Optimizing and Learning Diffusion Behaviors in Complex Network

Xiaojian Wu
School of Computer Science
University of Massachusetts Amherst
xiaojian@cs.umass.edu

Abstract

Many dynamic phenomena can be modeled as a diffusion process. For my dissertation, I study diffusion processes in the area of sustainability, such as how wildlife spreads over a fragmental landscape and how fish spread within a river network, and try to answer two important questions. 1) How to shape the diffusion by using a limited amount of resources, for example how to maximize the spread of birds by preserving a limited number of landscape units? 2) How to model the diffusion process and estimate the parameters of the model using incomplete and noisy observations? This document describes my current research progress and future research directions of answering these two important questions.

The Model of Diffusion Process

The dynamic phenomena such as the spread of information (Domingos and Richardson 2001; Kempe, Kleinberg, and Tardos 2003), the propagation of infectious disease among humans (Anderson and May 2002) are modeled as diffusion processes over underlying networks. The independent cascade model is widely used to model various types of diffusion processes (Kempe, Kleinberg, and Tardos 2003). The model is defined on a (directed) graph in which the vertices are either active or inactive. Each newly activated vertex v has one chance to activate each of its inactive neighboring vertices w independently with probability $p_{v,w}$, where (v, w) is an edge in the graph. The process is called progressive if vertices only go from inactive status to active status but not vice versa. It is called non-progressive process if vertices are allowed to return back to inactive status. Since any non-progressive process can be reduced into a progressive process using a layered graph (Kempe, Kleinberg, and Tardos 2003), we mainly study the progressive process. For simplicity, I assume that each activation takes one unit of time. Then, the progressive process can be simulated using the following procedure. At time step one, a subset of vertices called sources are active. At any time step, when a vertex v just becomes active, it attempts to activate its neighbor w succeeding with probabilities $p_{v,w}$. The outcome of this random event is sampled by flipping a coin of bias $p_{v,w}$. If the activation succeeds, the vertex w becomes active at the beginning of the next time step. The process becomes stable at certain time step when no new vertices become active.

Many natural behaviors of animals can be modeled as diffusion processes using the independent cascade model. An example is the fish migration process in which fish swim from the ocean upstream in a river network to access their historical habitats (O'Hanley and Tomberlin 2005). In the graph, each edge represents a stream segment and each vertex represents a junction of multiple streams. A vertex being active means that fish are able to access that junction point. The activation probability on the edge models the probability that fish can pass all barriers in that stream segment, such as dams, floodgates and culverts.

Planning and Optimization

To answer the first question, researchers created distinct optimization problems for different applications. For example, (Sheldon et al. 2010) formulate the problem of purchasing fragmental landscape to maximize the dispersal of Red-cockaded WoodPecker as an influence maximization problem in which a set of landscape units can be paid to be added into the networks to increase the connectivity. (O'Hanley and Tomberlin 2005) formulate the problem of removing instream barriers to maximize the spread of fish as an influence maximization problem in which the instreams barriers can be repaired by certain costs to increase the chance of fish to pass them. These optimization problems share a lot of similarities, but without being modeling in a consistent manner, it is hard to directly apply algorithm of one problem to another. Based on this motivation, we developed a uniform framework–stochastic network design–to model a broad class of network optimization problems (also including the above two problems) in the area of sustainability (Wu, Sheldon, and Zilberstein 2013b). In our framework, each edge of the graph is either present or absent which is stochastically determined by the probability on that edge. Being present means that the activation made by one of the vertex of this edge succeeds. Otherwise, it fails. Costly actions can be taken to raise or reduce these probabilities into various levels. Each vertex is associated with a reward or cost. A set of sources vertices are given as input. The objectives is to maximize the expected reward or minimize the expected cost or satisfy other more complicated optimization goals by deciding which actions to take subject to certain constraint on actions. For example, the costs of all actions being taken can not exceed a budget limit. In the barrier removal problem, the money to be paid to repair the barriers can not exceed the available funds.

My next goal is to know how to efficiently solve the various optimization problem instances within this general framework. It has been proved that the expected reward maximization problem is NP-hard even if the underlying graph is a tree. For general graphs, it is $\#P$-hard to calculate the probability that a vertex will end up being active given all the probabilities on edges. Therefore, we are looking for approximate algorithms.

In our paper (Wu, Sheldon, and Zilberstein 2013b), a sample average approximation based algorithm is given by extending the work of (Sheldon et al. 2010) to solve the expected reward maximization problem defined in our framework. The algorithm is faster than the previous algorithm but become unscalable if a large number of samples are used. We improved this algorithm by introducing a novel dual decomposition technique (Kumar, Wu, and Zilberstein 2012). The experimental results show that by using the dual decomposition technique, we can get a near optimal solution several times faster than the original algorithm. In (Wu, Sheldon, and Zilberstein 2014), we developed a much faster approximate algorithm for solving the stochastic network design problem in which the underlying network structure is a directed tree. The algorithm is proved to be FPTAS. Applying it to the barrier removal problem, we show that in practice our algorithm is much faster than an existing technique (O'Hanley and Tomberlin 2005).

In the next step, we would continue improving the algorithms that can work on general graphs. At the same time, we are looking for the efficient algorithm to solve stochastic network design problems with other types of objectives and constraints.

Learning Diffusion Models

In the area of social networks, many works have been done to model the diffusion process and the parameters of models have been estimated using the practical behaviors of the diffusion. For example, the behavior may involve the timing of a person posting some target information on their Twitter (Myers and Leskovec 2010; Gomez-Rodriguez, Leskovec, and Krause 2012). The learned models are accurate if a large amount of observations are available, which is not a problem in the area of social network, because, for example, it is easy to track the spread of rumor by using Facebook or Twitter. However, in the area of sustainability, the learning data is often incomplete and noisy because it is difficult and costly to monitor the animals' activities accurately. The defective observations make the traditional machine learning algorithms ineffective or even not applicable.

For my dissertation, I have been trying to derive effective learning algorithms to deal with those defective learning data. In our paper (Wu, Sheldon, and Zilberstein 2013a), we propose an algorithm to estimate the spread process of Red-cockaded Woodpecker using logistic regression model. In the data, where and when the birds are observed or unobserved are recorded. But for a lot of locations and time slots, no records are available. Our approach is able to estimate the parameters of the diffusion model accurately even with 80% missing data. I plan to continue to explore the methods of estimating other probabilistic models or to deal with the cases where no models are assumed. At the same time, I am interested in working on another associated problem: how to selectively monitor the diffusion process of species using a limited amount of resources such that the learned model is accurate. For example, how to deploy a limited number of sensors for monitoring birds' activities such that the observations collected by these sensors can be used to learn an accurate model.

References

Anderson, R. M., and May, R. M., eds. 2002. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.

Domingos, P., and Richardson, M. 2001. Mining the network value of customers. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 57–66.

Gomez-Rodriguez, M.; Leskovec, J.; and Krause, A. 2012. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery From Data* 5(4):21.

Kempe, D.; Kleinberg, J.; and Tardos, E. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 137–146.

Kumar, A.; Wu, X.; and Zilberstein, S. 2012. Lagrangian relaxation techniques for scalable spatial conservation planning. In *Proceedings of the 26th Conference on Artificial Intelligence*, 309–315.

Myers, S. A., and Leskovec, J. 2010. On the convexity of latent social network inference. In *Advances in Neural Information Processing Systems*, 1741–1749.

O'Hanley, J. R., and Tomberlin, D. 2005. Optimizing the removal of small fish passage barriers. *Environmental Modeling and Assessment* 10(2):85–98.

Sheldon, D.; Dilkina, B.; Elmachtoub, A.; Finseth, R.; Sabharwal, A.; Conrad, J.; Gomes, C.; Shmoys, D.; Allen, W.; Amundsen, O.; and Vaughan, W. 2010. Maximizing the spread of cascades using network design. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 517–526.

Wu, X.; Sheldon, D.; and Zilberstein, S. 2013a. Parameter learning for latent network diffusion. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligene*.

Wu, X.; Sheldon, D.; and Zilberstein, S. 2013b. Stochastic network design for river networks. In *NIPS Workshop on Machine Learning for Sustainability*.

Wu, X.; Sheldon, D.; and Zilberstein, S. 2014. Rounded dynamic programming for tree-structured stochastic network design. *Accepted in AAAI 2014*.