Lab Objective: Graph theory has many practical applications. A graph may represent a complex system or network, and analyzing the graph often reveals critical information about the network. In this lab we learn to store graphs as adjacency dictionaries, implement a breadth-first search to identify the shortest path between two nodes, then use the NetworkX package to solve the so-called “Kevin Bacon problem.”

Graphs in Python

Computers can represent mathematical graphs using various kinds of data structures. In previous labs, we stored graphs as trees and linked lists. For non-tree graphs, perhaps the most common data structure is an adjacency matrix, where each row of the matrix corresponds to a node in the graph and the entries of the row indicate which other nodes the current node is connected to. For more on adjacency matrices, see chapter 2 of the Volume II text.

Another common graph data structure is an adjacency dictionary, a Python dictionary with a key for each node in the graph. The dictionary values are lists containing the nodes that are connected to the key. For example, the following is an adjacency dictionary for the graph in Figure 4.1:

```python
# Python dictionaries are used to store adjacency dictionaries.
>>> adjacency_dict = {'A':['B', 'C', 'D', 'E'], 'B':['A', 'C'],
                   'C':['B', 'A', 'D'], 'D':['A', 'C'], 'E':['A']}

# The nodes are stored as the dictionary keys.
>>> print(adjacency_dict.keys())
['A', 'C', 'B', 'D', 'E']

# The values are the nodes that the key is connected to.
>>> print(adjacency_dict['A'])
['B', 'C', 'D', 'E']    # A is connected to B, C, D, and E.
```
Problem 1. Implement the \_\_str\_\_() method in the provided Graph class. Print each node in the graph, followed by a list of the neighboring nodes separated by semicolons. (Hint: consider using the join() method for strings.)

```python
>>> my_dictionary = {'A':['C', 'B'], 'C':['A'], 'B':['A']}
>>> graph = Graph(my_dictionary)
>>> print(graph)
A: C; B
C: A
B: A
```

Breadth-First Search

Many graph theory problems are solved by finding the shortest path between two nodes in the graph. To find the shortest path, we need a way to strategically search the graph. Two of the most common searches are depth-first search (DFS) and breadth-first search (BFS). In this lab, we focus on BFS.

A BFS traverses a graph as follows: begin at a starting node. If the starting node is not the target node, explore each of the starting node’s neighbors. If none of the neighbors are the target, explore the neighbors of the starting node’s neighbors. If none of those neighbors are the target, explore each of their neighbors. Continue the process until the target is found.

As an example, we will do a programmatic BFS on the graph in Figure 4.1 one step at a time. Suppose that we start at node C and we are searching for node E.

```python
# Start at node C
>>> start = 'C'
>>> current = start

# The current node is not the target, so check its neighbors
>>> adjacency_dictionary[current]
['B', 'A', 'D']

# None of these are E, so go to the first neighbor, B
>>> current = adjacency_dictionary[start][0]
>>> adjacency_dictionary[current]
```
You may have noticed that some problems in the previous approach that would arise in a more complicated graph. For example, what prevents us from entering a cycle? How can we algorithmically determine which nodes to visit as we explore deeper into the graph?

Implementing Breadth-First Search

We solve these problems using a queue. Recall that a queue is a type of limited-access list. Data is inserted to the back of the queue, but removed from the front. Refer to the end of the Data Structures I lab for more details.

A queue is helpful in a BFS to keep track of the order in which we will visit the nodes. At each level of the search, we add the neighbors of the current node to the queue. The collections module in the Python standard library has a deque object that we will use as our queue.
We have arrived at a new problem. The nodes A and C were added to the queue to be visited, even though C has already been visited and A is next in line.

We can prevent these nodes from being added to the queue again by creating a set of nodes to contain all nodes that have already been visited, or that are marked to be visited. Checking set membership is very fast, so this additional data structure has minimal impact on the program’s speed (and is faster than checking the deque).

In addition, we keep a list of the nodes that have actually been visited to track the order of the search. By checking the set at each step of the algorithm, the previous problems are avoided.

```python
deque(['A', 'D', 'A', 'C'])
```

Problem 2. Implement the `traverse()` method in the `Graph` class using a BFS. Start from a specified node and proceed until all nodes in the graph have been visited. Return the list of
visited nodes. If the starting node is not in the adjacency dictionary, raise a `ValueError`.

**Shortest Path**

In a BFS, as few neighborhoods are explored as possible before finding the target. Therefore, the path taken to get to the target must be the shortest path.

Examine again the graph in Figure 4.1. The shortest path from $C$ to $E$ is start at $C$, go to $A$, and end at $E$. During a BFS, $A$ is visited because it is one of $C$’s neighbors, and $E$ is visited because it is one of $A$’s neighbors. If we knew programmatically that $A$ was the node that visited $E$, and that $C$ was the node that visited $A$, we could retrace our steps to reconstruct the search path.

To implement this idea, initialize a new dictionary before starting the BFS. When a node is added to the visit queue, add a key-value pair to the dictionary where the key is the node that is visited and the value is the node that is visiting it. When the target node is found, step back through the dictionary until arriving at the starting node, recording each step.

**Problem 3.** Implement the `shortest_path()` method in the `Graph` class using a BFS. Begin at a specified starting node and proceed until a specified target is found. Return a list containing the node values in the shortest path from the start to the target (including the endpoints). If either of the inputs are not in the adjacency graph, raise a `ValueError`.

**Network X**

NetworkX is a Python package for creating, manipulating, and exploring large graphs. It contains a graph object constructor as well as methods for adding nodes and edges to the graph. It also has methods for recovering information about the graph and its structure.

A node can be an int, a string, or a Python object, and an edge can be weighted or unweighted. There are several ways to add nodes and edges to the graph, some of which are listed below.

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>add_node()</code></td>
<td>Add a single node to the graph.</td>
</tr>
<tr>
<td><code>add_nodes_from()</code></td>
<td>Add a list of nodes to the graph.</td>
</tr>
<tr>
<td><code>add_edge()</code></td>
<td>Add an edge between two nodes.</td>
</tr>
<tr>
<td><code>add_edges_from()</code></td>
<td>Add a list of edges to the graph.</td>
</tr>
</tbody>
</table>

```python
# Create a new graph object using networkX
>>> import networkx as nx
>>> nx_graph = nx.Graph()

>>> nx_graph.add_node('A')
>>> nx_graph.add_node('B')
>>> nx_graph.add_edge('A', 'B')
>>> nx_graph.add_edge('A', 'C') # Node 'C' is added to the graph
>>> nx_graph.add_edges_from([('A', 'D'), ('A', 'E'), ('B', 'C')])

# Graph edges can also have assigned weights.
```
>>> nx_graph.add_edge('C', 'D', weight=0.5)

# Access the nodes and edges.
>>> print(nx_graph.nodes())
['A', 'C', 'B', 'E', 'D']

>>> print(nx_graph.edges())
[('A', 'C'), ('A', 'B'), ('A', 'E'), ('A', 'D'), ('C', 'B'), ('C', 'D')]

>>> print(nx_graph.get_edge_data('C', 'D'))
{'weight': 0.5}

# Small graphs can be visualized with nx.draw().
>>> from matplotlib import pyplot as plt
>>> nx.draw(nx_graph)
>>> plt.show()
Achtung!

Because of the size of the dataset, do not attempt to visualize the graph with nx.draw(). The visualization tool in NetworkX is only effective on relatively small graphs. In fact, graph visualization in general remains a challenging and ongoing area of research.

NetworkX is equipped with a variety of methods to aid in the analysis of graphs. A selected few are listed below.

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjacency_matrix()</td>
<td>Returns the adjacency matrix as a SciPy sparse matrix.</td>
</tr>
<tr>
<td>dijkstra_path()</td>
<td>Returns the shortest path from a source to a target in a weighted graph.</td>
</tr>
<tr>
<td>has_path()</td>
<td>Returns True if the graph has a path from a source to a target.</td>
</tr>
<tr>
<td>prim_mst()</td>
<td>Returns a minimum spanning tree for a weighted graph.</td>
</tr>
<tr>
<td>shortest_path()</td>
<td>Returns the shortest path from a source to a target.</td>
</tr>
</tbody>
</table>

# Verify nx_graph has a path from 'C' to 'E'
>>> nx.has_path(nx_graph, 'C', 'E')
True

# The shortest_path method is implemented with a
# bidirectional BFS (starting from both ends)
>>> nx.shortest_path(nx_graph, 'C', 'E')
['C', 'A', 'E']
Problem 6. Use NetworkX to implement the `path_to_bacon()` method. Accept start and target values (actors’ names) and return a list with the shortest path from the start to the target. Set Kevin Bacon as the default target. If either of the inputs are not contained in the stored collection of dictionary values (if either input is not an actor’s name), raise a `ValueError`.

```python
>>> movie_graph = BaconSolver("movieData.txt")
>>> movie_graph.path_to_bacon("Jackson, Samuel L.")
['Jackson, Samuel L.', 'Captain America: The First Avenger', 'Stark, Peter', 'X-Men: First Class', 'Bacon, Kevin']
```

Problem 7. Implement the `bacon_number()` method in the `BaconSolver` class. Accept start and target values and return the number of actors in the shortest path from start to target. Note that this is different than the number of entries in the shortest path list, since movies do not contribute to an actor’s Bacon number.

Also implement the `average_bacon()` method. Compute the average Bacon number across all of the actors in the stored collection of actors. Exclude any actors that are not connected to Kevin Bacon (their theoretical Bacon number would be infinity). Return the average Bacon number and the number of actors not connected at all to Kevin Bacon.

As an aside, the prolific Paul Erdős is considered the Kevin Bacon of the mathematical community. Someone with an “Erdős number” of 2 co-authored a paper with someone who co-authored a paper with Paul Erdős.
Additional Material

Depth-First Search

A depth-first search (DFS) takes the opposite approach of a breadth-first search. Instead of checking all neighbors of a single node before moving on, it checks the first neighbor, then their first neighbor, then their first neighbor, and so on until reaching a leaf node. Then the algorithm backtracks to the previous node and checks its second neighbor. A DFS is sometimes more useful than a BFS, but a BFS is usually better\(^1\) at finding the shortest path.

Consider adding a keyword argument to the `traverse()` method of the `Graph` class that specifies whether to use a BFS (the default) or a DFS. This can be done by changing a single line of the BFS code.

Improvements to the BaconSolver Class

Consider adding a `plot_bacon()` method in the `BaconSolver` class that produces a simple histogram displaying the frequency of the Bacon numbers in the data set.

\(^1\)https://xkcd.com/761/