Finding Paths in Graphs

- Problem set #2 is due next Tuesday.
- A new version of Problem Set #1 is now on-line
From Last Time

We discussed how to turn a sequence into a graph

GACGCGGCGACGCGGCGCAA
GACGG
ACGGC
CGGCG
GGCGG
GCGGC
CGGCG
GGCGC
GCGCA
CGCAC
GCACG
CACGG
ACGCG
CGGCG
GGCGC
GCGCA
CGCAA

By placing edges connecting k-mers whose k-1 suffix matches a k-1 prefix

Our original sequence is just a path in this graph. How would you find it?
Parlor games

Once finding paths in graphs was a popular form of entertainment...
Graphs would be printed in newspapers, and people would try to find paths in them as a game.

The rules of our game

- Every node, k-mer, can be used exactly once
- The object is to find a path along edges that visits every node one time
- This game was invented in the mid 1800’s by a mathematician called Sir William Hamilton
Finding a Hamiltonian Path in our graph

For our desired sequence:

GACGGCGGCGCACGGCGCAA

is indeed a path in this graph.

How would you write a program To solve Hamilton’s puzzles?

Is the solution unique?
Another representation of k-mers in a graph

- Rather than making each k-mer a node, let's try making them an edge
- That seems odd, but it is related to the overlap idea
  - The 5-mer GACGG has a prefix GACG and a suffix ACGG
  - Think of the k-mer as the edge connecting a prefix to a suffix
  - This leads to a series of simple graphs

- Then combine all nodes with the same label
A De Bruijn Graph

This graph, like the previous one has the property that edges connect nodes where a k-1 suffix matches a k-1 prefix. Graphs of this type are called "De Bruijn" graphs, after a famous mathematician.

Recall that our original 5-mers are edges in this graph, whereas they were nodes in the previous one.

Now, how might you infer the original sequence using this graph?
This leads to a new game

The rules of our new game

- Every *edge*, k-mer, can be used exactly once
- The object is to find a path in the graph that uses each *edge* only one time
- This game was invented in the late 1700's by a mathematician called Leonhard Euler

A version of Euler's game:

Leonhard Euler

Bridges of Königsberg: Find a city tour that crosses every bridge just once
Two graphs, same problem

Two graphs representing 5-mers from the sequence "GACGCGCGGCACGGCGCAA"

**Hamiltonian Path:**

Each k-mer is a vertex. Find a path that passes through every vertex of this graph exactly once.

**Eulerian Path:**

Each k-mer is an edge. Find a path that passes through every edge of this graph exactly once.
De Bruijn's Problem

Nicolaas de Bruijn
(1918-2012)

A dutch mathematician noted for his many contributions in the fields of graph theory, number theory, combinatorics and logic.

Minimal Superstring Problem:

Find the shortest sequence that contains all $|\Sigma|^k$ strings of length $k$ from the alphabet $\Sigma$ as a substring.

Example: All strings of length 3 from the alphabet {'0','1'}.

$$\text{binary3} = \{\text{000}', \text{001}', \text{010}', \text{011}', \text{100}', \text{101}', \text{110}', \text{111}'\}$$

$$\begin{array}{cccc}
101 & 100 & & \\
001 & 111 & & \\
\text{Solution #1:} & \text{0001011100} & & \\
000 & 011 & & \\
010 & 110 & & \\
\end{array}$$

$$\begin{array}{cccc}
111 & 100 & & \\
001 & 101 & & \\
\text{Solution #2:} & \text{0001110100} & & \\
000 & 110 & & \\
011 & 010 & & \\
\end{array}$$

He solved this problem by mapping it to a graph. Note, this particular problem leads to cyclic sequence.
De Bruijn's Graphs

Minimal Superstrings can be constructed by finding a Hamiltonian path of an $k$-dimensional De Bruijn graph. Defined as a graph with $|\Sigma|^k$ nodes and edges from nodes whose $k-1$ suffix matches a node's $k-1$ prefix. Or, equivalently, an Eulerian cycle of in a $(k-1)$-dimensional De Bruijn graph. Here edges represent the $k$-length substrings.
Graph Representations

An example graph:

An Adjacency Matrix:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

An $n \times n$ matrix where $A_{ij}$ is 1 if there is an edge connecting the $i$th vertex to the $j$th vertex and 0 otherwise.

Adjacency Lists:

Edge = [(0,1), (0,4), (1,2), (1,3), (2,0), (3,0), (4,1), (4,2), (4,3)]

An array or list of vertex pairs $(i,j)$ indicating an edge from the $i$th vertex to the $j$th vertex.
An adjacency list graph object

```python
In [1]:
class BasicGraph:
    def __init__(self, vlist=[]):
        """ Initialize a Graph with an optional vertex list """
        self.index = {v:i for i,v in enumerate(vlist)}  # looks up index given name
        self.vertex = {i:v for i,v in enumerate(vlist)}  # looks up name given index
        self.edge = []
        self.edgelabel = []

    def addVertex(self, self, label):
        """ Add a labeled vertex to the graph """
        index = len(self.index)
        self.index[label] = index
        self.vertex[index] = label

    def addEdge(self, self, vsrc, vdst, label=' ', repeats=True):
        """ Add a directed edge to the graph, with an optional label.
        Repeated edges are distinct, unless repeats is set to False. """
        e = (self.index[vsrc], self.index[vdst])
        if (repeats or (e not in self.edge)):
            self.edge.append(e)
            self.edgelabel.append(label)
```
Let's generate the vertices needed to find De Bruijn's superstring of 4-bit binary strings... and create a graph object using them.

```python
In [2]: import itertools

binary = [''.join(t) for t in itertools.product('01', repeat=4)]
print(binary)

G1 = BasicGraph(binary)
for vsrc in binary:
    G1.addEdge(vsrc, vsrc[1:]+'0')
    G1.addEdge(vsrc, vsrc[1:]+'1')

print()
print("Vertex indices = ", G1.index)
print()
print("Index to Vertex = " , G1.vertex)
print()
print("Edges = ", G1.edge)

['0000', '0001', '0010', '0011', '0100', '0101', '0111', '1000', '1001', '1010', '1011', '1100', '1101', '1111']


Edges = [ (0, 0), (0, 1), (1, 2), (1, 3), (2, 4), (2, 5), (3, 6), (3, 7), (4, 8), (4, 9), (5, 10), (5, 11), (6, 12), (6, 13), (7, 14), (7, 15), (8, 6), (8, 1), (9, 2), (9, 3), (10, 4), (10, 5), (11, 6), (11, 7), (12, 8), (12, 9), (13, 10), (13, 11), (14, 12), (14, 13), (15, 14), (15, 15) ]
```
The resulting graph
The Hamiltonian Path Problem

Next, we need an algorithm to find a path in a graph that visits every node exactly once, if such a path exists.

How?

Approach:

- Enumerate every possible path (all permutations of N vertices). Python's `itertools.permutations()` does this.
- Verify that there is an edge connecting all N-1 pairs of adjacent vertices
All vertex permutations = every possible path

A simple graph with 4 vertices

![Graph Image]

In [5]:

```python
import itertools

start = 1
for path in itertools.permutations([1, 2, 3, 4]):
    if (path[0] != start):
        print()
        start = path[0]
    print(path, end=' ',

(1, 2, 3, 4), (1, 2, 4, 3), (1, 3, 2, 4), (1, 3, 4, 2), (1, 4, 2, 3), (1, 4, 3, 2),
(2, 1, 3, 4), (2, 1, 4, 3), (2, 3, 1, 4), (2, 3, 4, 1), (2, 4, 1, 3), (2, 4, 3, 1),
(3, 1, 2, 4), (3, 1, 4, 2), (3, 2, 1, 4), (3, 2, 4, 1), (3, 4, 1, 2), (3, 4, 2, 1),
(4, 1, 2, 3), (4, 1, 3, 2), (4, 2, 1, 3), (4, 2, 3, 1), (4, 3, 1, 2), (4, 3, 2, 1),
```
A Hamiltonian Path Algorithm

- Test each vertex permutation to see if it is a valid path
- Let's extend our BasicGraph into an EnhancedGraph class
- Create the superstring graph and find a Hamiltonian Path

```python
import itertools

class EnhancedGraph(BasicGraph):
    def hamiltonianPath(self):
        """ A Brute-force method for finding a Hamiltonian Path. 
        Basically, all possible N! paths are enumerated and checked 
        for edges. Since edges can be reused there are no distictions 
        made for "which" version of a repeated edge. """
        path = []
        for i in range(len(path)-1):
            if ((path[i], path[i+1]) not in self.edge):
                break
        else:
            return [self.vertex[i] for i in path]
        return []

G1 = EnhancedGraph(binary)
for vsrc in binary:
    G1.addEdge(vsrc, vsrc[1:]+"0")
G1.addEdge(vsrc, vsrc[1:]+"1")

# WARNING: takes about 20 mins
path = G1.hamiltonianPath()
print(path)
superstring = path[0] + ".".join([path[i][3] for i in range(1,len(path))])
print(superstring)

CPU times: user 18min 11s, sys: 52 ms, total: 18min 11s
Wall time: 18min 11s
['0000', '0001', '0010', '0100', '0011', '0110', '1000', '0101', '1010', '0111', '1111', '1110', '1100', '1000']
000010010101111111000
```
Visualizing the result
Is this solution unique?

How about the path = "000111101001011000"

- Our Hamiltonian path finder produces a single path, if one exists.
- How would you modify it to produce every valid Hamiltonian path?
- How long would that take?

One of De Bruijn's contributions is that there are:

\[
\frac{(\sigma!)^{\sigma^{k-1}}}{\sigma^k}
\]

paths leading to superstrings where \( \sigma = |\Sigma| \).

In our case \( \sigma = 2 \) and \( k = 4 \), so there should be \( 2^8 / 2^4 = 16 \) paths.
(ignoring those that are just different starting points on the same cycle)
Brute Force is slow!

- There are \( N! \) possible paths for \( N \) vertices.
- Our 16 vertices give \( 20,922,789,888,000 \) possible paths.

- There is a fairly simple Branch-and-Bound evaluation strategy
  - Grow the path using only valid edges
- Use recursion to extend paths along graph edges
- Trick is to maintain two lists:
  - The path so far, where each adjacent pair of vertices is connected by an edge
  - Unused vertices. When the unused list becomes empty we've found a path
A Branch-and-Bound Hamiltonian Path Finder

```python
import itertools

class ImprovedGraph(BasicGraph):

    def SearchTree(self, path, verticesLeft):
        """ A recursive Branch-and-Bound Hamiltonian Path search. Paths are extended one node at a time using only available edges from the graph. """
        if (len(verticesLeft) == 0):
            self.PathV2Result = [self.vertex[i] for i in path]
            return True
        for v in verticesLeft:
            if (len(path) == 0) or ((path[-1], v) in self.edge):
                if self.SearchTree(path+[v], [r for r in verticesLeft if r != v]):
                    return True
        return False

    def hamiltonianPath(self):
        """ A wrapper function for invoking the Branch-and-Bound Hamiltonian Path search. """
        self.PathV2result = []
        self.SearchTree([], sorted(self.index.values()))
        return self.PathV2result

G1 = ImprovedGraph(binary)
for vsrc in binary:
    G1.addEdge(vsrc, vsrctilde[1]+'0')
    G1.addEdge(vsrc, vsrctilde[1]+'1')
    t1 = timeit.timeit(G1.hamiltonianPath)
    path = G1.hamiltonianPath()
    print(path)
    superstring = path[0] + ''.join([path[i][3] for i in range(1, len(path))])
    print(superstring)
```

81 μs ± 684 ns per loop (mean ± std. dev. of 7 runs, 10000 loops each)

['0000', '0001', '0010', '0011', '0100', '0101', '0110', '0111', '1000', '1001', '1010', '1011', '1100', '1101', '1110', '1111']

0000100110111000
Is there a better Hamiltonian Path Algorithm?

- Better in what sense?
- Better = number of steps to find a solution are polynomial in either the number of edges or vertices
  - Polynomial: variable\(^\text{constant}\)
  - Exponential: constant\(^\text{variable}\) or worse, variable\(^\text{variable}\)
- For example our Brute-Force algorithm was \(O(V!) = O(V^V)\) where \(V\) is the number of vertices in our graph, a problem variable
- We can only practically solve only small problems if the algorithm for solving them takes a number of steps that grows exponentially with a problem variable (i.e. the number of vertices), or else be satisfied with heuristic or approximate solutions
- Can we prove that there is no algorithm that can find a Hamiltonian Path in a time that is polynomial in the number of vertices or edges in the graph?
  - No one has, and here is a million-dollar reward if you can!
  - If instead of a brute who just enumerates all possible answers we knew an oracle could just tell us the right answer (i.e. Nondeterministically)
  - It's easy to verify that an answer is correct in Polynomial time.
  - A lot of known similar problems will suddenly become solvable using your algorithm
What next?

Is there hope?

What if our k-mers are edges?