CSE 100: PA4 Report

Robin Heinonen

December 8, 2017

The union-find implementation of actorconnections is generally considerably faster than the BFS implementation. For example, for the full movie_casts.tsv and pair.tsv files, BFS averages about 8–9 seconds and union-find averages around 4–5 seconds. However, when a single actor pair is queried repeatedly many times in the same file, union-find vastly outperforms bfs, especially if that actor pair did not become connected until a relatively late year. For a specific pair that did not become connected until 2007 (PRIMUS, BARRY (I) and STRUFFOLINO, DAVID), union-find averages about 3.7 seconds — virtually identical to the runtime when the same actor pair is repeated just one time. However, BFS now performs much worse, taking well over a minute for the 100 copies (but under 5 seconds for the single copy).

The explanations for these differences are fairly straightforward. Union-find should be generally more efficient than BFS. Both algorithms loop over increasing years and actor pairs included in the input file, but the operations performed during the loops in union-find are generally faster. Union-find inserts each new first-order actor connections into the graph for each year (which involves a number of union-by-height operations), before checking if each pair is connected (which involves two find operations per pair). Union-by-height is an $O(\log N)$ (N number of actors) operation, while find takes an amortized time of (nearly) $O(1)$ due to the use of path compression. On the other hand, BFS must perform a full breadth-first search (using whatever edges exist for each year) for each actor pair for each year, which is an $O(N + |E|)$ (E being the set of edges) operation, considerably slower than simply running two find operations.

This difference is magnified when all the actor pairs are duplicates. Now, each find operation in union-find is (exactly!) a constant time operation (except, possibly, for the first find of each year) due to path compression. Hence there is essentially no difference between a single copy and 100 copies. However, BFS suffers tremendously because not only must the full BFS run every time for each actor pair (even though they are the same pairs), we don't gain any time by skipping actor pairs that are already done, which we can do when the actor pairs are distinct. Therefore, we must run a full BFS 100 times every year until the actors are found to be connected, which becomes especially expensive in later years when there is a very large number of edges in the graph.