Estimate Population Statistics via Crawling Networks

Fangwei Gao & Heting Gao

Mentor: Suhansanu Kumar

Overview

Objectives:
Estimate the statistics of social networks and search engine databases using only the provided query API

Approaches:
generate uniform samples
1) For networks: Random Walk sampler
2) For databases:
   a) Pool Based sampler
   b) Construct a network on database and sample using Random Walk sampler

Challenges:
1) Simple random walk generates non-uniform samples.
2) Graph on Search engine database should reflect degree of similarities between pages.

Random Walk Sampler

Mixing Time:
Number of steps for the distribution of samples generated by a random walk to converge within a given error.

General Random Walk Sampler:
Pick an arbitrary initial node \( u \)
For mixing-time number of steps
a) Randomly pick one node \( v \) from \( u \) and neighbors of \( u \) probability \( T \)
b) Set \( u = v \)
Return the node \( n \) as a sample

\( T \) represents the transition probability from node \( u \) to node \( v \) and is specified by different random walk sampler

Sampling Methods

Rejection Sampler (REJ):
For each sample, accept it with probability \( \frac{e}{\max|E|} \), where \( C \) is a constant.

\[ T(u, v) = \begin{cases} \frac{1}{d(u, v)} & \{u, v\} \in E \ \\
0 & \text{otherwise} \end{cases} \]

Metropolis Hasting Sampler (MH):

\[ T(u, v) = \begin{cases} \frac{1}{\max(d(u), d(v))} & \{u, v\} \in E \ \\
1 - \sum_{w \in \text{nei}(u)} \frac{1}{\max(d(u), d(v))} & u = v \ \\
0 & \text{otherwise} \end{cases} \]

Maximum Degree Sampler (MD):

\[ T(u, v) = \begin{cases} \frac{1}{d_{\text{max}}} & \{u, v\} \in E \ \\
1 - \frac{d(u)}{d_{\text{max}}} & u = v \ \\
0 & \text{otherwise} \end{cases} \]

Pool Based Sampler (PB)

Algorithm:
While(true):
Sample a document \( X \) using a sampler
\( W = \text{match}(p, X) \)
Toss a coin whose heads probability is \( \frac{1}{10} \)
If(coin comes up heads)
Break
Return \( X \)

\( \text{match}(p, X) \) is the number of queries in query pool \( p \) that \( X \) matches

Sample using PB with REJ in 1-term query pool

Results

Estimate graph size with uniform samples of 3 different samplers

Sample with MH sampler

Skills Gained & Future Plan

Skill Gained:
- Python with network structure packages and multiprocessing
- Multiple Random graph models
- Various network sampling algorithms

Future Plan:
- Perform real world experiments on online social networks such as Facebook and Twitter
- Compare and analyze the results of different sampling algorithms

Reference: