Lab 11: Data-Driven Advertising

Department of Electrical and Computer Engineering, Iowa State University
Software Tools for Large-Scale Data Analysis, Spring 2015

The goal of this lab is to design a data-driven method for targeted advertising of a product.

BACKGROUND

MacroSoft is trying to advertise its new wearable monocles to consumers. It is planning to make use of an implicit social network among consumers to get the word out to as many consumers as possible. Their strategy is as follows: Among the set of all consumers, U, select a subset T and deliver the ad to all users in T. Then each user X in T is asked to deliver the ad to all consumers that he/she knows, and for this, the user X receives a compensation c(X). The compensation c(X) is decided by the user X, but the greater c(X) is, the lesser the likelihood that X is selected to be delivered the ad (read further for details).

How does the company determine who knows whom among the consumers? It turns out the company also owns a communication infrastructure where users (the same set of consumers) talk to each other by sending messages. So, it has a log of all messages that were sent between users in U. you can assume that each consumer knows every consumer that she has a message to. Note that the only part of the log that is relevant and is used is the sender and recipient of each message.

YOUR TASK

You have a budget of $10,000 to spend on advertising. Your task is to spend this in such a way that the number of consumers who receive the ad is maximized. It doesn't make a difference whether a consumer receives an ad from one acquaintance or more than one.

DATA

There are N users, each with a numerical id ranging from 1 to N. You are given the following additional data. A file named "compensation.txt" is located on HDFS at **'/class/s15419x/lab11/compensation.txt'**, with the compensation that each user asks for. Each line in this file has two numbers separated by a comma, the first number is the user id and the second is the compensation that the user asks for.

There is also another file named "log.txt", which has the following information about all the communication calls exchanged between consumers. There is one line in this file for each communication action between a pair of users. You can assume that each user knows every user that she has communicated with. Since this is a very large file, it is stored on HDFS. The file is located at '**/class/s15419x/lab11/log.txt**'. Which tool, and which algorithm are all up to you.

A smaller dataset with N = 120 is at '**/class/s15419x/lab11/small/**'.

SUBMISSION

Create a zip (or tar) archive with the following and hand it in through blackboard.

- Commented Code for your programs. Include all source files needed for compilation.
- The set of users that are to be targeted, the amount that you will spend, and the number of consumers who will be eventually delivered the ad. This should be in the format of a clear list of users with all relevant data displayed(for example, their compensation price or how many people they can reach)
- A report describing your approach to the problem and a description of your algorithm. The optimal solution has been proven to be NP-hard, but there are many easy heuristics that lead to reasonable answers. We aren't looking for the optimal grouping, but do the best you can.