fan. The followed person can share stuff (e.g. videos) with the follower but not vice versa. The stylized fact is that the number of one's followers has a power distribution, i.e. most agents have very few followers while very few agents have very many followers. For the sake of simplicity, I assume that no person can be at the same time one's friend and follower.

The model uses an artificial social network that tries to replicate these stylized facts. Since there is no widely accepted algorithm to generate such a network, the two most widely used types of network, small world and power network, are mixed together to get a network with the suitable properties. The small world network is highly clustered and its average path length is small but its degree distribution is rather symmetric. The power network has the short average path length property and its nodes' degrees have a power distribution but its nodes are not clustered.

Each model network consists of 1 000 agents and its parameters are selected to resemble the properties of the empirical networks. Each network is created in two steps. First, a small world network is created by algorithm described by [15]; the code has been adapted from [17]. This network consists of symmetric (i.e. undirected) links that represent friendship. The agents are arranged into a circle. Each agent initially has 10 friends, 5 agents to the left of her and 5 agents to the right of her. Each link representing friendship is then rewired with probability 10 %, which creates the initial small world network. Second, a power network of directed links representing the following relationship is created over the friends' network. The algorithm has been adapted from [18]: one follower connection is added at a time, each agent is selected randomly as the follower with an equal probability and one other agent is selected randomly as the followed one with the probability proportional to each agent's number of friends and followers. 2 000 followers' links (i.e. 2 links per agent) are added. The mean clustering ratio of the networks is about 35 %. The distribution of the number of friends is symmetric, while the distribution of the number of followers has roughly a power distribution. The set of agent i's neighbors used in the activation mechanism described above is the union of the set of her friends and the set of her followers.

3.3 Marketer's seeding strategy and profit

The marketer's seeding strategy used in this paper is based on the approach developed by [12]. The agents are included into the seed because they have some desirable properties. Since there are multiple desirable properties, they are weighted. A seeding strategy (S, w) thus consists of two parts: the seed size S, i.e. how many agents the marketer initially infects, and weights w_j placed on measures of some desirable properties f_j that determine which agents are selected into the seed. An index $\sum w_j f_j$ is calculated for each agent and S agents with the highest value of the index are included into the seed. I use six measures f_j: 1) f_1 = the number of agent's friends divided by the maximal number of friends in the population, 2) f_2 = the number of agent's followers divided by the maximal number of followers in the population, 3) f_3 = the number of agent's friends and followers divided by the maximal number of friends and followers in the population, 4) $f_4 = 1-$ (agent's clustering ratio / the maximal clustering ratio in the population), 5) $f_5 = f_3 \cdot f_4$, and 6) f_6 is a random number drawn from $U(0, 1)$. The first three f_js are various measures of degree – the more connections an agent has, the better she can share the video. The f_4 is the agent's clustering ratio – the less likely an agent's neighbors are to be neighbors themselves, the better the video can spread through the population. The f_5 generalizes the notion of clustering and degree – the maximal degree and the minimal clustering are beneficial at the same time. The f_6 allows including agents into the seed randomly, which may be beneficial e.g. when the agents with the highest degree are connected together.

In contrast to the previous agent-based models of viral marketing, I explicitly assume that an agent's decisions to share a viral video and to adopt the product advertised by the video are independent. The expected revenue from the viral marketing campaign is then equal to $\rho\sigma N$, where N is the number of agents infected during the campaign, ρ is the probability that an infected agent adopts the product because of the campaign, and σ is the profit from one adopter. The cost of the campaign is $\gamma S + F$ where γ is a cost of seeding one agents, S is the seed size, and F is a fixed cost, e.g. the cost of creating the video. The marketer's problem is then to select the seeding strategy (S^*, w^*) that maximizes her expected profit from the campaign, $\Pi = \rho\sigma N - \gamma S - F$. The seeding strategy (S^*, w^*) maximizing Π also maximizes $\pi = N - cS$ where $c = \gamma/(\rho\sigma)$ is the relative cost of seeding one agent. (I do not explicitly address the problem of setting the optimal budget for the video creation. Thus F, ρ, and v are treated as constants.)

3.4 Simulation, implementation, and optimization

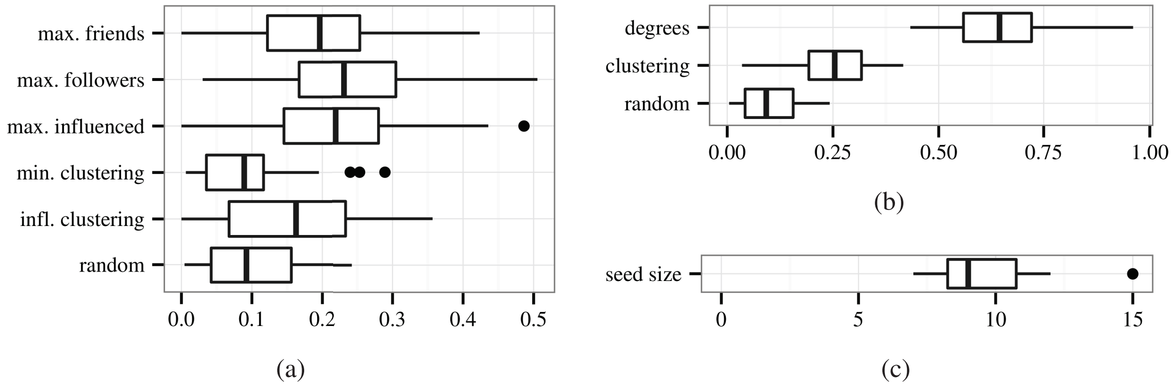
Each simulation consists of two parts: the initialization and the run. In the initialization, the agents are created and connected within the model network. Each agent i is assigned the probability p_i that she shares the video with her friends and followers. She then creates the list l_i of the agents which she shares the video with if infected. At the end of the initialization, the initial agents are infected. Each simulation run proceeds in discrete steps. In each step, each agent i infected in the previous step infects the agents in l_i and the total number of infected agents and the marketer's relative profit π are calculated. The run ends when there are no infected agents that have not yet shared the video. The model was simulated for 1 000 agents, the activation and network described in sections 3.1 and 3.2, maximal virality $v = 20$ %, and relative seeding cost $c = 10$. The model was implemented in NetLogo 5.0.5 [16]. The web interface of the model is available at http://www.econ.muni.cz/~qasar/english/models.html.

Since the parameter space of the problem is huge, the optimal seeding strategy (S^*, w^*) was searched by a genetic algorithm (GA). The search was performed in BehaviorSearch 1.0 [11]. The seed size S was searched on domain of $1, 2, \dots, 50$, each weight w_j on domain $0, 0.01, \dots, 1$. As in [12], all variables were encoded as Gray binary chromosomes. The initial population of strategies consisted of 50 randomly chosen strategies. The standard generational evolution steps (one-point crossover rate 0.7, mutation rate 0.03, and tournament selection with tournament size 3) were performed on the population of the strategies for 200 generations. The fitness of each individual strategy was evaluated by the mean relative profit π of 20 independent replications of the simulation. The best strategy was selected based on the recheck of 50 independent simulations. The individual weights were re-scaled to sum to unity. The whole search for the optimal parameters was repeated 30 times. In total, the search simulated roughly 6 million individual models, which took about 480 CPU-days.

To evaluate the relative performance of the GA optimal strategies, each optimal strategy and selected simple strategies were simulated on 100 additional instances of the model network. The simple strategies involved including $S = 1, \dots, 20$ agents with the highest number of friends, followers, or the sum of friends and followers into the initial seed. The resulting data were analyzed in R 3.1.0 [10].

4 Results

The GA optimal strategies found for our type of activation, social network, and profit specification resemble the results of [12]. The variability of the optimal weights is huge (see Figure 1a), while the expected profit is very similar for all optimal strategies (see Figure 2). This indicates that there is a partial substitutability between the measures of the desirable properties f_j and perhaps also that there is a plateau in the profit function. The first conjecture is supported by the fact that the variability of the sum of weights of all degree measures and the weights of the two clustering measures is much lower (see Figure 1b). In general, the optimal strategies place a great weight on the measures of degree and a smaller weight on the measures of clustering; surprising is the relatively high weight put on the random selection, which is in a stark contrast to the findings in [12]. This may further support the conjecture about the plateau in the profit function. The optimal seed sizes are relatively small: roughly equal to 1 % of the whole population of agents (see Figure 1c). This is probably caused by the relatively high relative cost c.



(a)

(b)

(c)

Figure 1 Distribution of the GA optimal seeding strategies found by the genetic algorithm. (a) Individual weights w_j. (b) Groups of weights: degrees are the sum of the weight of the maximal number of friends, followers, and their sum; clustering is the sum of the weights of the minimal clustering and product of maximal number of friends and followers and minimal clustering. (c) Seed sizes.

The expected relative profits π of all GA strategies are similar. However, the variance of the profits is huge (i.e. the viral video marketing campaign is risky); the same holds true for the simple seeding strategies (see Figure 2). Moreover, the GA optimal seeding strategies do not perform much better than the simple strategies with the appropriate seed size. This can be more clearly seen in Figure 3: even the seeding strategy with weights taken from the GA optimal strategy that performed the best on the sample of 100 networks does not perform better than the simple strategy of seeding agents with most followers or most friends and followers.

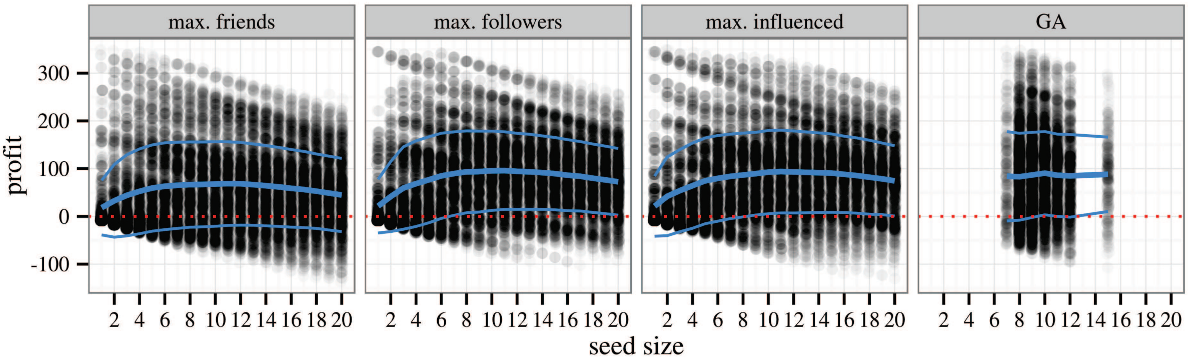


Figure 2 Comparison of performance of the GA optimal strategies and simple types of strategies (selecting agents with maximal number of friends, followers, or their sum). Individual semitransparent dots denote the relative profits π in individual simulations. The thick solid line denotes the average profit for each seeding type and seed size. The thin solid lines denote the average profit plus minus one standard deviation. The dotted line denotes the zero profit.

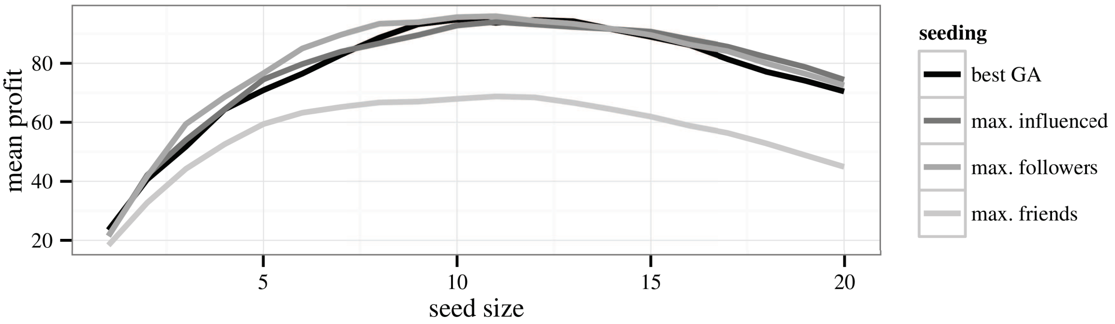


Figure 3 Comparison of expected relative profits π of the simple strategies and the best performing GA optimal strategy (GA strategy seed sizes were changed to be comparable with the simple strategies).

5 Conclusion

Our findings provide some external verification for the results of [12]. Even though I used a different type of a social network, activation mechanism, and profit specification, the optimal seeding strategy found by the genetic algorithm is similar to that of [12]: it puts a great weight on measures of degree and to a smaller extent on measures of clustering. Also similar to [12], the optimal GA strategy performed no better than a simple strategy of including agent with many connections into the initial seed. However, these results must be seen as preliminary since they were achieved only for one particular (rather low) virality of the video and one particular (rather high) relative cost of seeding. Their generalizability should be a subject of further research.

Acknowledgements

This paper has been created as a part of project of specific research no. MUNI/A/0781/2013 at the Masaryk University.

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the program “Projects of Large Infrastructure for Research, Development, and Innovations” (LM2010005), is greatly appreciated.

References

[1] Brudermann, T., Fenzl, T., 2010. Agent-based Modelling: A new approach in viral marketing research. In: *Advances in Advertising Research* (Terlutter, R., Diehl, S., Okazaki, S., eds.), vol. 1. Springer Gabler, 2010, 397–412.

[2] Cointet, J. P., and Roth, C.: How realistic should knowledge diffusion models be?. *Journal of Artificial Societies & Social Simulation* **10**(3), 2007.

[3] Delre, S. A., Jager, W., Bijmolt, T. H., and Janssen, M. A.: Will it spread or not? The effects of social influences and network topology on innovation diffusion. *Journal of Product Innovation Management* **27**(2), 2010, 267–282.

[4] Goldenberg, J., Han, S., Lehmann, D. R., and Hong, J. W.: The role of hubs in the adoption process. *Journal of Marketing* **73**(2), 2009, 1–13.

[5] 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.

[6] Goldenberg, J., Libai, B., Solomon, S., Jan, N., and Stauffer, D.: Marketing percolation. *Physica A: Statistical Mechanics and its Applications* 284(1), 2000, 335–347.

[7] Kempe, D., Kleinberg, J., 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*, ACM, 2003, 137–146.

[8] Krčál, O.: An agent-based model of price flexing by chain store retailers. In: *Proceedings of the 30th International Conference Mathematical Methods in Economics 2012* (Ramík, J., Stavárek, D., eds.), Silesian University, 2012, 461–466.

[9] Kurahashi, S., and Saito, M.: Word-of-mouth effects on social networks. In: *Knowledge-Based and Intelligent Information and Engineering Systems* (König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R. J., and Jain, L. C., eds.), part 3, Springer, 2011, 356–365.

[10] R Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. http://www.R-project.org/.

[11] Stonedahl, F. BehaviorSearch, computer software, 2010. http://www.behaviorsearch.org/.

[12] Stonedahl, F., Rand, W., and Wilensky, U.: Evolving viral marketing strategies. In: *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, ACM, 2010, 1195–1202.

[13] Tesfatsion, L., 2006. Agent-based computational economics: a constructive approach to economic theory. In: *Handbook of Computational Economics* (Tesfatsion, J., and Judd, K. L., eds.), vol. 2, 2006, Elsevier, 831–880.

[14] Watts, D. J., and Dodds, P. S.: Influentials, networks, and public opinion formation. *Journal of consumer research* **34**(4), 2007, 441–458.

[15] Watts, D. J., and Strogatz, S. H.: Collective dynamics of ‘small-world’ networks. *Nature* 393, 1998, 400–442.

[16] Wilensky, U.: NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL, 1999. http://ccl.northwestern.edu/netlogo/.

[17] Wilensky, U.: NetLogo Small Worlds model. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL, 2005. http://ccl.northwestern.edu/netlogo/models/SmallWorlds.

[18] Wilhite, A.: Economic activity on fixed networks. In: *Handbook of Computational Economics* (Tesfatsion, J., and Judd, K. L., eds.), vol. 2. Elsevier, 2006, 1013–1045.