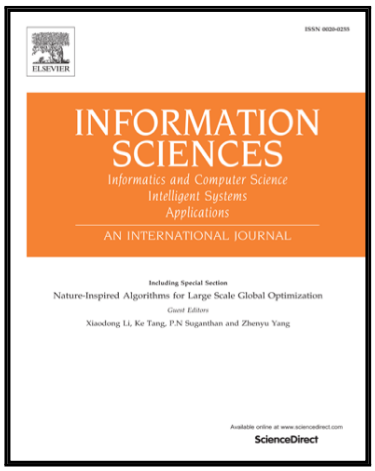

Uniform Random Sampling Not Recommended For Large Graph Size Estimation

Jianguo Lu, Hao Wang

PII: S0020-0255(16)30770-8
DOI: 10.1016/j.ins.2017.08.030
Reference: INS 13034

To appear in: *Information Sciences*

Received date: 6 September 2016
Revised date: 23 June 2017
Accepted date: 6 August 2017

Please cite this article as: Jianguo Lu, Hao Wang, Uniform Random Sampling Not Recommended For Large Graph Size Estimation, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.08.030

Uniform Random Sampling Not Recommended For Large Graph Size Estimation

Jianguo Lu, Hao Wang

School of Computer Science, University of Windsor, Canada

Abstract

The norm of data size estimation is to use uniform random samples whenever possible. There have been tremendous efforts in obtaining uniform random samples using methods such as Metropolis-Hasting random walk or importance sampling [2]. This paper shows that, on the contrary to the common practice, uniform random sampling should be avoided when PPS (probability proportional to size) sampling is available for large data.

To develop intuition of the sampling process, we discuss the sampling and estimation problem in the context of graph. The size is the number of nodes in the graph; uniform random sampling corresponds to uniform random node (RN) sampling; and PPS sampling is approximated by random edge (RE) sampling. In this setting, we show that for large graphs RE sampling outperforms RN sampling with a ratio proportional to the normalized graph degree variance. This result is particularly important in the era of big data, when data are typically large and scale-free [3], resulting in large degree variance.

We derive the result by giving the variances of RN and RE estimators. Each step of the derivation is supported and demonstrated by simulation studies assuming power law distributions. Then we use 18 real-world networks to verify the result. Furthermore, we show that the performance of random walk (RW) sampling is data dependent and can be significantly worse than RN and RE. More specifically, RW can estimate online social networks but not Web graphs due to the difference of the graph conductance.

1. Introduction

Size estimation is a classic problem that has many applications, ranging from the war time problem of finding out the number of German tanks [14], to the more recent challenge of gauging the size of the Web and search engines [20, 2, 6, 38] and online social networks [18, 15]. The direct calculation of data size is often not possible or desirable for several reasons. Quite often, data are hidden behind some searchable interfaces and programmable web APIs, such as online social networks and deep web data sources. The access is limited, and the data in its entirety are not available [37, 18]. The data can be distributed, and there is no central data repository such as in the case of peer-to-peer networks [30] or the Web [20]. Even when the data are available in one place, there are requirements for fast just-in-time analysis of the data [17]. Regardless of a large variety of application scenarios, a common approach to solving these problems is to use samples to have a fast estimation of the data size, instead of slow and direct counting of the data.

Many datasets can be viewed as graphs, especially the ones extracted from the Web and online social networks such as Twitter and Facebook. These graphs are large, often distributed and hidden behind searchable interfaces. The sampling process requires sending queries that occupy network traffic. In addition, most data sources impose daily quotas. In such cases, the sample size has to be far less than the data size, and it is paramount to choose an efficient sampling and estimation method.

For ease of discussion, sampling is modelled in the context of a graph, where uniform sampling corresponds to uniform random node (RN) sampling, PPS (probability proportional to size) sampling corresponds to random edge (RE) sampling. In this setting, we define the size as the number of nodes in the graph. Random walk (RW) sampling approximates PPS sampling in that the sampling probability is proportional to its degree asymptotically.

State of the art The norm of size estimation is to use uniform random samples whenever possible. Real data sources seldom provide uniform random samples directly. Therefore, there have been tremendous efforts to obtain uniform random samples from the Web [16], search engine indexes [2], and online social networks [12], to name a