coauthorship

January 23, 2020

[11]:

```
import matplotlib.pyplot as plt
import networkx as nx
import numpy as np
from fetcher3 import fetch_links
from networkx.algorithms.cluster import average_clustering, transitivity,␣
 ↪triangles
from networkx.algorithms.shortest_paths.generic import␣
 ↪average_shortest_path_length as avg_diameter
from networkx.algorithms.distance_measures import diameter as max_diameter
```

[2]:

```
# Create an undirected graph from the data

G = nx.Graph()

with open('gr_qc_coauthorships.txt') as edges:
    for edge in edges:
        G.add_edge(*edge.split())
```

Problem 2a

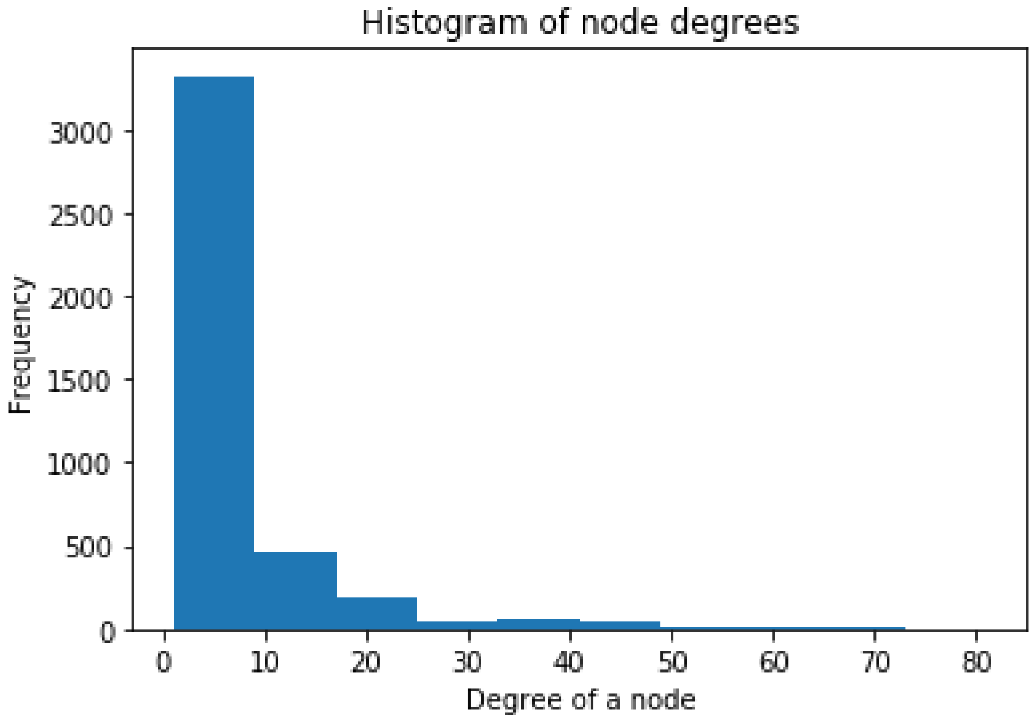
[14]:

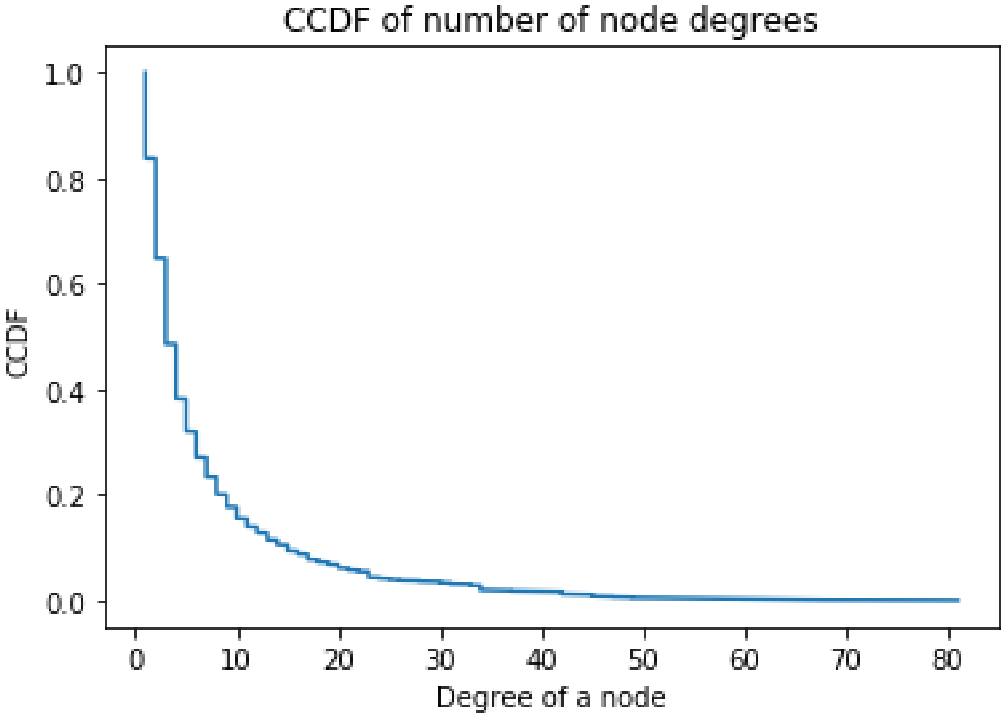
```
# Plot histogram
degrees = list(dict(G.degree).values())
plt.hist(degrees)
plt.title('Histogram of node degrees')
plt.xlabel('Degree of a node')
plt.ylabel('Frequency')
plt.show()

# Plot ccdf
plt.plot(np.sort(degrees), np.linspace(1, 0, len(degrees), endpoint=False))
plt.title('CCDF of number of node degrees')
plt.xlabel('Degree of a node')
plt.ylabel('CCDF')
plt.show()
```

```python
# Compute clustering coefficients
print('Average clustering coefficient = ', average_clustering(G))
print('Overall clustering coefficient = ', transitivity(G))

# Compute diameters
print('Average diamater = ', avg_diameter(G))
print('Maximal diamater = ', max_diameter(G))
```





```
Average clustering coefficient =  0.5568782161697919
Overall clustering coefficient =  0.6288944756689877
Average diamater =  6.049380016182999
Maximal diamater =  17
```

Problem 2b

[18]:

```python
T = sum(dict(triangles(G)).values()) / 3
print('T = ', T)

# We can use E[T] formula computed in problem 3a
n = len(G.nodes)
p = (6 * T / (n * (n - 1) * (n - 2))) ** (1 / 3)
print('p = ', p)
```

```
T =  47779.0
p =  0.015861688593415416
```

Problem 2c

The distribution of the node degrees of a Erdos-Renyi graph is Binomial(n - 1, p). The Erdos-Renyi model is not a good model for this graph, as the histogram of the node degrees in part a suggests that the degree distribution is heavy-tailed. We do not need the histogram to conclude this; we know that an Erdos-Renyi graph has that all edges form independently from each other, resulting in low clustering coefficients, while the coauthorship is most likely clustered, since the probability of authors a and c coauthoring is not independent from authors a and b coauthoring and b and c coauthering. This is especially because a paper is often coauthored by k > 2 authors, which creates a completely connected subgraph on k nodes, within which any triplet forms a triangle.