Physica A 505 (2018) 222–232

Contents lists available at ScienceDirect

Physica A

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

Inferring parent–child relationships by a node-remove centrality framework in online social networks

Xiao-Ke Xu [a,b], Xue Wang [a], Jing Xiao [a,*]

[a] College of Information and Communication Engineering, Dalian Minzu University, Dalian 116600, China
[b] Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China

HIGHLIGHTS

- Parent–child links are weak social embeddedness (weak tie) instead of strong embeddedness (strong tie).
- A novel node-remove framework is proposed to measure the node centrality considering its direct and indirect effect.
- The proposed algorithm is validated in online social networks and better than several representative algorithms.
- Our method may not only infer parent–child and familial relationships, but also detect other heterogeneous links.

ARTICLE INFO

Article history:
Received 30 October 2017
Received in revised form 16 February 2018
Available online 29 March 2018

Keywords:
Online social networks
Node centrality
Dispersion
Embeddedness
Strong ties

ABSTRACT

Online social networks can represent various kinds of relationships between users, so a significant task is to infer specific relationships, especially family members or romantic partners, by analyzing network topologies. In this study, we explored to infer a special kind of family links (i.e., parent–child links) in the QQ social network (the largest online social network of China) based on multiple node centrality algorithms. We found that most parent–child links are weak social embeddedness (weak tie) instead of strong embeddedness (strong tie), resulting that such kind of links are very difficult to detect in the user's ego social network. To get a high accurate detection, we firstly utilized user profile information such as age and gender to filter out the set of potential links and then combined network structure mining. We proposed a novel node-remove framework to measure a node centrality considering its direct and indirect effect on network topologies, and our method obtained a 11.3% higher performance than "dispersion" which is a new network centrality proposed in a recent study. We also found that calculating indirect effect by using a semi-local structure can obtain a better performance than using local or global structures, which suggests that parent–child links have a meso-scopic topology effect. Our method may not only infer parent–child relationships effectively, but could also detect other hierarchical relationships, such as manager–subordinate and advisor–advisee ties in online social networks.

1. Introduction

In the last ten years, more and more people have begun to embed themselves in online social networks and reorganize social relationships based on social network services. The heavy users of online social networks not only tend to connect to

* Corresponding author.
E-mail address: hrbeuxiaojing@aliyun.com (J. Xiao).

https://doi.org/10.1016/j.physa.2018.03.041
0378-4371/

their friends driven by homophily principle [1,2], but also link their family members to maintain intimate relationships by online interactions instead of offline face to face communications [3,4]. For example, Madden et al. found that two thirds of parents of teens aged 12–17 at least use a social networking site, and 80% of them have friended their children [5]. Furthermore, nearly 40% of Facebook users have either the parent or child on the site, and their communication frequencies do not decrease with the increasing of geographic distance [3]. Recently, using real-life data from a large-scale sample of Facebook users, Backstrom and Kleinberg found that romantic partnerships in social network neighborhoods can be accomplished with high accuracy according to the dispersion of social ties [6].

In this study, we are interested in how to infer the parent or child link based on a user's ego network structure [7]. Parent–child relationship is a singular type of social strong ties that plays a powerful role in social processes over a person's whole life course [8]. Parent–child links can be classified into two categories: parent-to-child and child-to-parent. Inferring parent-to-child links means that we have gotten the ego social network of a parent and we try to identify the hidden child in his/her friend list. Parent-to-child links include four sub-categories: father–son, father–daughter, mother–son, and mother–daughter. Here "father–son" means that the center user is a father, and he has tagged one of his QQ friend as his son. Child-to-parent relationship contains other four types of links: son–father, son–mother, daughter–father, and daughter–mother. Similar to "father–son", "son–father" means that the center user is a son, and he has tagged one of his QQ friend as his father.

Strong embeddedness [9,10] is the classic view to investigate the special importance of parent–child links. However, our results show that parent–child links are rarely found in the tie sets of the strong network embeddedness (strong ties). Especially, the results of the QQ dataset indicate that the ratio of children regarding their parents as the most strong embedding ties are less than that parents regarding their children as the most strong embedding ties. The embeddedness strengths of parent–child links are not as strong as we expect, which makes them difficult to be extracted from the center user's social network neighborhood. The reason for this is that there is a huge generation age difference between parent and child, so we try to utilize each user's age and gender as the filter to obtain a set of potential links before inferring parent–child links. By combining node centrality statistics, we can greatly improve the prediction accuracy for such kind of links. Our results suggest that a relative strong embeddedness instead of an absolute strong one exists between parents and children (the strongest tie in all the pairs of users who have an age gap of 15–40 years), because their social circles are not strongly overlapped induced by parent–child huge age difference.

Dispersion is another view to consider the importance of parent–child links. Dispersion was proposed in [6] to infer romantic relationships in Facebook by measuring whether two persons' mutual friends are well-connected. If the mutual friends of two users are not well connected to one another and have a long shortest path length, the two users have a high dispersion value. Compared with other node centrality statistics, dispersion centrality obtains the best performance for inferring parent–child links. However, several formations can be adopted to calculate dispersion, and it is difficult to know which formation is the best. In this study, we developed a novel node-remove framework to explain why the threshold-3 format of dispersion centrality can obtain the best performance. Furthermore, the proposed node-remove framework fuses the information of both embeddedness and dispersion centrality, which leads to a better detection performance. At last, the results suggest that parent–child links have a meso-scopic topology effect, which means calculating the indirect effect by using a semi-local structure can obtain a higher accuracy performance than using local or global structure information. Our findings can be used to not only infer parent–child links and other relative links, but also detect other heterogeneous types of links, such as manager–subordinate and Ph.D. advisor–advisee relationships in online social networks [11].

2. Dataset and problem description

2.1. Dataset description

Tencent QQ, popularly known as QQ in China, is an instant messaging software service. As of the first quarter of 2017, QQ had 861 million monthly active user accounts. At the highest peak, more than 266 million people were using QQ simultaneously. As a scientific research cooperation project with Tencent company, we have been permitted to use a large-scale dataset of randomly sampled QQ users who declare his/her child or parent in their friend lists. How to draw samples that can represent the QQ OSNs has remained a formidable task because of a number of conceptual and methodological reasons [12]. Here we take the uniform sampling method. Uniform sampling is a classic method based on probability theory that has been widely used in sampling of physical and human subjects for centuries [13]. When applied to network sampling, all nodes of any OSN under study are given an equal chance (hence the name of "uniform") to be sampled. Large number theorem ensures that a uniform sample, with a sufficiently large number of nodes, will represent accurately the underlying population (i.e., the QQ OSN). The dataset consists of 842 QQ users' ego network neighborhoods and 245,371 users, selected uniformly at random from all active users who list a parent or child in their friend lists. The neighborhood of each user has an average of 291 friends, which means that the accuracy of random selecting parent–child link is only about 0.3%. More detail information on this dataset can be found in Table 1. All the data in our analysis was used anonymously, and all the analysis was done in aggregate.

All the links between parent and child can be divided into two types: parent-to-child and child-to-parent. Parent-to-child links can be classified as four categories: father–son (fa–son), father–daughter (fa–dau), mother–son (mo–son), and mother–daughter (mo–dau). Here, the "father–son" relationship means that the center user is a father, and he has tagged one of his QQ friend as his son. In this study, our aim is to identify which user is his son based on the father's QQ ego social network.