Data Mining
Homework Assignment #12

Dmytro Fishman, Anna Leontjeva and Jaak Vilo

May 15, 2014

The goal for this homework is to get you started with basics of social network analysis.

Task 1

State why, for the integration of multiple heterogeneous information sources, many companies in industry prefer the update-driven approach (which constructs and uses data warehouses), rather than the query-driven approach (which applies wrappers and integrators). Describe situations where the query-driven approach is preferable over the update-driven approach.

Task 2

A data warehouse can be modeled by either a star schema or a snowake schema. Briey describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another. Give your opinion of which might be more empirically useful and state the reasons behind your answer.

Task 3

A data warehouse can be modeled by either a star schema or a snowake schema. Briey describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another. Give your opinion of which might be more empirically useful and state the reasons behind your answer.

Task 4

Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is

the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate. (a) Draw a star schema diagram for the data warehouse. (b) Starting with the base cuboid [date, spectator, location, game], what specic OLAP operations should one perform in order to list the total charge paid by student spectators at GM Place in 2004?

Task 5

In this task you have to come up with your own model for real-world networks generation. Come up with the algorithm and write down the pseudo code for your model. Implement the model and generate the graph of approximately the same size as in previous tasks. Compare degree distribution and clustering coefficient of your graph and other networks.
As an example, you may introduce a notion of nodes dying after some time. Note that without implementation only half of the point will be given.

Task 6

In this bonus task you will help biologists to fight the virus from the task one and two. Use data email_virus.txt, calculate the same components as in task 2b and the probability of emerging a 'large scale' epidemics as in task 2c. Hint: you may want to use BFS algorithm to estimate IN and OUT components.

This homework was inspired by Juri Leskovec stanford course "Social and Information Network Analysis".