The Caltech Graph has n = 2000 nodes and m = 124130 edges.
It was generated using the following selection policy:

Rules:
- Only visit pages within the Caltech domain.
- Visit each page only once.
- Ignore pages that are not HTML (ignoring multimedia/data files as well).
- Ignore the parameters in dynamic URLs to visit each dynamic page only once.
- Removes nodes that fail to be crawled due to interruptions/errors.
- Use breadth-first search algorithm to prioritize between urls.

Pros:
- Finds pages closer to the starting URL.
- Keeps diameter of the graph small via BFS.
- Saves time by visiting each page only once.

Cons:
- Potentially skips unique content of dynamic web pages.
- Misses any links within multimedia/data files.
- Misses any links that are not in the HTML source code.
- Misses any new or changed links that are added after each page is visited.

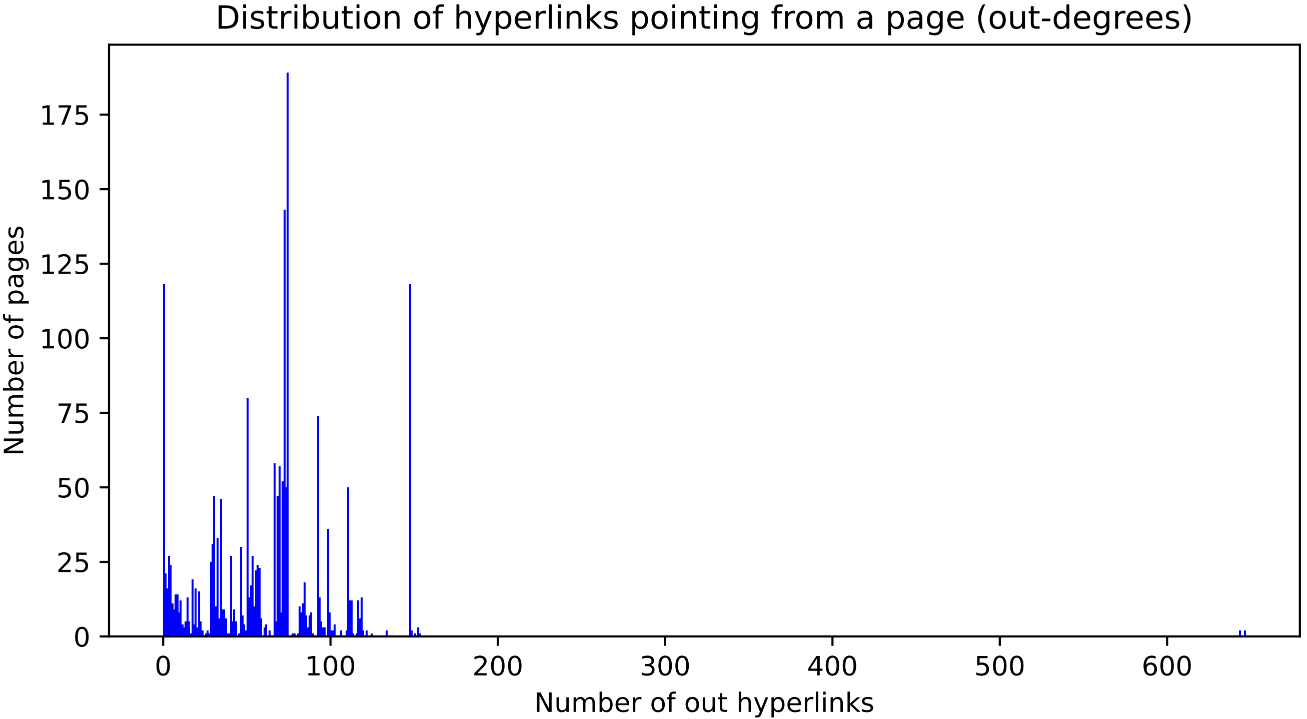
Clustering and Diameter analysis (treating the graph as undirected)

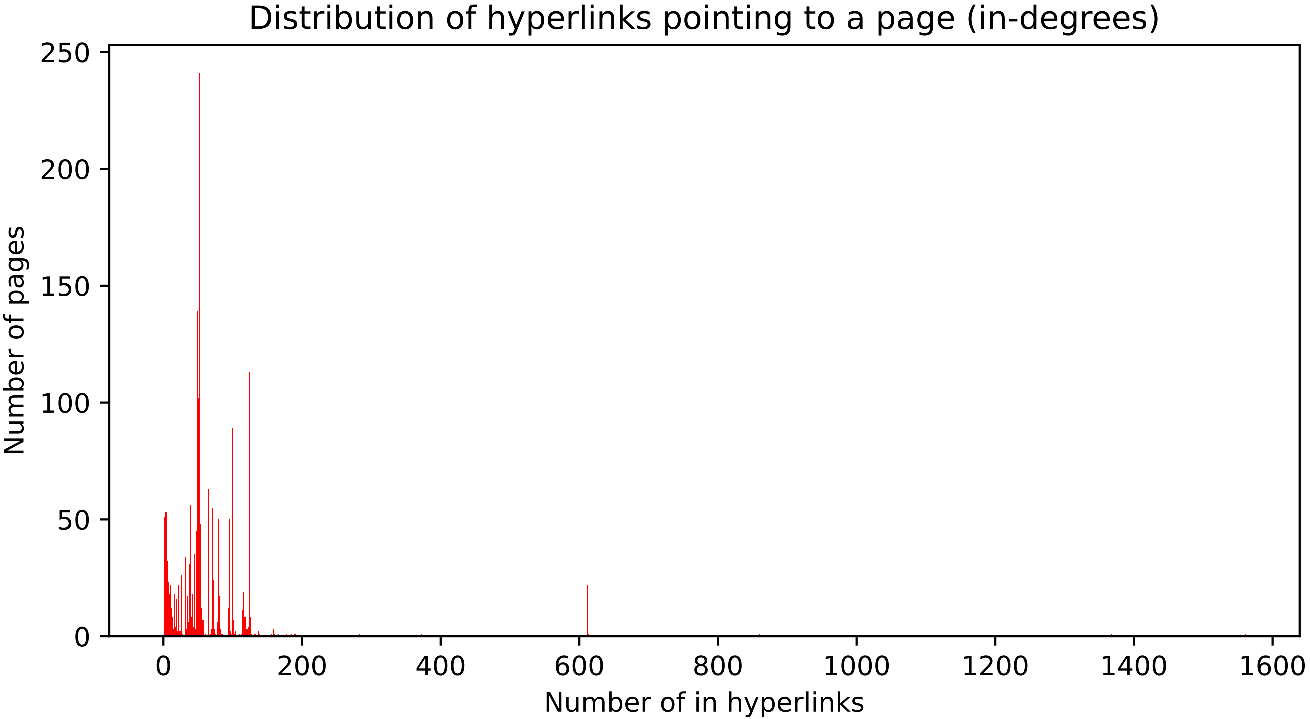
Global clustering coefficient: 0.4857047474903301

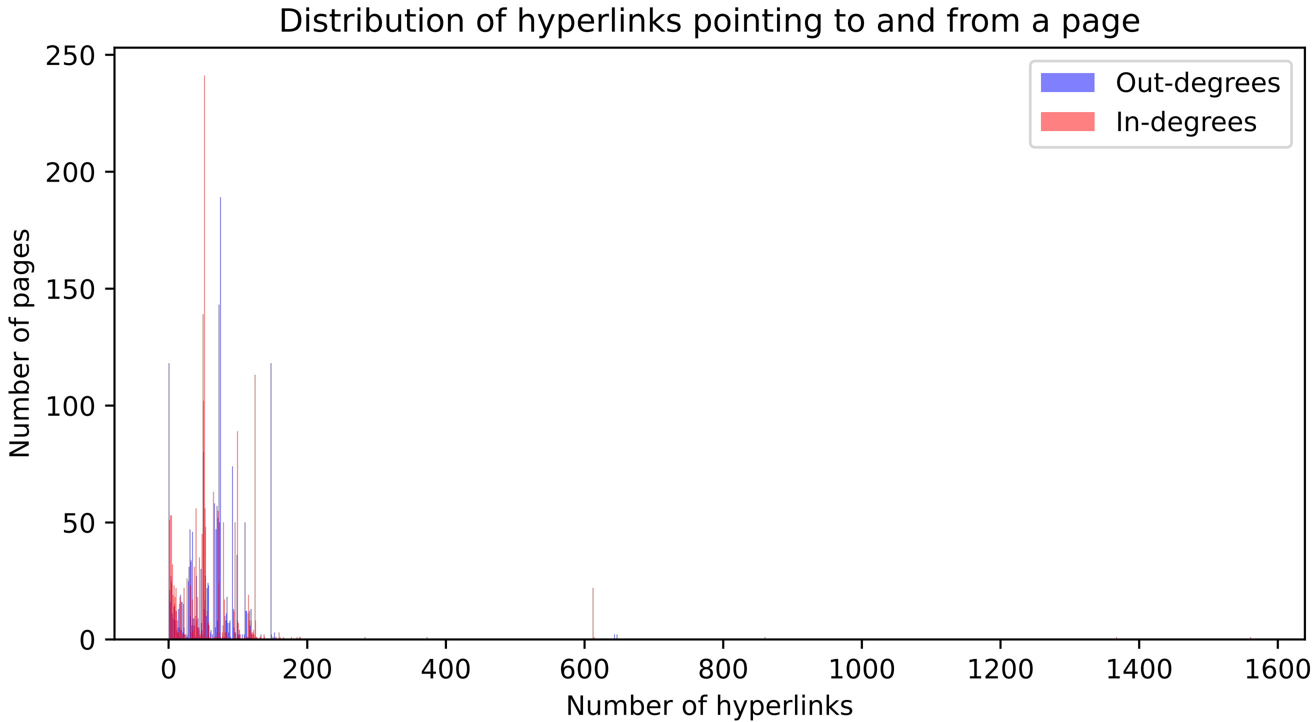
Average clustering coefficient: 0.898791351073389

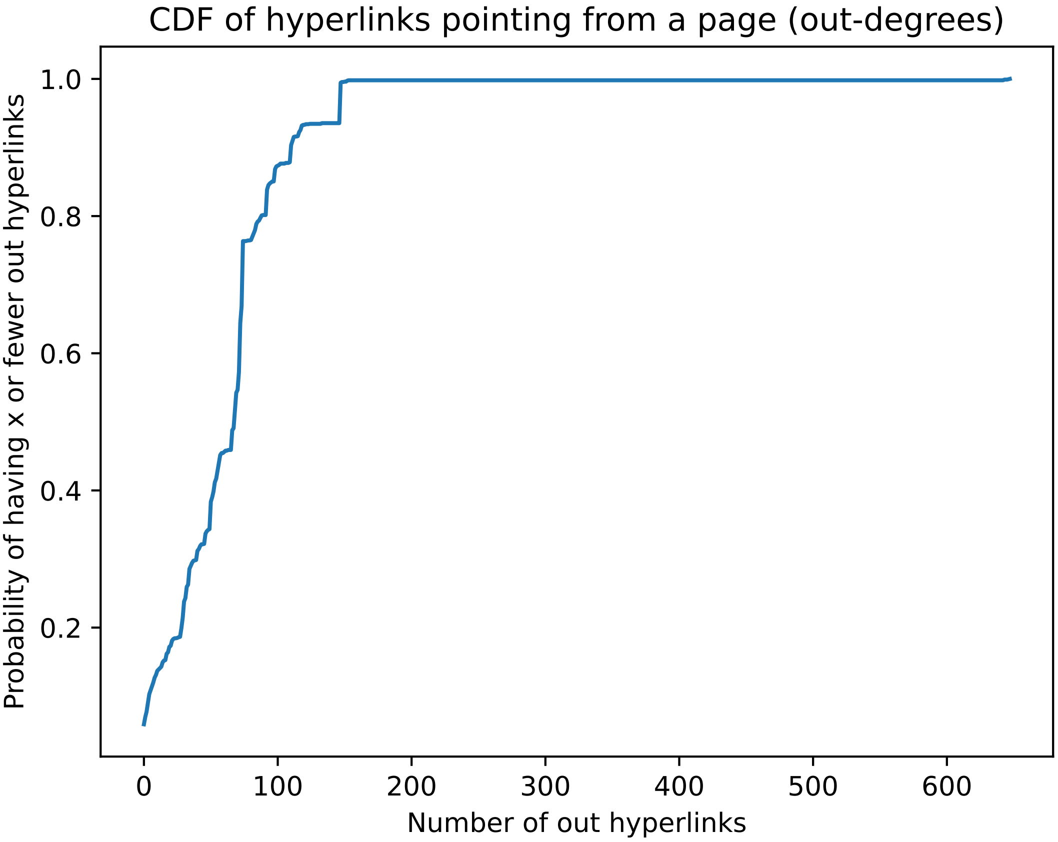
Maximum diameter: 4

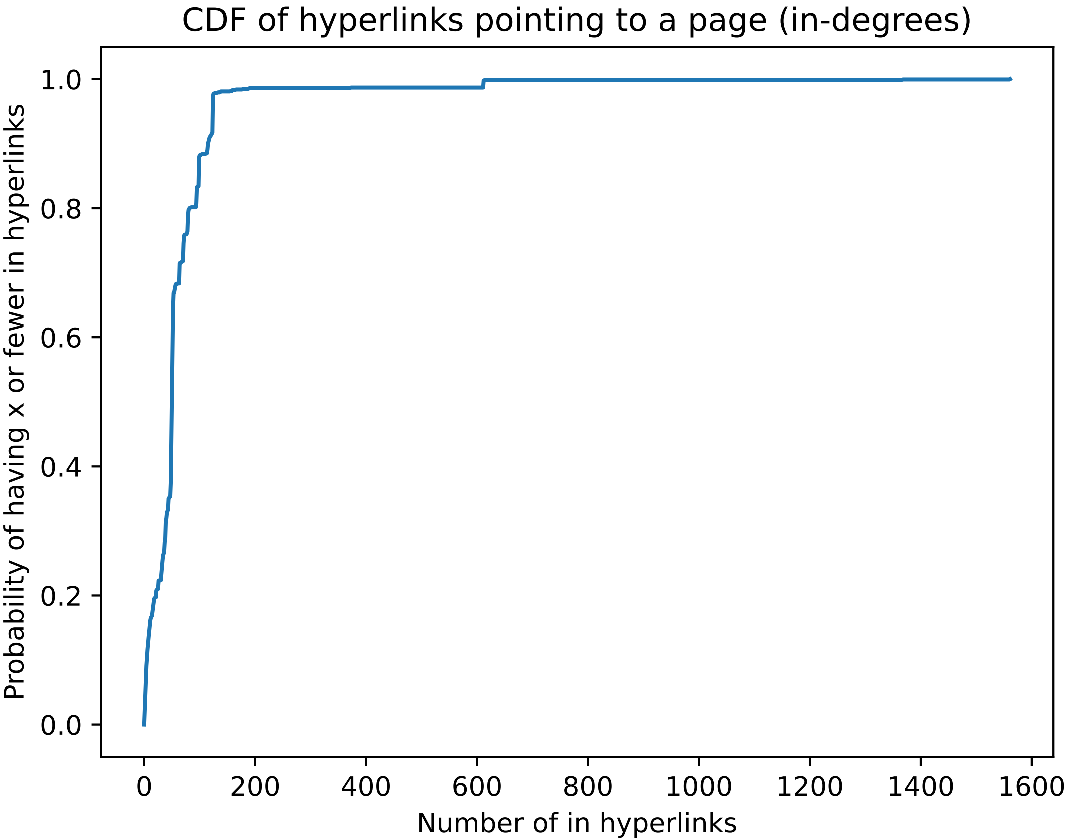
Average diameter: 2.257902451225613


Distribution of hyperlinks pointing from a page (out-degrees)
Number of pages
Number of out hyperlinks
0
25
50
75
100
125
150
175
0
100
200
300
400
500
600

Distribution of hyperlinks pointing to a page (in-degrees)
Number of pages
250
200
150
100
50
0
0
200
400
600
800
1000
1200
1400
1600
Number of in hyperlinks


Distribution of hyperlinks pointing to and from a page
Out-degrees
In-degrees
Number of pages
0
50
100
150
200
250
Number of hyperlinks
0
200
400
600
800
1000
1200
1400
1600


CDF of hyperlinks pointing from a page (out-degrees)
Probability of having x or fewer out hyperlinks
1.0
0.8
0.6
0.4
0.2
0
100
200
300
400
500
600
Number of out hyperlinks


CDF of hyperlinks pointing to a page (in-degrees)
Probability of having x or fewer in hyperlinks
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
1200
1400
1600
Number of in hyperlinks