Physica A xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Physica A

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

Q1 Identifying the main paths of information diffusion in online social networks

Q2 Hengmin Zhu [a], Xicheng Yin [a,*], Jing Ma [b], Wei Hu [a]

[a] School of Management, Nanjing University of Posts and Telecommunications, Nanjing 210023, People's Republic of China

[b] School of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, People's Republic of China

HIGHLIGHTS

- We propose a method to identify the main paths in online social networks.
- The weight of link is evaluated based on historical interaction records.
- Path importance is quantified by calculating the probability that messages spread via this path.
- Our method has taken weak ties' positive effects into account.
- The distribution of main paths shows significant community effect.

ARTICLE INFO

Article history:
Received 12 November 2015
Available online xxxx

Keywords:
Information diffusion
Social networks
Historical interactions
Main path

ABSTRACT

Recently, an increasing number of researches on relationship strength show that there are some socially active links in online social networks. Furthermore, it is likely that there exist main paths which play the most significant role in the process of information diffusion. Although much of previous work has focused on the pathway of a specific event, there are hardly any scholars that have extracted the main paths. To identify the main paths of online social networks, we proposed a method which calculates the weights of links based on historical interaction records. The importance of node is quantified and top-ranked nodes are selected as the influential users. The path importance is evaluated by calculating the probability that a message would spread via this path. Weak ties' positive effects have also been taken into account. We applied our method to a real-world network and found interesting insights. Each influential user can access another one via a short main path. The distribution of main paths shows significant community effect.

1. Introduction

On account of complicated structure and heterogeneous nodes of online social networks (OSNs), the mechanisms of information diffusion have become more complex. But ultimately message usually propagates along some specific pathways, for example, many online users tend to repost messages from friends' blog space [1,2], which brings about the information dissemination. Recently, an increasing number of researches on relationship strength show that there are some socially active links that have accelerated information spreading in OSNs. Maybe one user has many friends, but only a small proportion of his friends will interact with him frequently in usual social scenes. Furthermore, if we join several such active

* Corresponding author.
E-mail address: yxcooler@163.com (X. Yin).

http://dx.doi.org/10.1016/j.physa.2016.01.048
0378-4371/© 2016 Published by Elsevier B.V.

links together, it is likely that there exist main paths which play the most significant role in the process of information diffusion. This thought enlightens us that if we could understand the backbone network consisting of main paths, the breadth and depth of the diffusion may also be forecasted. This means the main path analysis will be beneficial to guiding or controlling the information dissemination from a global perspective.

Although much of previous work has focused on the epidemic models [3,4] and network topology [5–8] to study the dissemination of information, the actual mechanics of how any single piece of information spreads on a global scale have still been a mystery to be further solved. Based on definite cases, some scholars studied the propagation path of online messages [9–11]. Different from the pathway study on specific event, main path analysis aims to find the most significant paths because it is not an easy task to trace the messages in a large network. Main path analysis was first proposed by Hummon and Doreian in 1989 to trace the development trajectory of a research field based on the theory of depth-first search and traverse tree [12–14]. When used in knowledge diffusion network, the main paths are constructed by calculating the connectivity of the links in terms of their degree centrality and outlining the path formed by the nodes with the highest degree [15,16]. Doing main path analysis of citation network allows us to highlight which documents are central or have a major influence on the development of a scientific field. Likewise, analyzing the main paths of social network is also able to highlight the structural backbone of the relationship network. In general, main path analysis has been widely used in citation network and knowledge diffusion. Although much previous work has studied the information pathway of a specific case, there are hardly any scholars that have extracted the main paths from a social network.

It should be noted that the special qualities of social networks such as socialization, interactiveness and swiftness determine that the main-path methodology of it is totally different from that of other network, e.g., citation network. In fact, the dynamics of the spreading process of online information is explained by interactions between pairs of users and the topology of their interconnections [17]. And the information pathways are explained by interactions between nodes as well. Consequently, studying the historical interactions between each pair of users would be the basic premise of identifying the pathways.

Fortunately, OSNs often consist of more than just a record of social network ties. The historical interaction information is also contained within it [18,19]. In the social activities, since each user has a finite amount of resources (e.g., time) to use in the formation and maintenance of relationships, it is likely that they direct these resources towards the relationships that they deem more important [20]; which means, in social networks, ties with different strengths play different roles in maintaining the network connectivity [21–23]. Granovetter et al. firstly introduced the concept of tie strength in his paper when talking about weak ties [24]. Eric Gilbert et al. proposed several dimensions, e.g., intimacy, duration and interaction frequency to measure the strength of ties [25]. Theoretically, if we identify the strong ties of every node based on historical interaction records and then join several ties together, it is feasible to extract the main paths of the network.

In this paper, we proposed a quantitative method to identify the main paths of OSNs on the foundation of historical interaction data. Specifically, we calculated the weights of links based on historical interaction records, and the weight of link, i.e. tie strength, represents the historical interaction frequency between pairs of nodes. Then an influential node ranking algorithm, the LeaderRank [26,27] was adopted to identify the influential nodes. The top-ranked nodes were selected as the influential users, and then we evaluated the path importance by calculating the probability that a message would spread via this path. Finally we applied our method to a real-world network and found interesting insights.

2. Methodology

Our approach mainly consists of three steps: measurement of tie strength, identification of influential users, and identification of main paths. Tie strength, i.e. weight of the directed weighted social network, is measured by the intimacy between users and determines which follower is more likely to repost the message after one has posted his latest blog. Accordingly, the weight of link has a significant impact on the flow direction of information. Based on the tie strength, the identification of influential users helps us find the most influential users during the propagation of information. The main paths connecting influential nodes are relatively important paths those promote the information diffusion. Overall, the method is going to proceed step by step because the former step is the foundation of the latter one.

2.1. Measurement of tie strength

In many online social networking sites, users tend to follow other users to obtain more information. We represent the follow relations by a network with directed weighted links pointing from followers to their leaders. It is important to note that the direction of information spreading is opposite to that of link, i.e. the follow relation. Here we measure the tie strength with the dimension of interaction frequency. For example, the initial weight of directed edge from node A to B is the times user A reposted messages from user B during a certain period of time in the past. In real social scenes, the tie strength of edge AB actually indicates the probability that user A will repost from user B. Therefore, a standardization of the tie strength is necessary. The normalized weight w_{ij} is given by

$$w_{ij} = \frac{f_{ij}}{\sum f_i^{out}} \quad (1)$$

where f_{ij} is the initial weight, and $\sum f_i^{out}$ is the sum of initial weights of all out-links of node i (as shown in Fig. 1).