This assignment covers Modules 1+2; it is based in part on material that will be covered next Tue and Wed. It is due on Thu, 12 June 2008 at the beginning of class (no late hand-ins, please!). If possible, please send me a PDF file (which may be obtained by scanning handwritten pages) via e-mail to hoos@cs.ubc.ca and give me a hardcopy marked with your name.

Feel free to discuss the problems and solution ideas with other students, but you need to work out and write down the actual solutions on your own.

Keep all answers as concise as possible — all else being equal, short and precise answers will be scored higher.

Problem 1 (5+5+4+6=20 marks)

(a) Briefly describe the way in which empirical studies can advance knowledge in computing science (in your own words).

(b) Briefly outline the scientific method and explain how it relates to empirical algorithmics.

(c) Briefly discuss the difference between the relevance and the significance of an empirical result. Can you think of an example (which may be abstract) of a significant but irrelevant result?

(d) Briefly explain least three fundamental differences between empirical algorithmics and other empirical sciences and briefly outline their consequences.

Problem 2 (10 marks)

Briefly explain how the empirical analysis of algorithms can help in improving existing algorithms. (Your answer should be no longer than one paragraph.)

Problem 3 (5 marks) Briefly explain a conceptually simple decision problem. Choose a problem from an area you are interested in or working in; please do not use any of the examples discussed in class. (Your answer should be no longer than one paragraph.)

Problem 4 (3+5 marks)

(a) Draw an example of a cumulative distribution function (CDF) for a distribution with three distinct modes.

(b) Imagine that when analysing the performance of a deterministic decision algorithm on a large set of benchmark instances you obtain a solution cost distribution (SCD)

similar to the distribution from part (a). What does this observation suggest and how can you further explore that hypothesis? (Your answer should be no longer than one paragraph.)

Problem 5 (10+15+10+5 marks)

Download the performance data set 'uf250-1050-satz-scd.dat' from the course home page and perform the following analysis using R (http://www.r-project.org/) and/or Gnuplot (http://www.gnuplot.info/).

(a) Determine and report the following descriptive statistics for the second and third data columns (CPU time and nbranch): mean, median, variation coefficient, quartile ratio.

(b) Generate and show box plots (using R) for the distributions underlying the statistics from part (a). List the instance names (first column) for all outliers.

(c) Generate and show CDF plots for the distributions underlying the statistics from part (a).

(d) Do you see evidence for a heavy right tail in the distributions analysed in part (c)? Briefly explain your answer.

Problem 6 (Bonus problem; no marks, just good karma)

Draw the CDF plots from Problem 5(c) that would be obtained when running the *satz* algorithm (which generated the performance data studied here) on a machine that is effectively twice as fast.