Disjoint Word Chains

Sean A. Irvine
Hamilton, New Zealand
sairvin@gmail.com

In March 2015, Eric Angelini posted a message to the SeqFan mailing list associated with the On-line Encyclopedia of Integer Sequence (OEIS) [4, 6] asking about the longest sequence of distinct English number words where adjacent words share no letters. It turned out that the question had been asked previously in July 2011 and that Hans Havermann had made some observations that greatly simplified the problem by eliminating most integers from any solution. Following some discussion involving Eric, Hans, Bob Selcoe and myself, I implemented an exhaustive search and determined that the longest solution contains twenty-five numbers:

```
8, 4, 7, 30, 1, 36, 11, 40, 9, 42, 6, 0, 56, 100,
60, 101, 66, 111, 50, 700, 52, 707, 2, 5, 2000.
```

The search actually found 14 442 624 solutions of length 25; but the total number of solutions is considerably higher, because using the equivalences discovered by Havermann, the end-points can often be replaced with other numbers. For example, the 8 in the sequence above can be replaced with 16, 17, or any of twenty-four other numbers. The number sequence above was restricted to a particular systematic naming of numbers; names like 'one thousand one hundred' were used rather than 'eleven hundred'. Longer chains are possible if other number words are allowed such as 'half', 'pi', or 'googol'.

The purpose of this note is to investigate the longest disjoint word chains for various other sets of words and to explain the computer-assisted solution of such problems. As the number example shows, the length of the longest chain can be much less than the number of words in the set. So as well as finding long chains, it is of interest to determine upper-bounds on the length of a chain.

Consider finding the longest disjoint chain for the months of the year. Figure 1 is a graph representation of this problem where two words are linked if and only if they can follow one another in the sequence. Solving the problem is therefore equivalent to finding the longest path in the graph using each node at most once.

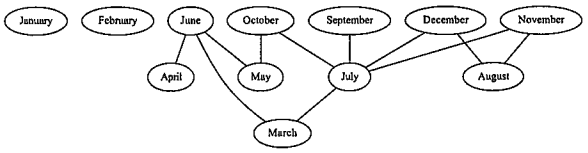

Figure 1: Graph for months.

Looking at Figure 1, it is immediately apparent that 'January' and 'February' will not appear in any long chain because they are not connected to the bulk of the graph. That is, 'January' and 'February' each share a letter with every other month. Except in special cases where a graph is completely disconnected (e.g. days of the week), all such isolated nodes can be pruned from the graph without affecting the solution. In this case, manual inspection of the remaining nodes yields a path of length eight: October, May, June, March, July, December, August, November. In fact, there are 12 possible paths of this length in the graph and no longer path exists.

As well as isolated nodes, a graph can contain redundant nodes. In the example of Figure 2, 'eight', 'sixteen', and 'seventeen' are leaf nodes connected to 'four' which is in turn connected to other nodes of the graph. Any longest solution will need at most one of these leaf nodes, so any two of them can be pruned without reducing the length of the longest solution. In the full numbers problem, it is this idea that gives the greatest reduction in computational difficulty.

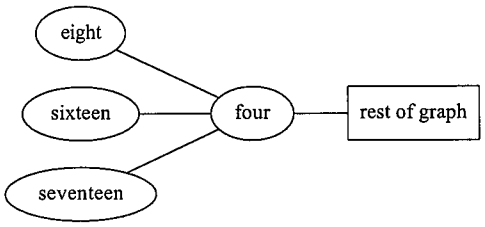

Figure 2: Example of redundant leaves.

Since at most two leaf nodes can occur in any longest path, if the graph contains more than two leaf nodes, then it will not be possible to get a path using all nodes. Thus, the presence of more than two leaf nodes can be used to reduce the upper bound on the length of the longest solution. However, it is often possible to get a stronger upper bound using a technique discussed later.

Another type of simplification allows the removal of certain edges that form triangles in the graph. In Figure 3, the edge between 'a' and 'b' can be removed,

because the path 'a'–'c'–'b' will always be better than 'a'–'b' alone. Note for this simplification to be valid, 'c' must have no other connections to the graph. In practice few simplifications of this type are possible.

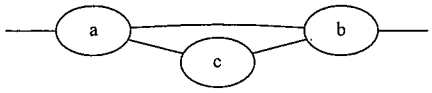

Figure 3: Example of a redundant edge.

The search amounts to a depth-first traversal of the graph. At each stage, a record is kept of nodes currently in the path. Each potential destination, not already in the path, from the current node is explored in turn. A record is kept of the longest observed path. In practice, long solutions appear to be found more quickly if nodes of lowest degree are visited first.

In general, the problem gets harder as the number of words increases. Consider solving the problem for the English names of the 24 letters of the modern Greek alphabet. The corresponding graph has 24 nodes and 106 edges (simplification removes only one edge) and is considerably more complex than the months example. Most nodes have multiple options of where to go next. Finding a solution by manual inspection is difficult.

The depth-first search just described found the following chain of length 23:

```
tau, epsilon, gamma, chi, zeta, phi, omega, nu, sigma, rho,
kappa, xi, delta, psi, theta, pi, eta, upsilon, beta, omicron,
                        alpha, mu, iota
```

This list is missing only '`lambda`'. But the first word in the chain contains an 'a' and thereafter every second word, including the last word, contains an 'a'. Therefore, it is not possible to include another word containing an 'a' without also introducing another word not containing an 'a'. Thus, 23 is the longest possible chain for the Greek letters.

Another difficult word list is the names of the chemical elements. Starting with the graph of 118 element names and applying the simplifications described above, reduces the graph to 49 nodes and 131 edges as depicted in Figure 4. However, 33 of the 49 remaining nodes contain the letter 'i', so the maximal chain can be no longer than $2(49 - 33) + 1 = 33$ by the alternating letter argument. In general, using this letter counting argument, the longest solution contains at most $\min\{n - \max\{0, l-2\}, 2(n - \max_i\{c_i\}) + 1\}$ words, where n is the total number of words in the simplified graph, l is the number of leaf nodes, and c_i is the number of words containing the letter i.

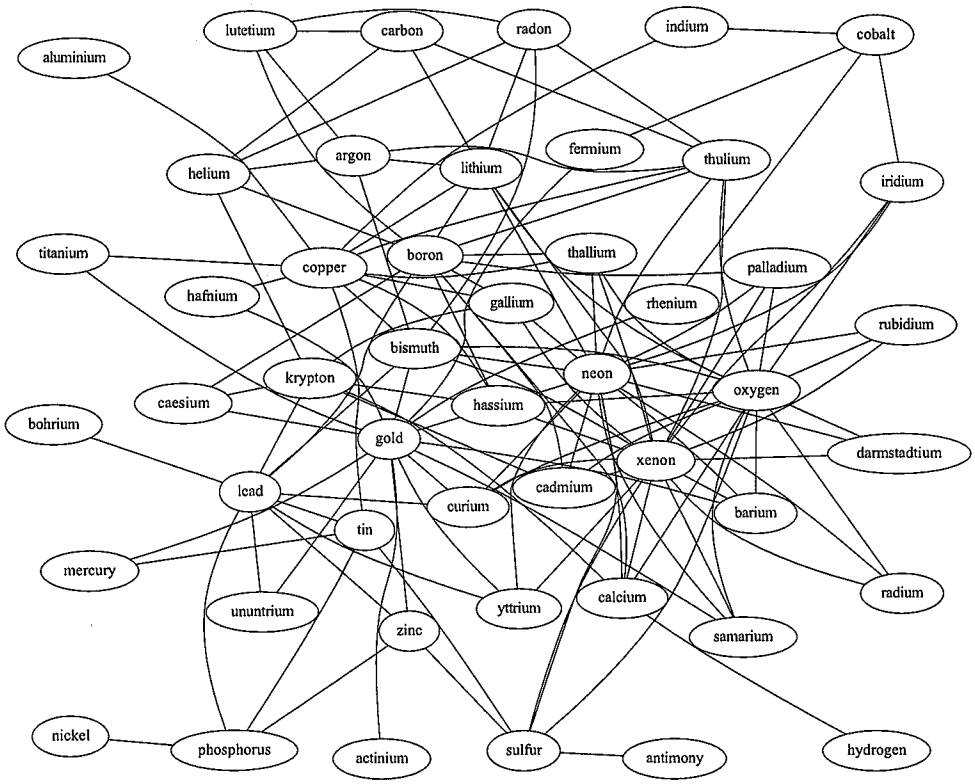

Figure 4: Reduced graph for the names of chemical elements.

Manual inspection of the simplified elements graph reveals that 'mercury' is connected to only 'tin' and 'gold'. Thus if the longest solution contains 'mercury', then the longest possible solution is reduced to at most 32.

Despite these simplifications the depth-first search procedure will not complete on the resulting graph in a reasonable time. However, it does quickly find some long solutions, including the following containing 31 words:

nickel, phosphorus, zinc, sulfur, tin, mercury, gold, hafnium, copper, indium, cobalt, iridium, oxygen, darmstadtium, xenon, rubidium, neon, yttrium, lead, bismuth, argon, lutetium, carbon, helium, radon, lithium, boron, caesium, krypton, calcium, hydrogen.

Mark Rickert [5] independently reported another solution with 31 words:

hydrogen, calcium, krypton, caesium, boron, palladium, oxygen, radium, neon, rubidium, xenon, iridium, cobalt, indium, copper,

thulium, carbon, lutetium, argon, helium, radon, bismuth, lead, ununtrium, gold, mercury, tin, phosphorus, zinc, sulfur, antimony.

Although not yet proved, a chain of 31 words is likely the longest possible since both manual and automated searches have failed to improve upon these solutions.

Sets of words can also be formed by considering all n-letter words for a given n. Here the SOWPODS official word list for Scrabble has been used. SOWPODS contains words of length 2 through 15 (words of length 1 or longer than 15 letters cannot be played in Scrabble). There is no print edition of SOWPODS, but it is effectively the same thing as *Collins Official Scrabble Words* [1]. Figure 5 shows the length of the best known solutions along with the size of the pruned graph and upper bound. The problem is completely solved for $n \leq 4$ where it is possible to put every word in the chain and $n = 6$ where the upper bound is achieved. As word length increases, the average number of shared letters between a pair of words also increases leading to a reduction in the upper bound relative to the number of words. Explicit chains for the cases in Figure 5 are available from the author.

n	total words	nodes	edges	upper bound	longest known
2	124	124	5720	124	124
3	1292	1292	501264	1292	1292
4	5454	5454	5808171	5454	5454
5	12478	12478	16194748	12478	12460
6	22157	22157	24697183	17739	17739
7	32909	32909	22962030	23083	23079
8	40161	40126	14489309	25895	25251
9	40727	39877	5048867	22795	19183
10	35529	30012	1281031	16887	10059
11	27893	15748	228996	9703	4584
12	20297	5559	36281	3853	1791
13	13857	1085	3436	863	346
14	9116	237	640	195	81
15	5757	61	164	41	34

Figure 5: Length of word chains for n-letter words taken from SOWPODS.

The problem of finding long disjoint chains of words is similar to other word games where chains are formed by changing one letter at a time or starting the next word with the last letter of the previous word. Many such examples can be

found in Chapter 4 of [3]. One such game is Geography where the names of cities are used, with each person in turn naming a city starting with the last letter of the previous city without repeating an already used city. Computationally, finding shortest chains is much easier than finding longest chains and is easily solvable in polynomial time (e.g. using Dijkstra's algorithm [2]). For example, allowing insertions and deletions, and using SOWPODS, the shortest chain from 'alpha' to 'omega' is

```
alpha, aloha, aroha, aroba, arba, alba, ala, mala, mela, mega,
                                                        omega.
```

while allowing anagrams at each step the shortest chain is

```
alpha, aleph, almeh, mahoe, omega.
```

If anagrams, insertions, and deletions are not allowed, then there is no solution.

References

[1] Collins. *Collins Official Scrabble Words*. Collins, 2012.

[2] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[3] Ross Eckler. *Making the Alphabet Dance*. Saint Martin's Press, New York, 1996.

[4] Olivier Gerard. SeqFan – Sequence Fanatics Discussion List. mailing list, see `http://list.seqfan.eu/cgi-bin/mailman/listinfo/seqfan`.

[5] Mark Rickert. Re: Chains with no shared letters between adjacent words. puzzles forum, `http://members2.boardhost.com/barryispuzzled/msg/1427490205.html`, March 2015.

[6] N. J. A. Sloane. The On-Line Encyclopedia of Integer Sequences. published electronically at `http://oeis.org/`.