Network Analysis Assignment - LLLL76

QUESTION ONE

A Group Graph is defined by four parameters m, k, p and q and is constructed as follows:

- Create m*k vertices. The vertices are partitioned into m groups each of size k.
- For each pair of vertices that belong to the same group, add an edge between them with probability p.
- For each pair of vertices that belong to different groups, add an edge between them with probability q.

We will choose arbitrary m and k to test the structure change of p and q, we choose values of p and q such that p + q = 0.5, p > q. The values of m and k will then be changed so as to cover three different test criteria over all values of p and q, they are then compared to the baseline. The three test criteria for values of m and k are such that m < k, m = k, and m > k.

In theory at higher values of p and lower values of q the structure of the graph will be groups of nodes that are highly interconnected with significantly fewer connections between these groups, at lower values of p and higher values of q as p and q come together we get closer to seeing a random graph with the probability of an edge being present inside a group and between two groups being similar, if the probability was exactly the same then we would see a random graph with a set probability of any two nodes being connected irrespective of what group they are in. The values of m and k are unlikely to change all that much with regards to the connections that we see as part of the structure to the graph. If we take the two extremes of our testing range, high p, low q and then lower p, higher q, in the instance of a high p, low q for m > k we will see a high number of small highly interconnected groups with few 'between group' edges. For the same values of m and k but with a lower p and higher q value we will see a structure that closely resembles a random graph with slightly more interconnects inside each group. In the final case of m < k for the two extremes of our p and q values, at high p and low q we would expect to see large highly connected groups with little in the way of inter group edges. Then for lower p and higher q we see the same trend towards a random graph but still with the same higher number of edges intra group as p != q.

In the first round of testing we want m and k to be equal, m = 20, k = 20, p = [0.45, 0.4, 0.35, 0.3], and q = [0.05, 0.1, 0.15, 0.2]. The degree distribution is shown in figure 0.1.

Second round of testing, m > k, m = 20, k = 5, p = [0.45, 0.4, 0.35, 0.3], and q = [0.05, 0.1, 0.15, 0.2]. The degree distribution is shown in figure 0.2.

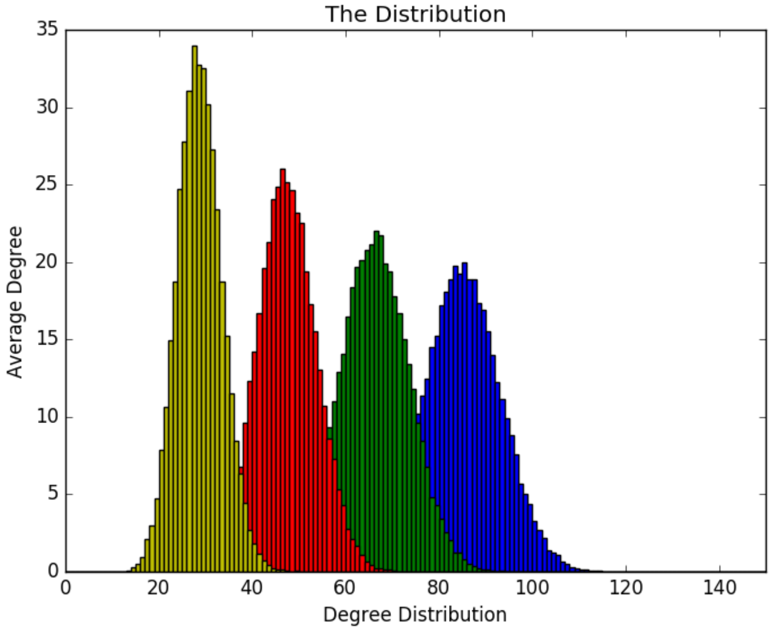
Third round of testing, m < k, m = 5, k = 20, p = [0.45, 0.4, 0.35, 0.3], and q = [0.05, 0.1, 0.15, 0.2]. The degree distribution is shown in figure 0.3.

DIAMETER

Using the values m = 50, k = 5, q = 0.1, run an average diameter over 100 graphs with variable p values gives the graph figure 0.5.

As the graph shows the average over all values of p is around 3 and it will converge to 3 for higher values of p. At higher P values there is a higher probability that there is an intra-group connection, whilst the Q value is fixed, meaning that there is a set probability of an inter-group connection. With higher values of probability for intra-group connections this will reduce the possible diameter due to

Figure 0.1: M = 20, K = 20, with [P,Q] varying: [0.3,0.2], [0.35,0.15], [0.4,0.1], [0.45,0.05] respectively.



the fact that if each group was strongly connected then the average path between any pair of nodes would be two that were in separate groups but unconnected, thus the travel to any node in the start group that is connected to any node in the target group which would then be connected to the target node, there will be a few exceptions such as two nodes in two groups that have no nodes connected, this is why the average for p = 1 is slightly above 3. Whilst we can see an obvious trend in the graph, due to the parameters chosen this is highly magnified, with most other values that were tested we see a much flatter result, shown in figure 0.6.

Question Two

A k-cycle is a set of k vertices 1, 2, . . . , k such that 1 is joined to 2, 2 is joined to 3, . . ., and k is joined to 1 (it does not matter whether or not other vertices are adjacent). For k = 4 and k = 5, find out for a random sample of the vertices, how many k-cycles each vertex belongs to and plot the distribution for the co-authorship, random, PA, and group graphs. All graphs are undirected.

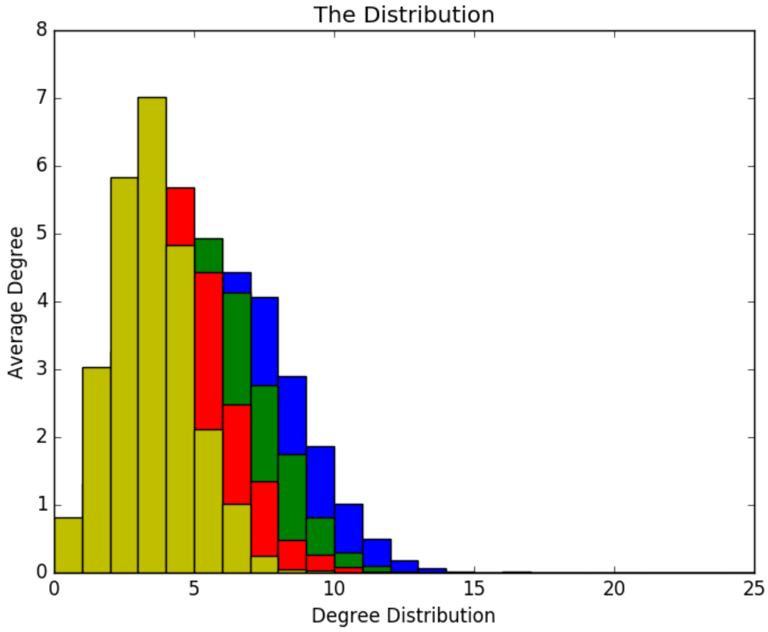
Four Cycles

As shown in figure 0.7 the co-authorship graph any given node could be in a significant range of 4-cycles, from 10 to 1000000. As shown from the graph none of the other graphs make an accurate model of the co-authorship graph.

The group graph, with parameters m = 40, k = 40, p = 0.5, and q = 0.02, shows a fairly condensed range of values, this would be due to the limited amount of inter-group connections that occur through a low q value. This would result in limiting the spread as there would be very few cycles that would be inter-group resulting in a cap being present in the number that can occur within a group.

The PA graph, with parameters nodes = 1600 and out degree = 35, shows two separate groups, with the higher being much smaller, this could be due to the much smaller collection of fully connected nodes that make up the core of the PA graph. The fully connected sub-graph would be present in many more 4-cycles than any of the other nodes that are added after this code group is created. This is due to the nature of a fully connected graph.

Figure 0.2: M = 5, K = 5, with [P,Q] varying: [0.3,0.2], [0.35,0.15], [0.4,0.1], [0.45,0.05] respectively.



Finally the random graph, parameters nodes = 1600 and probability = 0.02, shows a fairly small range of possible values, and is very similar to the group graph is size and range. This could be due to the similarity that the q value of the group graph and the probability of the random graph share.

Five Cycles

Figure 0.8 shows the number of 5-cycles that are present in each graph, this is very similar to the 4-cycle distribution, however there are obviously fewer due to the higher number of nodes required to make up a 5-cycle. The same parameters were used in each of the graphs and thus they show a very similar distribution for the same reasons as the 4-cycle distributions.

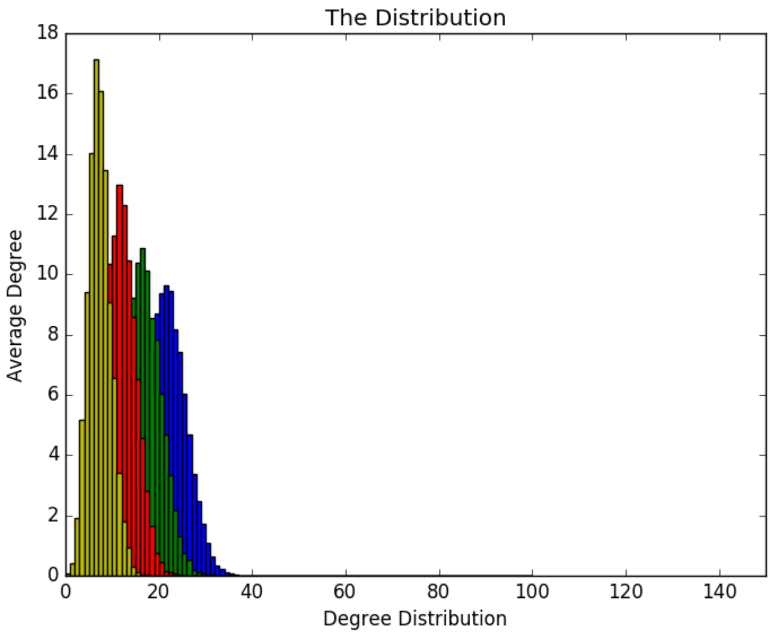
Question Three

Given a start and end vertex in a given graph find the search time, that being the average number of 'steps' or edges between two vertices in the given graph. For random, PA, and group graphs describe a search strategy, at each step using only local information which vertex to move to next. Explain why the strategy might be effective and implement and test it. Plot search time against the number of instances that achieve that time.

Random

By their definition random graphs are random and due to this the search strategy will incorporate a certain degree of randomisation. The logic here is to move around the graph randomly until a node is reached that is adjacent to the end node. A random graph means that we cannot infer any details about the structure of the graph, meaning we cannot add any more specialisation to this method.

Figure 0.3: M = 5, K = 20, with [P,Q] varying: [0.3,0.2], [0.35,0.15], [0.4,0.1], [0.45,0.05] respectively.



```
At start node;
while not at end node do
    if end node in neighbours[current node] then
        go to end node;
    else
        choose random unvisited neighbouring node;
        got to this node;
    end
end
```

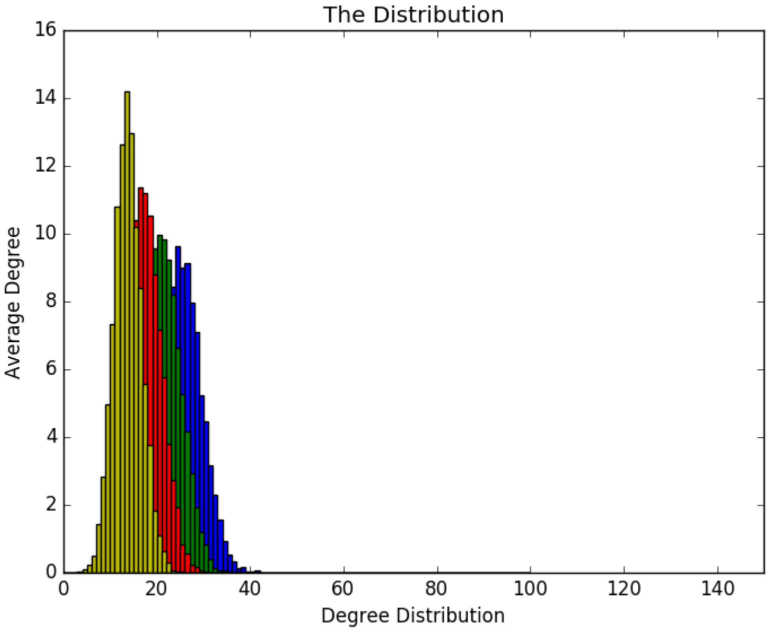
Algorithm 1: Random Strategy

For a random graph with 100 nodes and a connectivity probability of 0.05, as shown in figure 0.9, the most frequent search time is one. The average search time is 10.2.

Preferential Attachment

A PA graph is a fully connected graph that then has more nodes added to it. This means that the fully connected 'heart' of the graph is more likely to, firstly have more connections as well as having any 'extra nodes' connected to it. This will make up the core of the search strategy. The main focus of this search strategy is to get to the fully connected sub-graph which has a higher probability of being connected to the end node. If this fails then we visit nodes outside of this sub-graph and hope that they are connected to the end node. If we can avoid going to a node we have already been to then we will but if there is not other options then we will have to.

Figure 0.4: M = 20, K = 5, with [P,Q] varying: [0.3,0.2], [0.35,0.15], [0.4,0.1], [0.45,0.05] respectively.

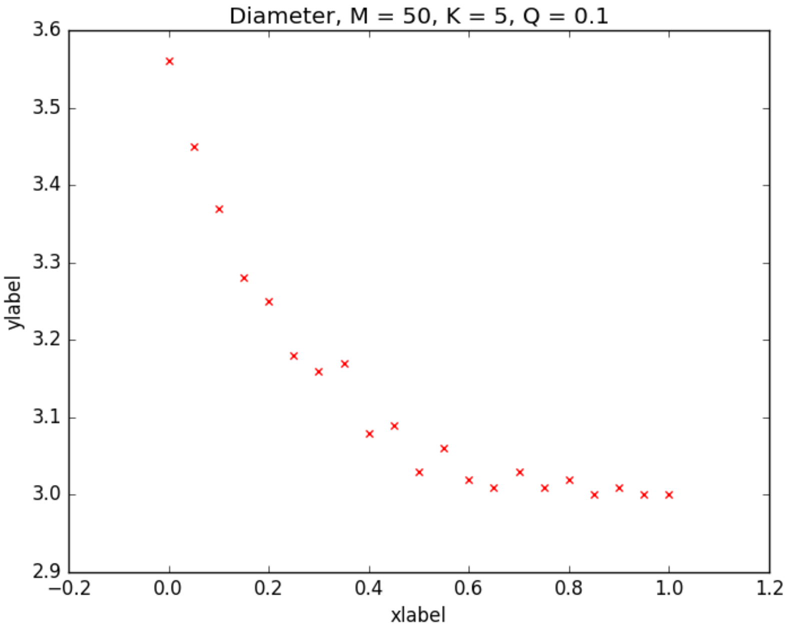


```
At start node;
while not at end node do
    if end node in neighbours[current node] then
        go to end node;
    else
        switch unvisted nodes in complete subgraph do
            choose one at random;
            go to it;
        end
        case unvisited nodes outside complete subgraph do
            choose one at random;
            go to it;
        end
        case visited nodes in complete subgraph do
            choose a random one;
            go to it;
        end
        otherwise do
            choose a random neighbour;
            go to it;
        end
    end
end
```

Algorithm 2: PA Strategy

For a PA graph with 100 nodes and put degree of 15, the most frequent search time is two and the average search time is 2.27.

Figure 0.5:

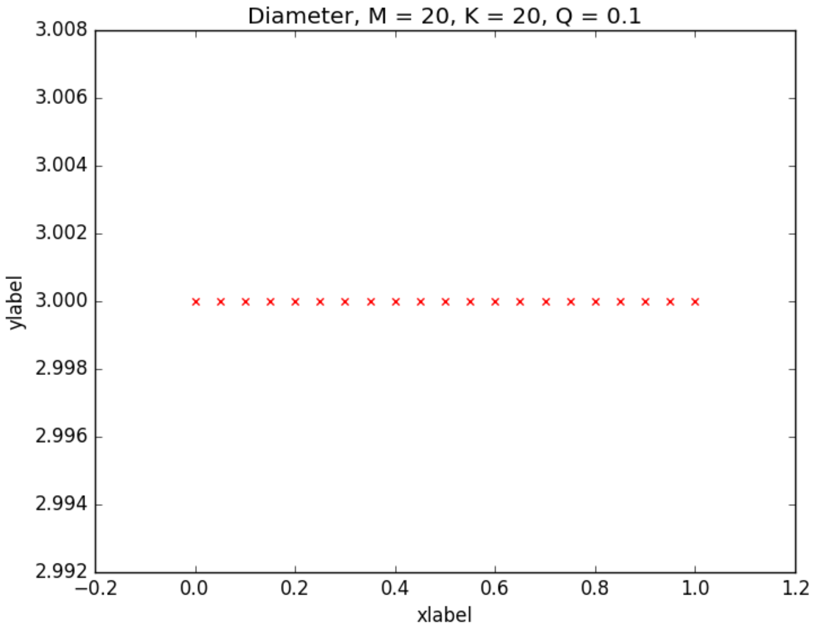


Group

The group graph takes in four parameters to decide is structure, m, the number of groups in the graph, k, the size of each group, the graph has m * k nodes, p, the probability of two nodes in the same group have an edge between them, q, the probability of two nodes not in the same group have an edge between them.

We will take into consideration the predetermined structure of the graph in the search strategy. For this search algorithm the parameters that are used to create the graph are known and thus we can have a more specialised search algorithm as we can infer from the parameters an insight into the structure of the graph. For a high P value it can be infered that there is a high probability of intra-group edges, meaning that it would be optimal to aim for the group that the end node is in as there is a high probability it is connected to any other node in its group. For a high Q value it can be infered that there is a high probability that there are inter-group edges and low probability that there exist intra-group edges so aiming for the group of the end node will be of very little help, so this will be avoided. The final case is where P and Q have very similar values and therefore the structure of the graph is similar to that of a random graph, thus we re-use the random graph algorithm.

Figure 0.6:



```
At start node;
switch High p value do
    while not at end node do
        if end node in neighbours[current node] then
            go to end node;
        else
            if any nodes in neighbours[current node] and in end group then
                choose one at random;
                go to it;
            else
                choose a random neighbour;
                go to it;
            end
        end
    end
end
case High q value do
    while not at end node do
        if end node in neighbours[current node] then
            go to end node;
        else
            if any nodes in neighbours[current node] but not in end group then
                choose one at random;
                go to it;
            else
                choose a random neighbour;
                go to it;
            end
        end
    end
end
otherwise do
    while not at end node do
        if end node in neighbours[current node] then
```

Figure 0.7:

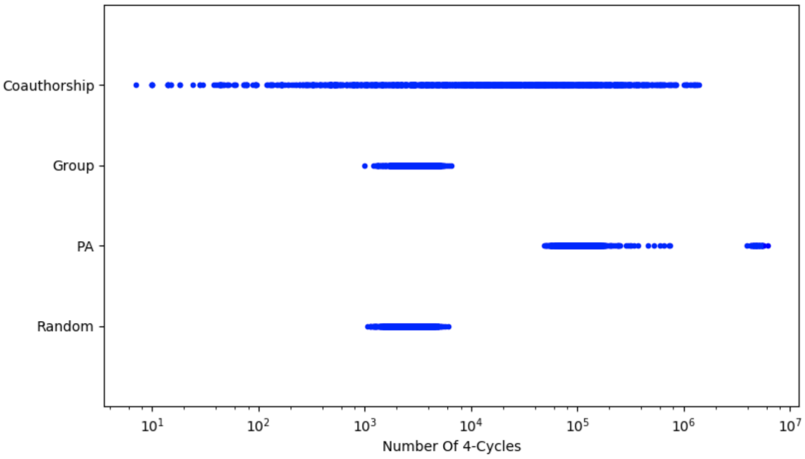
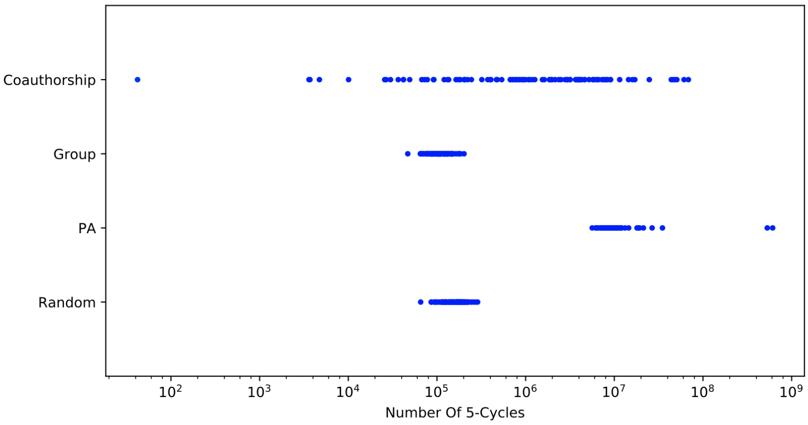


Figure 0.8:



For a group graph with m = 40, k = 40, p = 0.4, and q = 0.1, as shown in figure 0.11 the most frequent search time was one with the average search time being 5.2.

Figure 0.9:

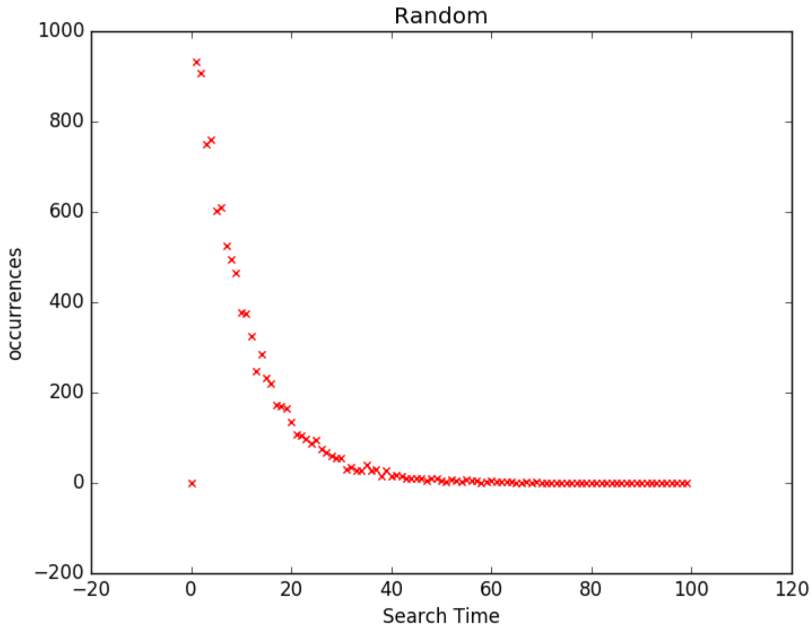


Figure 0.10:

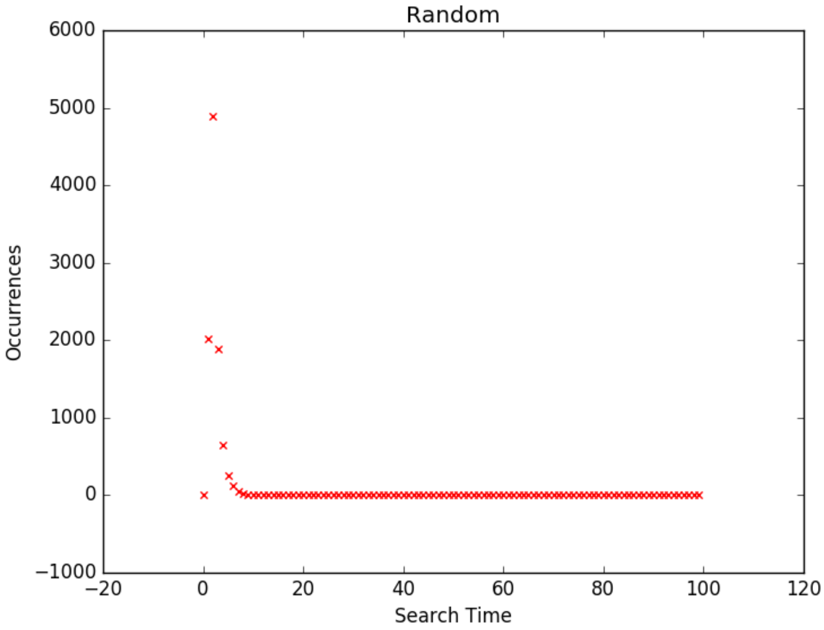
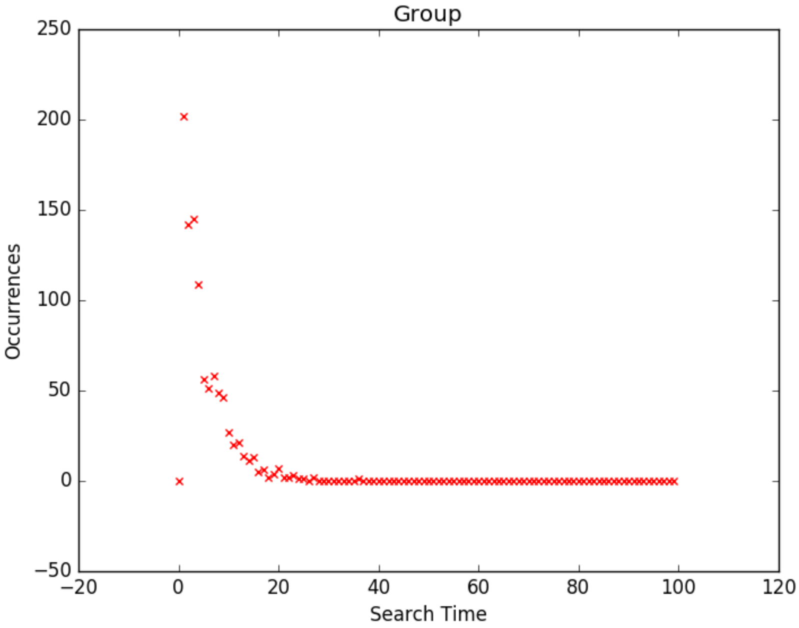


Figure 0.11:


Group
Occurrences
Search Time
250
200
150
100
50
0
−50
−20
0
20
40
60
80
100
120