Understanding and Managing Cascades on Large Graphs

B. Aditya Prakash
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA, USA
badityap@cs.cmu.edu

Christos Faloutsos
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA, USA
christos@cs.cmu.edu

ABSTRACT

How do contagions spread in population networks? Which group should we market to, for maximizing product penetration? Will a given YouTube video go viral? Who are the best people to vaccinate? What happens when two products compete? The objective of this tutorial is to provide an intuitive and concise overview of most important theoretical results and algorithms to help us understand and manipulate such propagation-style processes on large networks. The tutorial contains three parts: (a) Theoretical results on the behavior of fundamental models; (b) Scalable Algorithms for changing the behavior of these processes e.g., for immunization, marketing etc.; and (c) Empirical Studies of diffusion on blogs and on-line websites like Twitter.

The problems we focus on are central in surprisingly diverse areas: from computer science and engineering, epidemiology and public health, product marketing to information dissemination. Our emphasis is on intuition behind each topic, and guidelines for the practitioner.

1. INTRODUCTION

Graphs are ubiquitous and large-scale, from social networks, computer networks, mobile call networks, the World Wide Web, to protein interaction networks, and many more. In addition, propagation processes over them can give rise to astonishing macroscopic behavior, leading to challenging and exciting research problems in surprisingly diverse areas: from computer science and engineering, epidemiology and public health, product marketing to information dissemination. How do contagions spread in population networks? Which group should we market to, for maximizing product penetration? How stable is a predator-prey ecosystem, given intricate food webs? How do rumors spread on Twitter/Facebook? Questions such as how blackouts can spread on a nationwide scale, and how social systems evolve on the basis of individual interactions are all also related to propagation/cascade-like phenomena on networks.

Proceedings of the VLDB Endowment, Vol. 5, No. 12

The objective of this tutorial is to provide an intuitive and concise overview of most important theoretical results and algorithms to help us understand and manipulate such propagation-style processes on large networks. The target audience is data mining and data management researchers, who wish to learn more about models and tools for dealing with cascade-like processes on large datasets. There is special focus on the cross-disciplinary aspect of the concepts and tools involved. For maximum benefit, the expected prerequisite is an undergraduate degree in Computer Sc. or a related field. However, the tutorial's emphasis is on the intuition behind the material.

2. TUTORIAL OUTLINE

The tutorial contains three parts:

1. Theoretical results on the behavior of fundamental models abstracting such processes
2. Scalable Algorithms for changing the behavior of these processes e.g., for immunization, marketing etc.
3. Large-scale empirical studies of diffusion on blogs and on-line websites like Twitter

2.1 Theory

In this part, the we start by explaining the most common propagation models including the so-called SIS (Susceptible-Infected-Susceptible) "flu-like" model [1, 12], the IC (Independent Cascade) model [8], the so-called 'Bass' model [2] for product-adoption and so on [11, 7]. The goal is also to learn about fundamental properties of such processes in a variety of settings. Chakrabarti et al. [4] and Ganesh et al. [9] found that, for the flu-like SIS model, the epidemic threshold for any arbitrary graph depends on the leading eigenvalue of the adjacency matrix of the graph. Prakash et. al. [23] further discovered that the leading eigenvalue and a model-dependent constant are the only parameters that determine the epidemic threshold for almost all virus propagation models. Prakash et. al. gave the epidemic threshold for arbitrarily varying dynamic networks [24, 27]. We will also cover recent work on understanding models involving competition between multiple contagions ('iPhone vs Android') [20, 21, 22, 3]. We will also demonstrate how many such processes are similar and have the same core problems.

2.2 Algorithms

In this part, the aim is to leverage and utilize the understanding gained in Part 1, to actually manage such processes for our benefit - like algorithms for finding best people to immunize [5, 26, 24], finding the best people to market

to [13, 17, 10], algorithms to reverse-engineer epidemics like by finding the culprits [25, 14] etc. We also demonstrate how such optimization problems are closely related to the fundamental properties discussed before.

2.3 Empirical Studies

The goal in this part is to present large-scale studies on real-datasets and different scenarios, which will help one understand how to track the flow of pieces of information diffusing among the users (i.e. information cascades) [18, 15], how popularity of any 'meme' or contagion changes over time [6, 16, 28], how to use such patterns to improve models [19] etc. Due to an unprecedented availability of large datasets, the focus is more in the social media domain here.

3. BIOGRAPHICAL SKETCHES

B. Aditya Prakash is a PhD. student in the Computer Science Department, Carnegie Mellon University. He got his B.Tech (in CS) from the Indian Institute of Technology (IIT) - Bombay. He has published 16 refereed papers in major venues and holds two U.S. patents. His interests include Data Mining, Applied Machine Learning and Databases, with emphasis on large real-world networks and time-series. He will soon be joining the CS Department at Virginia Tech. as an Assistant Professor.

Christos Faloutsos is a Professor at Carnegie Mellon University. He is an ACM Fellow, he has published over 200 refereed articles and he has given over 30 tutorials in database and data mining venues. He has received the Presidential Young Investigator Award by NSF (1989), the Research Contributions Award in ICDM 2006, the SIGKDD Innovations Award (2010) and 18 best paper awards (including two "test of time awards). His research interests include data mining for graphs & streams, fractals, database performance, and indexing for multimedia & bio-informatics data.

4. ACKNOWLEDGMENTS

This material is based upon work supported by the Army Research Laboratory (ARL) under Cooperative Agreement Number W911NF-09-2-0053, the National Science Foundation (NSF) under Grants Number IIS-1017415 and CNS-0721736 and a Sprint gift. Any opinions, findings, and conclusions or recommendations in this material are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government, the ARL and the NSF, or other funding parties.

5. REFERENCES

[1] R. M. Anderson and R. M. May. *Infectious Diseases of Humans*. Oxford University Press, 1991.

[2] F. M. Bass. A new product growth for model consumer durables. *Management Science*, 15(5):215–227, 1969.

[3] A. Beutel, B. A. Prakash, R. Rosenfeld, and C. Faloutsos. Interacting viruses on a network: Can both survive? *SIGKDD*, 2012.

[4] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos. Epidemic thresholds in real networks. *ACM TISSEC*, 10(4), 2008.

[5] R. Cohen, S. Havlin, and D. ben Avraham. Efficient immunization strategies for computer networks and populations. *Physical Review Letters*, 91(24), Dec. 2003.

[6] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. In *PNAS*, 2008.

[7] P. S. Dodds and D. J. Watts. A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 232:587–604, September 2004.

[8] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[9] A. Ganesh, L. Massoulie, and D. Towsley. The effect of network topology in spread of epidemics. *IEEE INFOCOM*, 2005.

[10] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan. A data-based approach to social influence maximization. *Proc. VLDB Endow.*, pages 73–84, 2011.

[11] M. Granovetter. Threshold models of collective behavior. *Am. Journal of Sociology*, 83(6):1420–1443, 1978.

[12] H. W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42, 2000.

[13] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[14] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila. Finding effectors in social networks. In *SIGKDD*, pages 1059–1068, 2010.

[15] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *EC*, 2006.

[16] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. *ACM SIGKDD*, 2009.

[17] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. S. Glance. Cost-effective outbreak detection in networks. In *KDD*, pages 420–429, 2007.

[18] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs: Patterns and a model. In *SDM*, 2007.

[19] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: Model and implications. In *SIGKDD*, 2012.

[20] M. E. J. Newman. Threshold effects for two pathogens spreading on a network. *Physical Review Letters*, 95(10):108701, September 2005.

[21] N. Pathak, A. Banerjee, and J. Srivastava. A generalized linear threshold model for multiple cascades. *ICDM*, 2010.

[22] B. A. Prakash, A. Beutel, R. Rosenfeld, and C. Faloutsos. Winner takes all: Competiting viruses or ideas on fair-play networks. *WWW*, 2012.

[23] B. A. Prakash, D. Chakrabarti, M. Faloutsos, N. Valler, and C. Faloutsos. Threshold conditions for arbitrary cascade models on arbitrary networks. In *ICDM*, 2011.

[24] B. A. Prakash, H. Tong, N. Valler, M. Faloutsos, and C. Faloutsos. Virus propagation on time-varying networks: Theory and immunization algorithms. *ECML-PKDD*, 2010.

[25] D. Shah and T. Zaman. Detecting sources of computer viruses in networks: theory and experiment. In *SIGMETRICS*, pages 203–214, 2010.

[26] H. Tong, B. A. Prakash, C. E. Tsourakakis, T. Eliassi-Rad, C. Faloutsos, and D. H. Chau. On the vulnerability of large graphs. In *ICDM*, 2010.

[27] N. Valler, B. A. Prakash, H. Tong, M. Faloutsos, and C. Faloutsos. Epidemic spread in mobile ad hoc networks: Determining the tipping point. *IFIP NETWORKING*, 2011.

[28] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, pages 177–186, 2011.