Assignment 1 : Sampling from Twitter

Post Date: Thursday, January 24, 2013
Due Date: Thursday, January 31, 2013 11:59 PM
TA Office Hours: Monday, January 27, 2013 10:00 AM - 12:00 PM (North Building N021)

In this assignment, you will be estimating some properties of twitter users. You find a brief tutorial on how to get information about twitter users using python at (`https://www.cs.duke.edu/courses/spring13/compsci590.2/Assignment1/TwiterTut.pdf`). It describes APIs that allow you to query the twitter web site for various types of information including user profiles, tweets, timelines, etc. However, each user is only allowed to query the website 350 number of times each 60 minutes. You are advised to form groups of 3. You are also advised to start working on the assignment as soon as possible (reasons will be clear once you understand the assignment).

You are required to use this API to estimate the following properties for the first A Twitter users, for all $1 <= A <= 5$ million.

- "fraction of users from the US" (*hint: use time zone*)
- "fraction of users with more than 4,000 followers"
- "fraction of users with more than 500 friends"

You are given a file `https://www.cs.duke.edu/courses/spring13/compsci590.2/Assignment1/first_twiter_ids.txt` which contains the IDs of the first 5 million users in order of when they joined twitter. A trivial solution to the above problem would be to query twitter 5 million times. However, due to the limit on number of queries to the website this can take a few days. Your goal is to approximate it within a relative error of $\epsilon = 0.25$ with high probability ($\delta = 0.05$).

1. How many queries do you need to estimate the fraction of people (μ) satisfying the above properties among the first 5 million users within a relative error of $\epsilon = 0.25$ with high probability ($\delta = 0.05$)? (Hint: you may need to make some reasonable estimate on μ)
2. How many queries do you need to estimate the fraction of people (μ) satisfying the above properties among the first A users, for all $1 <= A <= 5$ million, within a relative error of $\epsilon = 0.25$ with high probability ($\delta = 0.05$)?
3. What is the answer to questions 1 and 2 for $\epsilon = 0.1$?
4. (BONUS) What epsilon error can we achieve if 10 groups pooled their data?

Please submit a report plotting the estimates you got for each of the properties, answering the 4 questions above, and outlining the method used to do the estimation.