PPS Sampling of Web Graph Using Preferential Jumping Strategy

Yi Li, Xiaoming Li
EECS, Peking University
Beijing, P.R.China
liyi@net.pku.edu.cn, lxm@net.pku.edu.cn

Jonathan J.H. Zhu
Web Mining Lab, Dept. of Media and Communication
City University of Hong Kong
Hong Kong, P.R.China
j.zhu@cityu.edu.hk

Abstract—Sampling is the most powerful tool for researchers to study important characteristics of the continuously growing Web. On Web page sampling problem, we collect a number of pages which are representative to the Web population. However, we believe Web sampling greatly differs from generic sampling problem. First of all, the randomness principle can not be applied to Web sampling mechanically; Secondly, randomness on page level should not be the only goal of Web sampling. We believe that there is still space to improve the randomness goal, and other than pursuing randomness on page level, new objectives should be set for host and domain levels.

In our work, we designed a new Web sampling method, called the Probability Proportional to the Size of Websites (PPSW for short) sampling. After certain preliminary experiments and analysis, we concluded that no former sampling methods took into account the host and domain level of the Web. Therefore we seek new Web sampling methods that can yield samples that are representative on host and domain level. With regard to the new objective, we redesigned the jumping strategy of the random walk while sampling. This preferential jumping strategy markedly increased the validity of random walk on host and domain level. More particularly, random walk based sampling methods have two configurations: whether the random walk has random jump probability, and whether the random walk is conducted on undirected Web graph with the help of search engine. Controlling these two configurations, together with our newly designed preferential jumping strategy, we conducted four kinds of new sampling experiments. Among the four groups of experiments, the directed one with random jump showed great performance improvement.

For evaluating our new PPSW sampling methods, we put forward new objectives, along with corresponding formula. The first two are coverage objectives. Comparatively speaking, the number of domains is several orders of magnitude smaller than the number of Web pages. Usually we are capable of handling this number data. Therefore, we wish the sample can cover as many hosts and domains as possible.

In addition to the two coverage objectives which are crude, we also proposed four proportion objectives. These four objectives tell us whether a sample reflects the sizes of hosts and domains from different angles: Domain Host Distribution, Domain Page Distribution, Host Page Distribution and Single Domain Page Distribution.

We conducted 150 comparison experiments for the three classical random walk based Web sampling methods and our PPSW sampling methods under a same environments that is as real as possible. By observing the process and results, we discussed their performances in the following aspects:

- Conventional Evaluations: e.g., out-, in-degree and PageR-ank distribution, and “Bucket Standard Deviation”.
- New Evaluations: by examining the two coverage and four proportion targets, we found that among all the sampling methods, our PPSW sampling methods has the best performance.
- Other Aspects: e.g., the length of walk, the stability and efficiency of sampling methods, the number of starting page set and search engines’ influences.

Keywords—Web sampling; Web sample; random walk; jumping strategy; PPS sampling; evaluation.

I. INTRODUCTION

Today’s Web is a continuously growing ocean of information. Such huge size of the oceanic information is far beyond our handling ability. In such a case, sampling is the most powerful tool for researchers to study important characteristics of the Web. Sampling is a kind of statistical inference method. Sampling is usually used to estimate the characteristics of the population through choosing and observing a small portion of it. The three main advantages of sampling are that the cost is lower, data collection is faster, and since the data set is smaller it is possible to ensure homogeneity and to improve the accuracy and quality of data.

On Web page sampling problem, we collect a number of pages which are representative to the Web population. However, we believe that Web sampling greatly differs from generic sampling problem. First of all, the randomness principle can not be applied to Web sampling mechanically; Secondly, randomness on page level should not be the only goal of Web sampling. We believe that there is still space to improve the randomness goal, and other than pursuing randomness on page level, new objectives should be set for host and domain levels.

Actually, random walk by following hyperlinks on Web pages is the only practical way to find new Web pages, thus most successful Web sampling methods are based on random walk. However, sampling by random walk on Web graph does not guarantee randomness. On the one hand, we can not assure that every page on the Web has a probability of access greater than 0. On the other hand, the access probability of all the Web pages does not equal to each other. In fact, the access probability of a page is related to its
degree and neighbors. How to improve random walk based sampling so that we can satisfy the randomness principle is the focus of recent researches. Several existing works have already proved in theory that their methods conform to the randomness principle, yet in our comparison experiments, their performances are flawed in some ways. Therefore, we seek an improved random walk based Web graph sampling method.

The second major problem that existing works did not notice, lies in the fact that Web is not only a set of pages, either is a Web sample.

II. RELATED WORK

A. Preliminaries

Sampling is that part of statistical practice concerned with the selection of an unbiased or random subset of individual observations within a population of individuals intended to yield some knowledge about the population of concern, especially for the purposes of making predictions based on statistical inference. Sampling is an important aspect of data collection. Researchers rarely survey the entire population for two reasons: the cost is too high, and the population is dynamic in that the individuals making up the population may change over time. The three main advantages of sampling are that the cost is lower, data collection is faster, and since the data set is smaller it is possible to ensure homogeneity and to improve the accuracy and quality of the data. Each observation measures one or more properties (such as weight, location, color) of observable bodies distinguished as independent objects or individuals. In survey sampling, survey weights can be applied to the data to adjust for the sample design. Results from probability theory and statistical theory are employed to guide practice. In business and medical research, sampling is widely used for gathering information about a population [1].

Here are some frequently used terminologies and their definitions. The set of objects whose characteristics are studied by researcher is called the population. Every object in the population is called an element. The set of chosen elements is called sample. In the most straightforward case, such as the sentencing of a batch of material from production (acceptance sampling by lots), it is possible to identify and measure every single item in the population and to include any one of them in our sample. However, in the more general case this is not possible. There is no way to identify all rats in the set of all rats. Where voting is not compulsory, there is no way to identify which people will actually vote at a forthcoming election (in advance of the election). These imprecise populations are not amenable to sampling in any of the ways below and to which we could apply statistical theory. As a remedy, we seek a sampling frame which has the property that we can identify every single element and include any in our sample. Not all frames explicitly list population elements. The sampling frame must be representative of the population and this is a question outside the scope of statistical theory demanding the judgment of experts in the particular subject matter being studied [1].

B. Classification of Sampling Methods

Sampling methods can be classified into two classes:

- A probability sampling scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection. Probability sampling includes: Simple Random Sampling, Systematic Sampling, Stratified Sampling, Probability Proportional to Size Sampling, and Cluster or Multistage Sampling.

- Nonprobability sampling is any sampling method where some elements of the population have no chance of selection (these are sometimes referred to as ‘out of coverage’/‘undercovered’), or where the probability of selection can not be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is nonrandom, nonprobability sampling does not allow the estimation of sampling errors. These conditions give rise to exclusion bias, placing limits on how much information a sample can provide about the population. Information about the relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population. Nonprobability Sampling includes: Accidental Sampling, Quota Sampling and Purposive Sampling.

C. Web Graph Sampling and Its Particularity

In Web sampling problem, the population is the set of all pages on the Web, every page is an element. In generic simple survey sampling works, people obtain a list of all the elements and are able to randomly access any element in the population. Unfortunately this is not the case in Web sampling. Good thing is, the Web can be abstracted as a graph, thus what we are capable of, is discovering new pages by following hyperlinks found on the visited pages. Therefore, random walk model is our fundamental tool in Web sampling work. Random walk on the Web is a relatively efficient way, and fit the Web sampling problem very well. Also, random walk is a relatively sophisticated model and is supported by theory. Moreover, we can use existing mature models to analyze Web sampling methods.

Actually, Web is not the only place where people found graph sampling problem, many other structures have graph abstractions, too. In P2P systems, peers and relationships within system are used to build graphs. Articles [2], [3] and
[4] are about sampling P2P systems. Is Web sampling a similar problem as P2P sampling? We doubt that.

Graph structure is not the only important characteristic of the Web. We notice that the Web is hierarchical. Refer to figure 1, the Web is divided into several levels in [5]: byte, character, word, block, page, sub-site, site, domain, top-level domain and national domain. The Web is also a multi-level structure. Although, we do not need to take care of that many levels in Web sampling problems, the author do believe that site (or host) and domain level need our attention.

Here we briefly introduce three classical random walk based Web sampling methods.

Rusmevichientong's method was proposed in [6]. We will call it method A for short from here on. Their method has two sub-types:

- Directed walk: Treat the Web as a directed graph. First, walk \( N \) steps as breaking in, then another \( M \) steps, and record the \( p \) pages that are visited. In the third step, \( p \) walks are conducted starting from each of the \( p \) visited pages, and their visit counts are recorded. After that the \( p \) visited are sampled with a probability reversely proportional to visit ratio.
- Undirected walk: The difference is that the random walk walks on undirected Web graph. And the visited pages are put into sample with a probability reversely proportional to their degrees. However, the paper did not mention how to change the Web into an undirected graph.

Bar-Yossef’s sampling method was proposed in [7]. We will call it method B from here on. The authors designed an algorithm to change the Web graph into an undirected graph using the in-link index of search engines.

Henzinger’s sampling method was proposed in [8]. We will call it method C in the future. In this work, the authors introduced random jump in the walk. Every time the walk need to choose a next page to visit, it jumps to a random page before choosing a neighbor. However, a true random jump is not practical in the Web environment, the algorithm randomly jumps to a visited page.


III. ALGORITHM DESIGN

In this section, we will introduce our self designed Web graph sampling, called the “Probability Proportional to the Sizes of Website” (PPSW for short).

A. Intuition

We intuitively request Web samples to be representative on host and domain levels. None of the existing Web graph sampling works has ever taken these requests into account. Also we notice that other than simply stick to randomness principle, auxiliary information are used in classical sampling methods, and we believe that such auxiliary information exist in Web sampling problems, too.

We hereby propose some specific intuitions that we think are reasonable.

- Coverage on host and domain level should be a primary goal of our new Web sampling method. One of the most important reasons why we need sampling is that the number of Web pages is far beyond our handling capability. Compare to the number of pages on the Web, the number of hosts and domains are several orders of magnitude smaller, and within our capability. Therefore, under ideal circumstance, we wish we can reach a 100% coverage on host and domain levels.
- The second intuition is that we wish the probability of every Website be sampled is proportional to the sizes of Website. We must concede even under a best condition, we may still not be able to reach a 100% coverage on host and domain levels, then the problem is which ones (hosts and domains) do we want to sample? Intuitively speaking, those that are more important and bigger in size should have higher probabilities of appearing in a sample. Obviously, some of the host and domains are more important than others. Similar to page level, Web is also a graph from the angle of host and domain, thus we can use PageRank (or HostRank, DomainRank) to mark the importance of a host/domain. For the size of a domain, there are two ways: the number of hosts within the domain and the number of pages within the domain. Similarly we can use the number pages with a host to denote its size. We notice that there is a certain type of sampling technique called probability proportional to size (a.k.a., PPS) sampling, in which the sample designer has access to an “auxiliary variable” or “size measure”, believed to be correlated to the variable of interest, for each element in the population. In PPS
sampling, the selection probability for each element is set to be proportional to its size measure. Although in Web sample design we do not have access to the importance and size of a host/domain, they are actually both related to the number of hyperlinks. Thus we can indirectly make use of hyperlinks to attain a PPS sampling on host and domain level.

- At last, we wish our sample can reflect the relative size of each host/domain. If two hosts have different sizes, host A is bigger than host B, our sample should correctly reflect this difference.

Moreover, we want our new sampling method more efficient and stable. Efficiency means sample more in less cost, and stability means that results do not vary much among different samplings.

B. Problem Standardization and Hypotheses

From now on, we treat the Web as a hierarchical object. As in figure 2, the Web consists of many distinct domains, each domain consists of many different hosts, and each host has many pages in it. Hyperlinks start from a page to another.

Here are some hypotheses, which confine our research within a certain area.

1) Random walk on page graph. Our sampling method is based on random walk, and we compare ours with other random walk based methods. Random walks will be conducted on page level Web graph.

2) Web sample frame. We do not have list of URLs on the Web, all we have is a set of certain number of starting pages. Then what is our sample frame? If the random walk is on directed graph, sample frame would be the union of all the strongly connected components which contain all the starting pages. If the random walk is on undirected graph, then the sample frame is the union of all the connected components which contain every starting page.

3) Starting pages. As discussed in 2, the selection of starting pages is important, in that it decides our sample frame. However we assume we have already had a set of starting pages. The selection of starting pages is beyond our research.

4) Restarting of random walk. It is highly possible that our random walk will encounter some dead ends. In such cases, we need to restart the walk at a new page, the selection of restarting pages is also out of our range. We assume that we have a set of restarting pages, and it is big enough.

5) The in-link index of search engines. Some times we will need the help of the in-link index of search engines. Therefore which pages are indexed is another problem. We will only assume there exists a search engine which indexed a number of random pages. However we will discuss the influence of the sizes of search engines (30%, 50% and 100% pages of the Web).

6) Random jump. Some sampling methods make use of random jump to avoid dead loops while walking. Before we selected next pages among the neighbor pages of current page, there is a certain probability the walk jumps to a random pages on the Web. But in real problems, we can not jump to whatever page we want. Fortunately, we can record all the visited pages, thus the random jump actually jumps to a random page in the visited set.

C. Preferential Jumping Strategy

Classical walk select a next node randomly from the neighbors of current node. Unfortunately, most hyperlinks on a Web page points to a page within current host/domain. If we map the page level walk up to the host and domain level, jumps within a host/domain only correspond to self loops on host/domain level graph. That makes the random walk very inefficient on host/domain level graph and easily stuck in a small number of hosts/domains. Random jump partially improves this flaw, but it is not a direct solution. What if we change the jumping strategy from equal probability among all the neighbors to “preferential selection”? 

Refer to figure 3, in our preferential jumping strategy neighbors are classified by current host/domain name into three groups on the first hand: out domain neighbors, out host neighbors and the rest neighbors reside in current host. In the second step, the algorithm treats all the neighbors from current host and current host as one self loop neighbor, because all the links point to current domain are self domain loops on domain level graph. Then randomly select a neighbor from the set of out domain neighbors plus the self domain loop neighbor. This way, the walk has a much higher probability to jump to a new domain.

If the former domain level selection chose the self domain loop neighbor, we lower our view to the host level graph. This time the algorithm treats all the neighbors from current
host as one self host loop neighbor, in that all the links point to current host are self host loops on host level graph. The randomly select a neighbor from the set of out host neighbors plus the self host loop neighbor. Thus our algorithm has a much higher probability to jump to a new host.

If the former host level selection chose the self host loop neighbor, the algorithm will select a neighbor from the current host neighbor group.

As in figure 3, the Web is a uniting of domain, host and page level graph. Before selecting any neighbor, we color all the links on page level graph. Those point to out domain pages are colored in blue, and those point to out host pages are colored in green, the rest current host links are colored in red. So if we move up to host level graph, all the red links degrade to a self loop. And if we move on to domain level graph, all the red and green links further degrade to a yellow self loop.

D. Probability Proportional to the Sizes of Websites Sampling and Analysis

The following algorithm 1 shows the framework of our Probability proportional to the sizes of Websites (PPSW) sampling. The framework can be divided into two stages: walk stage and sample stage. The walk stage visits pages, and sample stage put visited pages into a sample with certain strategy.

The overall algorithm has two configurations: whether the random walk has random jump probability, and whether the random walk is conducted on undirected Web graph with the help of search engine. Controlling these two configurations, together with our newly designed preferential jumping strategy, we designed four kinds of new sampling experiments: the directed walk without random jump (ND), the directed walk with random jump (RD), the undirected walk without random jump (NU), the undirected walk with random jump (RU).

Algorithm 2 shows the directed walk, while algorithm 3 shows the undirected walk. A directed walk treats all the pages pointed by the hyperlinks from current page as neighbors and no others, thus the walk is conducted on a directed graph. An undirected walk uses the help of search engine.

Algorithm 1: Sampling

**Input**: Web graph $G$, Search engine $E$, Length of walk $L$

**Output**: Sample $S$

1. Initialize $n$, $V \leftarrow \emptyset$;  
2. if undirected then Initialize trace $T$;  
3. while a walk of length $L$ do  
4. $V \leftarrow V \cup \{n\}$;  
5. if directed then $n \leftarrow \text{DIRECT\_VISIT}(n, G)$;  
6. else $n \leftarrow \text{UNDIRECT\_VISIT}(n, G, E, V, T)$;  
7. if random jump then $n \leftarrow \text{random node in } V$;  
8. $S \leftarrow \text{SAMPLE}(V)$;  
9. return $S$;

Algorithm 2: DIRECT\_VISIT

**Input**: node $n$, $G$

**Output**: next node $x$

1. $N \leftarrow G[n]$;  
2. if $N = \emptyset$ then return random node;  
3. $x \leftarrow \text{SELECT\_NEXT}(n, N)$;  
4. return $x$;

Algorithm 3: UNDIRECT\_VISIT

**Input**: node $n$, $G$, $E$, visited nodes $V$, trace $T$

**Output**: next node $x$

1. $O \leftarrow \Phi$, $I \leftarrow \Phi$, $N \leftarrow \Phi$;  
2. if $n \notin V$ then  
3. if $n \notin T$ then $T[n] \leftarrow \Phi$;  
4. $O \leftarrow G[n]$, $I \leftarrow$ at most 1000 nodes in $E[n]$;  
5. $N \leftarrow O \cup I \setminus T[n]$;  
6. for $p \in N$ do  
7. if $p \notin V$ then  
8. $T[n] \leftarrow T[n] \cup \{p\}$, $T[p] \leftarrow T[p] \cup \{n\}$;  
9. if $T[n] = \Phi$ then return random node;  
10. $x \leftarrow \text{SELECT\_NEXT}(n, T[n])$;  
11. return $x$;
domains contain 11 sample more from those bigger hosts/domains, that conforms while, the size of a host/domain is directly proportional to hyperlinks, this makes the probability of visiting these finally increase host/domain coverage. Secondly, because host/domain loop jump. Thus it is supposed to improve the subscribed above in section III-C.

Example, the Web consists of 6 levels are demonstrated by different grayscales. The areas domains, using different colors. And every domain contains shown in figure 4.

Let us image the Web as the circle shown in figure 4. It Here we try to explain our thoughts using an example to our intuition, too. In this section, we will discuss on how to evaluate a A. Sample Illustration

Here we try to explain our thoughts using an example shown in figure 4. It has three levels: domain, host and page levels respectively from inside to outside. The Web circle is divided into several domains, using different colors. And every domain contains several hosts, every host has several pages under it. Different levels are demonstrated by different grayscales. The areas that are sampled are marked with an asterisk. So in our example, the Web consists of $D = 8$ domains, each of the domains contain 1 to 3 hosts, add up to $H = 17$ hosts, every host has 3 to 9 pages, that is 94 pages in total. And in this example, we sampled 7 domains, 11 hosts and 23 pages.

B. Notations

In this section, we will give all the mathematic notations we use in the following sections. Our notations have two parts: the Web itself and sample. $W$ denotes the Web, it is a set of disjoint domains ($D$), and every $D$ consists of some disjoint hosts ($H$), and every $H$ is composed of a number of pages. Sample notations add bars to corresponding Web notations.

$W = \{D_i \mid i = 1, 2, \ldots, m\}$
$\bar{W} = \{\bar{D}_i \mid i = 1, 2, \ldots, m\}$
$D_i = \{H_{ij} \mid j = 1, 2, \ldots, n_i\}$
$\bar{D}_i = \{\bar{H}_{ij} \mid j = 1, 2, \ldots, n_i\}$
$H_{ij} = \{p_{ijk} \mid k = 1, 2, \ldots, l_{ij}\}$
$\bar{H}_{ij} \subset \bar{H}_{ij}$

And we define three functions for convenience. Function $e$ tells us whether a domain or host is sampled, and the value of function $f$ always equals to $y$. And function $s(p) = 1$ means page $p$ is sampled.

$e(\emptyset) = \begin{cases} 1 & \text{if } \emptyset \neq \Phi \text{ and } \exists o \in \emptyset, o \neq \Phi \\ 0 & \text{otherwise} \end{cases}$
$f(p_{ijk}) = 1$
$s(p_{ijk}) = \begin{cases} 1 & \text{if } p_{ijk} \in \bar{H}_{ij} \\ 0 & \text{otherwise} \end{cases}$

Here are some notations we are going to use frequently:
1) Number of domains \((D)\) and number of sampled domains \((\bar{D})\):

\[
D = m = |W| = \sum_{i=1}^{m} e(D_i), \quad \bar{D} = \sum_{i=1}^{m} e(\bar{D}_i)
\]

2) Number of hosts in domain \(i\) \((H_i)\), and number of sampled hosts in domain \(i\) \((\bar{H}_i)\):

\[
H_i = n_i = |D_i| = \sum_{j=1}^{n_i} e(H_{ij}), \quad \bar{H}_i = \sum_{j=1}^{n_i} e(\bar{H}_{ij})
\]

3) Number of hosts \((H)\) and number of sampled hosts \((\bar{H})\):

4) Number of pages in domain \(i\) host \(j\) \((P_{ij})\), and number of sampled pages in domain \(i\) host \(j\) \((\bar{P}_{ij})\):

\[
P_{ij} = l_{ij} = |H_{ij}| = \sum_{k=1}^{l_{ij}} f(p_{ijk}), \quad \bar{P}_{ij} = \sum_{k=1}^{l_{ij}} s(p_{ijk})
\]

5) Number of pages in domain \(i\) \((P_i)\), and number of sampled pages in domain \(i\) \((\bar{P}_i)\):

\[
P_i = \sum_{j=1}^{n_i} \sum_{k=1}^{l_{ij}} f(p_{ijk}), \quad \bar{P}_i = \sum_{j=1}^{n_i} \sum_{k=1}^{l_{ij}} s(p_{ijk})
\]

Also, we have the following relationship:

\[
H = \sum_{i=1}^{D} H_i, \quad \bar{H} = \sum_{i=1}^{\bar{D}} \bar{H}_i
\]

\[
P_i = \sum_{j=1}^{D} P_{ij}, \quad \bar{P}_i = \sum_{j=1}^{\bar{D}} \bar{P}_{ij}
\]

\[
P = \sum_{i=1}^{D} P_i = \sum_{j=1}^{D} H_i, \quad \bar{P} = \sum_{i=1}^{\bar{D}} \bar{P}_i = \sum_{i=1}^{\bar{D}} \bar{H}_i
\]

C. Conventional evaluations

1) Out-, In-Degree and PageRank Distribution: Out- and in-degree distribution of the Web is a focus of researchers’ attention, therefore we intent to evaluate samples from the angle of out- and in-degree distribution. Previous work showed that the out- and in-degree distribution of the Web conform to Power-Law distribution: \(P[d] = d^\lambda\). This distribution means that most pages have low degree, but a few pages have very high degree, it is called scale-free in sociology and physics. PageRank distribution is another important evaluation, previous works tell us it should conform to Power-Law, too.

In these three evaluations, we first care if the distribution conforms to Power-Law, then we will look at \(\lambda\) value of the distribution.

2) URL ID Bucket Standard Deviation: In this evaluation, the URLs are sorted alphabetically on the first hand, and then an ID is assigned to every URL in the Web. This way we group all the URLs by host names, URLs that have a same host name will be assigned consecutive IDs. Next we divide the URLs into a number of buckets (e.g. 20 buckets), and the sampled URLs are mapped into these buckets. The results show if the sample is equally distributed in these buckets. If the walk is flawed and gets stuck into few hosts, the distribution is supposed to show unbalance. We will use the following formula to calculate a “bucket standard deviation” for every sample:

\[
S = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( \frac{b_i}{\bar{P}} - \frac{1}{n} \right)^2}
\]

Here \(b_i\) denotes the number of pages in the \(i\)th bucket, and \(n\) is the number of buckets.

3) Other Aspects:

- Stability: here we examine if the results of a sampling method resembles each other. Stable method should yield similar results on different runs.
- Time efficiency: meaning the time cost by running the algorithm.
- Sample size: this evaluates how many pages a method can sample in a run. For a huge population as the Web, a sample must be big enough to represent it. We think that, within our handling ability, the sample size should be as big as possible.
- Size of starting page set: the random walk will need a new starting page when it encounters dead end, thus in real situation, a set of starting should be prepared in the first hand.

D. New Evaluations

In order to measure whether a sample is representative on host/domain level, we proposed following evaluations:

1) Coverage Objectives:

1) Domain coverage: sample as many domains as possible, that is:

\[
\frac{\bar{D}}{D} \rightarrow 1
\]

The corresponding domain coverage of the example in figure 4 is \(\overline{7} = 87.5\%\).

2) Host coverage: sample as many hosts as possible, that is:

\[
\frac{\bar{H}}{H} \rightarrow 1
\]

The corresponding host coverage of the example in figure 4 is \(\overline{17} = 64.7\%\).
2) Proportion Objectives:

3) Domain host distribution: here we believe that the number of hosts within a domain should be proportional to the real distribution, that is:

\[
\frac{1}{D} \sum_{i=1}^{D} \frac{1}{\delta_{H_i}} \cdot \left| \frac{\bar{H}_i}{H_i} - \frac{\bar{H}}{H} \right| \rightarrow 0
\]  

in which,

\[
\delta_{H_i} = \begin{cases} \frac{\bar{H}}{H} & \text{if } \frac{\bar{H}_i}{H_i} < \frac{\bar{H}}{H} \\ 1 - \frac{\bar{H}}{H} & \text{if } \frac{\bar{H}_i}{H_i} \geq \frac{\bar{H}}{H} \end{cases}
\]

The corresponding domain host distribution of the example in figure 4 can be calculated as:

\[
\begin{align*}
\frac{1}{8} \times & \left( 17 \times \left| \frac{0}{11} - \frac{11}{17} \right| + 17 \times \left| \frac{2}{6} - \frac{11}{17} \right| \\
+ & \left( 17 \times \left| \frac{2}{17} - \frac{11}{17} \right| + 17 \times \left| \frac{1}{17} - \frac{11}{17} \right| \\
+ & \left( 17 \times \left| \frac{0}{17} - \frac{11}{17} \right| + 17 \times \left| \frac{2}{17} - \frac{11}{17} \right| \right) \\
\end{align*}
\]

4) Domain page distribution: here we believe that the number of pages within a domain should be proportional to the real distribution, that is:

\[
\frac{1}{D} \sum_{i=1}^{D} \frac{1}{\delta_{P_i}} \cdot \left| \frac{\bar{P}_i}{P_i} - \frac{\bar{P}}{P} \right| \rightarrow 0
\]  

in which,

\[
\delta_{P_i} = \begin{cases} \frac{\bar{P}}{P} & \text{if } \frac{\bar{P}_i}{P_i} - \frac{\bar{P}}{P} < 0 \\ 1 - \frac{\bar{P}}{P} & \text{if } \frac{\bar{P}_i}{P_i} - \frac{\bar{P}}{P} \geq 0 \end{cases}
\]

The corresponding domain page distribution of the example in figure 4 can be calculated as:

\[
\begin{align*}
\frac{1}{8} \times & \left( 94 \times \left| \frac{0}{94} - \frac{23}{94} \right| + 94 \times \left| \frac{4}{11} - \frac{23}{94} \right| \\
+ & \left( \frac{94}{71} \times \left| \frac{3}{94} - \frac{23}{94} \right| + \frac{94}{71} \times \left| \frac{5}{16} - \frac{23}{94} \right| \\
+ & \left( \frac{94}{23} \times \left| \frac{2}{94} - \frac{23}{94} \right| + \frac{94}{23} \times \left| \frac{3}{18} - \frac{23}{94} \right| \\
+ & \left( \frac{94}{71} \times \left| \frac{3}{94} - \frac{23}{94} \right| + \frac{94}{71} \times \left| \frac{3}{12} - \frac{23}{94} \right| \right) \\
\end{align*}
\]

5) Host page distribution: here we believe that the number of pages within a host should be proportional to the real distribution, that is:

\[
\frac{1}{H} \sum_{i=1}^{H_i} \sum_{j=1}^{H_i} \frac{1}{\delta_{P_{ij}}} \cdot \left| \frac{\bar{P}_{ij}}{P_{ij}} - \frac{\bar{P}}{P} \right| \rightarrow 0
\]  

6) Single domain page distribution: here we believe that the number of pages within a host should be proportional to the real distribution in a single domain, that is:

\[
\forall 1 \leq i \leq D, \sum_{j=1}^{H_i} \left| \frac{\bar{P}_{ij}}{P_{ij}} - \frac{\bar{P}}{P} \right| \rightarrow 0
\]  

For the overall occasion, it is:

\[
\sqrt{\frac{1}{D} \sum_{i=1}^{D} \left( \frac{1}{H_i} \sum_{j=1}^{H_i} \frac{1}{\delta_{P_{ij}}} \cdot \left| \frac{\bar{P}_{ij}}{P_{ij}} - \frac{\bar{P}}{P} \right| \right)^2} \rightarrow 0
\]  

in which,

\[
\delta_{P_{ij}} = \begin{cases} \frac{\bar{P}}{P_{ij}} & \text{if } \frac{\bar{P}_{ij}}{P_{ij}} - \frac{\bar{P}}{P} < 0 \text{ and } \bar{P}_i > 0 \\ 1 - \frac{\bar{P}}{P_{ij}} & \text{if } \frac{\bar{P}_{ij}}{P_{ij}} - \frac{\bar{P}}{P} \geq 0 \text{ and } \bar{P}_i > 0 \\
1 & \text{if } \bar{P}_i = 0 \end{cases}
\]

The corresponding single domain page distribution of
the example in figure 4 can be calculated as:
\[
\left\{ \frac{1}{8} \times \left[ \frac{1}{12} \times \left( \frac{1}{1} \times \left[ \frac{1}{5} - \frac{0}{5} \right] \right)^2 + \frac{1}{22} \times \left( \frac{11}{7} \times \frac{2}{6} - \frac{4}{11} + \frac{11}{7} \times \frac{2}{5} - \frac{4}{11} \right)^2 + \frac{1}{22} \times \left( \frac{10}{3} \times \frac{1}{4} - \frac{3}{10} + \frac{10}{3} \times \frac{2}{6} - \frac{3}{10} \right)^2 + \frac{1}{32} \times \left( \frac{16}{5} \times \frac{1}{3} - \frac{5}{16} + \frac{16}{11} \times \frac{3}{7} - \frac{5}{16} \right)^2 + \frac{16}{11} \times \left( \frac{2}{6} - \frac{5}{16} \right)^2 + \frac{1}{22} \times \left( \frac{11}{2} \times \frac{2}{6} - \frac{2}{11} + \frac{11}{2} \times \frac{0}{5} - \frac{2}{11} \right)^2 + \frac{1}{32} \times \left( \frac{18}{3} \times \frac{0}{5} - \frac{3}{18} + \frac{18}{15} \times \frac{3}{9} - \frac{3}{18} \right)^2 + \frac{18}{3} \times \left( \frac{0}{4} - \frac{3}{18} \right)^2 + \frac{1}{22} \times \left( \frac{11}{8} \times \frac{3}{6} - \frac{3}{11} + \frac{11}{3} \times \frac{0}{5} - \frac{3}{11} \right)^2 + \frac{1}{22} \times \left( \frac{12}{3} \times \frac{1}{6} - \frac{3}{12} + \frac{12}{9} \times \frac{2}{6} - \frac{3}{12} \right)^2 \right\}^{\frac{1}{2}}
\]

Now let us take a look at these objectives. It is not difficult to notice and prove that there are following relationships among these six objectives:

- If the domain host distribution (objective 3) achieves optimal, then the domain coverage (objective 1) must have achieved optimal. And not vice versa.
- If the domain page distribution (objective 4) reaches optimal, then the host coverage (objective 2) must have achieved optimal. And not vice versa.
- If the host page distribution (objective 5) achieves optimal, then the domain coverage (objective 1) must have reached optimal. And not vice versa.
- If the host page distribution (objective 5) achieves optimal, then the domain page distribution (objective 4) must have reached optimal. And not vice versa.
- If the host page distribution objective (objective 5) reaches optimal, then the single domain page distribution (objective 6) must have achieved optimal. And not vice versa.

As we can see, the host page distribution looks like the highest level objective, and of course it is obviously very difficult to approach. Then we can ask two questions:

1) If sample A is worst than sample B in the sense of host page distribution objective, does it mean its other objectives will be worst, too?
2) Will two runs of one sampling method with different initial conditions result in similar host page distribu-

tion? This lead to the stability problem of sampling methods.

V. EXPERIMENTS

In this section, we will introduce our experiments and their results, and have some discussions. Our experiments compared our PPSW sampling with the three classical sampling methods.

A. Experiment Design

Table I demonstrates our control group and experiment plan. We will compare our PPSW sampling that has preferential jumping strategy with the three classical sampling methods. Random walk based sampling methods have two configurations: whether the random walk has random jump probability, and whether the random walk is conducted on undirected Web graph with the help of search engine. Method A contains two subtypes: the one walks on directed graph and the one walks on undirected graph. Method B walks on undirected graph, and method C walks on directed graph. Notice that all the undirected walks make use of the in-link index of search engine, thus we will discuss the influence of the size of the in-link index. We simulated three different sizes (30%, 50% and 100%) of in-link indices. Except method C, the walks of method A and B contains no random jump.

Controlling these two configurations, we plan to conduct eight groups of experiments for our PPSW sampling method, they are:

- ND: Directed walk without random jump.
- RD: Directed walk with random jump.
- NU3/NU5/NU: Undirected (with the help of 30%, 50% and 100% in-link index) walk without random jump.
- RU3/RU5/RU: Undirected (with the help of 30%, 50% and 100% in-link index) walk with random jump.

In order to measure the stability of every sampling method, we plan to conduct 10 experiments for every group. And other than the above experiment groups, we also yield a group of 10 true random samples (TRS) for comparison. Thus there will be 15 groups of experiments, and 150 samples in total.
The sampling methods can be divided into 2 stages: walk stage and sample stage. The walk stage visits pages, and sample stage put visited pages into a sample with certain strategy. Therefore we can not easily control the sample size, and then we make the length of all walk stages to 2000000 steps, (groups AD and AU are different because of their time cost and stability). This assures that our experiments are conducted under a same environment.

B. Environment
All the experiments are run on a single Dell x86_64 server. This server has two Intel(R) Xeon(R) L5335 2.0GHz quad-core CPUs, 4MB L3 cache, 16GB of memory, SATA hard drive, GNU/Linux 2.6.16.21 system. Most programs are written using Python and Shell script. We make sure that will be no more than one process sharing one core, and we time every experiment under a same environment.

All the walks surfs on a real graph with 105896555 pages, 114529 hosts and 98758 domains. The graph is stored on local disk using Berkeley DB.

C. Experiment Results
This section shows you the most important results of our experiments. For everyone of the 150 samples, we drew three diagrams for out-, in-degree and PageRank distribution, and a diagram for bucket evaluation, these will not be shown completely in this article because of space problem, please contact the author to get the complete results.

1) Conventional Evaluations: This section manifests the three conventional distribution evaluations: out-, in-degree and PageRank. Because the number of experiments is too big, and many samples are useless in some ways, here we only show you nine best sample results, they are AU, B3, B5, B, C, RD, RU3, RU5 and RU. If you need to see the complete results, please contact the author. Figure 5, 6 and 7 illustrate the comparisons of nine out-, in-degree and PageRank distributions of the nine samples with true distributions.

In figure 5, the true out-degree distribution (red plus sign) dose not show a Power-Law distribution until out-degree is larger than 50. And it has some noises at the high out-degree part, maybe because of spam pages. All the nine samples produce similar Power-Law shapes. Considering the slope of the Power-Law part, sample RU3 (red hollow triangle), RU5 (gray solid triangle) and RU (red hollow delta) show smaller slope, meaning that they have some biases to high out-degree pages. Sample B3 (deep blue asterisk), B5 (pink hollow square) and B (light blue solid square) have too many points at high out-degree range, and they all project at 10^3 to 10^4 out-degree range in various degrees. Sample C (yellow hollow circle) and RD (black solid circle) look smooth and stable.

Figure 6 is the comparison of in-degree distribution of the nine samples. The true distribution (red plus sign) is in standard Power-Law shape, the noises at high in-degree part may be because of spam pages. All nine shapes are in Power-Law, too. However, specifically speaking, sample B (light blue hollow square), C (yellow hollow circle) and RD (black solid circle) have relatively more noises at larger than 10^3 in-degree range, thus they have biases to high in-degree pages. From the slope angle, only the sample RU (red hollow delta) have a same slope as the true distribution, others all have biases to high in-degree pages. However, notice that sample AU made use of a 100% in-link index, which is impractical in real situation. Sample C and RD have the smallest slope, meaning that they have highest biases to high in-degree pages.

In figure 7, the true PageRank distribution (red plus sign) is a standard Power-Law distribution. Sample AU (green cross) and RU (red hollow delta) have the smallest sample size among the nine, and they are not as smooth as the other are. The other seven are in smooth Power-Law shapes, however, curved in varying range. From the slope angle, sample AU (green cross) and RU (red hollow delta) have closest slope to the real value, all others are smaller, meaning
that they have biases to high PageRank pages. Sample C (yellow hollow circle) and RD (black solid circle) have smallest slope values, these two have the most severe bias to high PageRank pages.

The “bucket standard deviation” value are shown in the table in next section.

2) New Evaluations: There are too many experiments and we can not show all the result tables because of space problem. Please contact the author to provide you the complete result tables. These two tables list the values of the biggest sample in each of the 15 groups, together with 10 true random samples’. In each line of tables, the number of sampled pages, hosts and domains are listed, together with the number visited pages, time cost length of walk, the number of no neighbor jump (when the walk encounters a dead end) and the bucket standard deviation and the four proportion objective values are listed: domain host distribution, domain page distribution, host page distribution and single domain page distribution.

D. Discussions

Let us first look the ten true random samples, our results show that a sample on our experiment Web graph should at least have a size of 15000 pages in order to gain an acceptable result.

1) Conventional Evaluations:

• AD: We designed two lengths of walk for method AD and AU: 250000 and 2500000 steps. However the results shows the AD sample sizes are all less than 15000 pages threshold, its sample sizes are too small to reflect even a nearly true Power-Law out-, in-degree and PageRank distribution. And the bucket evaluation diagrams show that AD samples all stuck on a small number of hosts.

• AU: AU samples are much better. The ones that have 2500000 steps of walk reached the 15000 pages threshold. The out-, in-degree and PageRank distribution diagrams reflect the Power-Law shape correctly, and sometimes theirs have the best slope value among all the samples, but notice that method AU used the 100% in-link index, which is impractical. Again the bucket diagram show that AU samples are very unbalanced, meaning that its walks still get stuck on few hosts.

• B3, B5 and B: The method B sample groups have the biggest sample sizes among all. The out-, in-degree and PageRank distributions of B3, B5 and B samples are in Power-Law shapes. However, the B5 samples have more noises than that of B3 sample, meaning that the help search engines have negative effect on Web sampling. In the bucket diagrams, all three groups of samples are not balanced enough. But B samples are more balanced than B5 samples, which are more balanced than B3 sample. All in all, the B sample groups have the biggest sample sizes, but the sample will have biases to high degree and PageRank pages, and will not be balanced enough.

• C: Method C stably generates samples that are big enough. The out-, in-degree and PageRank distribution are in Power-Law shapes, but the slope values are always the smallest compare to other samples’, thus its bias to high degree and PageRank pages are the most severe. The bucket standard deviation values tell us C samples are the most balanced ones, we believe this is because of the random jumps in the walks.

• ND: ND samples are too small in sizes, and are useless.

• RD: RD samples are among the best in all aspects. Theirs show smooth Power-Law shape, and are very balanced, although they have biases to high degree and PageRank pages. They are similar C samples.

• NU3, NU5 and NU: The NU sample groups are too small in sizes and are useless.

• RU3, RU5 and RU: The sizes of RU sample groups are all big enough, but the distribution performances are poor, the Power-Law shapes are not as smooth as others’, although the slope values are very close to real value. Bucket standard deviation values show that RU sample groups are not balanced.

2) New Evaluations: In the aspects of two coverage objectives, we can see that none of the samples can reach the corresponding true random sample. All the 3753 pages of AD sample are from a same host, in TRS samples, the sample that has 5000 pages have a 2.81% of host coverage and 1.85% of domain coverage. AU sample contains 26085 pages, with 1.98% and 1.85% of host and domain coverage, but the host and domain coverage of 30000-page TRS sample are 7.43% and 7.42%. B sample has 996545 pages in it, and covered 4.54% of hosts and 3.99% of domains, compare to the 1000000-page TRS sample, the TRS sample has a 33.39% of host coverage and a 33.11% of domain coverage. C sample 397862 pages, covered 11.35% hosts and 10.87%
domains, the corresponding TRS sample covered 21.98% hosts and 21.74%. RD sample contains 367625 pages, and it has highest host and domain coverage, 15.26% hosts and 14.95% domains.

RD sample have the smallest and best proportion objective values, the next is C sample. The others are not very different. Therefore, obviously method RD has the best performance among all the methods in the sense of the six new evaluations.

3) Other Aspects:

- Length of walk: It is easy to realize we can never access every page on the Web in one walk. Theoretically a walk of length N can only visit less than or equal to N pages. On the one hand, the Web is too gigantic. On the other hand, the Web graph maybe not connected, some of the pages may never be visited. Therefore in practical situation, the length of walk is much smaller than the size of Web. In our experiments, the Web graph has 105896555 pages, and we set the length of all walks to 2000000, a relatively small number.

- Stability: If a sampling method generates very different sample every time it is run, then this is an unstable sampling method. In our experiment, AD, AU and NU samples are very unstable, B sample groups are better, sample C and RD profit from random jumping and are the most stable samples.

- Time efficiency: Time efficiency means the time cost of a sampling method. As a matter of time cost, RU is the most inefficient one, it costs 30 hours every 2000000-step walk. Method AD and AU have very unstable time cost because of their algorithm. Method B seems like the most time efficient one, in that the algorithm stores a visited graph in memory for future access. Our PPSW sampling methods are not very time efficient, because the algorithm need to classify the entire neighbors set and the jumping strategy is more complicated. Method C costs four to five hours every walk.

- Sample size efficiency: This measures the number of elements sampled per unit walk. Usually method B samples most, 800000 to 900000 pages per 2000000 steps. Method C and RD samples 360000 pages in every 2000000 steps. Method RU usually samples 90000 to 100000 pages in every 2000000 steps. Method AD, ND and NU are very unstable, their sample sizes vary all the time, and the sizes themselves are usually very small.

- Number of starting pages: In our experiments, we randomly select a page for starting and restarting a walk, however in real situation, we will need to prepare a set of starting pages, thus the size of the set of starting pages become an important evaluation for sampling methods. Method C and RD use much bigger starting sets, while other sampling methods need much less ones.

- Influence of search engine: Search engine is need only when the walk is conducted on undirected graph, thus this is only for method B, NU and RU. We simulated three search engines in different sizes, indexed 30%, 50% and 100% of pages respectively. It is strange that method B3 performs better then B5, we have not found out the reason yet and it needs further study. The NU sample groups are worthless, it seems like search engine dose not help a bit. Observe that in RU3, RU5 and RU samples, we would expect some performance improvements with the increasing size of search engine. However the results show little improvements. Moreover, bigger search engine usually means more time cost, and our winner of the contest is RD, a method without the help of search engine, therefore we believe that search engines dose not play a positive role in Web sampling on the whole.

VI. Summary and Future Works

A. Summary

In this paper, we focus on Web page sampling problem. We think that Web sampling greatly differs from generic sampling problem. We believe that there is still space to improve Web sampling methods, and other than pursuing randomness on page level, new objectives should be set for host and domain levels.

In our work, we designed a new Web sampling method, called the Probability Proportional to the Size of Websites (PPSW for short) sampling. We found that no former sampling methods took into account the host and domain level of the Web. Therefore we seek new Web sampling methods that can yield samples that are representative on host and domain level. With regard to the new objective, we redesigned the jumping strategy of the random walk while sampling. This preferential jumping strategy markedly increased the validity of random walk on host and domain level. More particularly, random walk based sampling methods have two configurations: whether the random walk has random jump probability, and whether the random walk is conducted on undirected Web graph with the help of search engine. Controlling these two configurations, together with our newly designed preferential Web graph with the help of search engine. Controlling these two configurations, together with our newly designed preferential Web graph with the help of search engine. Controlling these two configurations, together with our newly designed preferential jumping strategy, we conducted four kinds of new sampling experiments. Among the four groups of experiments, the directed one with random jump (RD) showed great performance improvement.

For evaluating our new PPSW sampling methods, we put forward new objectives, along with corresponding formula. The first two are coverage objectives. Comparatively speaking, the number of domains is several orders of magnitude smaller than the number of Web pages. Usually we are capable of handling this number data. Therefore, we wish the sample can cover as many hosts and domains as possible.

In addition to the two coverage objectives which are crude, we also proposed four proportion objectives. These
four objectives tell us whether a sample reflects the sizes of hosts and domains from different angles: Domain Host Distribution, Domain Page Distribution, Host Page Distribution and Single Domain Page Distribution.

We conducted 150 comparison experiments for the three classical random walk based Web sampling methods and our PPSW sampling methods under a same environments that is as real as possible. By observing the process and results, we discussed their performances in the following aspects:

- Conventional Evaluations: e.g., out-, in-degree and PageRank distribution, and “Bucket Standard Deviation”.
- New Evaluations: by examining the two coverage and four proportion targets, we found that among all the sampling methods, our PPSW sampling methods has the best performance.
- Other Aspects: e.g., the length of walk, the stability and efficiency of sampling methods, the number of start-up pages and search engines’ influences.

B. Future Work

Because of time and energy, this paper did not cover several important aspects. We think the future direction would be the following:

- The selection of starting page set. Usually the starting page set will be chosen manually, therefore the selection of the pages would definitely affect the performance. We simply select these pages randomly in our work, but there are more to study. On the first hand, it is not possible to randomly choose a page in real situation; on the other hand, random selection may not be the best choice. The selection of starting page set. Usually the starting page set will be chosen manually, therefore the selection of the pages would definitely affect the performance. We simply select these pages randomly in our work, but there are more to study. On the first hand, it is not possible to randomly choose a page in real situation; on the other hand, random selection may not be the best choice.
- In our experiments, we used a fixed length of walk for every sampling run. However, we may need to ask another question: how does the length of walk affect the performance?
- Our PPSW sampling algorithm lacks theoretical analysis. Although the preferential jumping strategy showed great improvement in experiments, we did not prove its validity and correctness.
- The help of search engines has pros and cons. And it is strange it did not contribute as we expected it would. Moreover, maybe search engine is not the only mean to help converting the Web graph into an undirected one, the reciprocal sub-graph might help, too.

ACKNOWLEDGMENT

The authors would like to thank all the helps received. This work is partially supported by NSFC Grant 60933004. Thanks for the support of Strategic Research Grant (7002396) from City University of Hong Kong.

REFERENCES