Diffusion in Social and Information Networks: Research Problems, Probabilistic Models & Machine Learning Methods

Manuel Gomez-Rodriguez
MPI for Software Systems
manuelgr@mpi-sws.org

Le Song
Georgia Institute of Technology
lsong@cc.gatech.edu

ABSTRACT

In recent years, there has been an increasing effort on developing realistic models, and learning and inference algorithms to understand, predict, and influence diffusion over networks. This has been in part due to the increasing availability and granularity of large-scale diffusion data, which, in principle, allows for understanding and modeling not only macroscopic diffusion but also microscopic (node-level) diffusion. To this aim, a bottom-up approach has been typically considered, which starts by considering how particular ideas, pieces of information, products, or, more generally, *contagions* spread locally from node to node apparently at random to later produce global, macroscopic patterns at a network level. However, this bottom-up approach also raises significant modeling, algorithmic and computational challenges which require leveraging methods from machine learning, probabilistic modeling, event history analysis and graph theory, as well as the nascent field of network science. In this tutorial, we will present several diffusion models designed for fine-grained large-scale diffusion data, present some canonical research problem in the context of diffusion, and introduce state-of-the-art algorithms to solve some of these problems, in particular, network estimation, influence estimation and influence control.

1. TARGET AUDIENCE, AIMS AND PREREQUISITES

This tutorial is meant for a broad audience at WWW, including students and researchers specifically interested in models and machine learning methods in the context of diffusion over networks. It will provide an introduction to research problems, probabilistic models and machine learning methods in the context of diffusion of information, influence and behaviors over networks. No specific knowledge will be required beyond basic probability and graph theory; the tutorial is self-contained and most of the foundational concepts are introduced during the tutorial.

The participants will learn about the fundamental problems in diffusion networks, including network inference, influence estimation and influence control. We will present a concise overview of the main models and algorithms, the most recent theoretical results, and an overview of empirical results. We will provide pointers to code implementation of most of the algorithms, to allow participants to carry out a hands-on further investigation. Finally, the participants will also learn about open problems in the context of diffusion over networks.

2. FORMAT AND OUTLINE OF THE TUTORIAL

We will divide our **half-day tutorial** into four parts. In the first section, we will introduce different discrete time and continuous time models of diffusion over networks and introduce several canonical problems in the context of diffusion over networks. The remaining sections will be devoted to three of the above mentioned problems: network inference, influence estimation and influence control. For each problem, we will present and compare different algorithms, highlighting their assumptions and theoretical properties, as well as their performance in practice and limitations. We will focus on principled flexible methods, which can be adapted to different experimental setups and data sources.

2.1 Diffusion Models (45 minutes)

We will provide an introduction to modeling diffusion of information, influence, behaviors, products, or, more generally, *contagions* over networks. We will start with describing several idealized discrete-time models [18], and then continue introducing several continuous-time models, which allow to more realistically fit both non-recurrent [10, 13] and recurrent [17, 22] real diffusion data. We will conclude by introducing several canonical problems in the context of diffusion over networks, where we will highlight the importance of each problem and the progress that has been made so far in each of them. In the remaining parts of the tutorial, we will focus on three of these canonical problems: network inference, influence estimation and influence control.

2.2 Network Estimation (45 minutes)

Given a set of cascades and a diffusion model, the network inference problem consists of inferring the edges (and possibly model parameters) of an unobserved underlying network. The network inference problem has attracted significant attention in recent years, since it is essential to reconstruct and predict the paths over which information can spread, and to maximize sales of a product or stop infections. Here, we will first present several algorithms that allow us to infer the structure (edges) of a network [1, 12, 16], using a natural diminishing property of the problem, submodularity. Then, we will introduce a family of algorithms that allow us to infer not only the structure but also the model parameters [7, 10, 11], using convex programming, and will outline a theoretical analysis of their sampling complexity [4]. We will then show how to extend

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2741989.

the latter algorithms to support dynamic networks [14], topic modeling [8], and recurrent events [22, 23].

2.3 Influence Estimation (45 minutes)

Given a diffusion model and an arbitrary set of nodes that initially adopt a piece of information, idea or product, the influence estimation problem which is the basis for subsequent influence control consists of estimating the average number of follow-ups these nodes can trigger. These estimation problems in various diffusion models are often NP-hard [2], and hence call for efficient approximation algorithms with provable guarantees. Furthermore, the problem can be complicated by additional constraints such as estimating influence given a time windows. Here, we will formally define the problem, and briefly overview various algorithms for different diffusion models [2, 3, 5, 6, 15]. We will then focus on two algorithms for continuous-time diffusion models with provable guarantees [6, 15], and show how they provide solutions to influence estimation problem with additional timing constraints, and exploit randomization to scale to networks with million of nodes.

2.4 Influence Control (45 minutes)

Given a diffusion model, the influence control problem aims to manipulate either the information sources or the diffusion networks themselves in order to achieve certain outcome of the diffusion processes. For instance, in influence maximization, one aims to find a set of nodes whose initial adoptions of certain piece of information, idea or product can trigger the largest expected number of follow-ups. In influence blocking, one aims to remove a set of diffusion channels from a diffusion network such that rumor spread can be controlled, or, in influence facilitation, to add a set of diffusion channels such that an important announcement can reach people early. In influence shaping, one aims to drive the overall usage of a service to a certain level per user by incentivizing a small number of users to take more initiatives. Here, we will formally define the influence maximization, influence blocking and influence shaping problems and present several algorithms to solve them under different discrete-time diffusion models [2, 3, 18, 20, 21] and continuous-time diffusion models [6, 9, 15, 19].

3. TUTORS' BIO AND EXPERTISE

Manuel Gomez Rodriguez is an tenure-track independent research group leader at Max Planck for Software Systems. Manuel develops machine learning and large-scale data mining methods for the analysis and modeling of large real-world networks and processes that take place over them. He is particularly interested in problems motivated by the Web and social media and has received several recognitions for his research, including an Outstanding Paper Award at NIPS´13 and a Best Research Paper Honorable Mention at KDD´10. Manuel holds a PhD in Electrical Engineering from Stanford University and a BS in Electrical Engineering from Carlos III University in Madrid (Spain), and has been a Barrie de la Maza Fellow and a Caja Madrid Fellow.

Le Song is an assistant professor in the College of Computing, Georgia Institute of Technology. His principal research interests lie in the development of machine learning methodology, especially in kernel methods, probabilistic graphical models, temporal data and network analysis. Le Song received his Ph.D. in Computer Science from University of Sydney in 2008, and then conducted his post-doctoral research in the School of Computer Science, Carnegie Mellon University, between 2008 and 2011. Before he joined Georgia Institute of Technology, he worked briefly as a research scientist at Google. He is the winner of Outstanding Paper Award at NIPS´13 and Best Paper Award at ICML´10.

References

[1] B. Abrahao, F. Chierichetti, R. Kleinberg, and A. Panconesi. Trace complexity of network inference. In *KDD*, 2013.

[2] W. Chen, C. Wang, and Y. Wang. Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks. In *KDD*, 2010.

[3] W. Chen, Y. Wang, and S. Yang. Efficient Influence Maximization in Social Networks. In *KDD*, 2009.

[4] H. Daneshmand, M. Gomez-Rodriguez, L. Song, and B. Schölkopf. Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. In *ICML*, 2014.

[5] N. Du, Y. Liang, M. Balcan, and L. Song. Influence function learning in information diffusion networks. In *ICML*, 2014.

[6] N. Du, L. Song, M. Gomez-Rodriguez, and H. Zha. Scalable Influence Estimation in Continuous-Time Diffusion Networks. In *NIPS*, 2013.

[7] N. Du, L. Song, A. Smola, and M. Yuan. Learning Networks of Heterogeneous Influence. In *NIPS*, 2012.

[8] N. Du, L. Song, H. Woo, and H. Zha. Uncover topic-sensitive information diffusion networks. In *AISTATS*, 2013.

[9] M. Farajtabar, N. Du, M. Gomez-Rodriguez, I. Valera, H. Zha, and L. Song. Shaping social activity by incentivizing users. In *NIPS*, 2014.

[10] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the Temporal Dynamics of Diffusion Networks. In *ICML*, 2011.

[11] M. Gomez-Rodriguez, J. Leskovec, D. Balduzzi, and B. Schölkopf. Uncovering the Structure and Temporal Dynamics of Information Propagation. *Network Science*, 2014.

[12] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring Networks of Diffusion and Influence. In *KDD*, 2010.

[13] M. Gomez-Rodriguez, J. Leskovec, and B. Schölkopf. Modeling Information Propagation with Survival Theory. In *ICML*, 2013.

[14] M. Gomez-Rodriguez, J. Leskovec, and B. Schölkopf. Structure and Dynamics of Information Pathways in On-line Media. In *WSDM*, 2013.

[15] M. Gomez-Rodriguez and B. Schölkopf. Influence Maximization in Continuous Time Diffusion Networks. In *ICML*, 2012.

[16] M. Gomez-Rodriguez and B. Schölkopf. Submodular Inference of Diffusion Networks from Multiple Trees. In *ICML*, 2012.

[17] T. Iwata, A. Shah, and Z. Ghahramani. Discovering latent influence in online social activities via shared cascade poisson processes. In *KDD*, 2013.

[18] D. Kempe, J. M. Kleinberg, and E. Tardos. Maximizing the Spread of Influence Through a Social Network. In *KDD*, 2003.

[19] E. B. Khalil, B. Dilkina, and L. Song. Scalable diffusion-aware optimization of network topology. In *KDD*, 2014.

[20] M. Kimura, K. Saito, and H. Motoda. Minimizing the spread of contamination by blocking links in a network. In *AAAI*, 2008.

[21] H. Tong, B. A. Prakash, T. Eliassi-Rad, M. Faloutsos, and C. Faloutsos. Gelling, and melting, large graphs by edge manipulation. In *CIKM*, 2012.

[22] K. Zhou, H. Zha, and L. Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *AISTATS*, 2013.

[23] K. Zhou, H. Zha, and L. Song. Learning triggering kernels for multi-dimensional hawkes processes. In *ICML*, 2013.