CS505, Spring 2012, Project 2 description

In this project we look at the application of graph theory in studies of Social Networks (with security twist).

Human-resources and Security department (HRS) of *Barracuda National* (BN), a medium-size producer of luxury goods, intends to investigate the social connections within its work-force. The principal subject of studies are the emails exchanged between individuals. To conduct the study, the HRS randomly selected 1000 individuals on the work-force. To anonymize (to an extent!) the first names were masked, and only one person with a given surname was selected.

Subsequently, HRS collected one-month sample (December 2017) of emails sent by the individuals selected for the study. Moreover, to eliminate bias, only one communication between two individuals was considered. For obvious reasons, the emails were further anonymized by eliminating the content (only the metadata was retrieved). Next, the times of sending and reading were eliminated (again, to protect the privacy).

The resulting dataset consists of lines of strings of names. In a given line of the data the first string is the name of the sender. All subsequent strings in a line are recipients of mails (one mail per recipient) from the sender. For instance, if the first string in the line is DEXTER, and one of the strings in the line is MAREK, it means that during the December 2017, the employee *dexter* sent at least one email (on the company system) to the employee *marek*.

While the goals of HRS are quite ambitious, we will investigate a limited number of questions about the resulting social network. Of four questions below, three will be specified by the instructor. The fourth one must be formulated by the team and must involve meaningful problem about the BN social network.

1. The HRS of BN has a contract with the security company Kandella which established that the user *woods* recently sold a trade secret of HN (design of their new luxurious ladies handbag) on *dark Web*. It has further been established that the document containing the secret was provided to *woods* by someone within the company and probably

some of negotiations leading to this security incident were conducted by email in December 2017. Find the list of candidates for the leaker (i.e. person who may have signalled to *woods* that they will provide secret.

2. List all "cliques" of largest size (remember that in a directed graph a clique of size n involves $n(n-1)$ directed edges, not $\frac{n(n-1)}{2}$ edges, as in the undirected graph) . You will need to design an algorithm that computes maximal cliques. Hint: if C clique in our directed graph G then for every $C' \subseteq C$, the graph G restricted to C' is also a clique. Use this property to design a recursive algorithm for our task.

3. A *butterfly* in a digraph G is a subgraph B with five nodes $\{A, B, C, D, E\}$ having two cliques of size three $\{A, B, C\}$ and $\{C.D.E\}$ that are connected through common node C and have no additional edges. common node (and no additional edges). Check if the BN network contains a butterfly. If so, compute one.

4. Ask a meaningful question about the social network of BN employees, and provide an answer.

There are three parts to the project:

1. The dataset, as provided (http://www.cs.uky.edu/ marek/dataset) needs to be parsed, some representation of the set as a graph chosen, and implemented.

2. The answers to the three questions formulated by the instructor must be provided.

3. The additional reasonable question about BN social network (as provided in the dataset) must be devised and the answer provided.

The operating system for implementation (Linux, Windows) is up to the team. The implementation language is also up to the team. Python, Java etc. would be easier for me. Additionally, Python provides (more than one) module for graph processing (actually, it can handle very large graphs, with hundreds of thousands of nodes; ours has only 1000 nodes and less than 100,0000 arcs). Google provides a tool called *pregel* for processing of graphs. The enthusiasts of JAVA may also consider a graph database (well-suited to

programs in JAVA) called *Neo4J*. Of course, direct implementation is also possible.
Implementations and reporting requirements:

1. Students will form 2-person teams (there is plenty of things to do, and 2 is better than 1)

2. Each team will meet with the instructor on one of the dates: April 24, or April 26 for up to half-hour to discuss the results of the project. A signing sheet will be provided in the third week of April.

3. Each team will provide a word-processed report of their work on the project. A special attention should be paid to the answer to question (4), explaining the rationale for asking it.

4. It is expected that students will include in their report the information on *What was learned in this project.*