Information Processing and Management 50 (2014) 476–491

Contents lists available at ScienceDirect

Information Processing and Management

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

Variance reduction in large graph sampling

Jianguo Lu [*], Hao Wang

School of Computer Science, University of Windsor, 401 Sunset Avenue, Windsor, Ontario N9B 3P4, Canada

ARTICLE INFO

Article history:
Received 25 March 2013
Received in revised form 7 January 2014
Accepted 11 February 2014
Available online 15 March 2014

Keywords:
Uniform random sampling
Random walk
Graph sampling
Online social network
Scale-free network
Harmonic mean

ABSTRACT

The norm of practice in estimating graph properties is to use uniform random node (RN) samples whenever possible. Many graphs are large and scale-free, inducing large degree variance and estimator variance. This paper shows that random edge (RE) sampling and the corresponding harmonic mean estimator for average degree can reduce the estimation variance significantly. First, we demonstrate that the degree variance, and consequently the variance of the RN estimator, can grow almost linearly with data size for typical scale-free graphs. Then we prove that the RE estimator has a variance bounded from above. Therefore, the variance ratio between RN and RE samplings can be very large for big data. The analytical result is supported by both simulation studies and 18 real networks. We observe that the variance reduction ratio can be more than a hundred for some real networks such as Twitter. Furthermore, we show that random walk (RW) sampling is always worse than RE sampling, and it can reduce the variance of RN method only when its performance is close to that of RE sampling.

1. Introduction

The data on the Web or online social networks can be often viewed as a graph. The graph in its entirety may not be available for various reasons. It can be distributed over many machines (e.g., the Web and P2P networks), hidden behind searchable interfaces (e.g., search engines and online social networks), scattered among a larger graph (e.g., various communities in online social networks). Regardless of the causes, a common challenge is to reveal the properties of such graphs when we do not own the entire data. In the past, extensive research was carried out to explore the profile of search engines (Lawrence & Giles, 1998) and other data collections (Broder, 2006; Callan & Connell, 2001; Si & Callan, 2003). Most of them focused on obtaining uniform random node (RN) samples, such as uniform random web pages from the Web (Henzinger, Heydon, Mitzenmacher, & Najork, 2000) and search engines (Bar-Yossef & Gurevich, 2008), and uniform random bloggers from online social networks (Gjoka, Kurant, Butts, & Markopoulou, 2009). Once uniform random samples are obtained, network properties, in particular the attributes of the nodes including average degree, could be estimated with statistical guarantee.

In many cases, RN sampling works only in theory. The majority of real world networks are scale-free (Barabási & Albert, 1999), whose degree distributions follow a power law. Such scale-free networks often induce a large variance of the degrees. In theory, the variance does not exist when the exponent of the power law falls in certain range. In practice, the variance can be extremely high for very large networks. For instance, the coefficient of variation of the Twitter user network collected in 2009 (Kwak, Lee, Park, & Moon, 2010) is as high as 35.95. To understand the impact of such a high coefficient of variation, let us have a quick calculation for the sample size needed to reach 20% accuracy for its average degree 70.51. More precisely, to

[*] Corresponding author. Tel.: +1 519 253 3000.
E-mail addresses: jlu@uwindsor.ca (J. Lu), wang115o@uwindsor.ca (H. Wang).

http://dx.doi.org/10.1016/j.ipm.2014.02.003
0306-4573/Crown Copyright © 2014 Published by Elsevier Ltd. All rights reserved.

make sure that the estimation is within the range of 70.51 ± 14.10 with 95% confidence, the relative standard error RSE should be around 0.1, and the required sample size $n = 35.95^2 * /(0.1)^2 = 129,240$. In addition, uniform random samples are obtained with high cost because they are not provided directly by the data sources. Costly sampling methods, such as rejection sampling, have to be employed to obtain uniform samples. In the process, many samples are retrieved and rejected as invalid. The actual samples retrieved are many times larger than $129,240$, depending on the sampling methods allowed. Considering the network traffic involved and the daily quota imposed by the service provider, it is prohibitive to use uniform random sampling to obtain meaningful estimations. With increasingly more applications of big data analyses, there is an urgent need to find a method to reduce such a large variance.

Recent developments made empirical observations that simple random walk (RW) sampling or its extensions can improve degree estimator performance for P2P networks (Rasti, 2009), Facebook user network (Gjoka et al., 2009), Twitter user network (Lu & Li, 2012), and term-document bipartite graphs (Wang, Liang, & Lu, in press). Similar empirical observations are made for node size estimation (Katzir, Liberty, & Somekh, 2011; Kurant, Butts, & Markopoulou, 2012). We find that these observations are data dependent. Random walk can be much worse than uniform random sampling for other datasets, even when the graph is scale-free and the variance is very high as we will show in Section 4.3.

We find that it is random edge (RE) sampling, not RW sampling, that reduces the variance for graphs with large degree variation. In addition to this empirical observation, we explore the reason why RE outperforms RN sampling, and why RW does not. While it is easy to understand that uniform random sampling does not work well for scale-free networks, it was not clear whether RE sampling works better.

This paper shows that the variance of the RE estimator is bounded from above by a polynomial in the average degree and sample size. It implies that the performance of RE sampling does not deteriorate with the growth of degree variance of the graph, thereby it guarantees the superiority of RE sampling when degree variance is large. This result is particularly important for large graphs whose variance becomes larger compared with smaller data with the same distribution. Improvement ratios as high as 100 are observed on Twitter and other networks. Such a large gap has implications for both practitioners and researchers. Practitioners can greatly save the estimation effort and give a worst case error bound. Researchers can devise new sampling methods that approximate RE sampling when it is not directly supported by the data source under investigation. For instance, random walk with restart (Avrachenkov, Ribeiro, & Towsley, 2010) can succeed because it is similar to RE sampling and exploits the large gap between RN and RE sampling.

The major contribution of the paper is our development of the upper bound of the variance of RE estimator. The result holds independent of degree distribution and graph topology. We verify the result using both simulated datasets and 18 real world networks. A direct consequence of the upper bound is the improvement ratio between RN and RE methods. To illustrate that the ratio can be very large, we first demonstrate that degree variance (consequently RN estimator variance) can be in the order of $O(N/\ln^2 N)$ under the assumption of the power law distribution. Now that RE variance is upper bounded, the improvement ratio tends to be infinite when data size goes infinitely large. Finally, we show that random walk sampling can approximate the performance of RE sampling only when the conductance of the graph is not very small, or, when the graph is well-enmeshed.

In the following sections, we first introduce the background of the research in Section 2, including the sampling methods and their corresponding estimators, the related work, and its applications. Then in Section 3 we derive the variances of RN and RE estimators. By giving the upper bound of the variance of the RE estimator, we quantify the performance ratio between RN and RE methods. In Section 4, we verify our result on 18 real networks, and demonstrate that the performance of RW sampling depends on both degree variance and graph conductance.

2. Background and related work

2.1. RN, RE, and RW sampling

Given an undirected graph $G(V,E)$, where V is the set of nodes, and E the set of edges. Let $|V| = N$. Nodes are labeled as $1,2,\ldots,N$, and their corresponding degrees are $d_1,d_2,\ldots,d_N$. The volume of the graph is $\tau = \sum_{i=1}^{N} d_i$, the average degree is $\langle d \rangle = \frac{1}{N}\sum_{i=1}^{N} d_i = \tau/N$. The variance σ^2 of the degrees in the population is defined as:

$$\sigma^2 = \langle d^2 \rangle - \langle d \rangle^2, \tag{1}$$

where $\langle d^2 \rangle = \sum_{i=1}^{N} d_i^2/N$ is the second moment, i.e., the arithmetic mean of the square of the degrees in the total population. The coefficient of variation (denoted as γ) is defined as the standard deviation, or the square root of the variance, normalized by the mean of the degrees:

$$\gamma^2 = \frac{\sigma^2}{\langle d \rangle^2} = \frac{\langle d^2 \rangle}{\langle d \rangle^2} - 1. \tag{2}$$

Suppose that a sample of n elements $(d_{x_1},\ldots,d_{x_n})$ is taken from the population, where $x_i \in \{1,2,\ldots,N\}$ for $i = 1,2,\ldots,n$. Our task is to estimate the average degree $\langle d \rangle$ using the sample. Table 1 summarizes the notations used in this paper.

There are different ways to take the samples, notably by RN, RE, and RW samplings. In RN sampling, each node is sampled uniformly at random with replacement. In RE sampling, edges are selected with equal probability and two nodes incident to