Information Diffusion In Social Networks: Observing and Influencing Societal Interests*

Divyakant Agrawal Ceren Budak Amr El Abbadi
Department of Computer Science UCSB
Santa Barbara, CA 93106-5110, USA
{agrawal, cbudak, amr}@cs.ucsb.edu

1. INTRODUCTION

Social networks provide great opportunities for social connection, learning, political and social change, as well as individual entertainment and enhancement in a wide variety of forms. Because many social interactions currently take place in online networks, social scientists have access to unprecedented amounts of information about social interaction. This wealth of data can allow scientists to study social interactions on a scale and at a level of detail that has never been possible before.

In addition to providing a platform for scientists to observe social interactions at large scale, online social networks are also changing the very nature of social interactions. The process by which people locate individuals with shared interests, the number and nature of information sources, and the ability to share ideas across various topics have all undergone dramatic change. For instance, social networks have emerged as an important medium for the widespread distribution of news and instructions in events such as the 2008 U.S. Presidential Election and emergencies like the landfall of Hurricanes Ike and Gustav [12]. In light of these notable outcomes, understanding information diffusion in social networks is a critical research goal. This greater understanding can be achieved through data analysis, the development of reliable models that can predict outcomes of social processes, and ultimately the creation of applications that can shape the outcome of these processes. In this tutorial, we aim to provide an overview of such recent research based on a wide variety of techniques such as optimization algorithms, data mining, data streams covering a large number of problems such as influence spread maximization, misinformation limitation and study of trends in online social networks.

2. TUTORIAL OUTLINE

2.1 Characterizing Social Networks

A clear understanding of information diffusion in online social networks can not be achieved without a clear understanding of characteristics of social networks. Therefore, we start our tutorial with a discussion of the significance of social networks in today's world. Next we give a survey of measurement studies focusing on important characteristics that distinguish social networks from other networks [17]. At the end of this section, our goal is to formally identify the characteristics of social networks and their uses.

2.2 Diffusion of Information or Opinions

In this part of the tutorial, we outline the techniques used in optimizing or facilitating information diffusion in social networks. We identify two problem definitions through which a broad survey of techniques in recent research is provided. Namely, we explore the problems of maximizing the spread of influence and minimizing the spread of misinformation in social networks. As different as these problems are in terms of the motivation behind them, they both rely on sub-problems that are very similar. Through our study of these two problems, we delve into more detail about the sub-problems; (i) Diffusion model formation, (ii) formalization and optimization, (iii) large-scale data analysis and (iv) technique validation.

Diffusion Model Formation. Central to optimization problems relating to information diffusion is the problem of identifying the right diffusion model. Therefore, we provide a survey of available models and address the following questions: What are the necessary and sufficient parameters of an accurate model? How can we validate the use of a specific model? How can one obtain data about the parameters? Given the intricacy of human interactions, finding the right diffusion model is still an open problem, even in the presence of the large data sets available today. A large body of recent research on information diffusion explored a variety of diffusion models [4, 8, 10, 13, 16]. We will give an overview of such studies as well as studies validating some of these models on real data.

Formalization and Optimization. Depending on the exact definition of *influence* or the information diffusion model, employing an effective strategy for facilitating the spread of influence requires a different problem formalization. An intensively studied problem formulation is: Given a number k, which set S with $|S| \leq k$ will eventually infect as many nodes of graph G as possible? We give a summary of recent research on this and related optimization problems [1, 6, 7, 9, 11, 13] and discuss the advantages and disadvantages associated with each technique. We delve into more detail of one data-centric study that identifies different types of *influencers* [3]. This technique is orthogonal to the first set of studies discussed since it relies on data from real social networks and diffusion scenarios rather than assumed models.

A characteristic common to the studies discussed so far in the tutorial is the assumption that information cascades of campaigns happen in isolation. Next we discuss a group of problem formulations that capture the notion of competing campaigns [2, 4]. We also explore the problem of limiting the spread of misinformation

*This work is partially supported by NSF Grant IIS-1135389.

Proceedings of the VLDB Endowment, Vol. 4, No. 12

and discuss the desired characteristics of possible solutions. We specifically focus on [4] that studies this problem in various contexts and considers an extension of the problem with missing data showing that prediction in this setting is a supermodular problem.

Large-Scale Data Analysis. No matter which technique is used in studying information diffusion, large-scale data analysis is a significant aspect of the study as well as a significant challenge. We will explore such challenges and possible solutions. With the increase of studies in social networks, there are a number of data sets available to researchers [14, 15]. As obtaining data is crucial for studying information diffusion in social networks, we will give an overview of publicly available data sets and their characteristics.

Technique Validation. Without proper validation, it is impossible to see the impact of the theoretical research on real social networks. Therefore, we take a closer look into the validation techniques used by the recent research surveyed in this tutorial and aim to identify the proper validation methods.

2.3 Information Trends

In the second part of the tutorial we will focus on the interplay between different information campaigns and *trends* that result from such interplay. We will provide a summary of research on trend analysis in social networks and focus on the following questions: What is the nature of trends in social networks? What are important dimensions? How can one efficiently analyze trends in large scale social networks? Trends in social networks have recently been a major focus of interest among researchers and industry studying them from perspectives such as temporal and geographical dimensions or the sentiment embedded in the shared information [16, 18]. We will give a summary of such studies from both research and industry. Since information diffusion is a substantial part of the process that creates the information trends, properties that are defined in this context are of significant interest. We delve into one recent study that analyzes trends from the perspective of the structural properties of the graph [5] by identifying trends based on the number of topic discussions of connected (or disconnected) pairs of users. Since large-scale data analysis is an important challenge, we will focus on challenges associated with large-scale data analysis as well as possible solutions.

2.4 Other Research in Social Networks

Even though the focus of this tutorial is on information diffusion, we note that there are a large number of interesting studies relating to other aspects of social networks. In the last section of our tutorial, we will give an overview of studies about community formation and evolution and spam detection and elimination which have a great impact on information diffusion in social networks. Social network analytics will greatly benefit from the advances that have been made in the area of graph mining and graph data analysis. We will establish this connection through a high overview of examples of graph mining approaches to such problems.

3. GOALS OF THE TUTORIAL

3.1 Learning Outcomes

- Overview of important characteristics of social networks.
- Understanding various social interaction/diffusion models.
- Learning about state-of-the-art in data analysis and optimization problems in this area.
- A discussion about the appropriate validation methods and a list of open research challenges in social networks from the perspective of the applicability of research results to real life.
- A broad overview of other significant sub-categories of studies in social networks in addition to information diffusion.

3.2 Intended Audience

This tutorial is intended to benefit researchers and industry in the broad area of social networks. Our tutorial will provide a survey of the current research and applications relating to information diffusion in social networks and therefore provide essential knowledge for building new models, algorithms and systems in this area.

4. BIOGRAPHICAL SKETCHES

Divyakant Agrawal is a Professor of the Department of Computer Science at University of California, Santa Barbara. Prof. Agrawal's research expertise is in the areas of database systems, distributed computing, data warehousing, and large-scale information systems. **Ceren Budak** is a PhD Candidate at the Department of Computer Science, University of California Santa Barbara. Her research interests lie in the area of information diffusion in social networks. **Amr El Abbadi** is a Professor of the Department of Computer Science at University of California, Santa Barbara. His research interests lie in the area of scalable database and distributed systems.

5. REFERENCES

[1] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *WSDM*, pages 65–74, 2011.

[2] S. Bharathi, D. Kempe, and M. Salek. Competitive influence maximization in social networks. In *WINE*, pages 306–311, 2007.

[3] C. Budak, D. Agrawal, and A. El Abbadi. Where the blogs tip: connectors, mavens, salesmen and translators of the blogosphere. In *SIGKDD Workshop on Social Media Analytics*, 2010.

[4] C. Budak, D. Agrawal, and A. El Abbadi. Limiting the spread of misinformation in social networks. In *WWW*, 2011.

[5] C. Budak, D. Agrawal, and A. El Abbadi. Structural trend analysis for online social networks. In *VLDB*, 2011.

[6] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, 2010.

[7] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD*, pages 1029–1038, 2010.

[8] P. S. Dodds and D. J. Watts. A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 232:587–604, 2005.

[9] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD*, pages 57–66, 2001.

[10] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *KDD*, pages 1019–1028, 2010.

[11] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan. Discovering leaders from community actions. In *CIKM*, pages 499–508, 2008.

[12] A. L. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. In *ISCRAM'09*, 2009.

[13] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[14] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW '10*, pages 591–600, 2010.

[15] J. Leskovec. Stanford large network dataset collection. `http://snap.stanford.edu/data/index.html`, 2009.

[16] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD '09*, pages 497–506, 2009.

[17] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07*, pages 29–42, 2007.

[18] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *GIS*, pages 42–51, 2009.