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.
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

- 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.

**Amer 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


