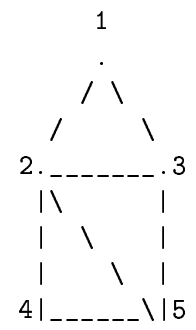
Theorem: Subset-sum is NP-complete

Proof: Clearly in NP.

For NP-hardness, reduce Hamilton cycle to subset sum.

Given a graph $G = (V, E)$, suppose that $|V| = n$ and $|E| = m$. The first step is figuring out how to represent an edge as a number. Lots of ways of doing this. Here's one. It's easiest to explain it by example. Consider the following graph:



In this example, $|V| = 5$. Then represent an edge (i, j) by a 10-digit number (5 to represent i and 5 to represent j).

- Example: represent $(4, 2)$ as the number 1011100010.
  - The first 5 digits 10111 have a 1 everywhere but at second position
  - The last 5 digits 00010 have a 1 only at the 4th position.
- The fact that (4,2) comes before (3,4) in a path will be encoded by the the 11101 that starts (4,2) matching up with the 00010 in (3,4).

In general, if there are n vertices, the edge (i, j) is represented as $2^n(2^0 + 2^1 + \cdots + 2^{j-2} + 2^j + \cdots + 2^{n-1}) + 2^{i-1}$. Call this number $e_{i,j}$. If $r_j = 2^0 + \cdots + 2^{j-2} + 2^j + \cdots + 2^{n-1}$ and $l_i = 2^{i-1}$ (r_j stands for "right j" and l_i stands for "left i"), then $e_{i,j} = 2^n r_j + l_i$. Note that if the graph is undirected, we need to represent both (i, j) and (j, i); that is, we'll use both $e_{i,j}$ and $e_{j,i}$.

That's not enough. We need to find a way of indicating the edge (i, j) is the kth edge in a Hamiltonian cycle. We do this using the number $2^{nk-1}e_{ij}$. That is, for the edge $(4, 2)$, we use 1011100010 to show that it comes first in the path, 101110001000000 to show that it comes second, 10111000100000000000 to show that it comes third in the path. There's only catch: if the edge comes in the last (nth) position in the cycle, it has to wrap around, so we represent that as $2^{n(n-1)}l_j + f_i$. For example, for the edge (4,2), we use 000100000000000000000010111. Let $e_{i,j,k}$ be the number that we use to represent edge $e_{i,j}$ in position k. Thus, for example, $e_{4,2,2}$ is 101110001000000.

What does all this buy us? Consider the cycle (1, 2, 4, 5, 3, 1) in the graph above. This consists of the edges (1,2), (2,4), (4,5), (5,3), and (3,1). That is, (1,2) is in the first position, (2,4) is in the second position, ..., and (3,1) is in the fifth position. Now look at the numbers representing (1,2) in the first position, (2,4) in the second position, etc. That is, consider the numbers $e_{1,2,1}, e_{2,4,2}, e_{4,5,3}, e_{5,3,4}, e_{3,1,5}$. Notice that $e_{1,2,1} + e_{2,4,2} + e_{4,5,3} + e_{5,3,4} + e_{3,1,5} =$

1111111111111111111111111 (that's 25 1's, if I did it right). This is easiest to see if I write the five numbers as a column:

$$\begin{aligned}
&1110100001\\
&111010001000000\\
&11110000100000000000\\
&1101100001000000000000000\\
&0010000000000000000011110
\end{aligned}$$

Notice how the "hole" at the left end of $e_{2,4,2}$ caused by the missing 1 in the fourth position is exactly filled by the 1 in the fourth position at the right end of $e_{4,5,3}$.

The bottom line is that if you take any cycle of length 5 (more generally, if you take any cycle with $|V|$ edges) and add up the numbers corresponding to the edges in the cycle in their position in the cycle, you get $2^0 + 2^1 + \cdots + 2^{n^2-1}$ (i.e., n^2 1's).

While finding a subset of numbers that adds up to n^2 1's will guarantee that there is a cycle, it won't guarantee that there is a *Hamiltonian* cycle. To make sure that the cycle is Hamiltonian, we need a way of keeping track of how many times a vertex appears in the cyle. Notice that a cycle is Hamiltonian iff a vertex appears in exactly two of the edges used. Thus, the trick will be two keep track of the vertices that appear in the path. We can encode this in the number used to represent an edge (i, j) too.

The idea is that the last 25 digits (in general, the last n^2 digits) of the number will be used to encode the path; the first nm digits will be used to encode which vertices appear and how often they appear in the path. Let $e'_{i,j,k} = e_{i,j,k} + 2^{n^2+m(i-1)} + 2^{n^2+m(j-1)}$. The key point is that each time vertex i occurs on an edge, it contributes $2^{n^2+m(i-1)}$ to the sum. If i occurs on two edges, it will contribute $2 \times 2^{n^2+m(i-1)} = 2^{n^2+m(i-1)+1}$ to the sum. Since i occurs on at most m edges (since that's the total number of edges), the most that i can contribute to the sum is

$$m \times 2^{n^2+m(i-1)} = 2^{\lg m} \times 2^{n^2+m(i-1)} = 2^{\lg(m)+n^2+m(i-1)} < 2^{n^2+mi)}.$$

That means we won't mix up the count of how many i's there are on the path with the number of j's on the path, for any other j.

Bottom line: Given a graph $G = (V, E)$, let S consist of all the numbers $e'_{i,j,k}, k = 1, \ldots, n-1$ for each edge $(i, j) \in E$. Notice that S consists of mn numbers (n numbers for each of the m edges in E). Moreover, each of the numbers has length $O(n^2 + nm)$. Let t (the desired sum) be $2^0 + 2^1 + \cdots + 2^{n^2-1} + 2^{n^2+1} + 2^{n^2+m+1} + \cdots + 2^{n^2+m(n-1)+1}$. Then G has a Hamiltonian path iff there is a subset of S that sums to t. That means we've reduced Hamiltonian cylce to subset sum, showing that subset sum is NP-complete.

(Whew!)