Algorithm Foundations of Data Science and Engineering Welcome Tutorial :-)

Tutorial 3

GAO Ming

DaSE @ ECNU

11 Mar., 2019

Tutorial 3

1. In count sketch for item frequency, the algorithm returns
$$\widehat{f}_a = \text{median}_{1\le i\le t} g_i(a)C[i][h_i(a)]$$
for a query a. Please give reason for $t = O(\log(1/\delta))$.
2. For the counting sketch algorithm, say the last line is changed from "On query a, report $\widehat{f}_a = \text{median}_{1\le i\le t} g_i(a)C[i][h_i(a)]$" to "On query a, report $\widehat{f}_a = \frac{\sum_{i=1}^{t} g_i(a)C[i][h_i(a)]}{t}$". The rest of the algorithm is kept as it is. Analyze the performance of this modified algorithm.
3. Given the input streaming $b, a, c, a, d, e, a, f, a, d$, and $k = 3$, i.e., three counters. Please write down the executing process step by step and find the result of the Misra-Gries summary.
4. From your opinion,
   - Is the Misra-Gries summary mergable? That is, two summaries of different inputs of size k can be combined together to obtain a new summary of size k that summarizes the union of the two inputs.
   - Is the Misra-Gries summary suitable to be used in distributed and parallel environments?