Understanding the Hit-rate Dynamics of a Large Website with an Agent-Based Model

Jane Moran
Mars Inc.
Kleine Kloosterstraat 8
1932 Sint-Stevens-Woluwe, Belgium
jane.moran@eu.effem.com

Francesco Cordaro
Mars Inc.
Kleine Kloosterstraat 8
1932 Sint-Stevens-Woluwe, Belgium
francesco.cordaro@eu.effem.com

ABSTRACT

We have studied the simultaneous effects of advertising and word of mouth communication on the hit rate dynamics of a major retail website. The website activity can be described as having a steady-state hit rate punctuated by sharp spikes that can rise to five or six times the steady-state value before returning to quiescent levels. The rise and fall of these spikes is very rapid, happening on a time-scale of about one day. We find that the activity can be reproduced well with a combination of exogenous stimuli and an endogenous communication model. Our communication model is an agent-based model; agents are organized in a hierarchical clustering structure that simulates a social environment. At each time step, an agent will visit the website with a probability given by a value equation $V(\rho)$, where ρ is the fraction of agents in the social cluster who have visited the website recently. Our results show that the effects of external media combined with word of mouth communication as described by the model result in an emergent behaviour that is in agreement with the observed hit rate dynamics. We measure the validity of our communication model by fitting results both to website data as well as to selected YouTube videos. The insights provided by this research will have an immediate business impact, as they provide a method for optimizing a media plan by exploiting the properties of pre-existing social networks. In the next phase of our research, we will integrate our results with information about purchases that occurred during visits to the website. This will allow us to measure the media plan ROI and to recommend the introduction of advertising that makes use of social interactions into the marketing mix.

Keywords

social networks, websites, viral marketing, agent-based modeling

1. INTRODUCTION

Understanding the response that consumers have to advertising is a difficult task, due to a variety of factors. First of all, advertising does not exist in a vacuum; it competes with many external factors such as seasonality, promotions and the activities of competitors. More importantly, there is a fundamental question of what kinds of measurements we should make in order to gauge consumer attention and engagement with the brand message.

Traditionally, advertising has been evaluated in several ways. Retail auditing tracks whether the absolute sales of the product increased during the period of advertisement. Econometric modeling generally performs non-linear regressions of total sales, taking into account factors such as media spend, promotional spend, price, competitor actions and seasonality. Brand equity trackers use panels and consumer surveys to measure how a product is viewed. Finally, many media companies offer specialized media data to their advertisers that describes the media consumption habits as well as the retail habits of their demographics, based primarily on data that they obtained from customer questionnaires.

Although these traditional methods all attempt to describe the response of consumers to advertising, none of them seeks to really understand the underlying causes of the responses that are seen, and they often rely on subjective data such as declarative statements that don't always reflect reality. Non-traditional methods of measuring advertising effectiveness have been discussed in the literature (eg.[6, 8]) but have not been widely implemented into businesses. In this paper, we examine the role that social networks play in the propagation of an advertising message, and we introduce a new variable by which to understand the consumption of advertising: the pervasiveness of word of mouth communication.

We have modeled the number of visits per day to a retail website using a combination of exogenous stimuli (simulating traditional media campaigns) and an endogenous agent-based communication model (simulating word-of-mouth communication). In addition to analyzing the effectiveness of propagating a message via word of mouth, this work will help us to understand advertising adstock, and will give us valuable information about the structure of social networks, which can lead to the introduction of smarter ad campaigns.

The paper is organized as follows: we begin by describing the data set in section one, followed by the model description in section two. Lastly, we discuss our results and outline how they will be used to promote the use of more effective advertising campaigns within our business.

2. THE DATA

Our first data set consists of a count of the number of visits each day to a retail website (the hit rate). The data covers nearly 2 years of website operation, starting on its

Cite as: Understanding the Hit-rate Dynamics of a Large Website with an Agent-Based Model. Jane Moran, Francesco Cordaro, *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10–15, 2009, Budapest, Hungary, pp. 105 – 109

launch date. For the first year of its operation, there were no consistent campaigns promoting the website. Beginning in the 12th month after the website launch however, a coordinated advertising campaign was launched, including promotion codes sent to mailing lists as well as online banner and offline press ads.

The website activity in the first year consists of a steady-state level punctuated by sharp spikes that rise to 5 or 6 times the steady-state value before relaxation. The spikes rise and fall in a single day, and are associated with media coverage of the website (one spike occurred on the day that the website was discussed on a radio program). In the second year, the website activity simultaneously increases and becomes noisier as the coordinated advertising campaign was launched. In order to simplify the modeling, the dataset has been normalized so that the hit rate is re-scaled into a $[0, 1]$ interval. A "hit" is in fact a unique visit to the website, meaning that when a browser with the same IP address visits the website more than once in a day, the visit is counted a single time only.

Our second data set consists of viewing trends for specific YouTube videos (www.YouTube.com). The videos were tracked using the TubeMogul^{TM} online tracking software (www.tubemogul.com). The time period covered by the tracking varies by video, but the counting is always done in the same way. YouTube counts a "view" as a full view - start to finish. Videos viewed in part are not counted as a view. Videos embedded in other websites can receive one view per IP address. The IP address restriction does not apply for videos being watched on YouTube. This data is aggregated at the daily level.

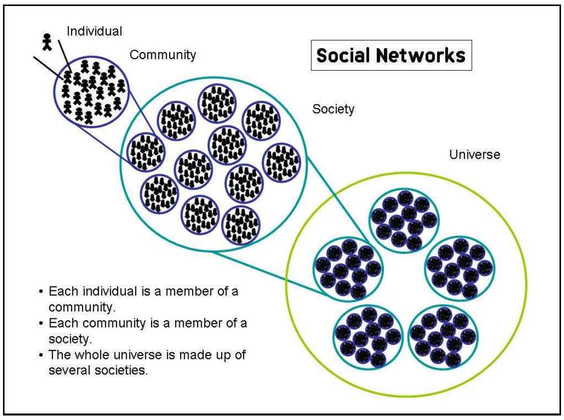
3. THE MODEL

Our model consists of two separate components: an agent-based communication model to study the effects of word of mouth communication, and a stimulus model to introduce the effects of external advertisement on the system. The individual effects of these separate components have been previously analyzed in the context of online book sales and video views in [5] and [4] where it was shown that both types of stimuli produce a power-law response in the system. Although we do not explicitly model this response function here, we see similar features throughout our data set.

3.1 Communication Model

In order to simulate communication, we have developed an agent-based model that allows a message to propagate through a social network. Our social network is a type of small-world network [9], and is constructed as a series of nested clusters of agents, with the most basic cluster consisting of 20 agents. A group contains 50 clusters, and there are 20 groups in the universe. This structure is illustrated in figure 1.

The cluster size was chosen to approximate the number of people with whom an average individual has regular interactions in his day-to-day activities. The extent to which an agent interacts with members of a cluster other than his own is given by the agent's social parameter s. The agent's social parameters are chosen from a pareto distribution to reflect the scale-free nature of social interactions [2]; $s \sim k x_{min}^k / x^{k+1}$, with $x_{min} = 0.0015$ and $k = 3$.



- Each individual is a member of a community.
- Each community is a member of a society.
- The whole universe is made up of several societies.

Figure 1: The agents' social structure.

During the execution of the program, each agent interacts either with members of his immediate cluster, or, with probability given by the social parameter s, another one of the clusters in his group. At each time step t in the model, the agent polls the other agents in his cluster to determine how many of them have visited the website recently. Based on the results of this poll, the agent assigns a value V to the information, calculated as in [3] as:

$$V(\rho) \equiv \frac{(1 + \theta^d)\,\rho^d}{\rho^d + \theta^d} \tag{1}$$

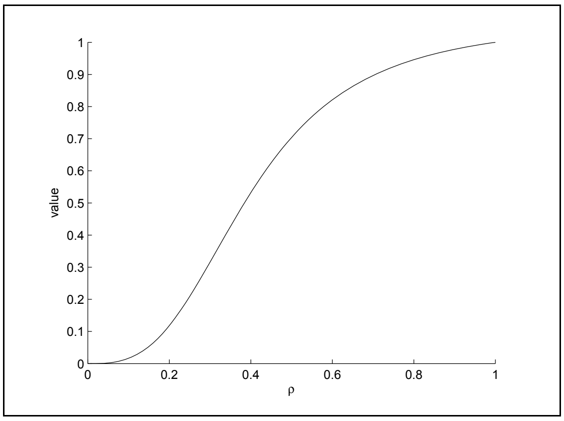


Figure 2: The shape of the value equation with parameters $\theta = 0.4$ and $d = 3$. The x-axis is ρ, the fraction of the cluster in the state of active communication, and the y-axis is the probability that a susceptible agent polling this cluster will enter the incubating state.

Here, ρ represents the fraction of the cluster population having visited the website at time-step $t - 1$ while θ and d are free parameters influencing the shape of the value curve (we use $\theta = 0.4$ and $d = 3$). The value is in the interval $[0, 1]$ and determines the probability that the agent will himself visit the website at a given time step $t + 1$. The shape of the value curve is shown in figure 2.

To model the overall behaviour of the individual agents we have adopted a simplified framework based on epidemiology [1, 7]. First, we define a parameter that describes the lag time between an agent first learning of the information being spread and his acting upon it. We call this the incubation parameter to reflect the fact that an agent in this state has been contacted, but is not yet communicating the information to other agents. The second parameter describes the time that an agent spends in a state of active communication with his neighbours; this is the communication parameter. Finally, the agents experience a stage of immunity immediately following the end of their communication. While in this state, an agent is "immune" to any further information that he may encounter. This stage is regulated by the recovery parameter. The three attention parameters are summarized below:

1. The incubation parameter, α_i, determines the number of time steps that elapse between the time that an agent becomes aware of the information and the time that he visits the web site.

2. The communication parameter, α_c, determines the number of time steps that an agent spends in a state of active communication with his cluster.

3. The recovery parameter, α_r, represents the number of time steps that an agent spends in a state of non-communication about the website, after his communication time steps are over.

The default state for agents not in one of the above states is susceptible. The attention parameters allow for a life-cycle of information flow, as each agent communicates according to the three-stage cycle of incubation, communication and recovery. In our model, the attention parameters for each agent are selected randomly from normal distributions centered at 0.3 days (α_i), 3.0 days (α_c), and 3.1 days (α_r) , with a standard deviation of 0.2 days. This life cycle allows a single cluster to simulate a changing social structure in which friendships are by turn forged and broken.

As an independent test of the validity of the communication model, we have used the communication model to fit the view rate of a recent viral YouTube video. The result, shown in figure 3, uses parameters that are identical to those used for the website model, with the exception that the values of the attention parameters were decreased from approximately 3 days to approximately 2 days.

3.2 Stimulus Model

The second aspect of the model is the stimulus model. This simulates the exogenous stimuli that influence the system (primarily in the form of deliberate advertising campaigns). In the optimal situation, we would re-construct this stimulus model directly from the actual media plan for the website. However, we have only partial information relating to the media plan, so our re-construction is estimated rather than exact.

In order to introduce an external stimulus into the model, we select a certain number of agents at random from the entire universe of agents, and for each one of those agents in a state of susceptibility (status given by α_s), his status is changed to incubating. The number of agents that are selected varies, from 1% of the total population for small stimuli, to 7.5% and 15% of the total population for moderate and large stimuli respectively.

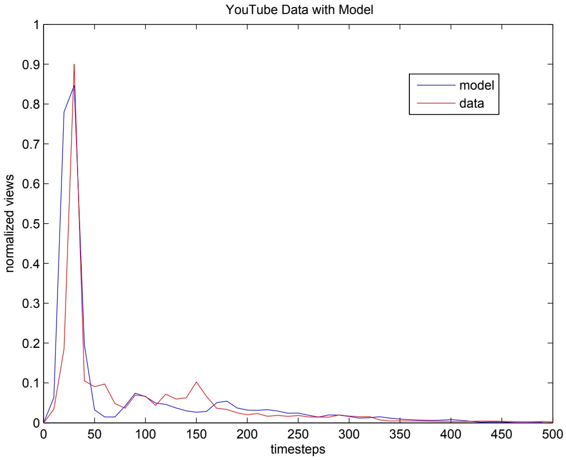


Figure 3: Modeling a YouTube video. The data from YouTube is shown in red, the results of our model are in blue.

In figure 4, the effects of the communication model and the stimulus model can be seen both separately and in combination. The communication model, when not enhanced by stimuli, exhibits a power-law decay that goes to zero. The stimulus model, on the other hand, produces local perturbations that are not propagated throughout the social network. It is only when the two aspects of the models are combined that we see an emergent behaviour that accurately reproduces the properties of the data.

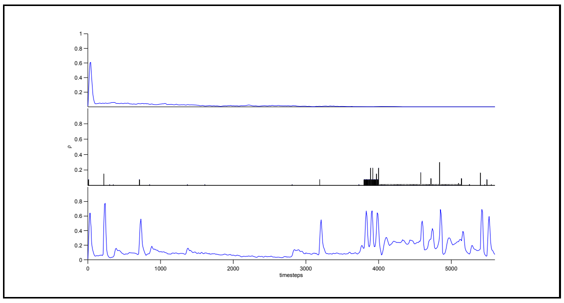


Figure 4: The effects of the communication model and the stimulus model combine to produce an emergent behaviour. The communication model is in the top pane (with an external stimulus introduced at timesteps 0, 25 and 50). The middle pane shows the stimulus model, and the bottom pane is the result when both models are used together.

4. RESULTS & DISCUSSION

The communication model has a total of 7 free parameters: the value equation parameters θ and d, the agent attention parameters α_i, α_c and α_r, and the social parameters x_{min} and k. The value equation parameters were set to allow for the existence of sharp spikes in the data. The agent attention parameters and the social parameters were set to allow for the continued propagation of the message at an appropriate level. The tuning of all of these parameters was partially accomplished with the help of a particle swarm optimization (PSO) routine. In the PSO routine, the optimal fitness was determined by minimizing the χ^2 values between the model and the data on a subsample of 21 points that were chosen from the first three quarters of the data set. Half of these points were representative of the steady-state behavior while the other half represented the rapid spikes in the data.

The optimal parameters, as determined by the PSO routine, were $\theta = 0.4$, $d = 3$, $\alpha_i = 3$, $\alpha_c = 30$, $\alpha_r = 31$, $x_{min} = 0.0015$, and $k = 3$. These parameters resulted in a χ^2 value of 20.1, less than the critical P-value of 23.68 at a 0.05 confidence level.

With the free parameters appropriately tuned, the simulation was run with 20,000 agents over 6450 timesteps. A plot showing both the data and the model is shown in figure 5. The model fits both the steady-state intervals as well as the rapidly fluctuating spikes. A characterization of the fit of the model to the data is shown in figure 6. The histogram is shown along with the best-fit normal distribution, which has $\mu = 0.0014 \pm 0.0093$ and $\sigma = 0.1195 \pm 0.0069$. The differences are evenly distributed about zero.

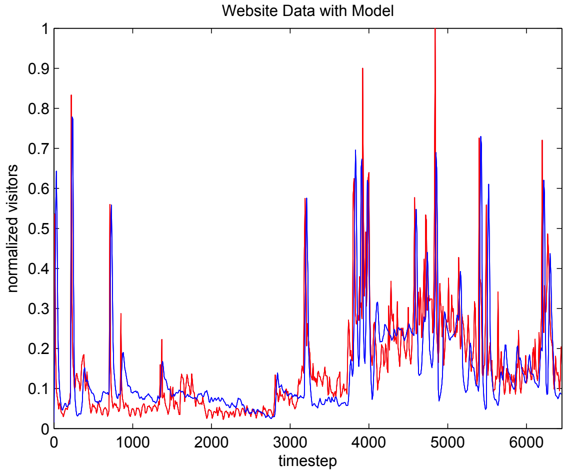


Figure 5: Website data and model. The data is shown in red, the results of our model are in blue. There are 10 timesteps in one day of real data.

Our results show that it is possible to reproduce the dynamics of a website by introducing exogenous and endogenous perturbations to a complex system. As a result of this work, we are able to recommend an optimal media plan that makes use of social networks in order to maximize the number of customers contacted while minimizing advertising costs.

Indeed, after the optimization of the α parameters we are able to infer how long an individual will remain in an incubation status; during this period we could act on the system by external stimuli to increase the overall awareness of the agents. Then, having found the average period of contagion, we can afford to lower the external media pressure (eventually to no stimuli at all) and let word-of-mouth dynamics play the role of advertiser. Finally, knowing the length of the silent period during which the communication is switched off, we can re-inject media into the system to prepare the agents for the next cycle. The overall result would be a fine tuned and calibrated series of media injections with targeted pressure in specific periods, i.e. a media plan optimized on the basis of the underlying network dynamics.

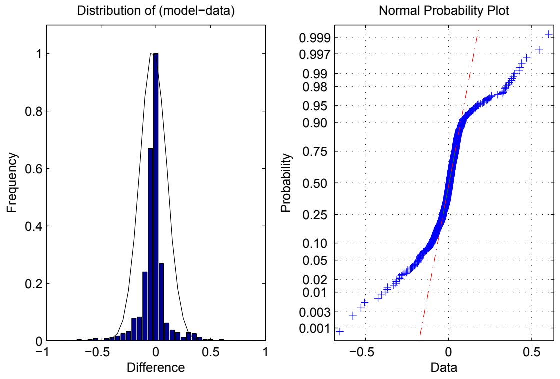


Figure 6: Measuring the goodness of fit. The panel on the left shows a distribution of the differences between the data and the model. The curve superimposed on the histogram is a normal distribution with $\mu = 0.0014$ and $\sigma = 0.1195$. The panel on the right shows the normal probability plot of the differences. Both plots show that although the differences are evenly distributed about zero, they do not strictly follow a normal distribution.

In other words, once we calibrate the model on the basis of a given real media plan (our training set) we would use the resulting α parameters to run several other simulations where now the parameters subject to optimization are the frequencies and intensities of stimuli (the optimal media plan). This would be a optimization constrained by the total cost of the stimuli.

In the next steps of this research program, we plan to improve the stimulus model by introducing more complete media plan data. Following this stage, we will investigate the relationship between the method by which a customer becomes aware of the product and the probability that the website visit results in a sale. This information will bring additional value to the business, as it will identify the advertising methods that result in the best consumer engagement with our products.

5. REFERENCES

[1] R. M. Anderson and R. M. May. Population biology of infectious diseases: Part I. *Nature*, 280(5721):361–367, 1979.

[2] A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999.

[3] E. Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America*, 99(Suppl 3):7280–7287, 2002.

[4] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[5] F. Deschâtres and D. Sornette. Dynamics of book sales: Endogenous versus exogenous shocks in complex networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 72(1):016112, 2005.

[6] R. D. Groot. Consumers don't play dice, influence of social networks and advertisements. *Physica A*

Statistical Mechanics and its Applications, 363:446–458, May 2006.

[7] W. O. Kermack and A. G. McKendrick. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society of London*, 115(772):700–721, Aug. 1927.

[8] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1):5, 2007.

[9] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, June 1998.