The Genetic Algorithm as a General Diffusion Model for Social Networks

Mayank Lahiri
Department of Computer Science
University of Illinois at Chicago
mlahir2@uic.edu

Manuel Cebrian
The Media Laboratory
Massachusetts Institute of Technology
cebrian@media.mit.edu

Abstract

Diffusion processes taking place in social networks are used to model a number of phenomena, such as the spread of human or computer viruses, and the adoption of products in 'viral marketing' campaigns. It is generally difficult to obtain accurate information about how such spreads actually occur, so a variety of stochastic diffusion models are used to simulate spreading processes in networks instead. We show that a canonical genetic algorithm with a spatially distributed population, when paired with specific forms of Holland's synthetic hyperplane-defined objective functions, can simulate a large and rich class of diffusion models for social networks. These include standard diffusion models, such as the *independent cascade* and *competing processes* models. In addition, our *genetic algorithm diffusion model* (GADM) can also model complex phenomena such as information diffusion. We demonstrate an application of the GADM to modeling information flow in a large, dynamic social network derived from e-mail headers.

Introduction

Modeling the spread of infectious diseases through person-to-person contacts is an important problem in mathematical epidemiology (Hethcote 1989). Many mathematical models of disease transmission, however, make the simplifying assumption that individuals in a population mix uniformly at random. In order to relax this assumption, the spread of infections are studied on network-structured populations (Keeling 1999; Newman 2002), also called *social networks*, where person-to-person contacts are constrained by a graph structure. Since real social network data is readily available, the general idea of an infectious disease spreading through a network has also been used to model more general *diffusion processes*, such as the diffusion of innovations (Valente 1996), so-called 'viral' marketing that spreads knowledge of a product through word-of-mouth recommendations (Kempe, Kleinberg, and Tardos 2003), and the spread of electronic worms through a computer network (Aspnes, Rustagi, and Saia 2007), among others. Although the ultimate purpose of modeling each type of diffusion process is generally different, they all fundamentally rely on a *diffusion model*, which probabilistically specifies how a process – such as a disease, computer virus, or adoption of a product – starts and spreads through a network of individuals. A variety of diffusion models have been proposed to model various spreading processes in networks (Kempe, Kleinberg, and Tardos 2003; Kleinberg 2007).

In this paper, we show that a form of the canonical Genetic Algorithm (GA) (Goldberg 1989), when paired with Holland's recently developed *hyperplane-defined* objective functions (Holland 2000), comprise a large, rich class of diffusion models for social networks, with some interesting properties. In particular, the *GA Diffusion Model (GADM)* (a) contains several popular diffusion models from economics, sociology, and mathematical epidemiology as special cases, such as the susceptible-infected and competing process models, (b) allows the computation of diffusion over a large class of diverse models as a direct result of the previous point, which in some cases may be preferable to choosing a single model for a particular application (Cointet and Roth 2007), and (c) allows the modeling of complex, non-linear phenomena such as information exchange. In order to demonstrate the final point, we use the GADM to model information flow in a large social network derived from e-mail traffic. Our results indicate that some individuals in the network are intrinsically, and non-trivially, better connected in terms of receiving information as it propagates from person to person.

Our work here is, to the best of our knowledge, the first to explicitly and formally unite purely optimization-oriented studies of genetic algorithms with spatially structured populations (Min et al. 2006; Payne and Eppstein 2006), and diffusion processes in social networks (Kempe, Kleinberg, and Tardos 2003; Kleinberg 2007). The genetic diffusion model presented here also draws from work in collective optimization (Scardovi and Sepulchre 2006) and parallel genetic algorithms (Cantu-Paz 1998).We feel that these are valuable connections, and by framing diffusion in social networks in terms of the genetic algorithm, we hope to initiate some crossover in the considerable theoretical developments in each of these fields.

Diffusion in Social Networks

A social network is a graph-theoretic representation of the interconnections or interactions between a set of unique en-

tities. The vertices of the graph represent individual humans, animals, or networked computers (for example), and an edge connects two vertices that are related or interacting in some way. Some modern examples of social networks include online social networks, where vertices are user accounts and edges represent 'friendships' between accounts, and communications networks, where vertices represent e-mail addresses or telephone numbers, and edges represent e-mails sent or telephone calls placed between vertices.

Definition 1. A *social network* is a directed or undirected graph $G = (V, E)$, where a labeled vertex $v \in V$ represents an individual in some physical system, and an edge $(u, v) \in E$ represents an interaction between two individuals. A *dynamic social network* is a multigraph $G = (V, E)$, where E is a bag of edges, and each timestamped edge $(u, v)_t \in E$ represents an interaction (u, v) that occurred at time $t \in \mathbb{Z}^+$.

In a typical 'susceptible-infected' setting for a diffusion process, a set of vertices in the network is initially marked as *activated*. The diffusion process proceeds in discrete timesteps. At each timestep, an activated vertex might come into contact with inactive vertices. The inactive vertex then has a chance of being activated through this contact, according to the probabilistic rules specified in the diffusion model. This process repeats until a pre-specified stopping criterion is satisfied, such as a particular fraction of vertices being in an active state, or for a specified number of timesteps.

Within this framework, the following are some typical questions associated with diffusion processes:

1. **(Extent)** Given a specific subset of initially activated vertices in a network and a diffusion model, how many vertices are expected to become activated after a specified period of time?
2. **(Targeting)** Which vertices should be targeted as initiators to result in the maximum extent of spread (Domingos and Richardson 2001)? This problem is NP-hard to solve optimally, regardless of the diffusion model used (Kempe, Kleinberg, and Tardos 2003).
3. **(Blocking)** Which vertices should be targeted for immunization to minimize the expected number of activated vertices (Habiba et al. 2008; Anshelevich et al. 2009)?

The targeting and blocking problems both depend on the extent of a diffusion process, which in turn depends on a specific diffusion model.

Definition 2. A *diffusion model* accepts as input a graph structure $G = (V, E)$, a *state vector* $S_v^{(t)}$ for every vertex $v \in V$ at time t, and an optional vector of internal parameters P. Based on the state of all interacting individuals, it outputs a new state vector $S_v^{(t+1)}$ for every vertex at time $t + 1$. For a static (non-dynamic) social network, the graph structure is the same at each timestep. For a dynamic social network, the graph at each timestep is defined as $G_t = (V, E_t)$, where $E_t = \{(u, v) : (u, v)_t \in E\}$ is the set of edges marked with timestamp t.

As an example, consider the *independent cascade* model (Kempe, Kleinberg, and Tardos 2003; Kleinberg 2007), which is closely related to the Susceptible-Infected-Susceptible (SIS) and Susceptible-Infected-Removed (SIR) models in mathematical epidemiology (Hethcote 1989).

Definition 3. In the *independent cascade* diffusion model, the state vector $S_v^{(t)}$ of each vertex v at time t is of dimension 1, *i.e.*, a single binary variable, where a value of 1 means that the vertex is activated. A global parameter p represents the probability of an inactive vertex becoming activated on each contact with an activated vertex. In a static social network, each freshly activated vertex is allowed to activate each of its neighbors exactly once, with probability p each time (Kempe, Kleinberg, and Tardos 2003). In a dynamic network, each contact at a particular timestep can result in an activation, with probability p each time (Lahiri et al. 2008).

The GA Diffusion Model

We now show how the canonical genetic algorithm (Goldberg 1989), *i.e.*, one with binary string chromosomes and one-point crossover, can be used as the framework for a large, rich class of diffusion models. The description of the model in this section applies to both static and dynamic social networks.

Recall that a dynamic social network consists of a set of individuals $V = \{v_1, ..., v_n\}$ interacting over a period of T discrete timesteps. We begin by mapping every individual in the social network to a chromosome in a GA population. Let the state of each individual S_v be a binary chromosome of length β in the GA population. This can initially be set to all zeros, or chosen according to some random distribution. We also define an objective function $f(\mathbf{x})$ that assigns an objective score to any state string, the exact mechanism of which will be described shortly. For every edge (u, v) in the social network at timestep t, we apply the logic of the canonical genetic algorithm to the corresponding chromosomes in the GA population:

1. At time t, for edge (u, v), let the corresponding state strings be $S_u^{(t)}$ and $S_v^{(t)}$ respectively. Initially, let $S_u^{(t+1)} = S_u^{(t)}$ and $S_v^{(t+1)} = S_v^{(t)}$.
2. A *crossover point* c is selected uniformly at random from the integer range $[1, \beta]$. Two new state strings are created by swapping the tails of $S_u^{(t)}$ and $S_v^{(t)}$, where the tail is defined as all positions including and after index c. Let these two new state strings be $\mathbf{y}_1$ and $\mathbf{y}_2$.
3. The objective score of each new state string is then evaluated according to the objective function $f(\mathbf{x})$. If any of them have a greater objective score than either of their parents, the corresponding parent's state string is replaced for the next iteration.

$$S_u^{(t+1)} = \underset{x \in \{S_u^{(t)}, S_u^{(t+1)}, \mathbf{y}_1, \mathbf{y}_2\}}{\arg\max} f(x)$$

$$S_v^{(t+1)} = \underset{x \in \{S_v^{(t)}, S_v^{(t+1)}, \mathbf{y}_1, \mathbf{y}_2\}}{\arg\max} f(x)$$

In the case of ties in the objective scores of the original and a new string, the original state string is retained.

The algorithm above is similar to a GA with a spatially distributed population (Min et al. 2006; Payne and Eppstein 2006), except that the GA's selection operator is replaced with real social network data that dictates the sequence of mating operations. The 'diffusion' in the GADM occurs as state vectors are modified using the crossover operator and subsequently adopted based on their objective scores. Clearly, the missing component to add meaning to this mapping, in terms of a diffusion process, is the choice of objective function $f(\mathbf{x})$, which we have not defined so far.

Holland's *hyperplane-defined functions* (HDFs) are a class of synthetic objective functions with some interesting properties that fill this gap (Holland 2000). An HDF is constructed by defining *schemata* (short substrings with wildcards that start at a specific position) randomly and hierarchically, starting with relatively short schemata of order 1 occurring at random starting positions within the string. Pairs of such schemata are concatenated to generate schemata of order 2, and so on, with each schema receiving an individual positive or negative score, also specified randomly from some range. HDFs take a binary string as input and return an objective score that is the sum of the scores of all the individual schemata contained in it. HDFs can be used to generate objective functions of arbitrary difficulty (*e.g.*, nonlinear, nonseparable, nonsymmetric, etc. (Holland 2000)). For our purposes, however, the construction above is sufficient.

The following example illustrates the generation of a simple HDF $f(\mathbf{x}) \mapsto \mathbb{R}$ that takes a binary string of length $n = 10$ as input. The asterisk character ('*') denotes a 'don't care' character that matches both a zero and a one.

`*01*******`	*score 2, order 1 schema*
`*****110**`	*score 2, order 1 schema*
`********10`	*score 3, order 1 schema*
`*01**110**`	*score -4, order 2 schema*
`*01*****10`	*score 4, order 2 schema*

The following binary strings are now evaluated using the HDF above:

`1010000011`	*score:* 2
`0010011000`	*score:* $2 + 2 + (-4) = 0$
`1010011010`	*score:* $2 + 2 + 3 + (-4) + 4 = 7$

A natural analogy for a GADM paired with a HDF is information diffusion. Each schema corresponds to a 'unit' of information, with different schemata carrying different values, as in the example above. Depending on the initial state strings, every individual in the network knows certain pieces of information. Some combinations of schemata might be detrimental when combined, such as the negative-valued order 2 schemata in the example above, while others could have a value greater than the sum of their parts. Each time a pair of individuals interact, they randomly exchange information, which is modeled by the crossover operator. This process could result in neither gaining any benefit from the interaction, or one (directed network) or both (undirected network) gaining more information than they had before.

We note that information diffusion in a network has been analyzed using epidemic models, in the specific instances of a blog network (Gruhl et al. 2004) and wireless networks (Khelil et al. 2002), but the notion of 'information' is restricted to a binary variable, or a single 'unit' of information in our terminology. The GADM overcomes this restriction by using schema to model units of information. The connection to diffusion models in general also raises some interesting theoretical questions beyond the scope of this paper, such as the relationship between the rate of diffusion of schema in genetic algorithms with network-structured populations, and the *epidemic threshold* of the network (Chakrabarti et al. 2008).

Special Cases

The primary reason for using a HDF as the objective function in the GADM is the flexibility it offers in terms of modeling diffusion processes, specifically in the fact that carefully defined initial states and HDFs result in commonly used diffusion models for social networks. Thus, reductions of existing diffusion models to special cases of the GADM establish a formal connection between the two. We now describe a few such reductions.

Proposition 1. *The independent cascade diffusion model for dynamic social networks, in the susceptible-infected (SI) case, is a special case of the GADM, when the probability of activation* $0 < p < 1$ *is a rational number. For irrational* p*, an arbitrarily precise approximation can be obtained.*

Proof. We define the HDF by creating a single schema of length α at an arbitrary position within the state string, with all bits equal to 1. This schema has value 1, so that any state that contains the entire schema will evaluate to objective score 1, while all other states will evaluate to objective score 0. The initially activated individuals receive the schema in their states, while all other individuals receive a zero string for their state. When an activated individual comes into contact with an inactive one, the probability that the activation schema will be transfered to the inactive individual is equal to the probability that the crossover point c falls outside the schema. Let the schema begin at position i and end at position j, inclusive of both positions. Thus,

$$P(activation) = P(c \leq i) + P(c > j)$$

If these conditions on c are satisfied, then one of the child state strings will contain the schema. However, since c is chosen uniformly at random from the interval $[1, \beta]$, we can rewrite this probability in terms of the relative lengths of the schema and state string.

$$P(activation) = 1 - \frac{\alpha - 1}{\beta} = \frac{\beta - \alpha + 1}{\beta}$$

Thus, to model the SI independent cascade process with a probability p, we fix β at an arbitrary (but sufficiently large) value and create a HDF with a single schema of value 1 and length α, where

$$\alpha = \beta(1 - p) + 1$$

To ensure that we are replicating the exact probabilities associated with the independent cascade process, we note that since p is a rational number, $(1-p) = \frac{a}{b}$, where a and b are integers. We can then choose β to be a multiple of b, so that the resulting value of α is an integer. This replicates the probability of infection precisely, with the side-effect that the sum of objective values in the GA's population is equal to the number of activated individuals. □

Another example of a special case is that of two or more processes spreading on the same network, such as recruiting for competing clubs (Bharathi, Kempe, and Salek 2007), or a virus and an immunizing alert spreading through a computer network (Aspnes, Rustagi, and Saia 2007). Each individual starts out either by belonging to a specific club or being unaffiliated. As unaffiliated individuals interact with their neighbors, they are recruited into their neighbors' affiliations with possibly different, but fixed and independent, probabilities, but an individual once affiliated does not switch affiliations (hence, they are mutually exclusive).

Definition 4. The *mutually exclusive competing processes* diffusion model for dynamic networks is defined as follows.

1. Each vertex v is initially assigned to one of n categorical states $s(v)_0 = \{m_1, ..., m_n\}$, or an extra unaffiliated state.
2. Individuals who are affiliated with a state do not change states. When an unaffiliated individual v_i comes into contact with an affiliated individual v_j who is in state m_j, v_i switches to affiliation m_j with probability $p(m_j)$.

Proposition 2. *The* mutually exclusive competing processes *diffusion model for dynamic networks is a special case of the genetic diffusion model, when $p(m_j)$ is a rational number for all $j \in [1, n]$.*

Proof. We create an HDF with n overlapping schema, one for each affiliation, all starting at position 1 and bounded by 1 bits at their endpoints. As in Proposition 1, each schema has value 1 and lengths $\alpha_1, ..., \alpha_n$, and the length of the state vector is β. Let the probabilities of an unaffiliated vertex taking on each affiliation be $p_1 = \frac{a_1}{b}, ..., p_n = \frac{a_n}{b}$, where the rational probabilities have been scaled to have common denominator b. Similar to Proposition 1, we derive the following expressions for α_i:

$$\alpha_i = \frac{\beta}{b} \cdot (b - a_i) + 1 \quad 1 \leq i \leq n$$

We also scale all lengths uniformly such that $\min_i(\alpha_i) \geq n + 2$. Since each schema starts at position 1, an individual cannot have more than one affiliation at a time. Unaffiliated individuals receive state strings of all zeros. However, with multiple overlapping schemata of varying length, if the schemata are not chosen carefully, a crossover between two affiliated individuals could cause one or both to switch affiliations. The following solves this issue: each schema is bounded by 1 bits, with the middle bits being comprised of the binary representation of 2^i, where i is the index of the schema ordered by length. Each schema therefore has exactly 3 bits set.

It is easy to show that any two schemata in such a set, when spliced together at an arbitrary location, will result in a string that contains at most one valid schema from the set. This is because all schemata start at position 1 in the string, and only the relative spacing between the first three 1 bits define a valid schema according to our definition. Thus, at most one schema in the set will be matched. Since the GA retains the original string in cases of objective score ties, and splicing the strings of two affiliated individuals will not result in a child string with a *higher* objective score, affiliated individuals will never switch affiliations. □

As an example of the construction in Proposition 2, when $\beta = 20$, the following schemata define affiliation probabilities of 0.75, 0.65, and 0.45 respectively:

```
110001**************
10100001************
100100000001********
```

Experimental Evaluation

We stated earlier that the GADM can be used to model complex, non-linear phenomena. To illustrate this point, we now present a short case study applying the GADM to simulate information diffusion through a real dynamic social network. We assume that individuals in the network communicate with each other to exchange information, with the goal of selfishly maximizing the amount of information that each possesses. A key difference from earlier studies on information diffusion, such as (Gruhl et al. 2004; Khelil et al. 2002), is that we model multiple 'units' of information propagating from person to person, where different units can interact in complex, non-linear ways, according to a randomly generated HDF, to affect the total 'information value' of each person. The objective is to estimate the information value of each vertex over multiple random state initializations and HDFs, in order to determine if everyone in the network is positioned to receive the same amount of information, on average, as a result of their interactions with other individuals. Social networks are known to generally be structurally dominated by a relatively small set of high-degree vertices (Newman 2003); does a similar stratification exist for information flow?

Dataset. The Enron e-mail dataset is a publicly available repository of e-mails sent by and to executives of the former Enron Corporation[1]. It is one of the largest public datasets of a corporate e-mail environment, and is naturally represented as a dynamic social network. Each vertex in the network is an e-mail address, and a directed and timestamped edge represents an e-mail sent between two addresses. We parsed the headers of all e-mails and obtained 1,326,771 timestamped edges corresponding to individual e-mails sent between 84,716 e-mail addresses. There were 215,841 unique timestamps, non-uniformly

[1] A reference C++ implementation of the GADM, the Enron dataset, and other supporting material may be found online at `http://code.google.com/p/gadm/`

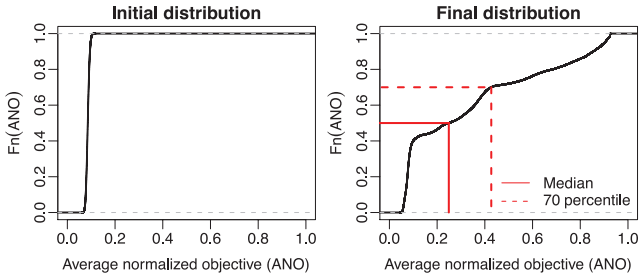

Figure 1: Cumulative distribution of initial (left) and final (right) ANO values for the Enron e-mail dataset.

covering a period of approximately 4 years.

Setup. We start by generating a random HDF of arbitrary length, and random initial state strings for each individual. After the GADM processes all timestamped edges, the final objective value of all vertices are normalized relative to the maximum objective value in the population. Intuitively, a vertex with a high final objective value has accumulated the most valuable combination of units of information, which could be the result of its inherent position in the network, or a chance occurrence based on its randomly chosen initial state string. To account for the latter bias, we run multiple trials with different HDFs and random initial state strings. Over many trials, this gives us the *average normalized objective* (ANO) value for each vertex, where a vertex with a high ANO ends up with valuable information relative to other vertices, regardless of what it starts with. This is the 'information value' of a vertex that was referred to earlier.

Results. All ANO values were found to converge in under 250 random trials. Figure 1 (left) shows the *initial* cumulative distribution of the ANO values of all individuals in the network, which are strongly clustered and have very low dispersion. We would expect a similar final ANO distribution if all individuals were, on average, comparable in terms of their network position, *i.e.*, no vertex would consistently end with high normalized objective values from different starting states. Figure 1 (right) shows that this is clearly not the case. The final distribution of ANO values exhibits stratification; there are a small number of vertices that consistently end with the highest-scoring schemata, regardless of their initial state. The highest scoring individual, for example, ends with an objective value that is on average equal to 92% of the best objective value in the population, regardless of initial state.

Is the 'information elitism' we see in Figure 1 a genuine feature of the network, or the result of some trivial network property like degree? In order to determine the answer to this question, Figure 2 shows a correlation matrix of the final ANO values of the top 2% of vertices compared to several simple network features of each node, such as the in-degree (number of incoming email partners), the total number of emails received from all neighbors, and the time at which a vertex was first observed in the dataset. The cell $(i, j < i)$

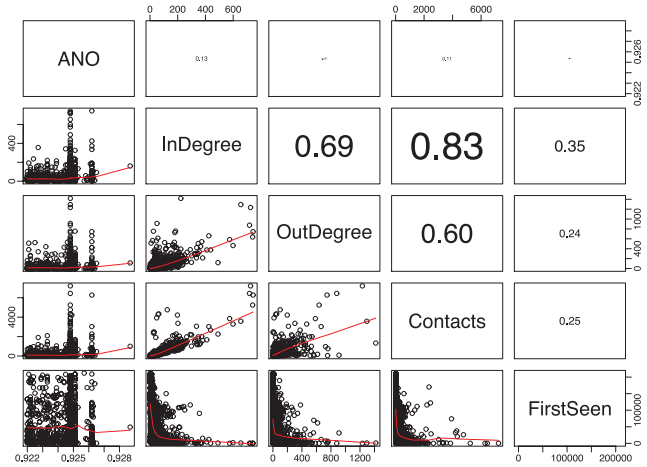

Figure 2: A correlation matrix comparing the final average normalized objective (ANO) score to four network properties, for individuals in top 2% of ANO. Numerals represent correlation coefficients between the corresponding pairs of variables; lines are fitted using locally weighted regression.

in Figure 2 plots feature i against feature j, and the numeral in cell (j, i) is the corresponding correlation coefficient.

Figure 2 shows that high ANO values cannot be explained by trivial graph features like degree. Intuition might suggest that a person who receives e-mail from many people, *i.e.*, a vertex with high in-degree, would be an accumulator of information, with correspondingly high ANO scores, but the correlation between ANO and in-degree is just 0.13 for the top 2% of vertices, and 0.34 for the entire dataset. The correlations are similarly low for out-degree. The high pairwise correlations between in-degree, out-degree, and total volume of incoming and outgoing e-mails traffic (the *Contacts* feature) for the top 2% of vertices is to be expected.

The high ANO values also cannot be explained by *dynamic* features of the vertices, such as the total number of e-mails sent and received by a vertex, or the length of time that a vertex has been present in the dataset (the *FirstSeen* feature). The latter accounts for the possible bias against vertices introduced later in the dataset, which have less time to accumulate information. The correlation coefficients for these two features are also extremely low for both the top 2% of vertices as well as the entire dataset.

Thus, we have shown that there is a stratification in final ANO values, indicating a corresponding stratification in the network position of individuals with regard to information flow, and that simple network features in isolation do not explain this stratification. Whether a more complex nonlinear relationship exists between the ANO and these or other, more complicated, network features is an interesting question for future research, as is the utility of the ANO as a ranking measure for nodes in the network.

Conclusion

We have shown that a form of the canonical genetic algorithm paired with Holland's hyperplane-defined functions

can be used as a large, rich class of diffusion models for static and dynamic social networks. We described formal reductions from two diffusion models in the literature: the independent cascade model, and a similar model in which two or more mutually exclusive processes are competing with each other to affect nodes in a network.

We also presented a case study on the Enron e-mail dataset, modeling probabilistic information flow between people as they exchange e-mails. The results indicated that a small section of vertices in the dataset are privileged in terms of the structure and dynamics of the network, and consistently receive more information than other vertices, regardless of how much they start with. We also showed that this phenomenon in the Enron dataset is not a trivial result that can be explained by simple network properties. The relationship between a vertex's non-trivial network properties (*e.g.,* PageRank or betweenness centrality), and its level of 'information value' is a relationship that warrants further study. It would also be interesting to see if this phenomena, like the commonly observed skew in degree distribution (Newman 2003), exists across different networks.

Finally, we note that the GADM presented here utilizes a very basic form of the genetic algorithm. Like the profusion of different types of GAs themselves, it is simple to extend the model presented here to model even more complex phenomena. For instance, the incorporation of mutation into the GADM loop would encompass a richer class of diffusion models, as would the introduction of more complex state replacement rules after crossover, instead of the retain-the-best method described here. Another interesting avenue of research is to investigate the possible use of genetic programming to learn a diffusion model that matches an observed spread (difficult as they are to obtain), while incorporating external properties of vertices.

Acknowledgements. Parts of this research were performed over a *café con leche* while both authors were at Telefonica Research in Madrid, Spain. M.L. is supported by a UIC Dean's Scholar Fellowship.

References

Anshelevich, E.; Chakrabarty, D.; Hate, A.; and Swamy, C. 2009. Approximation algorithms for the Firefighter problem: Cuts over time and submodularity. *Algorithms and Computation* 974–983.

Aspnes, J.; Rustagi, N.; and Saia, J. 2007. Worm versus alert: Who wins in a battle for control of a large-scale network? *Principles of Distributed Systems* 443–456.

Bharathi, S.; Kempe, D.; and Salek, M. 2007. Competitive influence maximization in social networks. *LNCS* 4858:306–311.

Cantu-Paz, E. 1998. A survey of parallel genetic algorithms. *Calculateurs Paralleles, Reseaux et Systems Repartis* 10(2):141–171.

Chakrabarti, D.; Wang, Y.; Wang, C.; Leskovec, J.; and Faloutsos, C. 2008. Epidemic thresholds in real networks. *Trans. Inf. Syst. Secur.* 10(4):1–26.

Cointet, J. P., and Roth, C. 2007. How realistic should knowledge diffusion models be? *J. Artificial Societies and Social Simulation* 10(3):5.

Domingos, P., and Richardson, M. 2001. Mining the network value of customers. In *Proc. 7th KDD*, 57–66. ACM.

Goldberg, D. E. 1989. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. 2004. Information diffusion through blogspace. In *Proc. 13th Conf. on WWW*, 491–501.

Habiba; Yu, Y.; Berger-Wolf, T.; and Saia, J. 2008. Finding spread blockers in dynamic networks. In *Proceedings of the 2nd SNA-KDD Workshop*.

Hethcote, H. 1989. Three basic epidemiological models. *Biomathematics* 18:119–144.

Holland, J. H. 2000. Building blocks, cohort genetic algorithms, and hyperplane-defined functions. *Evol. comp.* 8(4):373–391.

Keeling, M. 1999. The effects of local spatial structure on epidemiological invasions. *Proceedings of the Royal Society B: Biological Sciences* 266(1421):859.

Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *Proc. 9th ACM SIGKDD*, 137–146.

Khelil, A.; Becker, C.; Tian, J.; and Rothermel, K. 2002. An epidemic model for information diffusion in MANETs. In *Proc. 5th ACM workshop on modeling, analysis and simulation of wireless and mobile systems*, 54–60. ACM New York, NY, USA.

Kleinberg, J. 2007. Cascading behavior in networks: Algorithmic and economic issues. *Algorithmic game theory*.

Lahiri, M.; Maiya, A. S.; Sulo, R.; Habiba; and Berger-Wolf, T. Y. 2008. The impact of structural changes on predictions of diffusion in networks. In *IEEE ICDM Workshops*, 939–948.

Min, Y.; Jin, X.; Su, X.; and Peng, B. 2006. Empirical analysis of the spatial genetic algorithm on small-world networks. *Lecture Notes in Computer Science* 3993:1032.

Newman, M. 2002. Spread of epidemic disease on networks. *Physical Review E* 66(1):16128.

Newman, M. E. J. 2003. The structure and function of complex networks. *SIAM Rev.* 45(2):167–256.

Payne, J., and Eppstein, M. 2006. Emergent mating topologies in spatially structured genetic algorithms. In *Proc. 8th GECCO*, 207–214.

Scardovi, L., and Sepulchre, R. 2006. Collective optimization over average quantities. In *Proc. IEEE Conf. on Decision and Control*, 3369–3374.

Valente, T. 1996. Network models of the diffusion of innovations. *Computational & Mathematical Organization Theory* 2(2):163–164.