Computer Communications 73 (2016) 12–21

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Phantom cascades: The effect of hidden nodes on information diffusion

Václav Belák[a], Afra Mashhadi[b], Alessandra Sala[b], Donn Morrison[c,*]

[a] Insight Centre for Data Analytics Galway, Ireland
[b] Bell Labs, Dublin, Ireland
[c] Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway

ARTICLE INFO

Article history:
Received 10 February 2015
Revised 19 May 2015
Accepted 9 July 2015
Available online 17 July 2015

Keywords:
Information cascades
Information diffusion
Online social networks
Independent cascade model

ABSTRACT

Research on information diffusion generally assumes complete knowledge of the underlying network. However, in the presence of factors such as increasing privacy awareness, restrictions on application programming interfaces (APIs) and sampling strategies, this assumption rarely holds in the real world which in turn leads to an underestimation of the size of information cascades. In this work we study the effect of hidden network structure on information diffusion processes. We characterise information cascades through activation paths traversing visible and hidden parts of the network. We quantify diffusion estimation error while varying the amount of hidden structure in five empirical and synthetic network datasets and demonstrate the effect of topological properties on this error. Finally, we suggest practical recommendations for practitioners and propose a model to predict the cascade size with minimal information regarding the underlying network.

1. Introduction

Simulating information diffusion processes is a critical aspect of understanding how information spreads in real world networks. A word-of-mouth viral marketing campaign is a well known example where a company wishes to estimate the spread of an advertisement or uptake of a product over a social network. A prerequisite for such studies is access to real network data over which the processes of diffusion can be studied. Online social networks (OSNs) are a natural choice for this purpose as they are readily available and in many cases constitute the desired diffusion medium (e.g., Facebook and Twitter). Consequently such datasets have been widely used in previous research [8,20,23,27].

However, access to the complete network in question is rarely possible. Factors such as privacy settings, application programming interface (API) restrictions and sampling strategies contribute to missing network structure. For example, it is well known that users of OSNs are growing increasingly privacy-aware. A recent large-scale study of 1.4 million Facebook users [6] revealed an increase in privacy-enabled profiles over 15 months from 17% to more than 50%, effectively rendering those users hidden from study yet still actively connected to their friends. A possible implication of this trend is that as these networks become more and more *partial*, research using this data to simulate the spread of information may become less accurate. The problem is illustrated by the example in Fig. 1 showing a hypothetical information cascade over two networks: a complete network referred to as the *oracle*, where information about all nodes and edges is known, and a partial view, where some portion is hidden.

This leads to an important question: how does partial network knowledge affect models of information diffusion? In other words, *how can we quantify the error introduced as we move away from the completeness assumption of the underlying network?, and how can we correct for it?* These are the central motivating questions we aim to address in this work. We focus on the *Independent Cascade Model* (ICM) [4,12,17,20] as a classic example of an information diffusion algorithm and measure the differences between information cascades on oracle and partial networks. To this end we formulate a three-step approach that begins with a complete network, samples a set of nodes of increasing size to be marked *hidden* (i.e., modelling users who have privacy enabled), and finally simulates information spreading across the resulting networks.

With the proposed methodology we are able to quantify the error introduced due to network partiality based on a theoretical oracle scenario (i.e., how *would have* the information spread with full knowledge). We pay particular attention to the paths that lead to the activation of visible nodes in the oracle network by distinguishing whether or not they were activated via hidden nodes. We refer to this class of activated nodes as the *phantom cascade* and set out to understand how the size of this cascade changes as a function of the percentage of hidden nodes.

As we explain in Section 2, there is limited prior work studying the effect of partial information on diffusion models [3,23,28] from which our work differs in two fundamental aspects: (i) by distinguishing

[*] Corresponding author.
E-mail addresses: vaclav@belak.net (V. Belák), afra.mashhadi@alcatel-lucent.com (A. Mashhadi), alessandra.sala@alcatel-lucent.com (A. Sala), donn.morrison@idi.ntnu.no, donn.morrison@deri.org (D. Morrison).

http://dx.doi.org/10.1016/j.comcom.2015.07.012
0140-3664/© 2015 Elsevier B.V. All rights reserved.

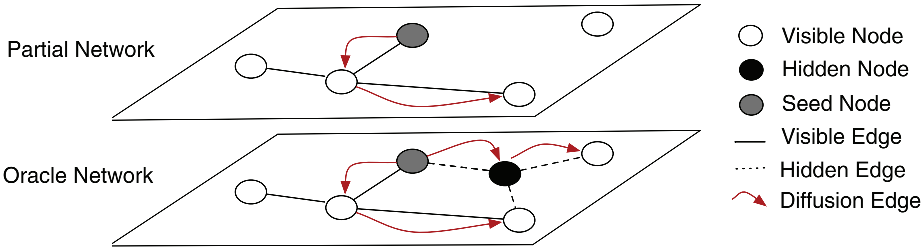

Fig. 1. An information cascade on an oracle (complete) network and its partially observed counterpart.

between cascade types we are able to compare the observed cascade on the partial network to a theoretical cascade on the oracle network, (ii) we focus on missing data at the network level (as opposed to the cascade level [23,28]) and study how information cascades over the incomplete network. To the best of our knowledge, this is the first study to quantify the error related to the diffusion cascade caused by nodes hidden at the network level.

The remainder of this paper is organised as follows. In Section 2 we discuss related work on missing data in information cascades. In Section 3 we present the datasets and the methodology we use to measure the effect of partial knowledge on the diffusion model. Section 4 presents the results in which we demonstrate the magnitude of the problem and its relation to topological properties specific to different networks. In Section 5 we introduce a model correcting for the effect of missing data. In Section 6.1 we discuss the implication and limitations of this work. Finally we envisage our future directions and conclude this study in Section 7.

2. Related work

The problem of missing data in networks has been addressed from different perspectives ranging from network sampling [5,10,22], where the aim is to obtain a representative subset of the network, to network reconstruction [8,14], where nodes and edges are inferred to recreate the original network. Other works, such as [5,15], have examined the effect of missing nodes and edges on topological metrics of the network (e.g., average node degree, diameter, clustering coefficient). These and other studies [19,21] have helped to build an understanding of the implications of research conducted on sampled or otherwise incomplete network data.

In the context of information diffusion, however, very little research has been conducted and thus the effect of missing data on the information cascades themselves is not well understood. Of the vast amount of research on information spreading and diffusion, we are aware of only four studies partially addressing this problem [3,23,27,28]. In [3], the authors uncovered a logarithmic error as a function of the amount of missing data for diffusion simulations on a small telecommunications call graph. Diffusions are simulated and compared on both an oracle and partial graph. However, the partial graph is created by removing nodes, not hiding them, so the role they play in the oracle is not studied. On the other hand, our approach permits comparison between theoretical spread on the oracle and observed spread on the partial network through a characterisation of cascades based on activation paths, yielding insight into how different cascades contribute to diffusion error.

The other three approaches attempt to infer properties of the total information cascade from a partially observed cascade [23,27,28]. In [23], the authors address cascade distortions under missing data and propose fitting k-trees to correct the distortion. In [28] a similar approach is taken but the authors incorporate node activation time to constrain the fit of *consistent* trees to the information cascades, while [27] exploits node activation time in a model that predicts the final number of activated nodes without knowledge of the network. The key differences between these works and ours are (i) they focus on missing data at the cascade level, whereas we focus on missing data at the network level, (ii) they examine cascades that have already occurred, while we aim to estimate the error under missing data prior to an attempted real world diffusion (e.g., a viral marketing campaign) and propose a model to correct for it. Furthermore, in [28], the authors assume *complete* knowledge of the underlying network in order to fit the cascade trees, which is rarely the case in the real world.

3. Methodology

In this section we aim to study the effect of partial network knowledge on the diffusion process using five datasets representing both empirical and synthetic social networks. We describe the creation of partial networks and propose a novel methodology to quantify the error due to missing information by characterising the different paths through which a node can become activated during the diffusion process.

3.1. Datasets

Three factors drove the choices in datasets. First, our aim is to use graphs that are arguably as *complete* as possible to serve as *oracle* networks, i.e., where all nodes and edges are known. We refer to them as oracle networks because they represent *relatively* complete (in the case of empirical) or theoretically complete (in the case of synthetic) social networks.[1] Secondly, to account for the effect of network size we chose networks of different scale. Finally, the graphs should have properties consistent with social networks (i.e. have community structure, exponential degree distribution, small diameter, etc.) because we are simulating a process from the social network domain, that is, where a message is spread by word-of-mouth.

Empirical. We selected two empirical networks: the DBLP co-authorship network [18] and the ArXiv Astro Physics co-authorship network [16]. These networks were chosen because they represent the flow of knowledge through scientific collaboration and are examples of relatively complete networks for those communities.[2]

Synthetic. In addition to the empirical datasets, we generated random networks with different topological properties. Generating synthetic graphs has the benefit of yielding networks of desired size and topological properties, providing a benchmark comparison for the real networks and a more complete picture of the results. Furthermore, the resultant graphs are by definition complete, as they are realised from a generative process. We selected three graph models: 2.5K series [11], TOSHK [25], and the Erdős–Rényi random graph [9].

[1] The notion of completeness could indeed be argued for a known sampled network, given that it could play the role of oracle compared to a partial network derived by sampling it. However, for simplicity we define a complete network as one that is, to the best of our knowledge, complete in that it encompasses all possible nodes and edges (i.e., it is not sampled from a larger network).

[2] The DBLP network is a multigraph with parallel edges between authors denoting multiple co-authored papers. However, in this work we ignore parallel edges in order to leverage the seed selection strategy proposed by [4] and to maintain consistency across datasets.