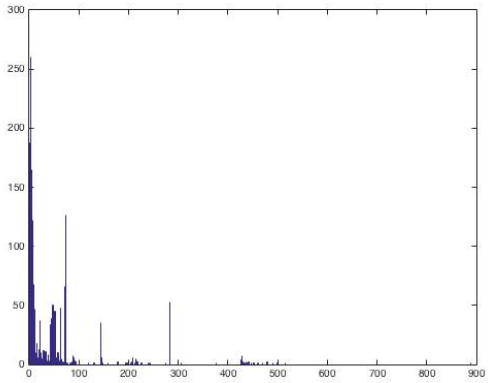
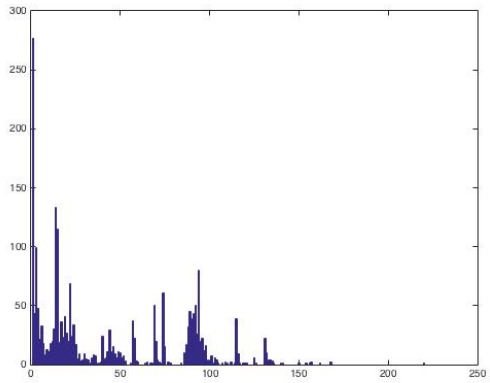
Problem 3:

1. Code in crawler.py, calculate.py and combine.py
2. I choose the breadth-first searching policy. The advantage of this searching policy is that it will not lead to a endless entry in one particular node, which as its name, it will first go through the node in same level, and then go dive to a deeper level. This will prevent the crawler from fetching all 2000 in one same root.
3. Two histograms.

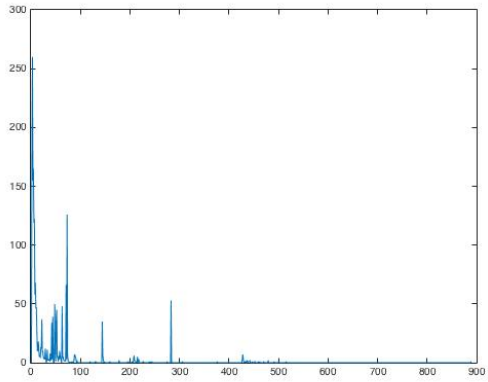


Numbers of hyperlinks which point to each page

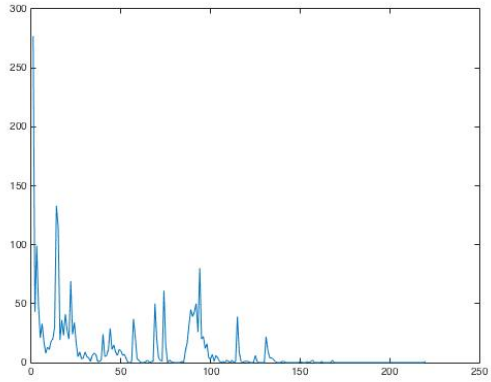


Numbers of hyperlinks per page

4. Two ccdf functions.



Numbers of hyperlinks which point to each page



Numbers of hyperlinks per page

5. The average clustering coefficient of the graph is: 0.7289
   The overall clustering coefficient of the graph is: 0.4688

6. The average diameter of the graph is: 2.522
   The maximal diameter of the graph is 4

7. The comparison between collaboration network and the web graph:
   In collaboration network, degree distribution mainly concentrates at the mean value while in web graph, the degree distribution seems more randomly. There exist many pages which have large degree values indicating that this page has many links with other links. In comparison, in collaboration network, due to the natural property of people, it is rarely possible to many such a person who has many “links” with others.

   When considering the clustering coefficient, we find that both average and overall clustering coefficient are very similar.

   When considering the diameter, diameter in web graph is much smaller than the diameter in collaboration network. This makes sense, because web is much more connected for its convenience and fast properties. While collaboration network do not need such property, and collaboration between researchers are more complex and may be stymied by various reasons: country, subject, language and etc.