CSE 521
Assignment 1
Due Tuesday, April 8, 2003

The purpose of this problem is to practice modeling real world problems as abstract algorithms problems. This problem has been investigated by the Windows software distribution division at Microsoft. A customer requests (and pays for) delivery of a set of n upgrade or patch files. Based on the file names, the delivery system requests from the customer names of related files that the customer might already have. The customer sends the names of its related files to the delivery system. The name of the empty file is always among the names of the files sent by the customer. Naturally, the delivery system already has the customers files. To minimize the delivery time and bandwidth usage, the delivery system computes the length of diff(A, B) for all files A that the customer has, and files B that are intended for delivery. It also computes the length of diff(A, B) for all pairs of files A and B that are intended for delivery. Be sure to note that diff(A, B) does not normally equal diff(B, A). The system is restricted to send only diff's: if the client has to get a new file B, it might send diff(A, B), but only if the client already has file A either because it is among the files it had in the beginning or it was sent earlier. After receiving diff(A, B), the client can reconstruct the new file B, which can later be used as the basis for reconstructing other files. In the end, the delivery system will create a sequence of files $B_1, B_2, \ldots B_m$ and a mapping $\sigma : \{1, \ldots, m\} \rightarrow \{0, 1, \ldots, m-1\}$ with the properties:

i. All the files to be delivered are in the sequence.

ii. The sequence only contains files that are to be delivered and files that the customer already has.

iii. For $1 \leq i \leq m$, $\sigma(i) < i$.

iv. $\sigma(i) = 0$ if B_i is a file the customer already has and $\sigma(i) > 0$ if B_i is a file to be delivered.

v. The sum of the lengths of diff$(B_{\sigma(i)}, B_i)$, for B_i in the the set of files to be delivered, is minimized.

With this information in hand the delivery system can package a set of difference files to send to the customer. Naturally, this set would be compressed. The customer can then decompress, unpackage the differences, and reconstruct the files it requested in the first place.

1. Show how to model this problem as a problem in graph theory. Describe the vertices, edges, and other features of the graph representing this file delivery problem. Describe the optimization problem on graphs that models this file delivery problem.

2. Find an polynomial time algorithm in the theory literature. Give a complete citation to where the solution can be found. Give a brief description (maximum one page) of the algorithm and its time and storage complexity.