Journal of Network and Computer Applications 41 (2014) 126–134

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Real time enhanced random sampling of online social networks

Giannis Haralabopoulos, Ioannis Anagnostopoulos *

Department of Computer Science and Biomedical Informatics, University of Central Greece, Lamia 35100, Greece

ARTICLE INFO

Article history:
Received 17 April 2013
Received in revised form
22 August 2013
Accepted 21 October 2013
Available online 20 November 2013

Keywords:
Crawling
Graph sampling
Random node sampling
Social networks
Twitter

ABSTRACT

Social graphs can be easily extracted from Online Social Networks (OSNs). However, as the size and evolution of this kind of networks increases over time, conventional sampling methods used to evaluate large graph information cannot accurately project network properties. Furthermore, in an attempt to deal with ever increasing access and possible malicious incidents (e.g. Denial of Services), OSNs introduce access limitations for their data, making the crawling/sampling process even harder. A novel approach on random sampling is proposed, considering both limitations set from OSNs and resources available. We evaluate our proposal with 4 different settings on 14 different test graphs, crawled directly from Twitter. Additionally, we test our methods in various graphs from Stanford Network Analysis Project Collection. Results show that every scenario needs a different approach. Conventional Random Node Sampling is better used for small sampling sizes, while Enhanced Random Node Sampling provides quicker and better results in larger samples. Still many questions arise from this work that can be considered as future research topics.

1. Introduction

Online Social Networks (OSNs) provide a great source for social graph analysis, but due to their dynamic nature, advanced methods and lots of resources are required for processing their statistical properties. This dynamic behaviour is demonstrated in Twitter, which climbed 22 spots in the world's most visited sites within a year[1], and currently is the second most visited OSN[2]. As such, every attempt to estimate its graph properties is getting even harder. Various sampling methods used throughout the years are not able to accurately describe the structure and properties of such a huge social graph. Most of these methods analyse and evaluate subgraphs, which only consists of the sampled nodes. For example, although random sampling is widely popular in graph sampling processes, it fails to accurately project most graph properties (Stumpf et al., 2005).

In addition, OSNs pose access limitations in their data, in order to protect their operation and deal with increasing access, overload or spamming incidents. For example, Twitter API version 1, imposes a limit of 350 requests per hour (for authenticated users), thus making data acquisition a costly procedure, resource wise. Still, if we effectively use these requests, various sampling methods can still be employed and evaluated. So, the issues that should be addressed are summarised in the following questions:

- how can we fully utilise our limited graph crawling requests?
- can sampling methods provide sufficient data about an OSN graph?
- can we enhance existing sampling methods to further improve results?

The answer to all three questions is our proposed real time enhanced sampling process. In other words, a selective process that takes request limiting into consideration, coupled with enhancements to improve the derived results. This work aims to test two enhancements over the random sampling method. Our proposals make use of all the information available in a single node of an OSN. In addition, it is executed in real time and parallel to the crawling process, thus saving valuable resources. Lastly, the proposed enhancements can be incorporated, with little effort, in every other graph sampling method.

The OSN on which we tested our proposals is Twitter. Within a 25-day time period we crawled data (called from this point herein as Test Graphs – TGs). To be more specific, the crawling process, started on 5th of March 2012 and ended in the 30th of the same month. As seeds, we used the top thirty greek Twitter users, based on followers count. The method applied was Breadth First Crawling, and the minimum number of nodes to be fully analysed was equal to 25 K per seed. Our calculations showed that we fully analysed the second and up to the fourth degree neighbours, for

* Corresponding author. Tel.: +30 22310 66937.
E-mail address: janag@ucg.gr (I. Anagnostopoulos).

[1] Compete.com, last accessed 04/04/2013, http://siteanalytics.compete.com/twitter.com/.

[2] Social Networks: Facebook Takes Over Top Spot. Twitter Climbs, Compete.com, 2009, last accessed 04/04/2013, http://blog.compete.com/2009/02/09/facebook-myspace-twitter-social-network/.

1084-8045/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jnca.2013.10.016

each seed. As far as the terminology is concerned, a user in Twitter corresponds to a graph node and the user's followers/following linkage corresponds to in/out edges of that node.

The first enhancement we propose, is based on the fact that social networks tend to follow a Pareto Distribution (Pareto et al., 1964). We tested three different degree distributions, which could improve sampling results. Since we do not have any knowledge of graph properties, we are investigating the area surrounding Pareto 80-20 distribution. We observe that our implementation is not bound to any a-priori known properties, since it is applied on the fly along with the crawling process.

The second improvement is based on the observation that when we sample a graph and we only use the resulting subgraph for analysis, we lose a vital information considering the hard request limitations imposed. Thus, we propose to add all the neighbouring nodes of a sampled node to the subgraph. This in fact is the combination of two techniques, random and neighbourhood sampling.

In our work, the evaluated graph properties are the Number of Edges, Mean Degree, Clustering Coefficient, Assortativity, Number of Components (NoC) and the time required for the process. We compare the properties of each TG with sampled mean values of multiple iterations, and in order to get a better idea on time handling, we provide another comparison of results along with time required to produce them.

The remainder of the paper is organized as follows. In the next section, we provide a thorough review in respect to the related literature for large graph sampling in OSNs, combined with the issues arose from web crawling. Section 3 describes our proposed methodology in relation to the datasets and the resources used, as well as the proposed algorithmic enhancements of our proposed sampling scheme called "Real Time Enhanced Random Sampling". Following the description of our proposal, Section 4 presents the experimentations conducted for evaluation purposes. More specifically, this section presents the results and the related discussion of experimentations over various graphs sampled by us (our TGs), as well as over different graphs from the publicly available Stanford Large Network Dataset Collection. Finally, the last section concludes our work and introduces some thoughts we have for the future.

2. Related work

Graph analysis has been the subject of many essays, while being a significant topic in many real-life applications in both technological and social field. From simple metrics to advanced trend prediction, graphs can provide valuable information through OSN analysis. However, the amount of disseminated information is vast, thus suggesting graph sampling an important process for analysing OSNs structure and properties.

One of the first deep graph analysis was authored by Broder et al. (2000). This research conducted with data acquired by Altavista crawls nearly 14 years ago, setting up the fundamental macroscopic elements and attributes of a large evolving graph as the World Wide Web (WWW). The authors verified strong theoretical concepts and observations (e.g. in/out-degree power law distributions), which still appear either in a macroscopic scale (web graph), or in a microscopic scale (OSN unique user cluster).

An early attempt to describe some solvable models in respect to the structure of OSN based on random graphs with arbitrary degree distributions, appears in Newman et al. (2002), in which the authors provide models for different graphs that are applied later to real OSN. The authors observed that in several OSNs of those years, the existing models provided sampling instances capable of describing the whole graph, whereas in others OSNs the same methods were not applicable. Perhaps indicating a different social structure in the network that is not captured by the selected random graph sampling scheme.

Later on, the authors in Leskovec and Faloutsos (2006) provided an extensive analysis and evaluation in respect to large graph sampling. Up to those years, the literature related to sampling techniques from undirected graphs, concluded that some graph properties could be preserved through random-node selection with less than 30% samples in comparison to the original graph size (Krishnamurthy et al., 2005). In Leskovec and Faloutsos (2006), the authors claimed that simple uniform random node selection methods, outperforms edge selection-based sampling strategies, in both static or highly evolving graph patterns, even having sample sizes less than 15% of the initial graph.

In Ahn et al. (2007), the authors firstly analyse the structure of a huge graph based on its properties and degree distributions, and secondly its users behaviour towards the belief that OSN resembles real human-based social network attributes. The authors compare the structures of three online social networking services, namely Cyworld, MySpace, and Orkut, each with more than 10 million users – at the time of their research. Having analysed the complete Cyworld network, and parts of MySpace and Orkut obtained with Snowball Sampling, they ended up in the conclusion that different types of OSN users are strongly related to large graph attributes. Such as clustering coefficient distribution, assortativity/disassortativity, network size, average path length, and effective diameter. Additionally, the authors evaluate the snowball sampling method as a breadth first search method on a 12M node/190M edge graph.

The authors in Dasgupta et al. (2008), deal with community identification of huge information networks. Many different databases were analysed with the graph conductance – as described in Chung (1997) – playing a key role in conclusions. An important outcome of the work was the fact that, community size is proportional to "blending" with the whole network. The work also deals with the implications and metaphors in the case of graph partitioning algorithms, in real-world networks and communities detection in them.

In the work described in Zou and Holder (2010), the authors – by accepting that frequent pattern mining plays a significant role in sampling large graphs – validated the concept of subgraph mining. Through their technique, called "Random Areas Selection Sampling", they handled sampled graphs along with the initial graph and then compared the results. The authors mainly used efficient sampling method for estimating subgraph concentration and detecting network motifs, as well as sampling approaches in order to reveal reliable subgraphs from large probabilistic graphs as described in Kashtan et al. (2004) and Hintsanen and Toivonen (2008). Zou and Holder (2010) claimed that their proposal had the highest accuracy among all other graph-sampling methods, by performing experimentations in large graphs, such as Citation Graph, Amazon Graph, and WWW conference series graph. Another important work in subgraph mining is described in Leskovec and Faloutsos (2006), where the authors evaluate several sampling schemes (such as random node, random edge, random jump, etc.) that require some or full knowledge of the original graph.

Random walks (RWs) in graphs has been an established sampling scheme in large OSNs. The works in Krishnamurthy et al. (2008), Rasti et al. (2008) and Gjoka et al. (2010) employs RWs in order to sample user entities in large OSNs such as Friendster, Twitter and Facebook. The authors in Ribeiro and Towsley (2010) proposed a sampling scheme capable of exploring multiple dependent RWs to further improve the sampling procedure in loosely connected subgraphs. Similarly to that, the authors in Gjoka et al. (2011) introduced the concept of multigraph