Physica A 463 (2016) 356–365

Contents lists available at ScienceDirect

Physica A

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

Flexible sampling large-scale social networks by self-adjustable random walk

Xiao-Ke Xu [a,*], Jonathan J.H. Zhu [b]

[a] College of Information and Communication Engineering, Dalian Minzu University, Dalian 116600, China
[b] Web Mining Lab, Department of Media and Communication, City University of Hong Kong, Hong Kong, China

HIGHLIGHTS

- A sampling method called Self-Adjustable Random Walk (SARW) is proposed.
- The advantage of SARW sampling in comparison with four prevailing methods.
- Mixing both induced-edge and external-edge information for calculating network measures.
- SARW sampling can generate unbiased samples with maximal precision and minimal cost.

ARTICLE INFO

Article history:
Received 5 April 2016
Received in revised form 9 July 2016
Available online 26 July 2016

Keywords:
Online social networks
Random walk
External-edge
Network sampling

ABSTRACT

Online social networks (OSNs) have become an increasingly attractive gold mine for academic and commercial researchers. However, research on OSNs faces a number of difficult challenges. One bottleneck lies in the massive quantity and often unavailability of OSN population data. Sampling perhaps becomes the only feasible solution to the problems. How to draw samples that can represent the underlying OSNs has remained a formidable task because of a number of conceptual and methodological reasons. Especially, most of the empirically-driven studies on network sampling are confined to simulated data or sub-graph data, which are fundamentally different from real and complete-graph OSNs. In the current study, we propose a flexible sampling method, called Self-Adjustable Random Walk (SARW), and test it against with the population data of a real large-scale OSN. We evaluate the strengths of the sampling method in comparison with four prevailing methods, including uniform, breadth-first search (BFS), random walk (RW), and revised RW (i.e., MHRW) sampling. We try to mix both induced-edge and external-edge information of sampled nodes together in the same sampling process. Our results show that the SARW sampling method has been able to generate unbiased samples of OSNs with maximal precision and minimal cost. The study is helpful for the practice of OSN research by providing a highly needed sampling tools, for the methodological development of large-scale network sampling by comparative evaluations of existing sampling methods, and for the theoretical understanding of human networks by highlighting discrepancies and contradictions between existing knowledge/assumptions of large-scale real OSN data.

© 2016 Published by Elsevier B.V.

* Corresponding author.
E-mail addresses: xiaokeeie@gmail.com (X.-K. Xu), j.zhu@cityu.edu.hk (J.J.H. Zhu).

http://dx.doi.org/10.1016/j.physa.2016.07.055
0378-4371/© 2016 Published by Elsevier B.V.

1. Introduction

Online social networks (OSNs) have been widely investigated by scientists on structure, function, or evolution dynamics of complex systems [1,2], by engineering researchers on control, synchronization, or optimization of networks [3,4], and by scholars of social sciences and business on interpersonal relationship, group pressure, or community politics [5]. Although attractive, research on OSNs faces a number of difficult challenges. One of the bottlenecks lies in the accessibility of OSN data, which are not only in massive quantity (usually involving 10 million–1 billion nodes) but also unavailable (except for the owners of the OSNs or their commissioned consultants) [6]. Sampling, then, becomes one and perhaps the only one feasible method to solve problems of quantity and inaccessibility of OSN data.

Sampling of large-scale online social networks is a new issue, on which the research community has so far made limited progress [7,8]. In the classic studies of offline social networks, the researchers usually investigated the network population, rather than a sample of the network, because offline social networks were (and still are) of a small size (e.g., with hundred individuals) [9,10]. When sampling was occasionally called for, the quality of resulting samples employed by convenient methods (e.g., snowball sampling [11,12]) was barely evaluated.

After the emergence of OSNs, scholars from computer science and other IT domains have joined, at an ever-growing rate, in research on OSNs. They have brought in machine-oriented sampling tools (e.g., breadth-first search [7,13,14] and random walks [8,15,16]) that are easy to implement and highly efficient, but problematic in the quality of resulting samples. On the other hand, the classic probability sampling methods (e.g., uniform sampling [17–19]) require prior knowledge of the physical location of all nodes in any OSN under study, which is usually not available to third-party researchers. These prior works provide useful insights into the subgraph sampling problem, but samples generated by these methods do not meet three key performance indicators: **validity** (i.e., representative of the underlying network population), **reliability** (yielding precise and stable measurement), and **practicality** (requiring minimal prior knowledge and manageable sample size).

In this study we aim to fill the gap by designing and testing a flexible sampling method, called Self-Adjustable Random Walk (SARW), based on real data of large-scale OSNs. The proposed method is expected to contribute to the research on OSNs in three ways. First, we aim to contribute to the research practice by devising a new sampling method that is easy to implement, flexible to accommodate a variety of networks, and likely to generate less biased results than the existing methods do. Second, we intend to contribute to the methodological development of networking sampling by providing results from formal tests of competing sampling methods over the three most commonly used network measures based on real data. Third, although the study is evidently methodologically-oriented, we believe that it will also contribute to the theoretical understanding of the nature and structure of OSNs. As we have observed in this study, our rigorously controlled experiments produce a number of results (e.g., uniform samples do not always provide unbiased estimates of the underlying network population) that are contradictory to some of the existing knowledge (which is, strictly speaking, untested assumptions). Instead of blaming the methods, we think that a systematic exploration of the discrepancies may prove to be far more informative than one usually expects from methodological studies. The strengths and weaknesses of the sampling method will be evaluated against four prevailing methods, including uniform, breadth-first search (BFS), random walk (RW), and revised RW (e.g., MHRW) sampling.

2. Dataset and methods description

2.1. Dataset description

Compared with previous evaluation studies of network sampling that were based on simulated data [15,20,21], a unique strength of the proposed study is that we will use a real dataset of a large-scale OSN (involving 10M nodes and 200M edges) for the empirical tests. We have worked out a special arrangement with the OSN (which remains anonymous at this stage to avoid similar requests for data from other parties) to have access to the needed information (i.e., nodes and their links) through a dedicated API. The identification (e.g., real name, user ID, Web address, etc.) of all users has been removed to protect the anonymity of the users. As far as we know, up to now there is very few sampling studies based on a complete large-scale OSNs (except [22]). Here we show the comparison results of statistical measures between our real OSN and Cyworld (see Table 1).

We will use the same data as described above to evaluate several popular sampling methods (i.e., uniform, BFS, RW, MHRW, and BRW). In the study, we will carry out a more comprehensive evaluation of the existing sampling methods by involving the three most commonly used network measures. Furthermore, the evaluation will be performed to obtain the differences between sample estimates and population parameters. We expect the results not only can offer more consistent conclusions about the relative strengths and weaknesses of the methods, but also can provide additional insights to the development of our proposed SARW method.

2.2. Reviews of sampling methods

In previous studies of OSNs, various sampling methods have been employed. Despite different names used, the methods come essentially from three branches of the sampling family (see Fig. 1): probability (or uniform) sampling, breadth-first search (BFS) sampling, and random walk (RW) sampling. However, the methods differ significantly in the underlying

Download English Version:

https://daneshyari.com/en/article/973973

Download Persian Version:

https://daneshyari.com/article/973973

Daneshyari.com