# Homework 5

Deadline: 1:30pm, May 6, 2013
Available points: 100. Perfect score: 100.

You will receive 10% extra credit points if you submit your answers as a typeset PDF (preferably using LaTeX, in which case you can also submit electronically your source code). There will be a 5% bonus for typewritten but not typesetted answers. Resources on how to use LaTeX are available on the course's website. **Do not submit Word documents, raw text, etc.** Make sure to generate and submit a PDF if you want to get the extra credit points. In this case you can submit your solutions electronically through `sakai.rutgers.edu`.

If you choose to submit handwritten answers and we are not able to read them, you will not be awarded any points for the part of the solution that is unreadable. Handwritten answer-sheets can be submitted during the office hours of the TA Amr Naguib Bakry (Hill 206).

Try to be precise. Have in mind that you are trying to convince a very skeptical reader (and computer scientists are the worst kind...) that your answers are correct.

Each pair of students must write its solutions **independently from other teams**, i.e., without using common notes or worksheets with other students. Each pair of students need to submit only one copy of their solutions. You must indicate at the top of your homework who you worked with. You must also indicate any external sources you have used in the preparation of your solution. **Do not plagiarize online sources and in general make sure you do not violate any of the academic standards of the course, the department or of the university** (the standards are available through the course's website).

**Problem 1 (30 points):** You are operating a company that provides on-demand Internet media. You have $n$ customers and you are interested in dividing your customer base into $d$ communities based on their past common preferences. You are planning to use this information to predict a user's preferences based on the preferences of other users in the same community.

For each pair of members $m_i$ and $m_j$ you can define a similarity value $s(m_i, m_j)$ (e.g., the number of common media they have downloaded in the past), where $s(m_i, m_i) = 0$, $s(m_i, m_j) > 0$ for $m_i \neq m_j$ and $s(m_i, m_j) = s(m_j, m_i)$.

One way to optimize the classification of the customer base into $d$ communities $\{C_1, C_2, \ldots, C_d\}$ is to minimize the following inter-cluster similarity metric:

$$ics(C_1, C_2, \ldots, C_d) = \max_{\forall\ 1 \leq i,j \leq d,\ m_i \in C_i, m_j \in C_j} s(m_i, m_j)$$

i.e., you want to minimize the similarity of the most similar pair of customers that have been assigned to different communities.

a) Provide an efficient algorithm that returns a classification of your customer base into $d$ communities that achieves the minimum $ics$ value over all possible classifications into $d$ communities.

b) What data structures do you assume for the implementation of your algorithm? What is the running time of your approach given the data structures you employ?

c) Prove the correctness of your algorithm.

**Problem 2 (20 points):** Consider the following problem. You are given an undirected graph $G(V, E)$ and you need to compute a subgraph $G'(V, E' \subset E)$ with the following properties:

- every vertex in $G'$ has at most $d$ incident edges,
- $G'$ has one connected component,
- $G'$ has $|V| - 1$ edges.

If such a graph $G'$ does not exist, you should be able to report that. Show that this problem is NP-complete.

**Problem 3 (25 points):** Julie is playing a strategy game where she has the capability of bombarding the road network of her opponent Jack. Each road connects a pair of cities among $n$ total cities belonging to Jack. In $m$ of these cities, there is also an army located. Denote Jack's cities with armies as $\{c_1, \ldots, c_m\}$.

Julie wants to bombard Jack's road network so that it becomes a set of disconnected components, where each army lies in a different connected component, i.e., there is no sequence of roads that connects $c_i$ and $c_j$, for all $i, j \in [1, m]$. In this way, when Julie starts attacking each army of Jack one by one, it will be difficult for Jack to provide support to his units. Since Julie does not have an infinite supply of airforce units, she wants to identify the minimum number of roads that she will bombard so as to achieve this objective.

(a) Provide a polynomial time algorithm that solves Julie's problem exactly when Jack has only two armies.

(b) Provide a solution that achieves an approximation ratio of at most 2 when Jack has three armies.

(c) Propose a local search approach for the general case where Jack has $m$ armies.

**Problem 4 (25 points):** Consider Jack's approach in the above strategy game. Jack has realized Julie's plans because she has recently produced a significant number of airforce units. Jack is interested in deploying anti-aircraft guns along his road network so as to be able to defend it against Julie's attack.

But Jack's resources are limited so he cannot deploy anti-aircraft units along his entire road network. The cost of deploying a sufficient number of anti-aircraft units for defending a road directly depends on the length of that road. The length of a road corresponds to the distance between the cities it connects. Therefore, the lengths of the roads satisfy the properties of a metric function. The assumption is that if Jack deploys a sufficient number of such units along a road, Julie's bombardment will be ineffective.

So Jack is trying to identify the minimum set of roads that need to be protected so that all of his armies remain in the same connected component (i.e., the cities $\{c_1, \ldots, c_m\}$ will be connected given this set of roads) and he has to spend the minimum amount of resources in purchasing the anti-aircraft guns (which directly relates to the length of the corresponding roads that need to be protected).

Assume that the road network is a complete graph. Provide a polynomial time approximation algorithm for Jack's problem that has an approximation ratio of 2.