Data Mining

Homework 3

Due: 22/1/2017, 23:59.

Instructions—Read carefully!

You must hand in the homeworks electronically and before the due date and time.
You should do this homework individually, not in groups.

Handing in: You must hand in the homeworks by the due date and time by an email to Aris that will contain as attachment (not links!) a .zip or .tar.gz file with all your answers and subject
`[Data Mining class] Homework #`
where # is the homework number. In the text you must also mention the team members. After you submit, you will receive an acknowledgement email that your homework has been received and at what date and time. If you have not received an acknowledgement email within 2 days after you submit then contact the Aris.

The solutions for the theoretical exercises must contain your answers either typed up or hand written clearly and scanned.

The solutions for the programming assignments must contain the source code, instructions to run it, and the output generated (to the screen or to files).

For information about collaboration, and about being late check the web page.

Most of the questions are not very hard but require time and thought. **You are advised to start as early as possible, to work in groups, and to ask the instructor in case of questions.**

Problem 1. In this problem you are requested to implement the streaming algorithms that we did in class. We will implement them on data generated by twitter. Your algorithms should be able to obtain the data and compute their estimates online.

1. Implement the Flajolet–Martin estimate for counting distinct elements (F_0). Obtain ℓ independent estimates and combine them using the median of the average technique.
2. Implement the Alon–Matias–Szegedy algorithm for estimating the second moment (F_2). Obtain ℓ independent estimates and combine the by taking the average.
3. Create programs for calculating F_0 and F_2 without the streaming model (shell commands can be useful).

First test your algorithms by trying on the dataset at:
`ftp://ita.ee.lbl.gov/traces/NASA_access_log_Jul95.gz`
by considering the frequencies of the various IPs. Experiment and report results for different values of ℓ and different group sizes (for the Flajolet–Martin schema).

Then apply your algorithms on twitter data obtained by the twitter streaming API, as they are generated. Save also the data on disk so that you can verify your algorithms.

For each case, for the different values of ℓ and group sizes (for F_0) you should report in two tables, one for F_0 and one for F_2:

- the number of records

- the values of F_0 (or F_2) returned by your streaming algorithm
- the true values F_0 (or F_2)
- the absolute and relative errors
- the value of ℓ
- the group size (for the Flajolet–Martin schema)

Problem 2 (optional, extra credit). Consider a graph $G = (V,E)$, with $V = V_1 \cup V_2 \cup V_3 \cup V_4$, each $|V_i| = n'$, and all the V_is disjoint, so that $|V| = n = 4n'$. Assume that for each V_i a fraction p' of the edges exist, that is, $m' = p' \cdot \binom{n'}{2}$ out of the $\binom{n'}{2}$ edges within V_i exist, and that for each V_i,V_j with $i \neq j$ a fraction p'' of edges exist (i.e., $m'' = p'' \cdot n'^2$ out of the n'^2 possible edges exist). Consider the following two partitionings of the graph:

1. There are four partitions, with every V_i in each own partition
2. There are two partitions, V_1 and V_2 in one and V_3 and V_4 in the other.

Show that the modularity of the first partition is asymptotically higher from the modularity of the second if and only if $p' > p''$. Explain whether this result is expected.