HW1: Network Construction

Kunlun Zhu

Reading: https://www.cs.cornell.edu/home/kleinber/networks-book/networks-book-ch02.pdf

Problem

Constructing Networks: Suggesting Similar Papers

The citation network is a directed network where the vertices are academic papers and there is a directed edge from paper A to paper B if paper A cites paper B in its bibliography. Google Scholar performs automated citation indexing and has a useful feature to find similar papers. In the following, we analyze two approaches for measuring similarity between papers.

(a) Co-citation network: Two papers are said to be cocited if they are both cited by the same third paper. The edge weights in the cocitation network correspond to the number of cocitations. How do you compute the (weighted) adjacency matrix of the cocitation network from the adjacency matrix of the citation network?

(b) Bibliographic coupling: Two papers are said to be bibliographically coupled if they cite the same other papers. The edge weights in a bibliographic coupling correspond to the number of common citations between two papers. How do you compute the (weighted) adjacency matrix of the bibliographic coupling from the adjacency matrix of the citation network?

(c) Bibliographic coupling and cocitation can both be taken as an indicator that papers deal with related material. However, they can in practice give noticeably different results. Why? Which measure is more appropriate as an indicator for similarity between papers?

Some notes and summarization for the reading:

In the reading we review the definition of Graph as well as its path and connectivity, and then we overview the definition of components especially the understanding of giant components. **The Small-World Phenomenon** is a very typical phenomenon which indicate the 6 degrees of separation.

Answer

a) Given the adjacency matrix of size $n \times n$ (where n = |V|) with

$$A_{ij} = \begin{cases} 1 \ if \ (i,j) \epsilon E \\ 0 \ otherwise \end{cases} \quad (1)$$

Let's denote the adjacency matrix of the co-citation matrix as below:

$$Co_{ij} = c_{i,j} \ if \ the \ i_{th} \ paper \ and \ the \ j_{th} \ paper \ has \ c_{i,j} \ number \ of \ cocitation$$

We will prove the following conclusion:

$$c_{i,j} = \overrightarrow{a_{.,I}} \cdot \overrightarrow{a_{.,j}} \quad (2)$$

Where $\overrightarrow{a_{.,I}}$ denotes the i-th column of the adjacent matrix $\boldsymbol{A}$ as a vector.

The reason is simple, the k-th element in the vector $\overrightarrow{a_{\cdot,I}}$ represent if the k-th paper had cited the i-th paper.

$$\overrightarrow{a_{\cdot,I}}[k] = \begin{cases} 1 & \text{if the } k_{th} \text{ paper had cited the } i_{th} \text{ paper} \\ 0 & otherwise \end{cases} \quad (3)$$

So we have $\overrightarrow{a_{\cdot,I}}[k] \times \overrightarrow{a_{\cdot,J}}[k] = 1$ if the k_{th} had cited both i_{th} paper and j_{th}, so the dot product of the two vector $c_{i,j} = \overrightarrow{a_{\cdot,I}} \cdot \overrightarrow{a_{\cdot,J}}$ is exactly what we want as the weight for the Adjacency matrix of the co-citation network.
Thus, we have following results according to the function (2).

$$Co = A^T A \quad (4)$$

This is same as function (2) according to the matrix multiplication.

b) Given the adjacency matrix of size $n \times n$ (where n = |V|) with

$$A_{ij} = \begin{cases} 1 & if\ (i,j) \epsilon E \\ 0 & otherwise \end{cases} \quad (5)$$

Let's denote the adjacency matrix of the bibliographic coupling as below:

$$B_{ij} = b_{i,j} \quad (6)$$

$$if\ the\ i_{th}\ paper\ and\ the\ j_{th}\ paper\ has\ b_{i,j}\ number\ of\ common\ citations$$

Similarly, I will prove the following conclusion:

$$b_{i,j} = \overrightarrow{a_{I,\cdot}} \cdot \overrightarrow{a_{J,\cdot}} \quad (7)$$

Where $\overrightarrow{a_{I,\cdot}}$ denotes the i-th row of the adjacent matrix $\boldsymbol{A}$ as a vector.
The reason is simple, the k-th element in the vector $\overrightarrow{a_{I,\cdot}}$ represent if the k-th paper had been cited the i-th paper.
So, we have $\overrightarrow{a_{I,\cdot}}[k] \times \overrightarrow{a_{J,\cdot}}[k] = 1$ if the k_{th} had been cited by both i_{th} paper and j_{th}, so the dot product of the two vector $b_{i,j} = \overrightarrow{a_{I,\cdot}} \cdot \overrightarrow{a_{J,\cdot}}$ is exactly what we want as the weight for the Adjacency matrix of the bibliographic coupling network.

Similarly, we have

$$B = AA^T \quad (8)$$

This is same as function (7) according to the matrix multiplication.

c) The reason is that the two perspectives 'Co-citation' and 'Bibliographic coupling' have no causal relationship and no proof of strong correlation, so two results may have huge difference.

I believe the Co-citation measurement is more appropriate for the following reasons. The Co-citation network is dynamic and it can change over time, and the bibliographic coupling will be fixed once both papers had been written. According to the 'Law of large number', we will have stronger indication if two paper are similar according to the co-citation changes.

Simulation from python (Kunlun Zhu)

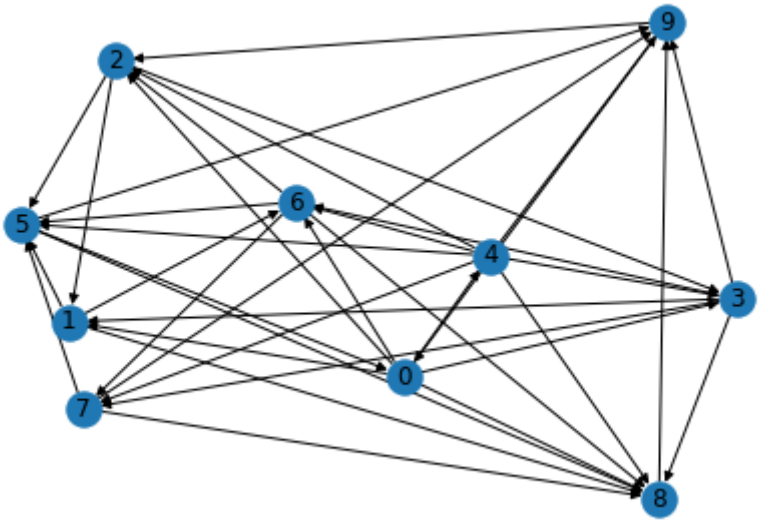
In [1]:
```
#!/usr/bin/python3.9
# -*- coding: utf-8 -*-
#author Kunlun Zhu 2022/6/23
#setup the python environment
import numpy as np
import networkx as nx
import matplotlib.pyplot as plt
```

In [2]:
```
# Let's suppose a 10 nodes network
A = np.random.randint(2, size=(10,10)) #Adjacency matrix
for index, a in np.ndenumerate(A):
    if a == 1:
        A[index[1], index[0]] = 0 # To make sure A is a one-way directed matrix

link_list = []
for index, a in np.ndenumerate(A):
    if a == 1:
        link_list.append((index[0], index[1]))

print('Adjacency Matrix')
print(A)
print('link_list:')
print(link_list)
```

```
Adjacency Matrix
[[0 1 1 1 1 0 1 0 1 0]
 [0 0 0 0 0 1 1 0 1 0]
 [0 1 0 1 0 1 0 0 0 0]
 [0 1 0 0 0 0 1 1 1 1]
 [0 0 1 1 0 1 1 1 1 1]
 [1 0 0 0 0 0 0 0 1 1]
 [0 0 1 0 0 1 0 1 1 0]
 [0 0 0 0 0 1 0 0 1 1]
 [0 0 0 0 0 0 0 0 0 1]
 [1 0 1 0 0 0 0 0 0 0]]
link_list:
[(0, 1), (0, 2), (0, 3), (0, 4), (0, 6), (0, 8), (1, 5), (1, 6), (1, 8), (2, 1), (2,
3), (2, 5), (3, 1), (3, 6), (3, 7), (3, 8), (3, 9), (4, 2), (4, 3), (4, 5), (4, 6),
(4, 7), (4, 8), (4, 9), (5, 0), (5, 8), (5, 9), (6, 2), (6, 5), (6, 7), (6, 8), (7,
5), (7, 8), (7, 9), (8, 9), (9, 0), (9, 2)]
```

In [3]:
```
G = nx.DiGraph()
G.add_nodes_from(range(10))
G.add_edges_from(link_list)
nx.draw(G, with_labels=True)
print('Citation Network')
plt.show()
```

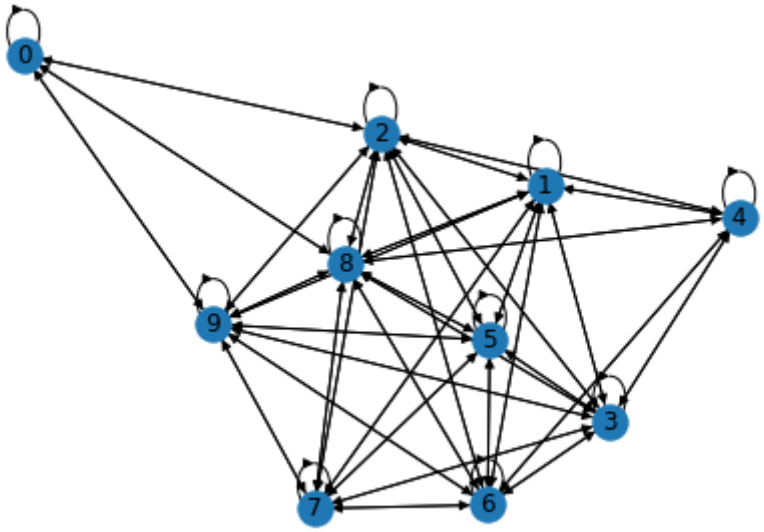
```
Citation Network
```



In [4]:

```
A_t = A.transpose()
Co = np.matmul(A_t, A)
Cocitation_G = nx.DiGraph()
Cocitation_G.add_nodes_from(range(10))
link_list_c = []
for index, c in np.ndenumerate(Co):
    if c >= 1:
        for k in range(c):
            link_list_c.append((index[0], index[1]))
Cocitation_G.add_edges_from(link_list_c)
nx.draw(Cocitation_G, with_labels=True)
print('Cocitation_matrix')
print(Co)
print('Co-citation Network')
plt.show(Cocitation_G)
```

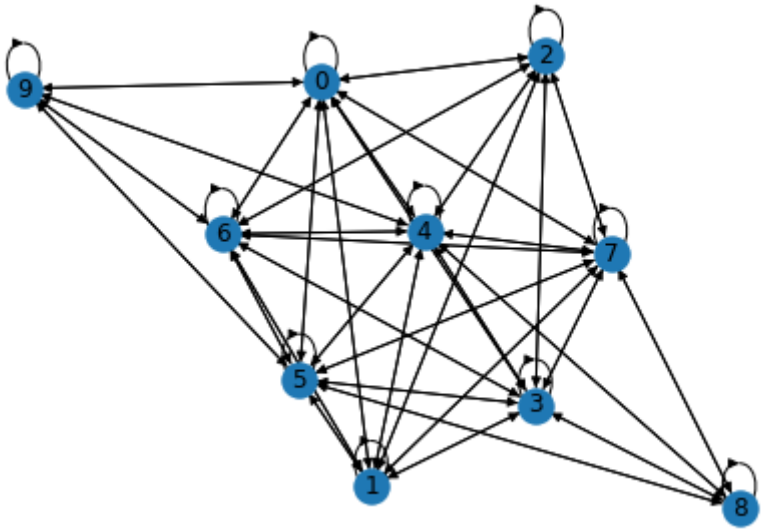
```
Cocitation_matrix
[[2 0 1 0 0 0 0 0 1 1]
 [0 3 1 2 1 1 2 1 2 1]
 [1 1 4 2 1 2 2 2 3 1]
 [0 2 2 3 1 2 2 1 2 1]
 [0 1 1 1 1 0 1 0 1 0]
 [0 1 2 2 0 5 2 2 4 2]
 [0 2 2 2 1 2 4 2 4 2]
 [0 1 2 1 0 2 2 3 3 2]
 [1 2 3 2 1 4 4 3 7 4]
 [1 1 1 1 0 2 2 2 4 5]]
Co-citation Network
```



In [5]:

```
A_t = A.transpose()
Bi = np.matmul(A, A_t)
Bibliographic_G = nx.DiGraph()
Bibliographic_G.add_nodes_from(range(10))
link_list_b = []
for index, b in np.ndenumerate(Bi):
    if b >= 1:
        for k in range(b):
            link_list_b.append((index[0], index[1]))
Bibliographic_G.add_edges_from(link_list_b)
nx.draw(Bibliographic_G, with_labels=True)
print('Bibliographic_matrix')
print(Bi)
print('Bibliographic Coupling Network')
plt.show(Bibliographic_G)
```

```
Bibliographic_matrix
[[6 2 2 3 4 1 2 1 0 1]
 [2 3 1 2 3 1 2 2 0 0]
 [2 1 3 1 2 0 1 1 0 0]
 [3 2 1 5 4 2 2 2 1 0]
 [4 3 2 4 7 2 4 3 1 1]
 [1 1 0 2 2 3 1 2 1 1]
 [2 2 1 2 4 1 4 2 0 1]
 [1 2 1 2 3 2 2 3 1 0]
 [0 0 0 1 1 1 0 1 1 0]
 [1 0 0 0 1 1 1 0 0 2]]
Bibliographic Coupling Network
```


9
0
2
6
4
7
5
3
1
8