Typical Distance between Two Randomly Selected Vertices of a Erdős-Rényi Binomial Random Graph : A Simulation Study

Aditya Ghosh, Sayak Chatterjee

Abstract

Consider the Erdős-Rényi binomial random graph model. We studied through simulations the typical graph distance between two randomly selected vertices, when p is above the connectivity threshold or in the sparse but super-critical regime. We also studied the typical distance in square lattice percolation where p is constant and greater than $1/2$.

1 Introduction

Consider the Erdős-Rényi binomial random graph model, which we shall denote by ER(n, p). The object of our study is the *typical distance* in ER(n, p), which is defined as the graph-distance between any two randomly selected vertex. We shall denote the typical distance in ER(n, p) by H_n, following the notation in van der Hofstad [2].

Suppose that $G \sim$ ER(n, p). If G is connected, the *average distance* of G is defined as the average of all distances $d(u, v)$ for u and v in G. If G is not connected, we define the average distance to be the average of all distances $d(u, v)$ for pairs of u and v both belonging to the same connected component. Clearly, the average distance is the expected value of the typical distance H_n, conditioned upon the event that H_n is finite. Following is a famous result on the average distance of $G \sim$ ER(n, p).

Theorem. (Chung and Lu, [1]) If $np \geq c >1$ for some constant c, then almost surely the average distance of ER(n, p) is $(1 + o(1))(\log n/ \log np)$, provided $(\log n/ \log np)$ goes to infinity as $n \to \infty$.

It is well-known that the graph $G \sim$ ER(n, p) is connected with high probability when p is above the connectivity threshold $\frac{\log n}{n}$. In the sparse but super-critical regime ($p = c/n$ where $c > 1$), the graph has a giant cluster of size $O(n)$ and second largest cluster of size $O(\log n)$. We considered p to be mainly in these two regimes, but also had a look at the case when p is constant.

We simulated ER(n, p) for different choices of n and p and discarded the simulated distances that were infinite, so that we could compare our simulated results with the above theorem. What we found is that the $o(1)$ term in the theorem is quite ambiguous and shows some strange behaviour when we look at the simulations. We also studied the standard deviation of the typical distance and performed tests for normality and symmetry of the distribution of the typical distance.

Finally we also studied the typical distance in the square lattice percolation, which is defined as follows. We consider the square lattice with vertices $\{(i, j) : -n \leq i, j \leq n\}$ and the edges connecting each pair of points that are exactly one unit of distance apart. We join each edge with a constant probability p and hence obtain a random graph. It is known that the connectivity threshold for this graph is $p = 1/2$. We studied the typical distance for this random graph when $p > 1/2$.

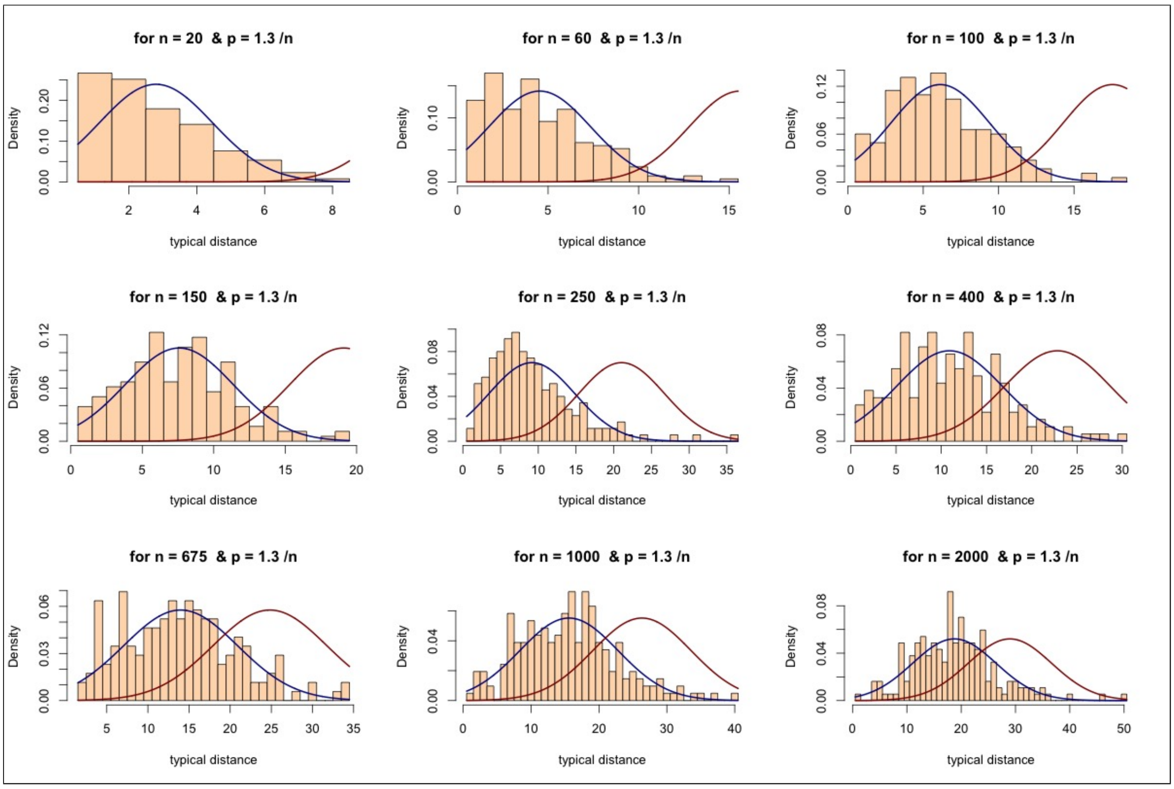
The simulations were done in **R**, mainly using the igraph package. The R code used for these simulations is available in the following Github repository: https://github.com/ghoshadi/random-graphs/.

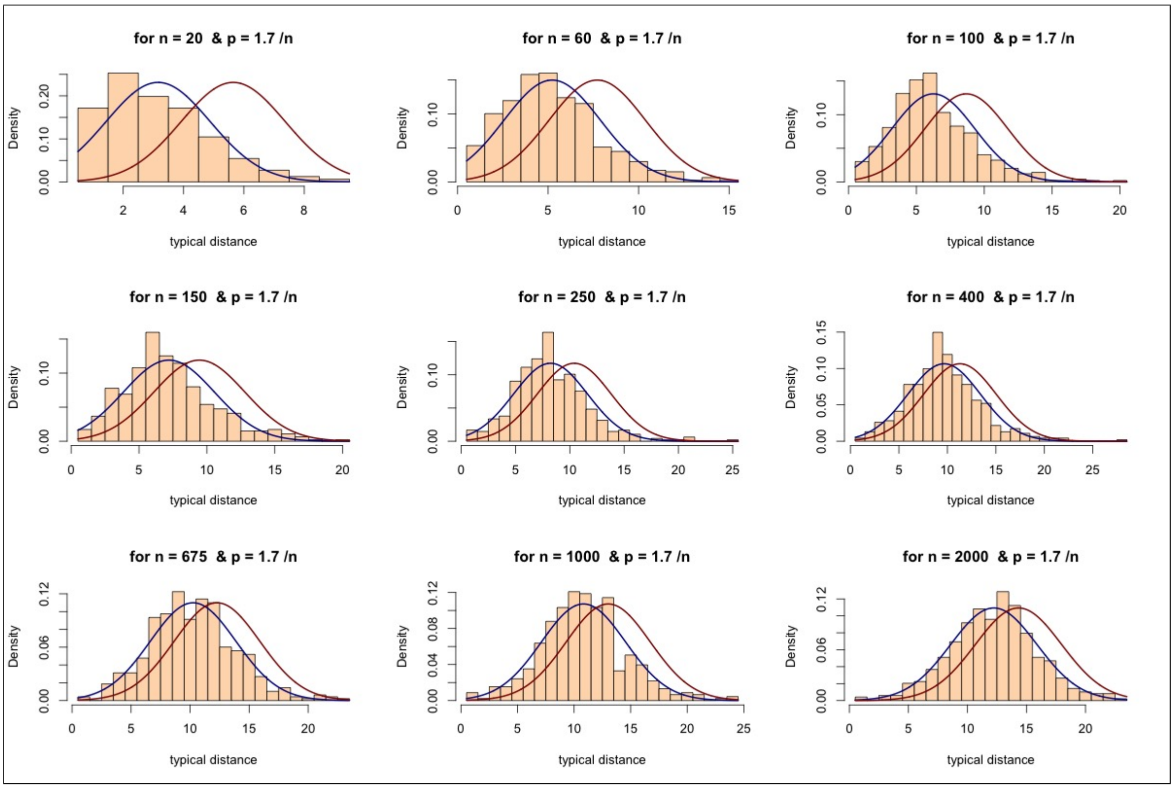
2 Sparse but super-critical regime

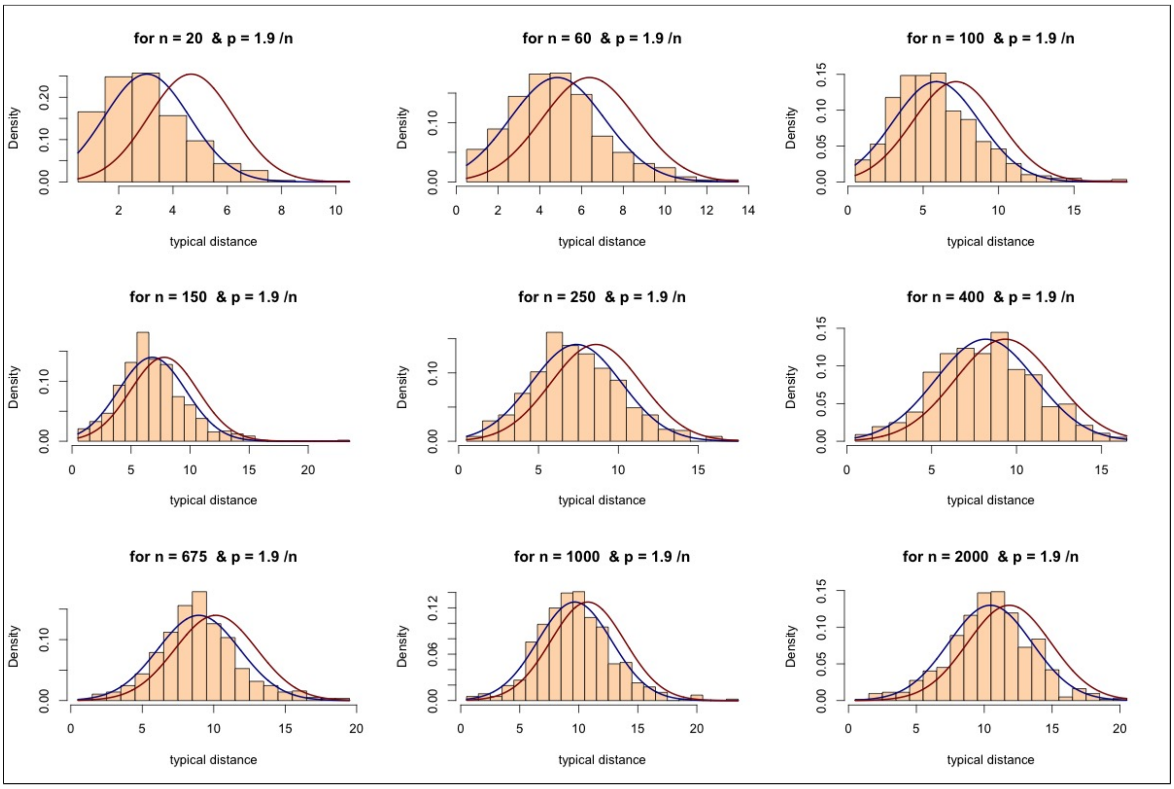
We simulated ER$(n, c/n)$ for 1000 times for each pair of (c, n) where $c = 1.1, 1.3, 1.7, \ldots, 2.5$ and $n = 20, 60, 100, 150, 250, 400, 675, 1000, 2000$. These graph sizes are chosen so that $\log n$ varies almost linearly, which may help us revealing some patterns. Not all distances we observed were finite, but we threw away the infinite ones during our analysis.

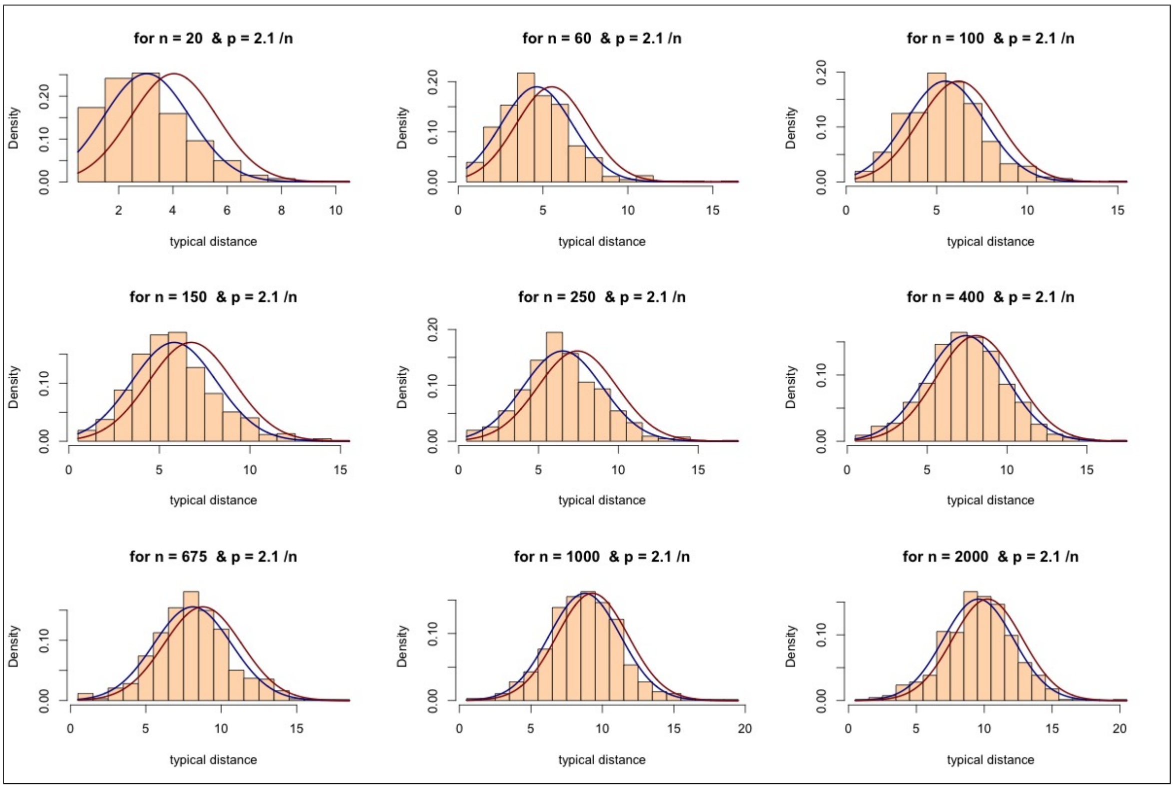
2.1 Histograms

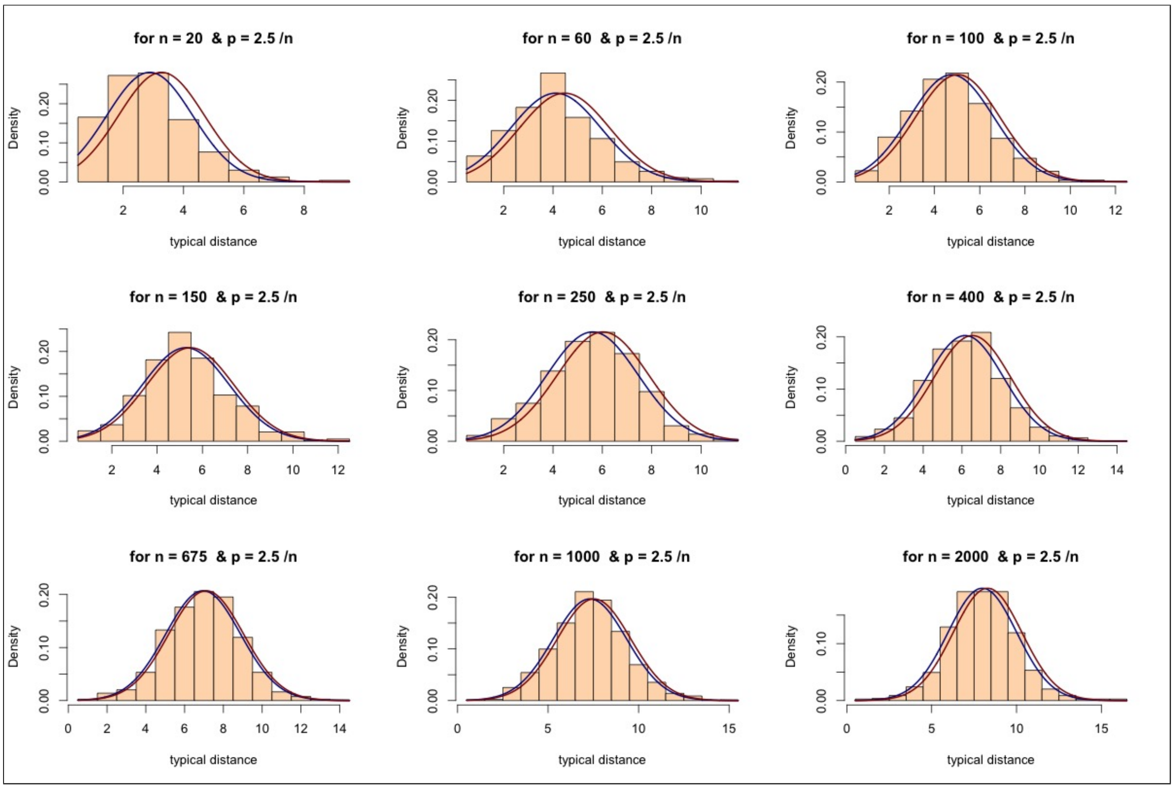
For each of the values $np = c = 1.3, 1.7, 1.9, 2.1, 2.5$, the histograms of the observed (finite) typical distances in ER$(n, c/n)$ for each choice of the graph size n are given below (histograms for the other values of c show similar pattern, those are omitted here). The blue line shows the normal density with mean and s.d. estimated by the sample mean and sample s.d. The red line shows the normal density with mean equal to $\log n/ \log c$ and s.d. estimated by the sample s.d.




for n = 20 & p = 1.7 /n
for n = 60 & p = 1.7 /n
for n = 100 & p = 1.7 /n
for n = 150 & p = 1.7 /n
for n = 250 & p = 1.7 /n
for n = 400 & p = 1.7 /n
for n = 675 & p = 1.7 /n
for n = 1000 & p = 1.7 /n
for n = 2000 & p = 1.7 /n
Density
typical distance


for n = 20 & p = 1.9 /n
for n = 60 & p = 1.9 /n
for n = 100 & p = 1.9 /n
for n = 150 & p = 1.9 /n
for n = 250 & p = 1.9 /n
for n = 400 & p = 1.9 /n
for n = 675 & p = 1.9 /n
for n = 1000 & p = 1.9 /n
for n = 2000 & p = 1.9 /n
Density
typical distance

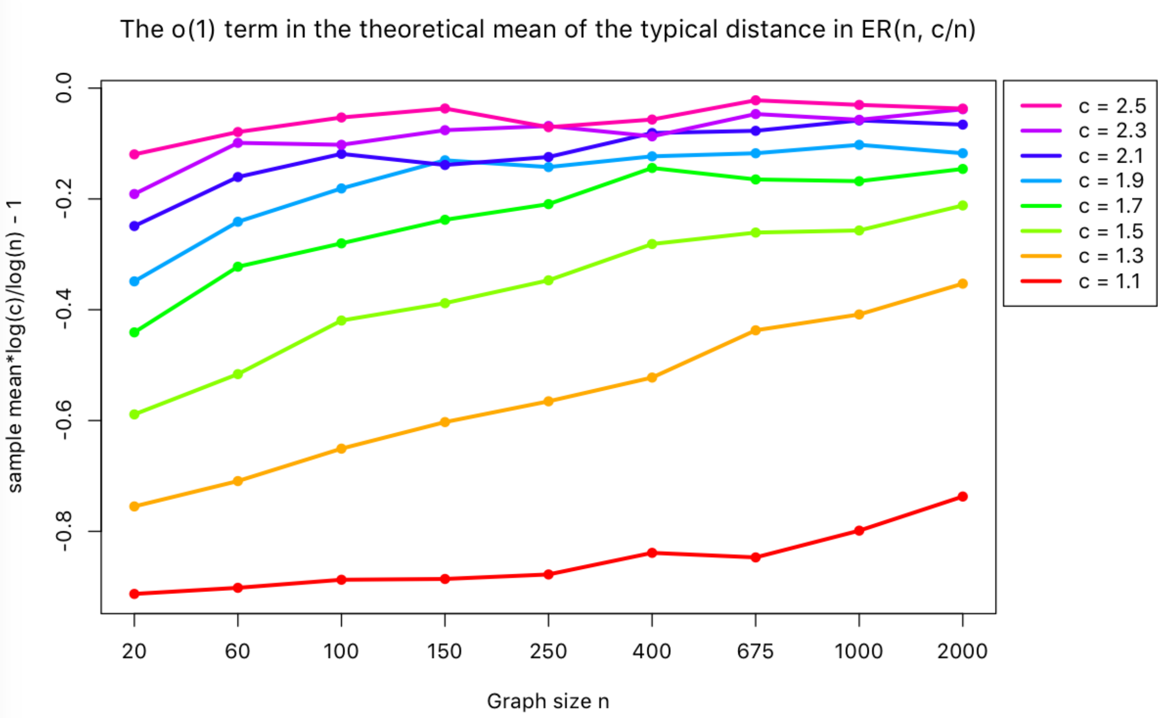




Observations: For the smaller values of c and n we notice that the difference between sample mean and $\log n/\log np$ is so high that the central part of the red curve is missing from the plot. We also see that this difference between sample mean and $\log n/\log np$ decreases gradually as c and n both increases. This difference we observed suggests that we should plot the $o(1)$ term $=$ (the sample mean $-\log n/\log c) \cdot \log c/\log n$ and see how fast does it fall.

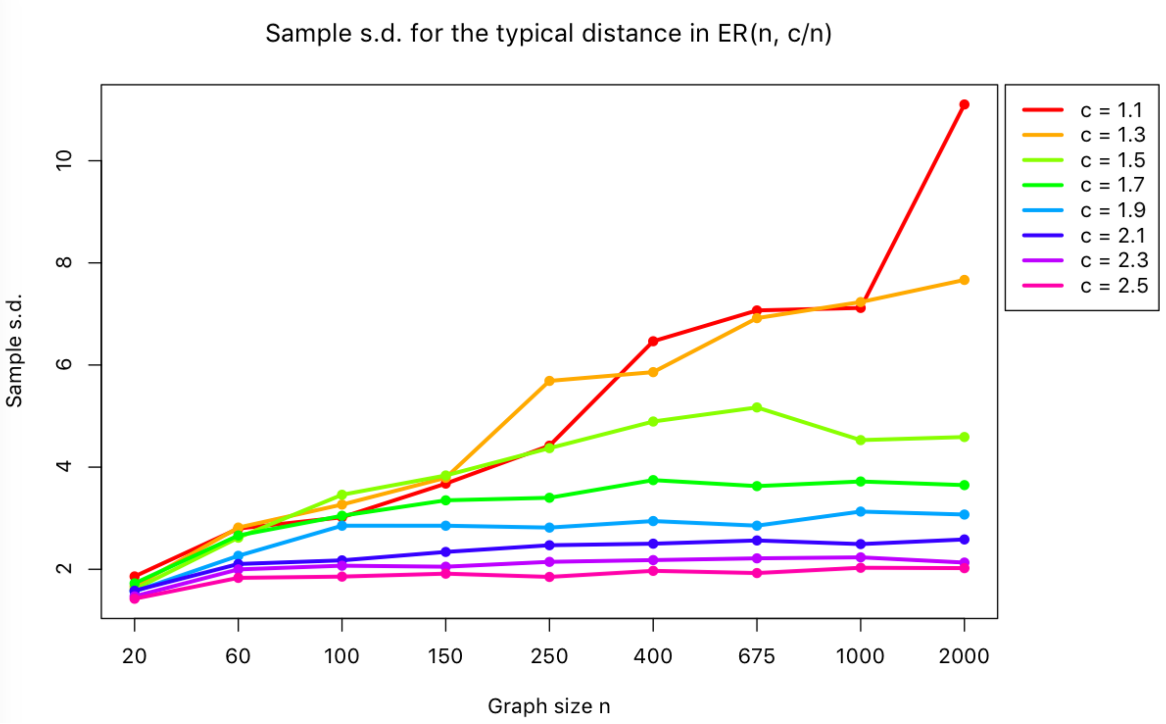
2.2 Studying the $o(1)$ term

Plot of the $o(1)$ term $=$ (the sample mean $\cdot \log c / \log n) - 1$ is given below. Different colors are used for indicating different values of $c = np$.



Observe that the $o(1)$ term falls at a faster rate when the value of c is more than 2. For smaller values of c, the error term is much larger and falls at a slower rate.

2.3 Studying the standard deviations



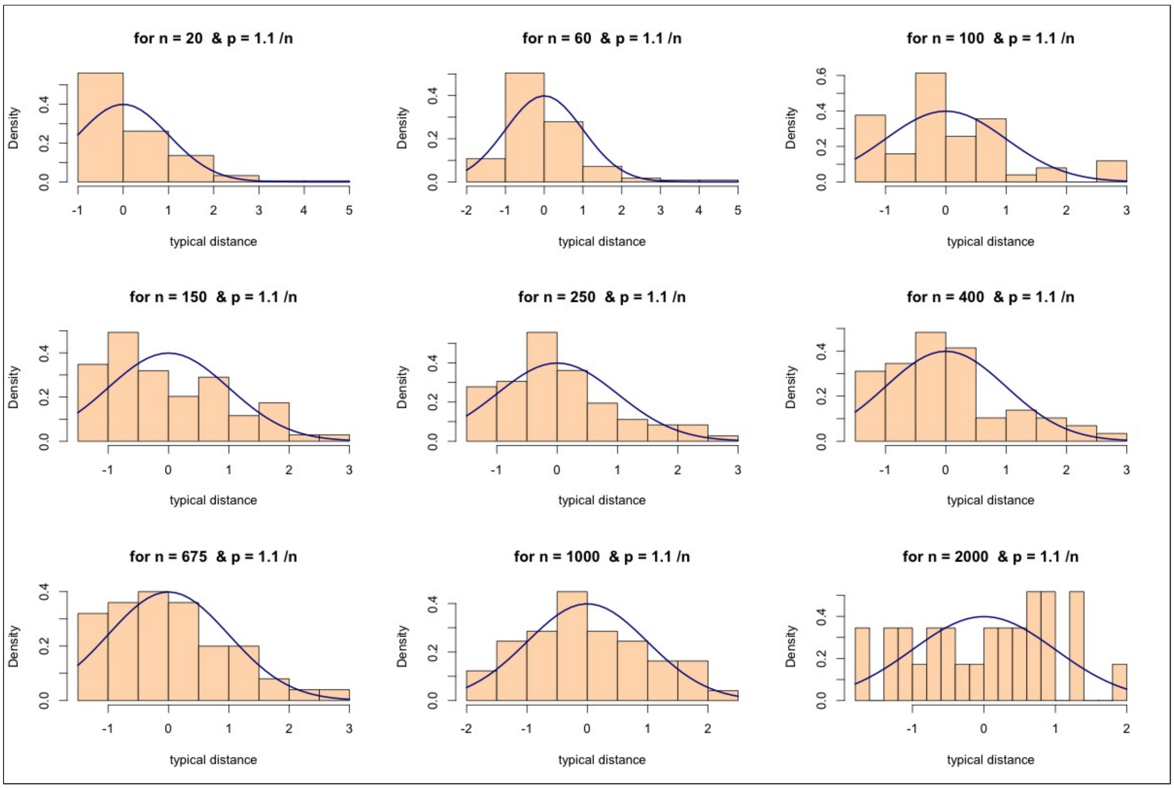
We can see a steady increase for $c = np > 2$, and for the values of c near 1, the s.d. in fact increases rapidly. Note that here the sample s.d. is calculated without any scaling. But even if after scaling the s.d. for smaller values of c will go to zero, the plot suggests that this decay would happen at a much slower rate than the decay for the higher values of c.

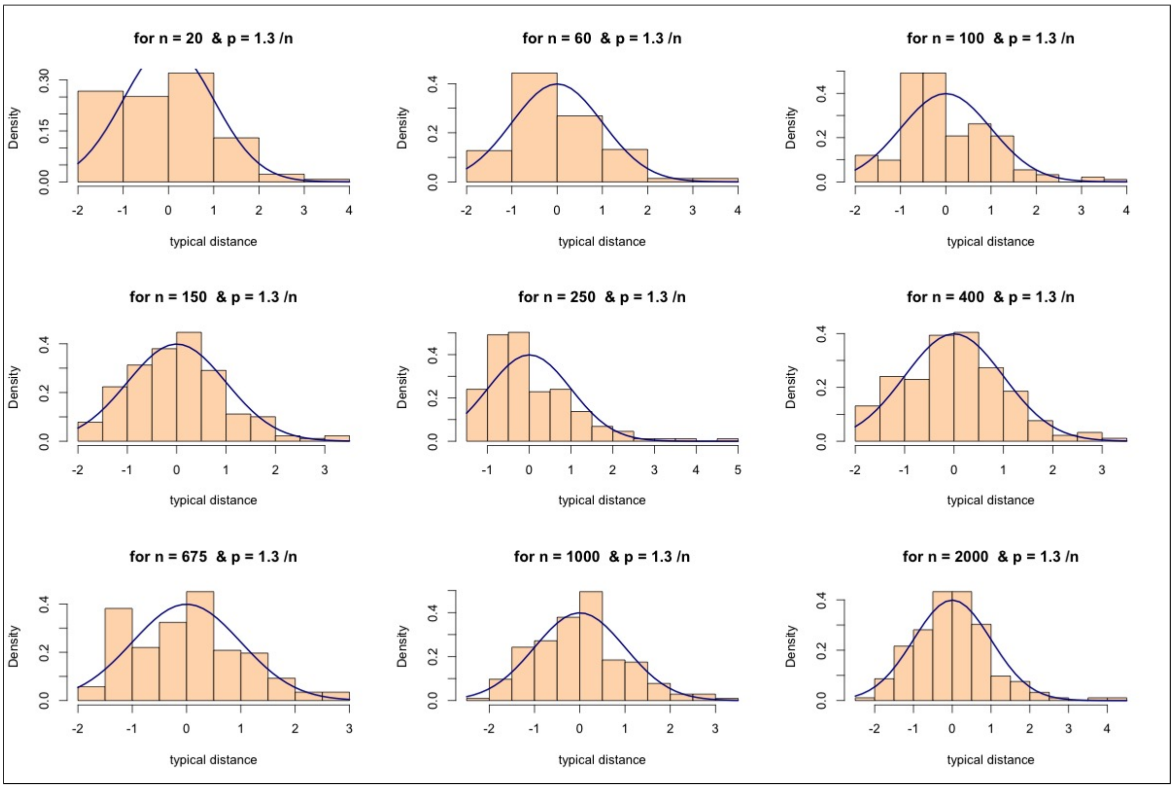
2.4 Testing normality

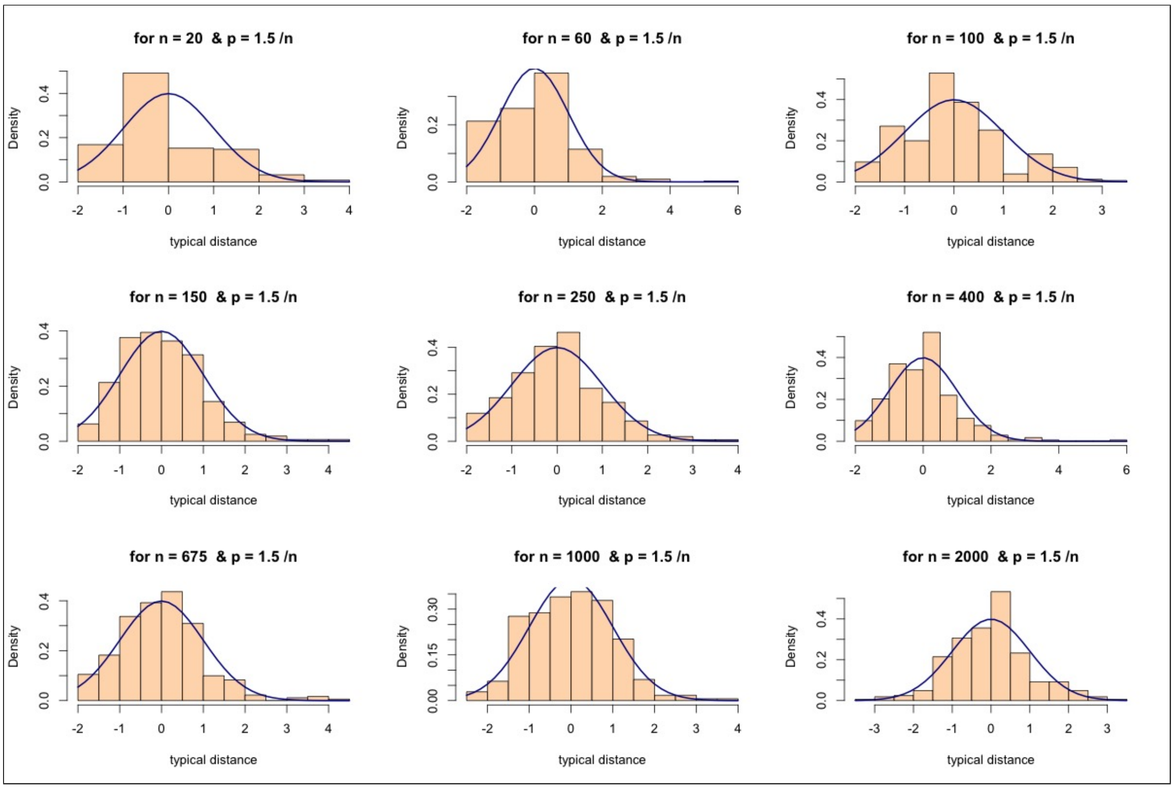
Before the formal tests for normality, we had a look at the histograms for the standardized data and Q-Q plots, for a visual assessment. The histograms for standardized samples are given below, but the Q-Q plots are omitted.

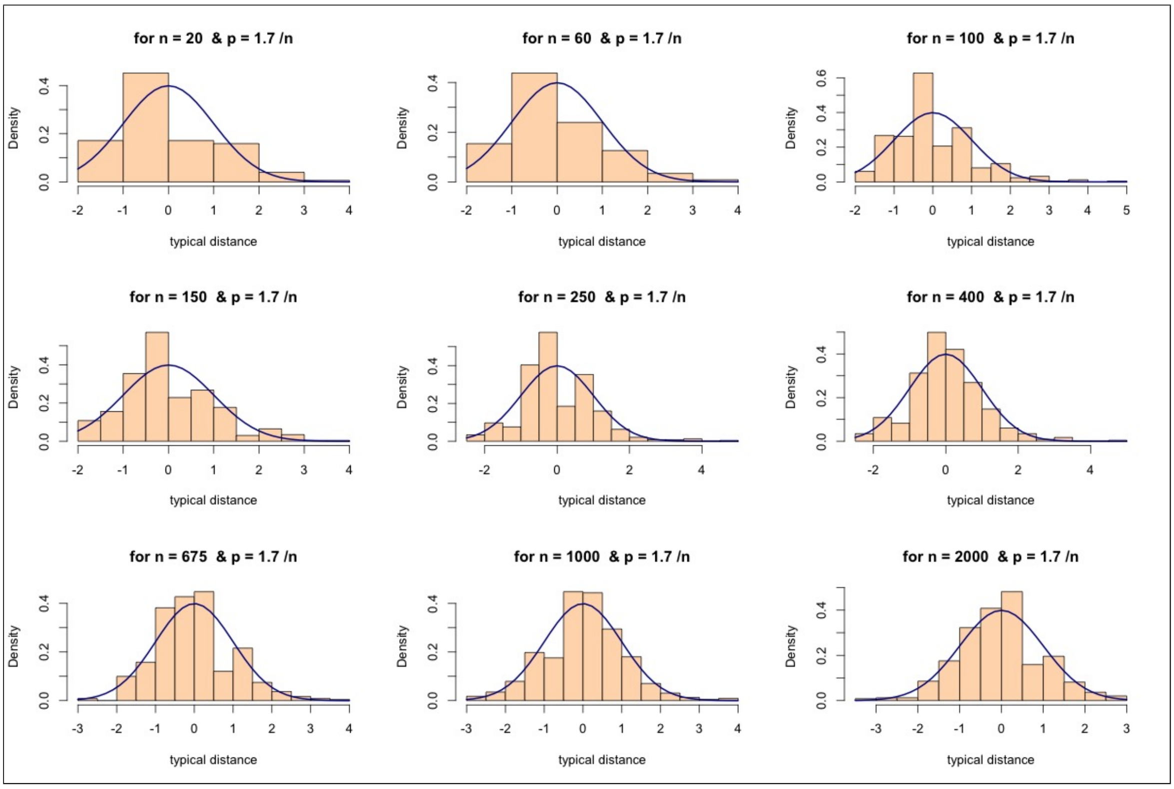
2.4.1 Histogram for the standardized samples

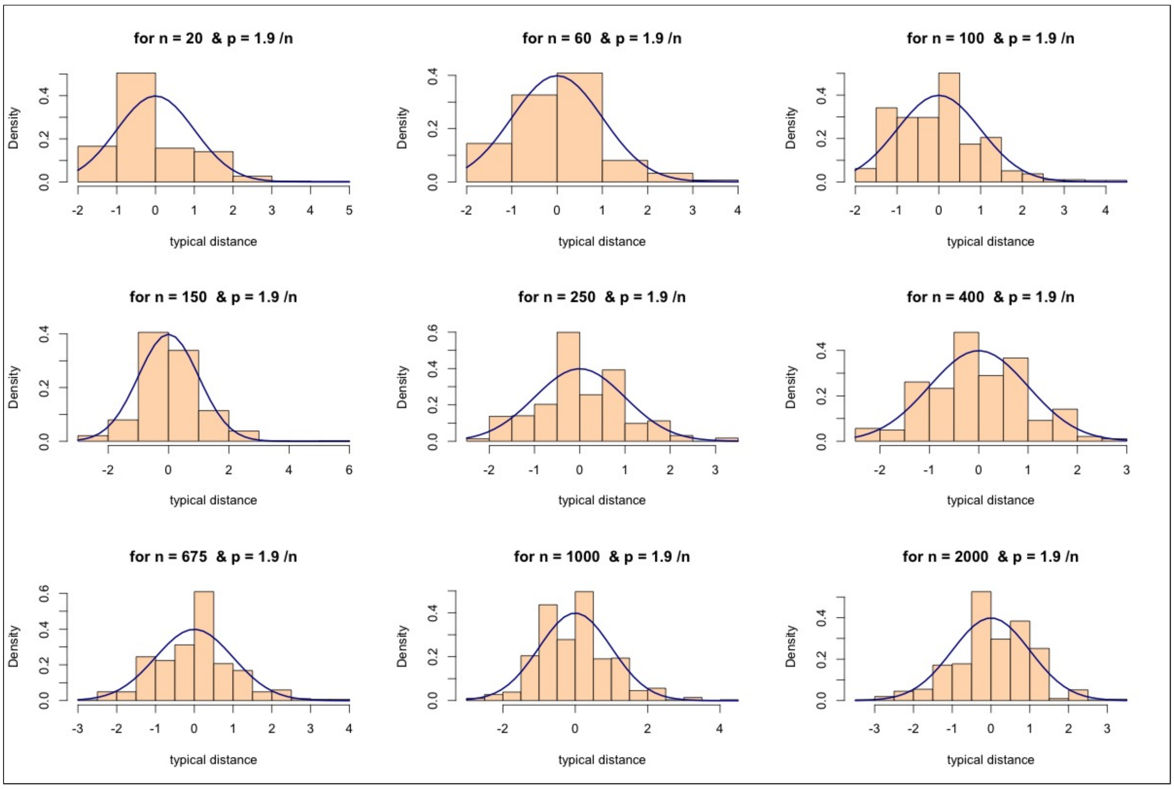
Each sample is standardized using sample mean and s.d. The blue line is the standard normal density. Here we include the histograms for every pair of c and n that we simulated, because that will accompany us while comprehending the p-values of the tests of normality and symmetry which will be performed next.

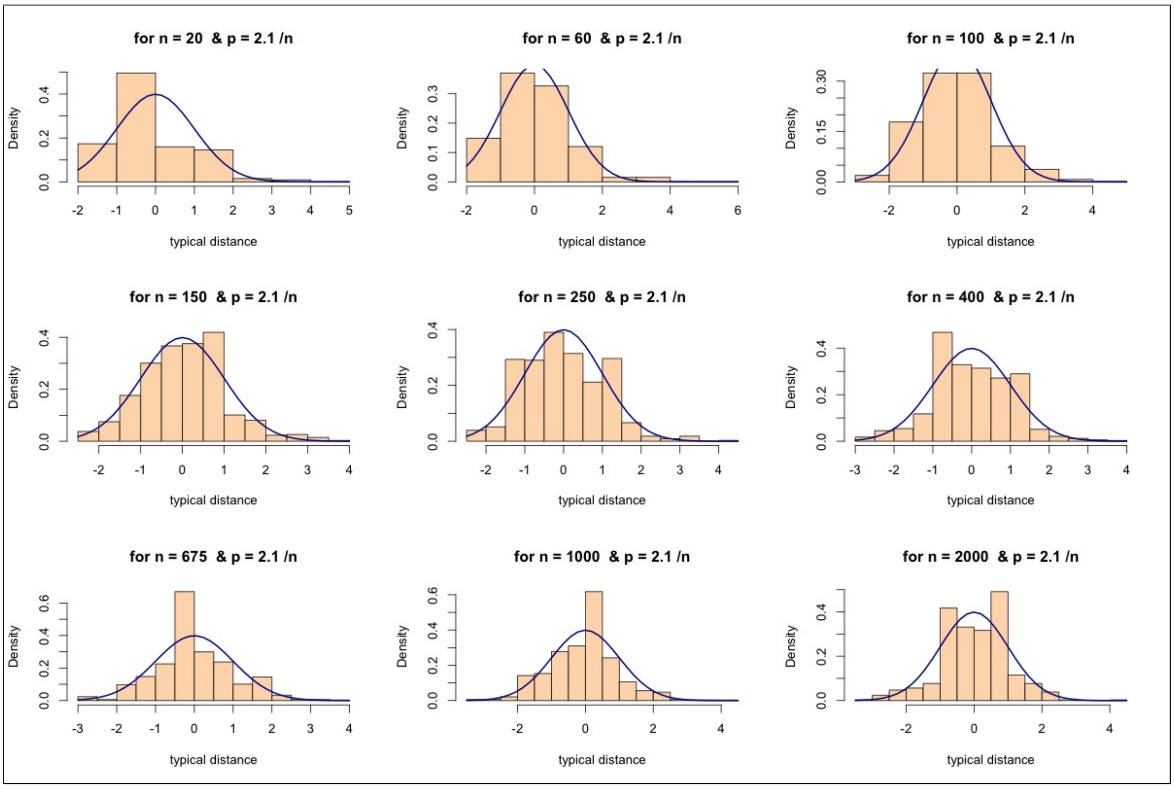


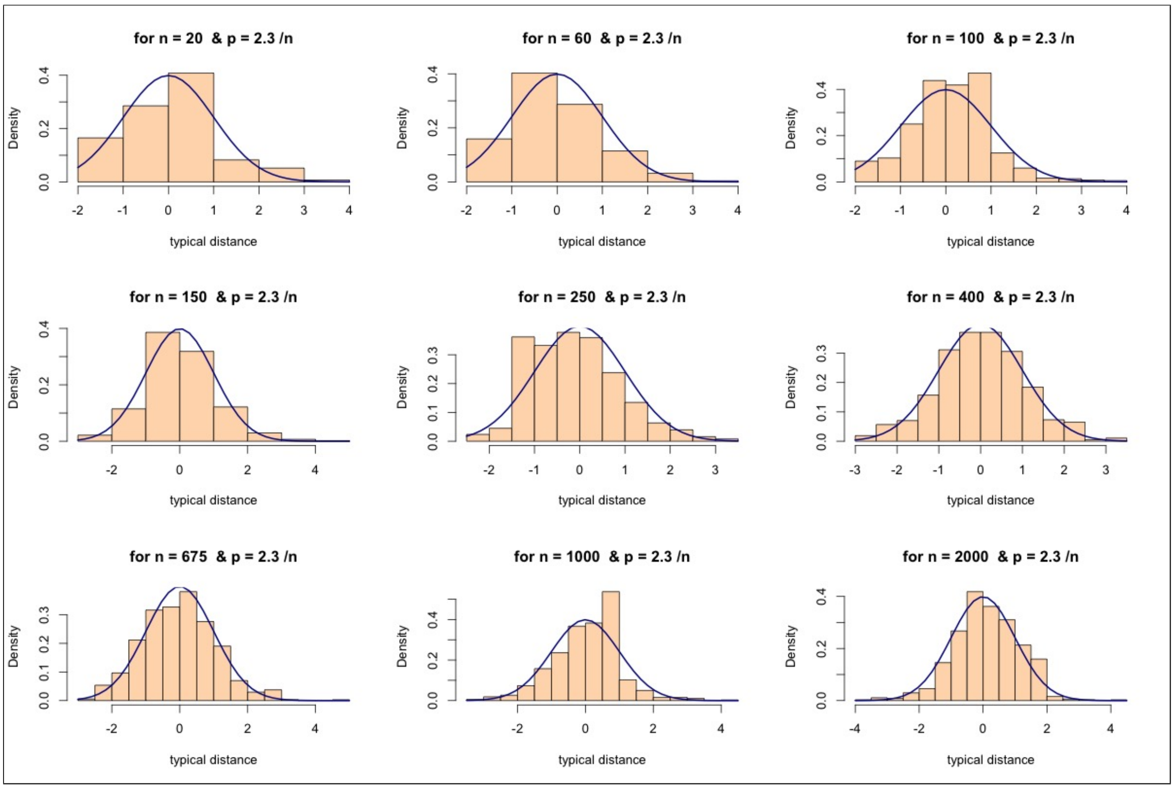

for n = 20 & p = 1.3 /n
for n = 60 & p = 1.3 /n
for n = 100 & p = 1.3 /n
for n = 150 & p = 1.3 /n
for n = 250 & p = 1.3 /n
for n = 400 & p = 1.3 /n
for n = 675 & p = 1.3 /n
for n = 1000 & p = 1.3 /n
for n = 2000 & p = 1.3 /n
Density
typical distance

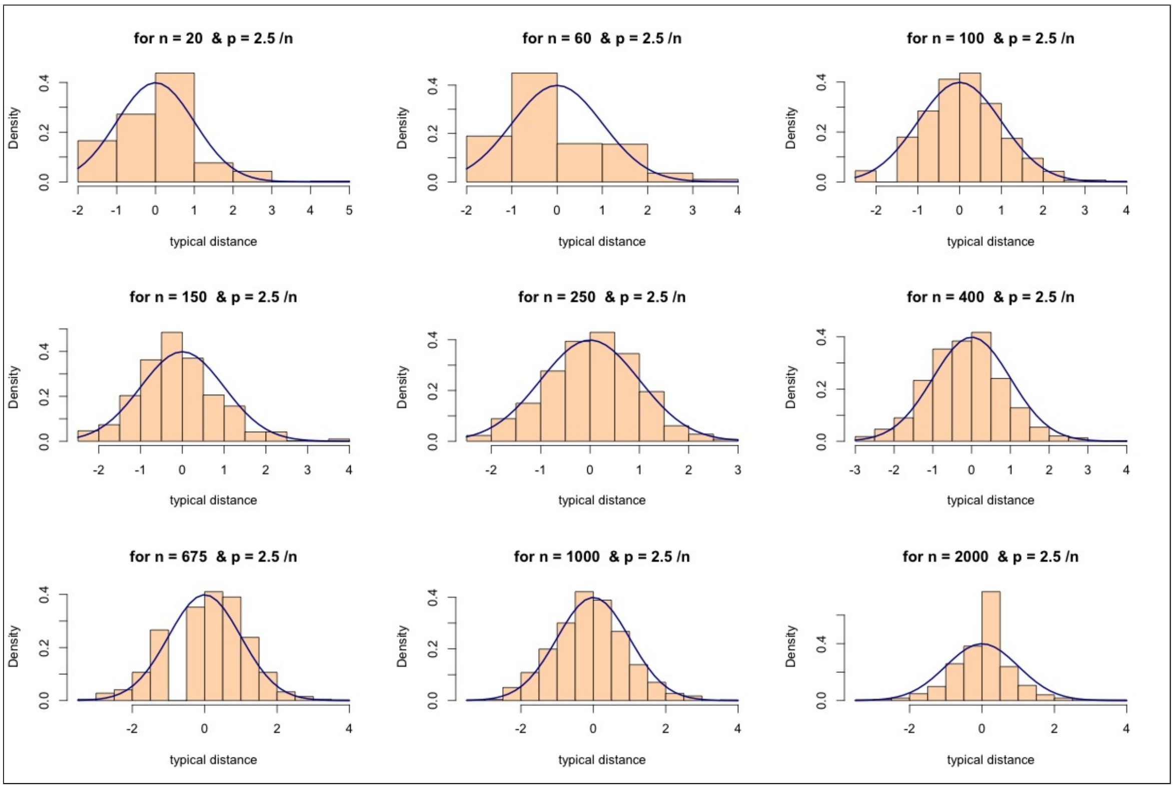

for n = 20 & p = 1.5 /n
for n = 60 & p = 1.5 /n
for n = 100 & p = 1.5 /n
for n = 150 & p = 1.5 /n
for n = 250 & p = 1.5 /n
for n = 400 & p = 1.5 /n
for n = 675 & p = 1.5 /n
for n = 1000 & p = 1.5 /n
for n = 2000 & p = 1.5 /n
Density
typical distance


for n = 20 & p = 1.7 /n
for n = 60 & p = 1.7 /n
for n = 100 & p = 1.7 /n
for n = 150 & p = 1.7 /n
for n = 250 & p = 1.7 /n
for n = 400 & p = 1.7 /n
for n = 675 & p = 1.7 /n
for n = 1000 & p = 1.7 /n
for n = 2000 & p = 1.7 /n
Density
typical distance


for n = 20 & p = 1.9 /n
for n = 60 & p = 1.9 /n
for n = 100 & p = 1.9 /n
for n = 150 & p = 1.9 /n
for n = 250 & p = 1.9 /n
for n = 400 & p = 1.9 /n
for n = 675 & p = 1.9 /n
for n = 1000 & p = 1.9 /n
for n = 2000 & p = 1.9 /n
Density
typical distance

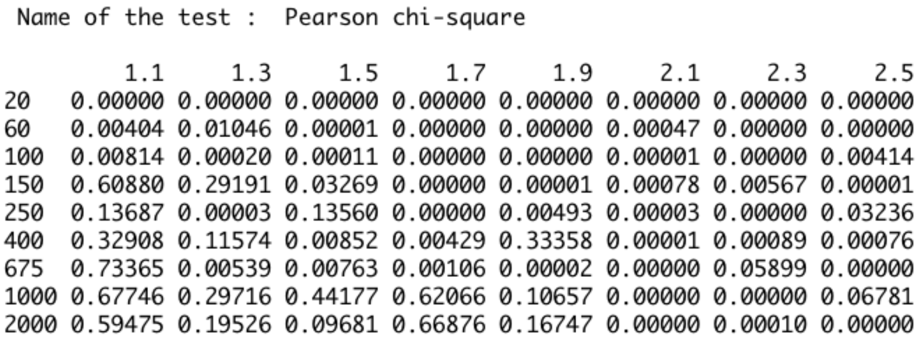

for n = 20 & p = 2.1 /n
Density
0.0 0.2 0.4
-2 -1 0 1 2 3 4 5
typical distance
for n = 60 & p = 2.1 /n
Density
0.0 0.1 0.2 0.3
-2 0 2 4 6
typical distance
for n = 100 & p = 2.1 /n
Density
0.00 0.15 0.30
-2 0 2 4
typical distance
for n = 150 & p = 2.1 /n
Density
0.0 0.2 0.4
-2 -1 0 1 2 3 4
typical distance
for n = 250 & p = 2.1 /n
Density
0.0 0.2 0.4
-2 -1 0 1 2 3 4
typical distance
for n = 400 & p = 2.1 /n
Density
0.0 0.2 0.4
-3 -2 -1 0 1 2 3 4
typical distance
for n = 675 & p = 2.1 /n
Density
0.0 0.2 0.4 0.6
-3 -2 -1 0 1 2 3 4
typical distance
for n = 1000 & p = 2.1 /n
Density
0.0 0.2 0.4 0.6
-2 0 2 4
typical distance
for n = 2000 & p = 2.1 /n
Density
0.0 0.2 0.4
-2 0 2 4
typical distance


for n = 20 & p = 2.3 /n
Density
0.0 0.2 0.4
-2 -1 0 1 2 3 4
typical distance
for n = 60 & p = 2.3 /n
Density
0.0 0.2 0.4
-2 -1 0 1 2 3 4
typical distance
for n = 100 & p = 2.3 /n
Density
0.0 0.2 0.4
-2 -1 0 1 2 3 4
typical distance
for n = 150 & p = 2.3 /n
Density
0.0 0.2 0.4
-2 0 2 4
typical distance
for n = 250 & p = 2.3 /n
Density
0.0 0.1 0.2 0.3
-2 -1 0 1 2 3
typical distance
for n = 400 & p = 2.3 /n
Density
0.0 0.1 0.2 0.3
-3 -2 -1 0 1 2 3
typical distance
for n = 675 & p = 2.3 /n
Density
0.0 0.1 0.2 0.3
-2 0 2 4
typical distance
for n = 1000 & p = 2.3 /n
Density
0.0 0.2 0.4
-2 0 2 4
typical distance
for n = 2000 & p = 2.3 /n
Density
0.0 0.2 0.4
-4 -2 0 2 4
typical distance

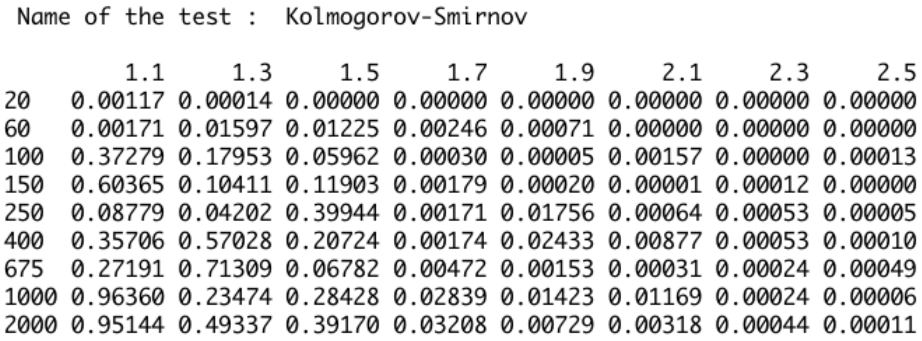


2.4.2 Table of p-values for different tests

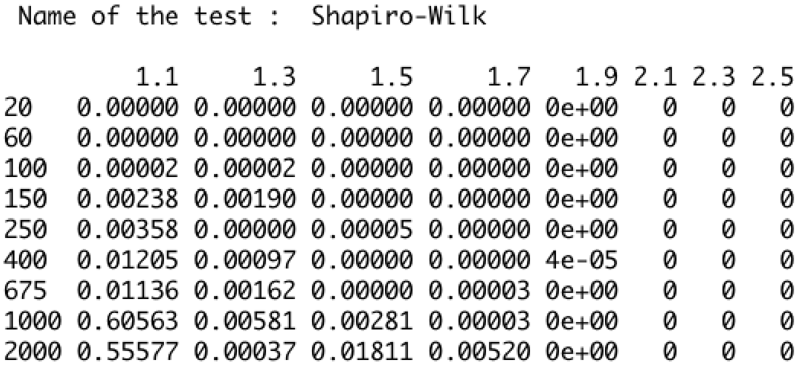
On the standardized samples, we perform Pearson's chi-square goodness of fit test, Kolmogorv-Smirnov test, and Shapiro-Wilk test for each pair of c and n. The table of the p-values for these tests are given below.



```
 Name of the test :  Pearson chi-square

         1.1     1.3     1.5     1.7     1.9     2.1     2.3     2.5
20    0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
60    0.00404 0.01046 0.00001 0.00000 0.00000 0.00047 0.00000 0.00000
100   0.00814 0.00020 0.00011 0.00000 0.00000 0.00001 0.00000 0.00414
150   0.60880 0.29191 0.03269 0.00000 0.00001 0.00078 0.00567 0.00001
250   0.13687 0.00003 0.13560 0.00000 0.00493 0.00003 0.00000 0.03236
400   0.32908 0.11574 0.00852 0.00429 0.33358 0.00001 0.00089 0.00076
675   0.73365 0.00539 0.00763 0.00106 0.00002 0.00000 0.05899 0.00000
1000  0.67746 0.29716 0.44177 0.62066 0.10657 0.00000 0.00000 0.06781
2000  0.59475 0.19526 0.09681 0.66876 0.16747 0.00000 0.00010 0.00000
```



```
 Name of the test :  Kolmogorov-Smirnov

         1.1     1.3     1.5     1.7     1.9     2.1     2.3     2.5
20    0.00117 0.00014 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
60    0.00171 0.01597 0.01225 0.00246 0.00071 0.00000 0.00000 0.00000
100   0.37279 0.17953 0.05962 0.00030 0.00005 0.00157 0.00000 0.00013
150   0.60365 0.10411 0.11903 0.00179 0.00020 0.00001 0.00012 0.00000
250   0.08779 0.04202 0.39944 0.00171 0.01756 0.00064 0.00053 0.00005
400   0.35706 0.57028 0.20724 0.00174 0.02433 0.00877 0.00053 0.00010
675   0.27191 0.71309 0.06782 0.00472 0.00153 0.00031 0.00024 0.00049
1000  0.96360 0.23474 0.28428 0.02839 0.01423 0.01169 0.00024 0.00006
2000  0.95144 0.49337 0.39170 0.03208 0.00729 0.00318 0.00044 0.00011
```



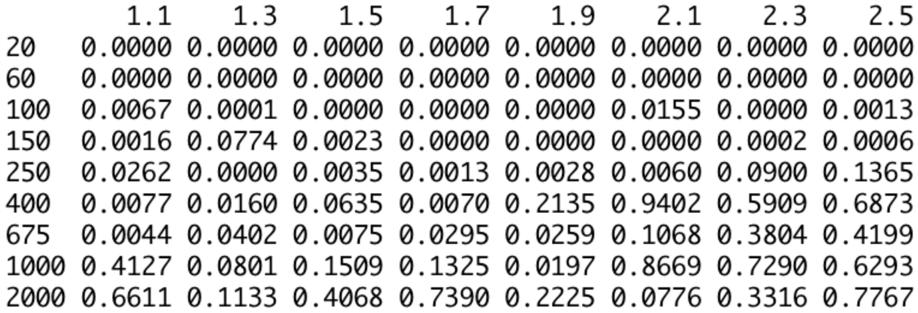
```
 Name of the test :  Shapiro-Wilk

           1.1     1.3     1.5     1.7   1.9 2.1 2.3 2.5
20   0.00000 0.00000 0.00000 0.00000 0e+00   0   0   0
60   0.00000 0.00000 0.00000 0.00000 0e+00   0   0   0
100  0.00002 0.00002 0.00000 0.00000 0e+00   0   0   0
150  0.00238 0.00190 0.00000 0.00000 0e+00   0   0   0
250  0.00358 0.00000 0.00005 0.00000 0e+00   0   0   0
400  0.01205 0.00097 0.00000 0.00000 4e-05   0   0   0
675  0.01136 0.00162 0.00000 0.00003 0e+00   0   0   0
1000 0.60563 0.00581 0.00281 0.00003 0e+00   0   0   0
2000 0.55577 0.00037 0.01811 0.00520 0e+00   0   0   0
```

Observations: We observe that the p-values get bigger towards the lower left corner of the table. This means that for small values of c and large values of n, our data passes the test of normality with great confidence! On the other hand, the cases for n and c both large are quite surprizing. Despite the standardized histograms being close enough (at least visually) to the standard normal density (e.g. look at the standardized histogram for $n = 400, c = 2.3$, or for $n = 1000, c = 2.5$), the p-values corresponding to them are quite low. A possible reason for this might be that the number of observations being large, the tests for normality become more sensitive to the data, in the sense that even very small deviations from the standard lead to much decrease in the p-value.

2.5 Testing symmetry

We perform the Randles-Fligner-Policello-Wolfe test of symmetry on the standardized data and record the p-values in the following table.

Table of the p-values

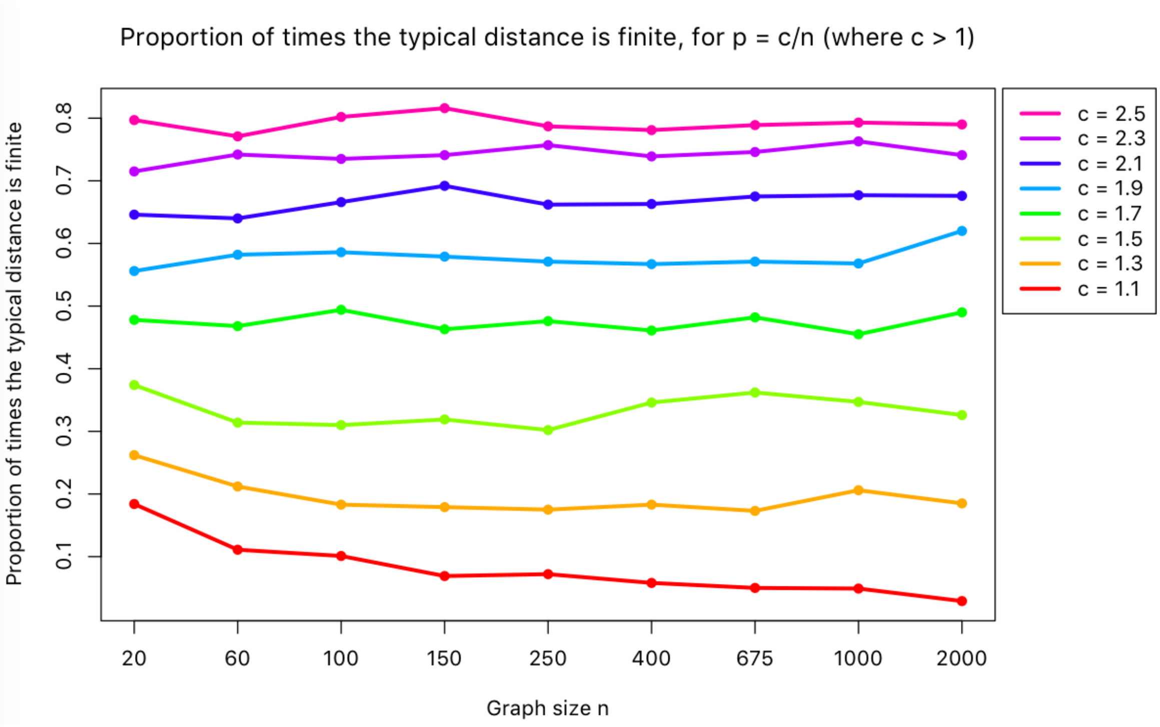


```
        1.1    1.3    1.5    1.7    1.9    2.1    2.3    2.5
20   0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
60   0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
100  0.0067 0.0001 0.0000 0.0000 0.0000 0.0155 0.0000 0.0013
150  0.0016 0.0774 0.0023 0.0000 0.0000 0.0000 0.0002 0.0006
250  0.0262 0.0000 0.0035 0.0013 0.0028 0.0060 0.0900 0.1365
400  0.0077 0.0160 0.0635 0.0070 0.2135 0.9402 0.5909 0.6873
675  0.0044 0.0402 0.0075 0.0295 0.0259 0.1068 0.3804 0.4199
1000 0.4127 0.0801 0.1509 0.1325 0.0197 0.8669 0.7290 0.6293
2000 0.6611 0.1133 0.4068 0.7390 0.2225 0.0776 0.3316 0.7767
```

Comments: A general pattern in the above table is that the p-values increase gradually as we move towards the lower right corner. This pattern is also visible from the standardized histogram plots – those histograms are more symmetric around the mean for larger values of n and c. Note that this is somewhat contradictory to the table of p-values of the normality tests, where we saw the p-values to increase towards the lower left corner. It is really strange that the samples for larger n and smaller c are accepted with high p-values in the tests of normality but rejected poorly in tests of symmetry.

2.6 Proportion of times the observed distance is finite

Plot of the proportion of times the observed typical distance is finite is given below.



Comments: It is clear from the above plot that the proportion of times the typical distance is finite depends more on the value of $c = np$ than on the graph size n itself.

Remark

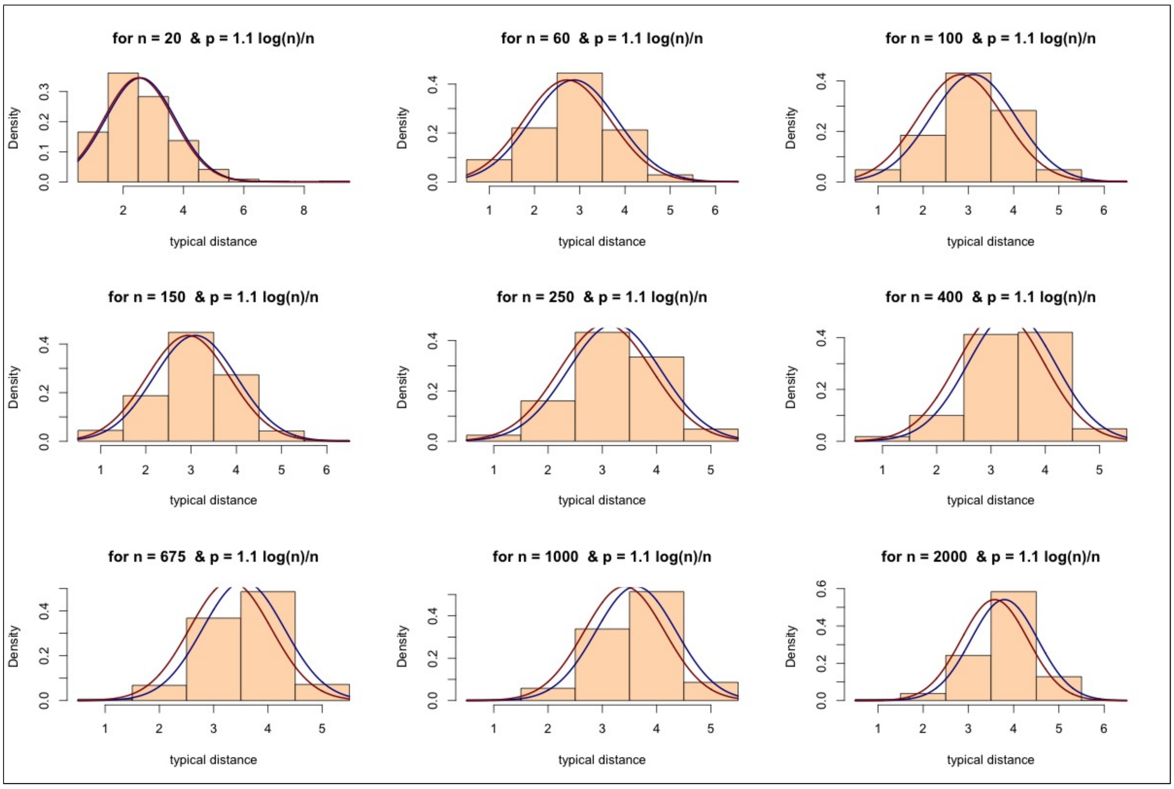
We repeated all the above calculations also for the graph distance in ER(n, p) between a fixed vertex (say vertex 1) and another randomly chosen vertex. The results we found were similar to the case for typical distance, as expected. Hence we do not include them in this report.

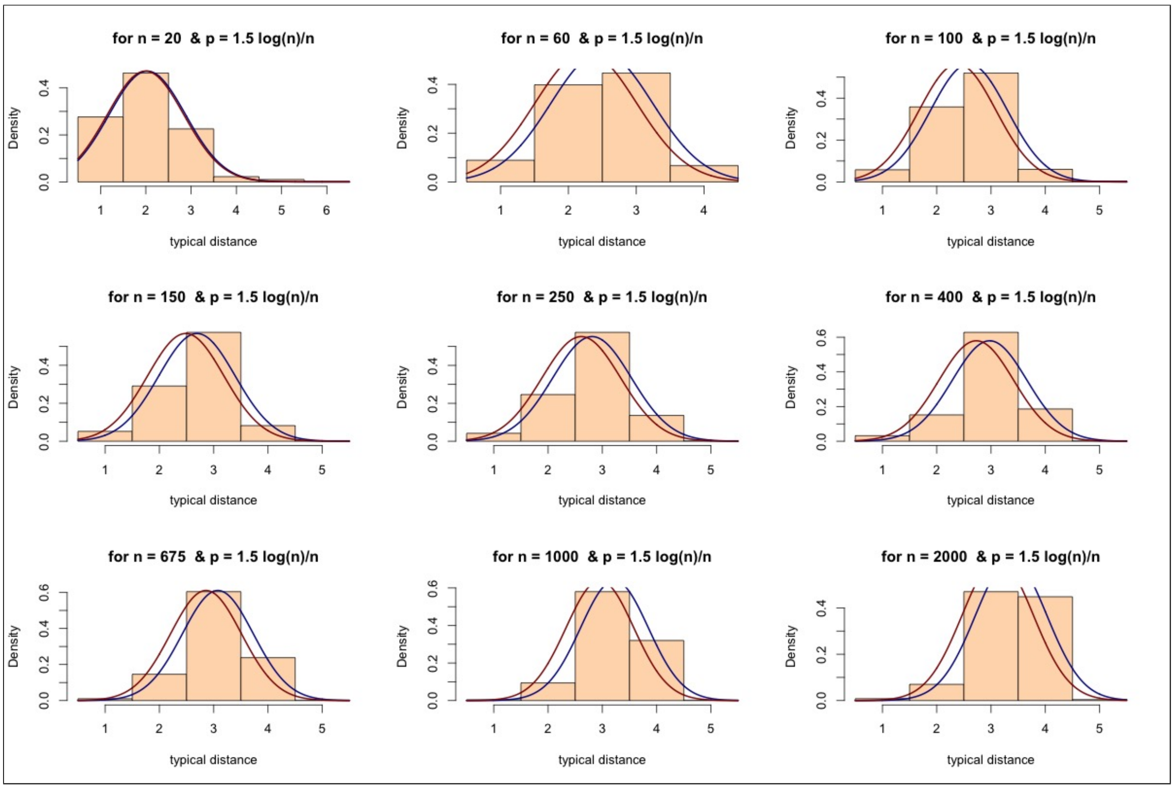
3 Connectivity regime

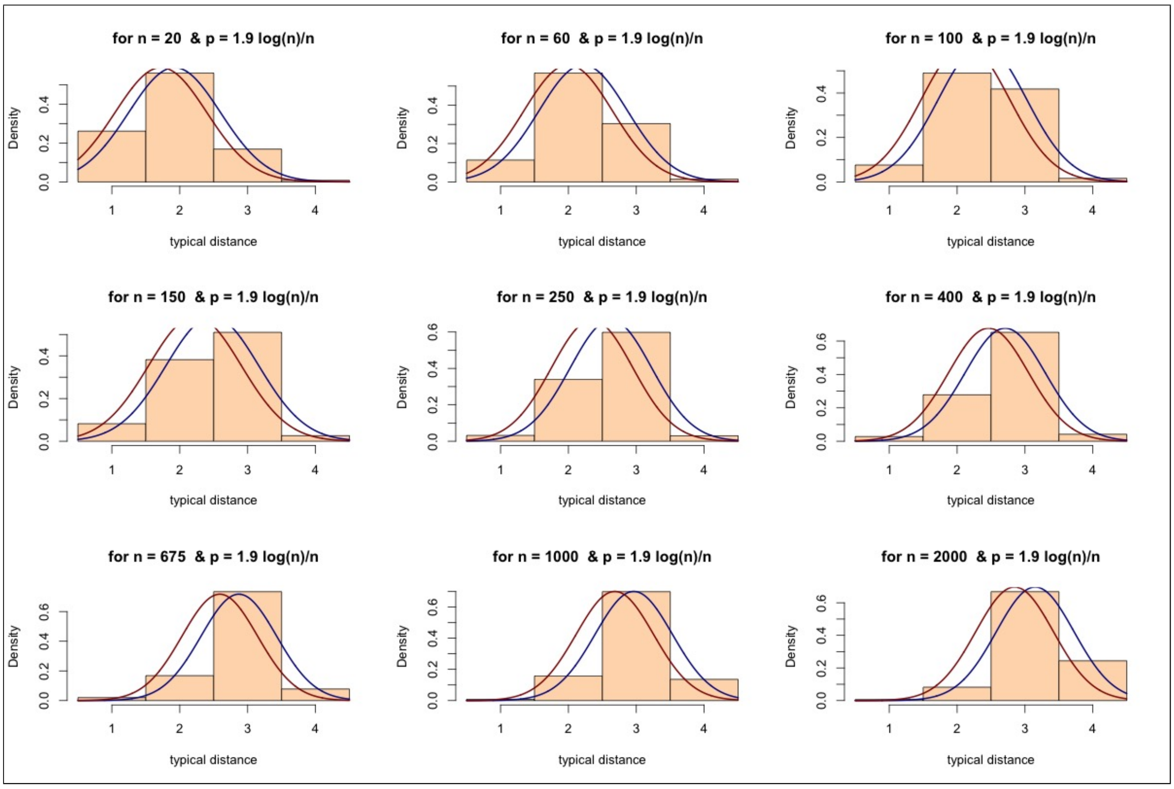
We simulated ER$(n, c \log n/n)$ for 500 times for each pair of (c, n) where $c = 1.1, 1.3, 1.7, \dots, 2.5$ and $n = 20, 60, 100, 150, 250, 400, 675, 1000, 2000$. As noted earlier, these graph sizes are chosen so that $\log n$ varies almost linearly. Since these graphs are connected with high probabilities, we got only a few infinite ones, which were thrown away during our analysis.

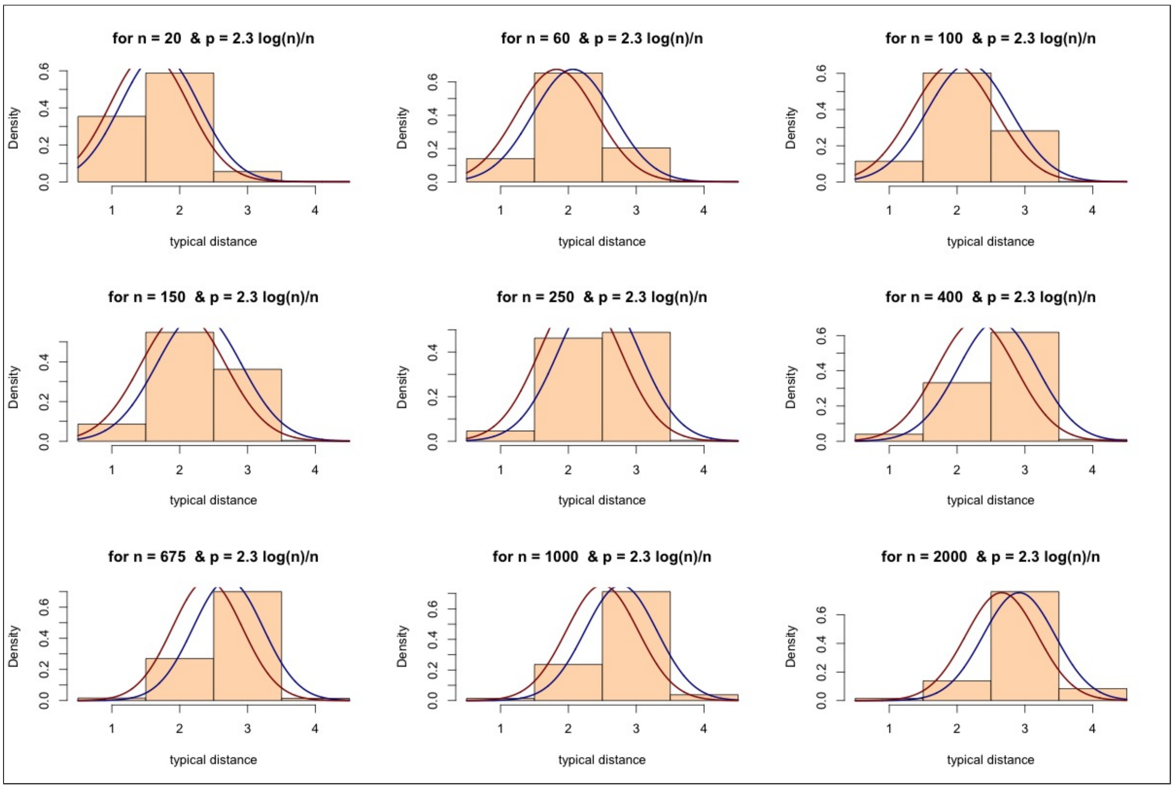
3.1 Histograms

For each of the values $np = c = 1.1, 1.5, 1.9, 2.3$, the histograms of the observed (finite) typical distances in ER$(n, c \log n/n)$ for each choice of the graph size n are given below (histograms for the other values of c show similar pattern, those are omitted here). The blue line shows the normal density with mean and s.d. estimated by the sample mean and sample s.d. The red line shows the normal density with mean equal to $\log n/ \log(c \log n)$ and s.d. estimated by the sample s.d.


for n = 20 & p = 1.1 log(n)/n
Density
0.0 0.1 0.2 0.3
2 4 6 8
typical distance
for n = 60 & p = 1.1 log(n)/n
Density
0.0 0.2 0.4
1 2 3 4 5 6
typical distance
for n = 100 & p = 1.1 log(n)/n
Density
0.0 0.2 0.4
1 2 3 4 5 6
typical distance
for n = 150 & p = 1.1 log(n)/n
Density
0.0 0.2 0.4
1 2 3 4 5 6
typical distance
for n = 250 & p = 1.1 log(n)/n
Density
0.0 0.2 0.4
1 2 3 4 5
typical distance
for n = 400 & p = 1.1 log(n)/n
Density
0.0 0.2 0.4
1 2 3 4 5
typical distance
for n = 675 & p = 1.1 log(n)/n
Density
0.0 0.2 0.4
1 2 3 4 5
typical distance
for n = 1000 & p = 1.1 log(n)/n
Density
0.0 0.2 0.4
1 2 3 4 5
typical distance
for n = 2000 & p = 1.1 log(n)/n
Density
0.0 0.2 0.4 0.6
1 2 3 4 5 6
typical distance


for n = 20 & p = 1.5 log(n)/n
Density
0.0 0.2 0.4
1 2 3 4 5 6
typical distance
for n = 60 & p = 1.5 log(n)/n
Density
0.0 0.2 0.4
1 2 3 4
typical distance
for n = 100 & p = 1.5 log(n)/n
Density
0.0 0.2 0.4
1 2 3 4 5
typical distance
for n = 150 & p = 1.5 log(n)/n
Density
0.0 0.2 0.4
1 2 3 4 5
typical distance
for n = 250 & p = 1.5 log(n)/n
Density
0.0 0.2 0.4
1 2 3 4 5
typical distance
for n = 400 & p = 1.5 log(n)/n
Density
0.0 0.2 0.4 0.6
1 2 3 4 5
typical distance
for n = 675 & p = 1.5 log(n)/n
Density
0.0 0.2 0.4 0.6
1 2 3 4 5
typical distance
for n = 1000 & p = 1.5 log(n)/n
Density
0.0 0.2 0.4 0.6
1 2 3 4 5
typical distance
for n = 2000 & p = 1.5 log(n)/n
Density
0.0 0.2 0.4
1 2 3 4 5
typical distance

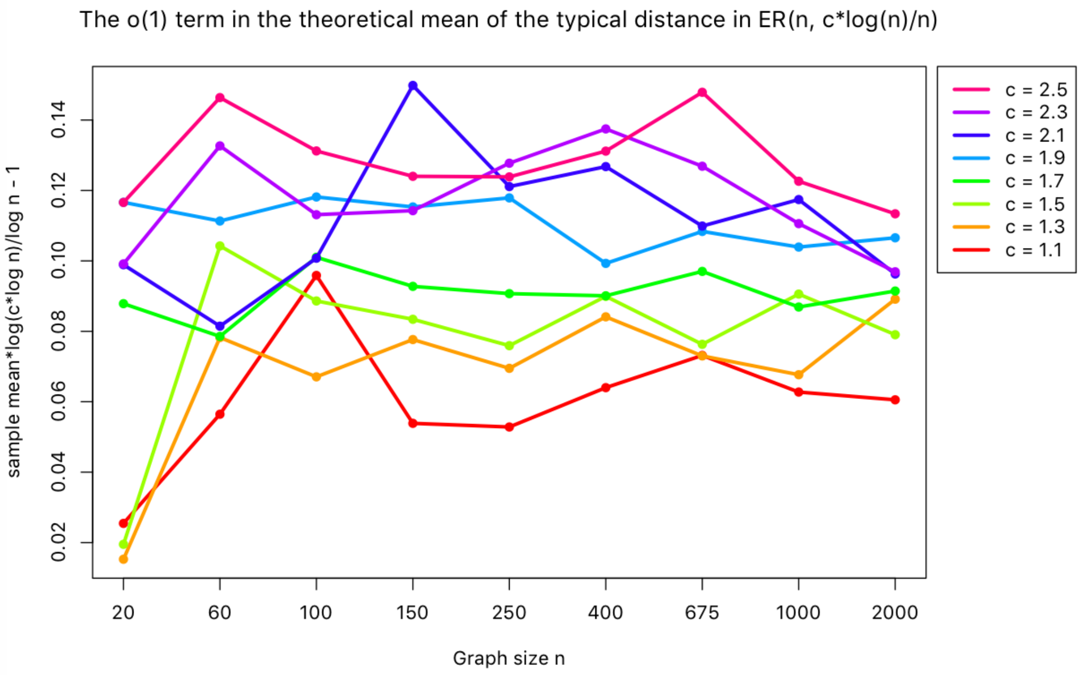




Observations: We notice that the red curve is always behind the blue curve, indicating that the sample mean is always more than $\log n/\log np$, which is exactly opposite to the previous case of $p = c/n, c > 1$. Another observation here is that the difference between sample mean and $\log n/\log(c\log n)$ is not much even for smaller values of c and n. In fact, this difference slightly increases as c or n increases, which it should, because this difference is not just the $o(1)$ term, it equals the sample mean $-\log n / \log(c \log n) =$ (the $o(1)$ term) $\times \log n / \log(c \log n)$.

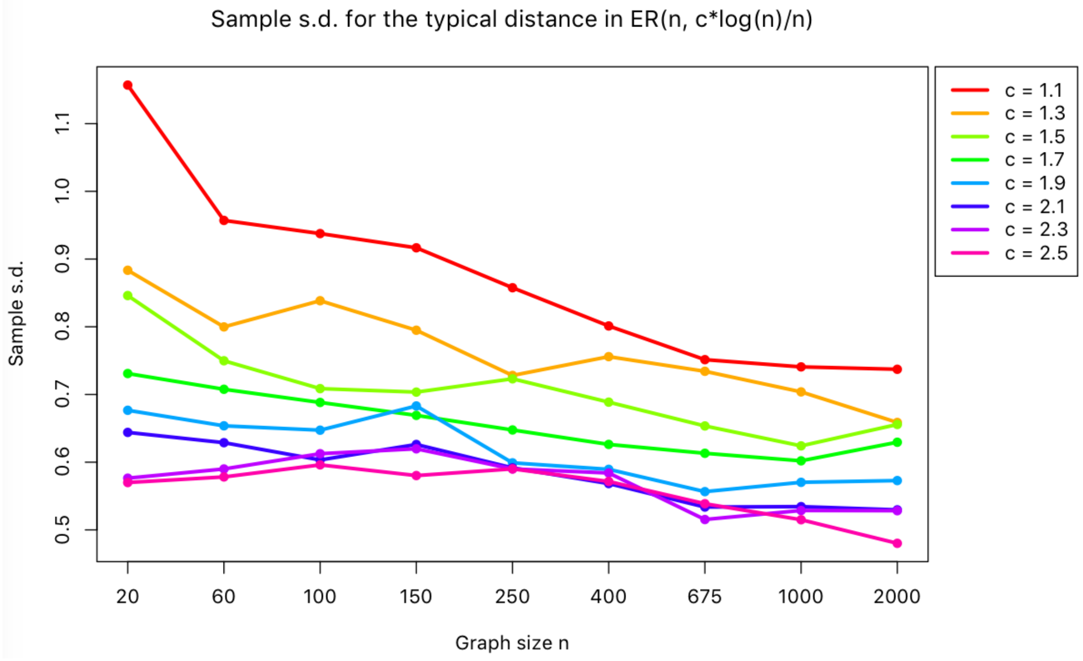
3.2 Studying the $o(1)$ term

By '$o(1)$ term' we mean (the sample mean $\times \log(c \log n)/\log n) - 1$. Just as earlier, we use different colors in the following plot of the $o(1)$ term for indicating different values of c.



Note that this plot does not exhibit any prominent difference among the different values of c. Comparing this plot with the plot of the o(1) term in the $p = c/n$ case, we can see that not only the sign of the o(1) term is changed, but its magnitude is also much less than the $p = c/n$ case.

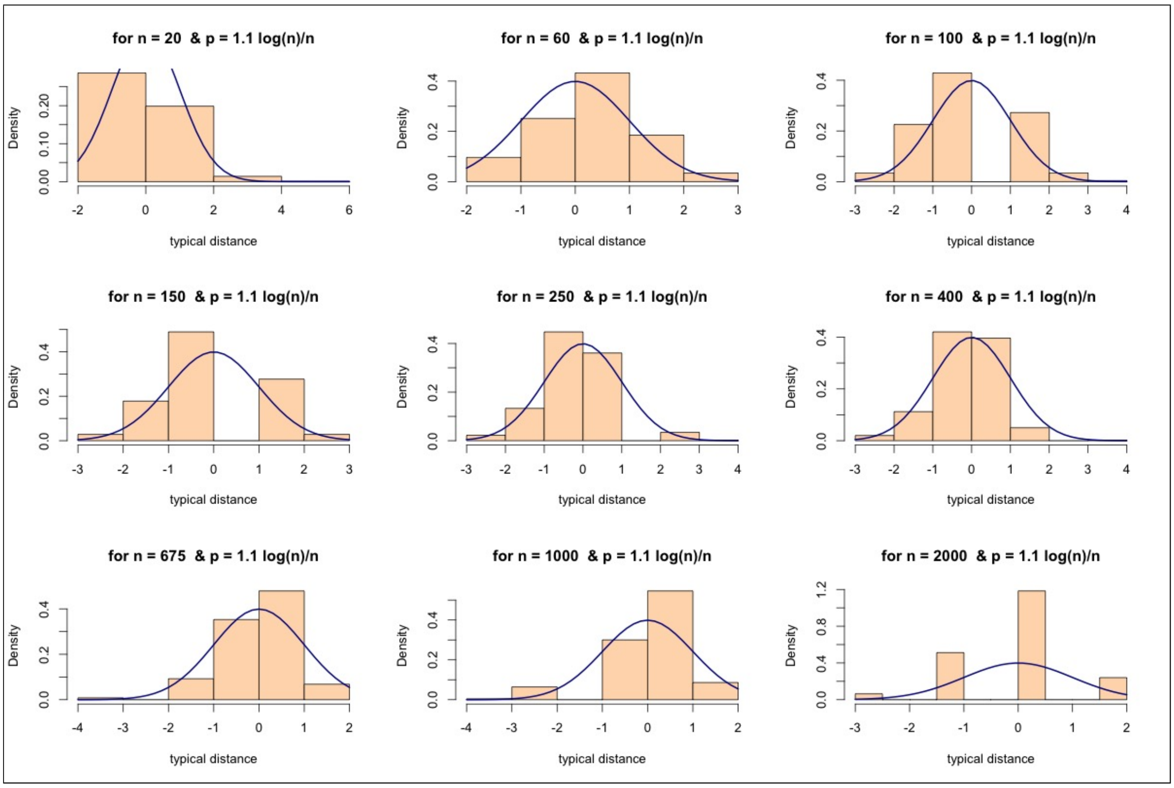
3.3 Studying the standard deviations

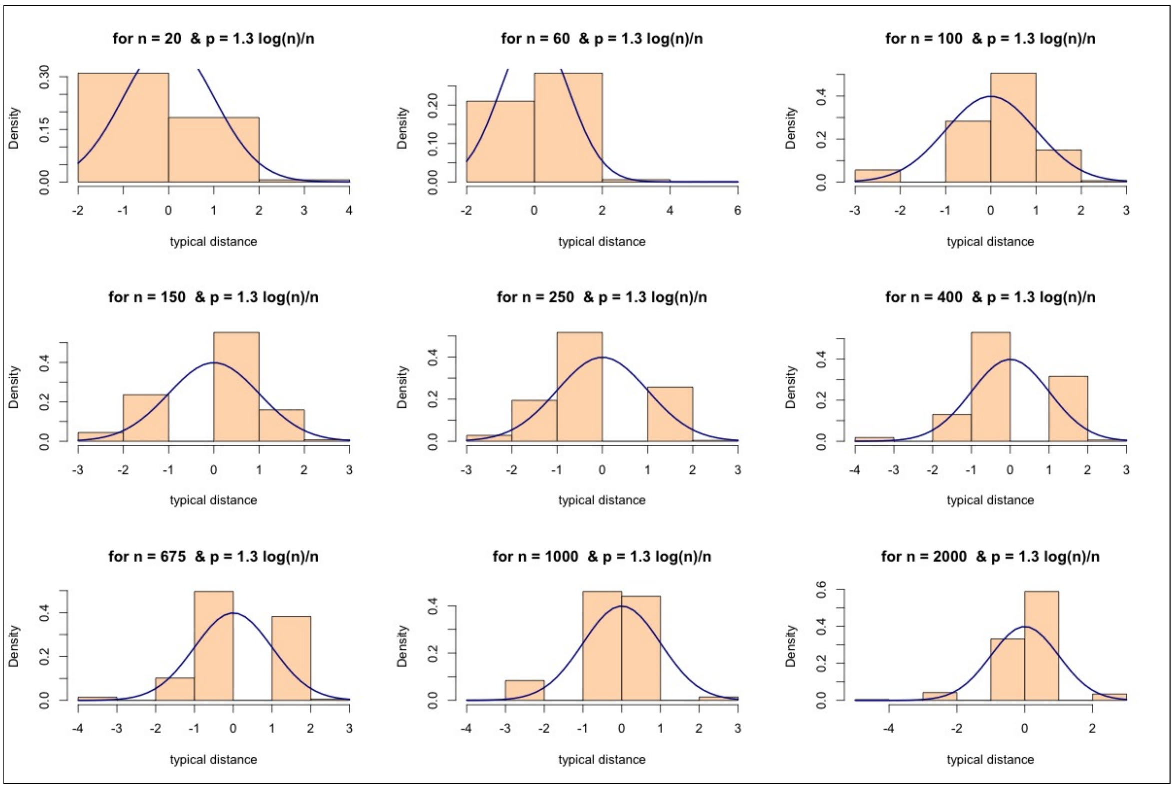


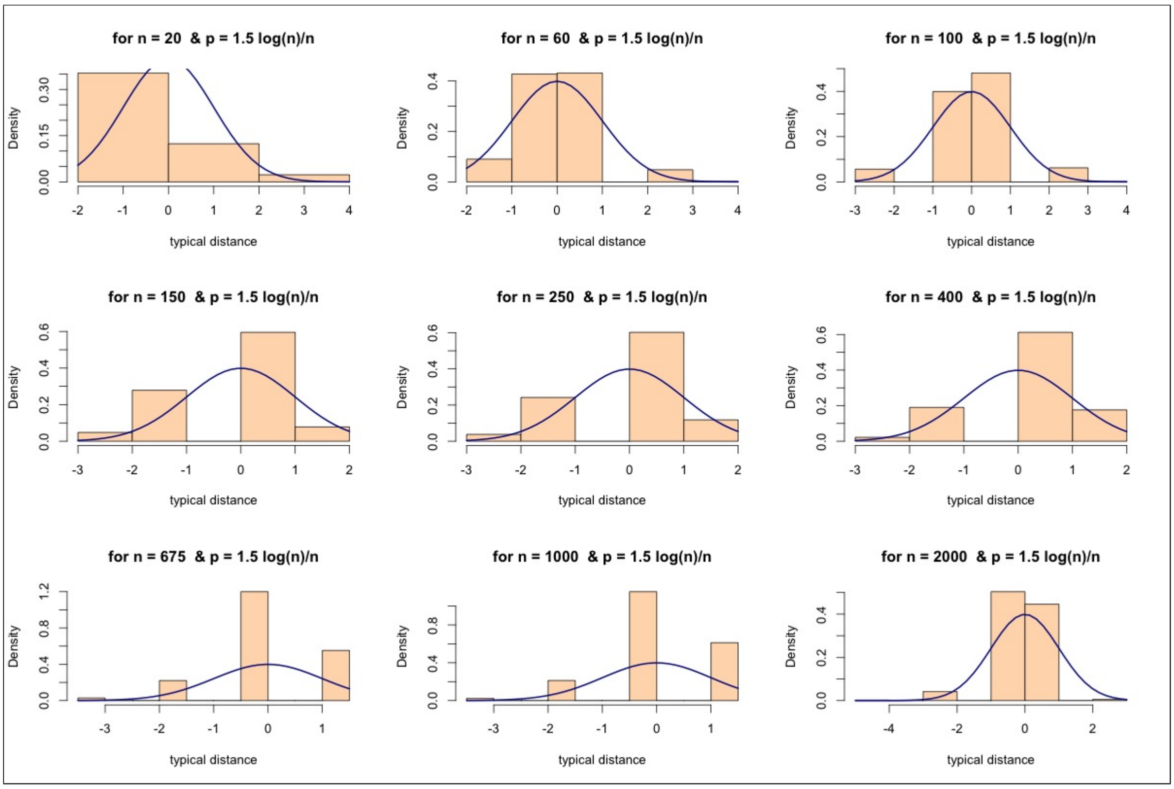
Observations: It is notable that the standard deviations are decreasing, and the rate of its decay is much faster for the higher values of c.

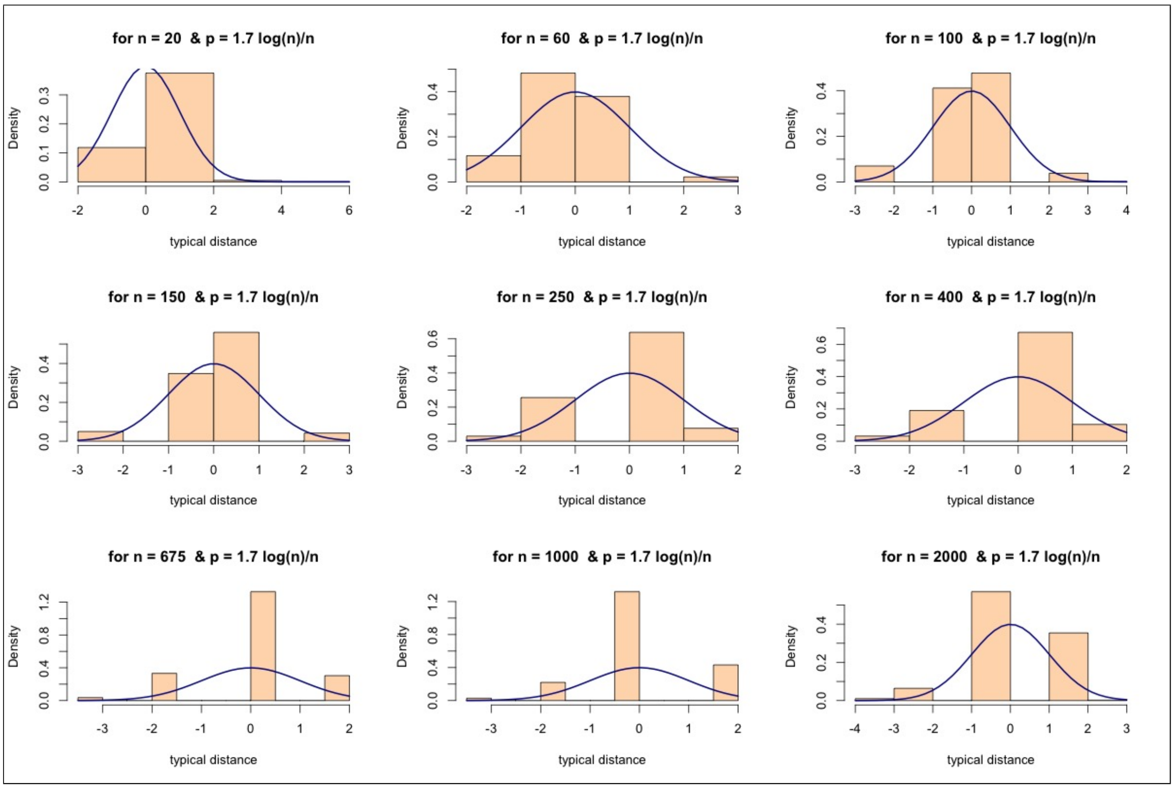
3.4 Testing normality

Each sample is standardized using sample mean and s.d. and the histograms for the standardized samples are shown below. The blue line is the standard normal density. Here we include the histograms for every pair of c and n that we simulated, because that will accompany us while comprehending the p-values of the tests of normality and symmetry which will be performed next.

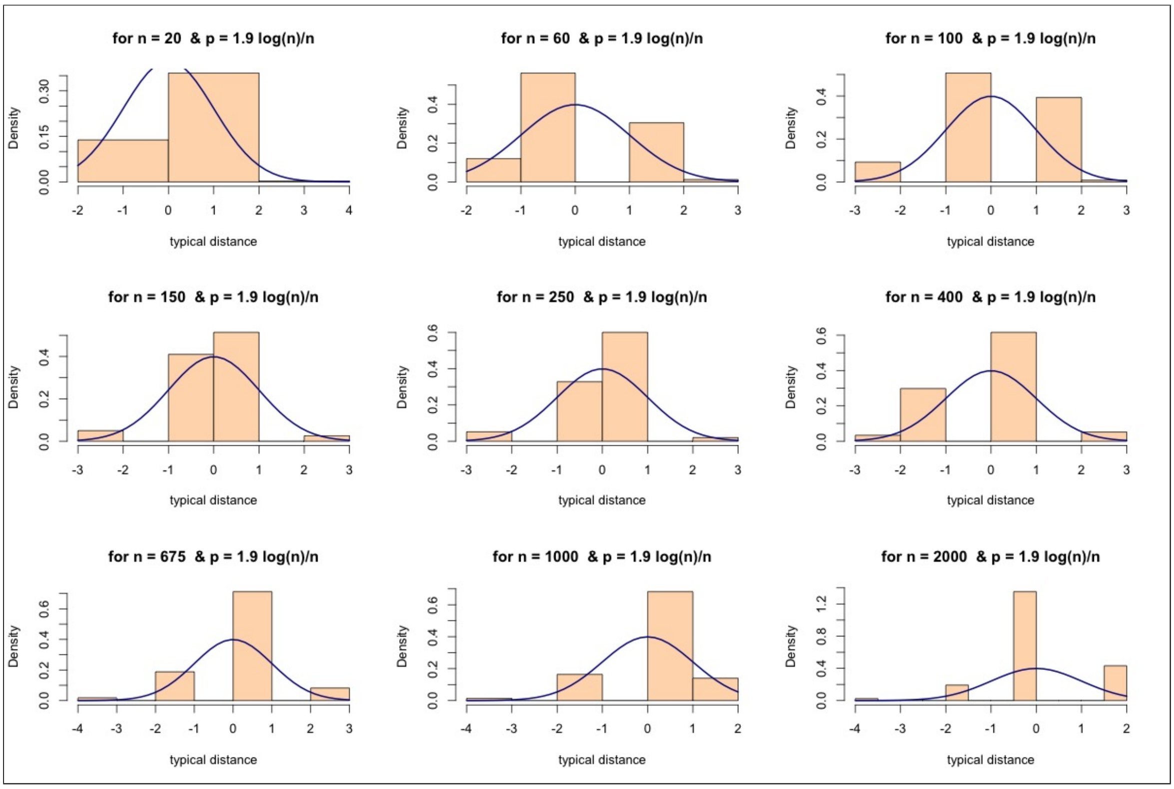



for n = 20 & p = 1.3 log(n)/n
for n = 60 & p = 1.3 log(n)/n
for n = 100 & p = 1.3 log(n)/n
for n = 150 & p = 1.3 log(n)/n
for n = 250 & p = 1.3 log(n)/n
for n = 400 & p = 1.3 log(n)/n
for n = 675 & p = 1.3 log(n)/n
for n = 1000 & p = 1.3 log(n)/n
for n = 2000 & p = 1.3 log(n)/n
Density
typical distance

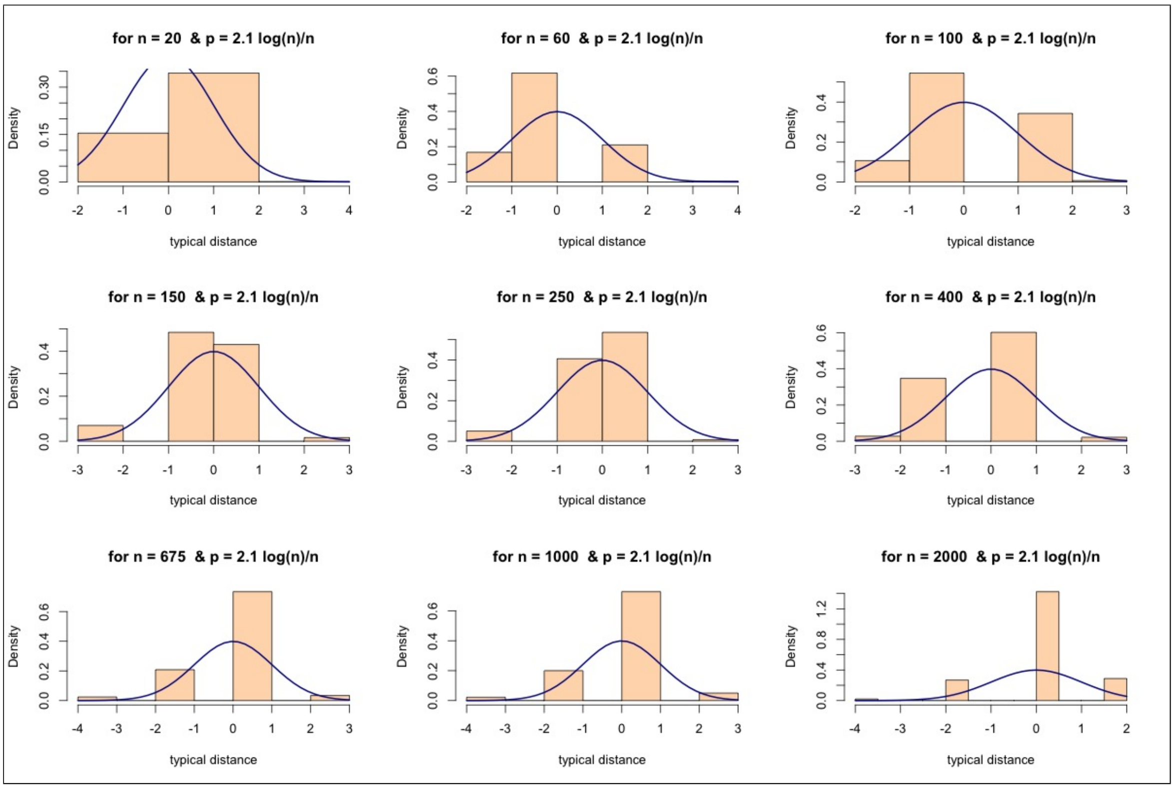

for n = 20 & p = 1.5 log(n)/n
for n = 60 & p = 1.5 log(n)/n
for n = 100 & p = 1.5 log(n)/n
for n = 150 & p = 1.5 log(n)/n
for n = 250 & p = 1.5 log(n)/n
for n = 400 & p = 1.5 log(n)/n
for n = 675 & p = 1.5 log(n)/n
for n = 1000 & p = 1.5 log(n)/n
for n = 2000 & p = 1.5 log(n)/n
Density
typical distance

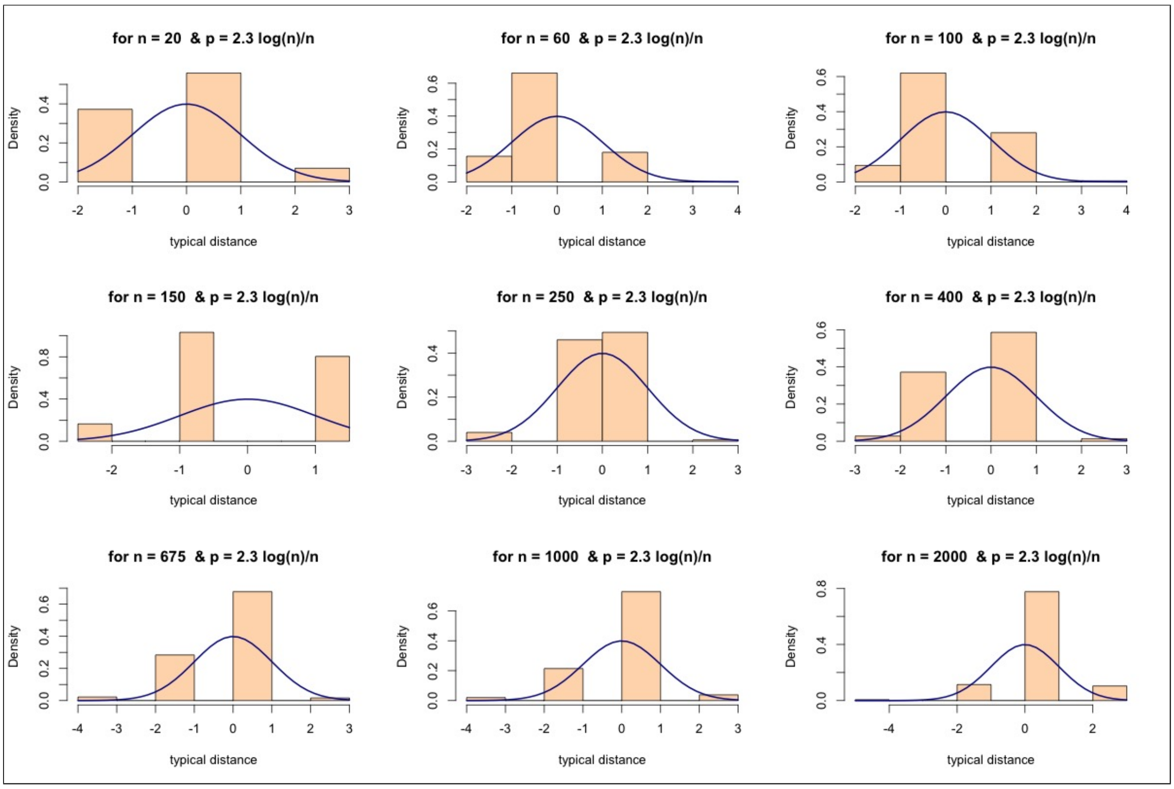


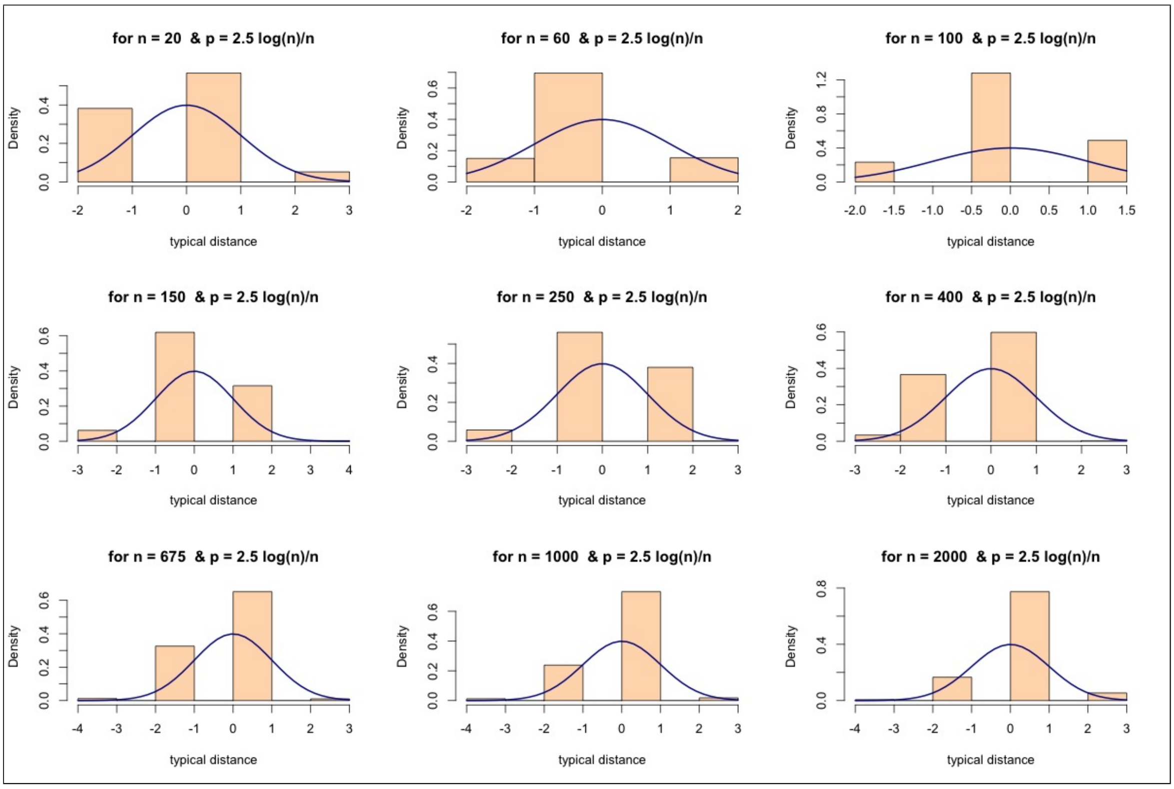
for n = 20 & p = 1.7 log(n)/n
for n = 60 & p = 1.7 log(n)/n
for n = 100 & p = 1.7 log(n)/n
for n = 150 & p = 1.7 log(n)/n
for n = 250 & p = 1.7 log(n)/n
for n = 400 & p = 1.7 log(n)/n
for n = 675 & p = 1.7 log(n)/n
for n = 1000 & p = 1.7 log(n)/n
for n = 2000 & p = 1.7 log(n)/n
Density
typical distance



for n = 20 & p = 1.9 log(n)/n
for n = 60 & p = 1.9 log(n)/n
for n = 100 & p = 1.9 log(n)/n
for n = 150 & p = 1.9 log(n)/n
for n = 250 & p = 1.9 log(n)/n
for n = 400 & p = 1.9 log(n)/n
for n = 675 & p = 1.9 log(n)/n
for n = 1000 & p = 1.9 log(n)/n
for n = 2000 & p = 1.9 log(n)/n
Density
typical distance


for n = 20 & p = 2.1 log(n)/n
Density
0.00 0.15 0.30
-2 -1 0 1 2 3 4
typical distance
for n = 60 & p = 2.1 log(n)/n
Density
0.0 0.2 0.4 0.6
-2 -1 0 1 2 3 4
typical distance
for n = 100 & p = 2.1 log(n)/n
Density
0.0 0.2 0.4
-2 -1 0 1 2 3
typical distance
for n = 150 & p = 2.1 log(n)/n
Density
0.0 0.2 0.4
-3 -2 -1 0 1 2 3
typical distance
for n = 250 & p = 2.1 log(n)/n
Density
0.0 0.2 0.4
-3 -2 -1 0 1 2 3
typical distance
for n = 400 & p = 2.1 log(n)/n
Density
0.0 0.2 0.4 0.6
-3 -2 -1 0 1 2 3
typical distance
for n = 675 & p = 2.1 log(n)/n
Density
0.0 0.2 0.4 0.6
-4 -3 -2 -1 0 1 2 3
typical distance
for n = 1000 & p = 2.1 log(n)/n
Density
0.0 0.2 0.4 0.6
-4 -3 -2 -1 0 1 2 3
typical distance
for n = 2000 & p = 2.1 log(n)/n
Density
0.0 0.4 0.8 1.2
-4 -3 -2 -1 0 1 2
typical distance


for n = 20 & p = 2.3 log(n)/n
Density
0.0 0.2 0.4
-2 -1 0 1 2 3
typical distance
for n = 60 & p = 2.3 log(n)/n
Density
0.0 0.2 0.4 0.6
-2 -1 0 1 2 3 4
typical distance
for n = 100 & p = 2.3 log(n)/n
Density
0.0 0.2 0.4 0.6
-2 -1 0 1 2 3 4
typical distance
for n = 150 & p = 2.3 log(n)/n
Density
0.0 0.4 0.8
-2 -1 0 1
typical distance
for n = 250 & p = 2.3 log(n)/n
Density
0.0 0.2 0.4
-3 -2 -1 0 1 2 3
typical distance
for n = 400 & p = 2.3 log(n)/n
Density
0.0 0.2 0.4 0.6
-3 -2 -1 0 1 2 3
typical distance
for n = 675 & p = 2.3 log(n)/n
Density
0.0 0.2 0.4 0.6
-4 -3 -2 -1 0 1 2 3
typical distance
for n = 1000 & p = 2.3 log(n)/n
Density
0.0 0.2 0.4 0.6
-4 -3 -2 -1 0 1 2 3
typical distance
for n = 2000 & p = 2.3 log(n)/n
Density
0.0 0.4 0.8
-4 -2 0 2
typical distance



for n = 20 & p = 2.5 log(n)/n

for n = 60 & p = 2.5 log(n)/n

for n = 100 & p = 2.5 log(n)/n

for n = 150 & p = 2.5 log(n)/n

for n = 250 & p = 2.5 log(n)/n

for n = 400 & p = 2.5 log(n)/n

for n = 675 & p = 2.5 log(n)/n

for n = 1000 & p = 2.5 log(n)/n

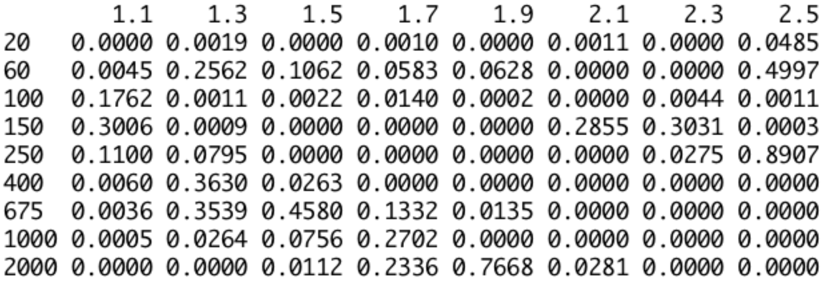
for n = 2000 & p = 2.5 log(n)/n

Comments: It is clear from the above histograms that the distribution of the typical distance in the connectivity regime does not come any close to the normal distribution, because of its discrete nature. This is also reflected by the tests for normality: when we performed the tests (same list as in the previous section), all the p-values we got were very much close to 0. (So small that when we printed them for 5 places of decimal, all of them were reported as 0; hence we do not include those tables here.)

3.5 Testing symmetry

Just as earlier, we perform the Randles-Fligner-Policello-Wolfe test of symmetry on the standardized data and record the p-values in the following table.

Table of the p-values



```
        1.1    1.3    1.5    1.7    1.9    2.1    2.3    2.5
20   0.0000 0.0019 0.0000 0.0010 0.0000 0.0011 0.0000 0.0485
60   0.0045 0.2562 0.1062 0.0583 0.0628 0.0000 0.0000 0.4997
100  0.1762 0.0011 0.0022 0.0140 0.0002 0.0000 0.0044 0.0011
150  0.3006 0.0009 0.0000 0.0000 0.0000 0.2855 0.3031 0.0003
250  0.1100 0.0795 0.0000 0.0000 0.0000 0.0000 0.0275 0.8907
400  0.0060 0.3630 0.0263 0.0000 0.0000 0.0000 0.0000 0.0000
675  0.0036 0.3539 0.4580 0.1332 0.0135 0.0000 0.0000 0.0000
1000 0.0005 0.0264 0.0756 0.2702 0.0000 0.0000 0.0000 0.0000
2000 0.0000 0.0000 0.0112 0.2336 0.7668 0.0281 0.0000 0.0000
```

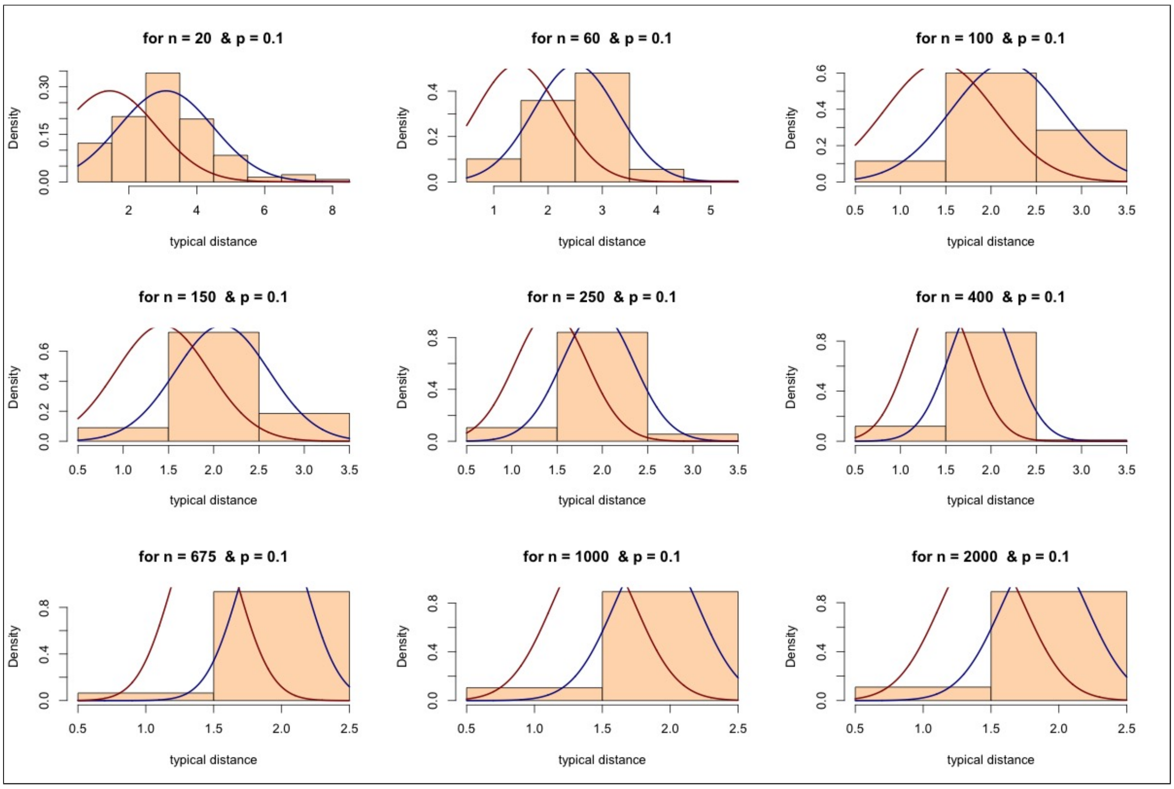
Comments: No major pattern is observed in this table of p-values, except for that the p-values along the 'diagonal' are very small. We suspect that the absence of any pattern is due to the discrete nature of the data. For instance, consider the case $n = 250, c = 2.5$. We don't find the histogram to be so symmetric that it should produce the p-value 0.8907. But looking closely, we note that there are only three values that are observed, namely $1, 2$ and 3, and the frequencies of 2 and 3 are nearly same, while the frequency of 1 is negligible to them. Perhaps this is the reason why we get such a large p-value in this case.

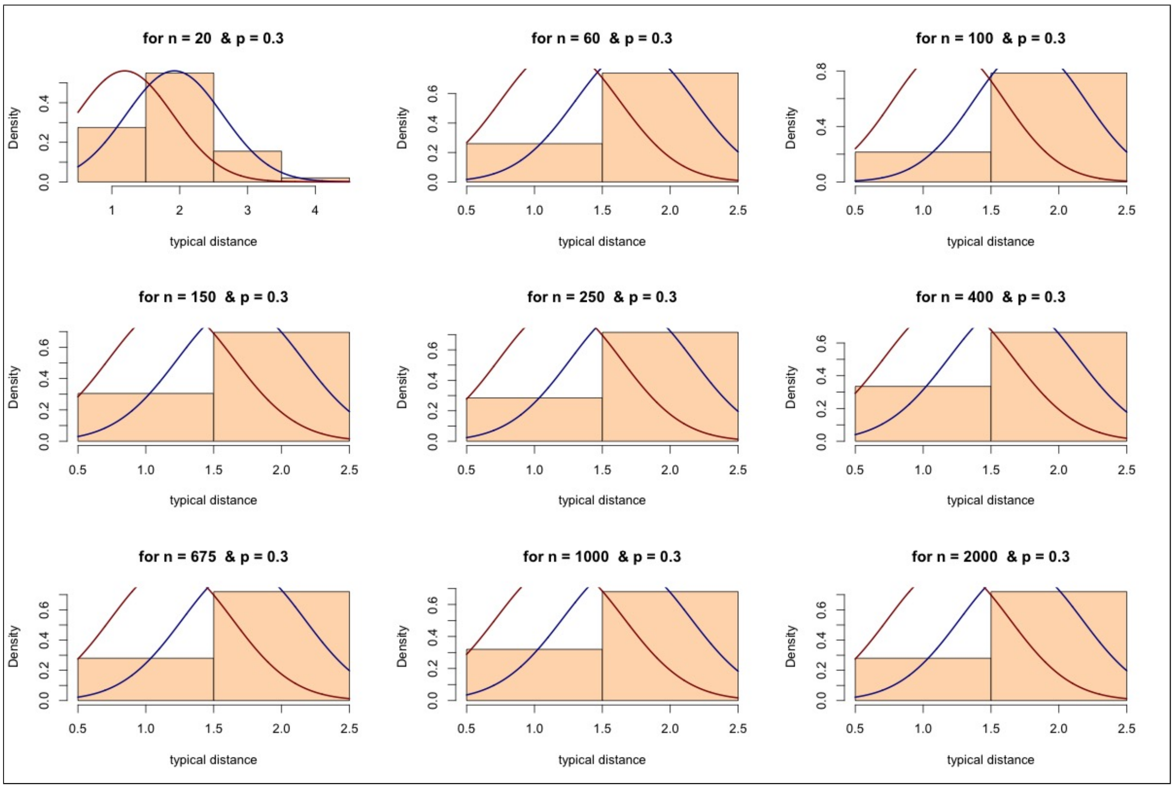
4 Constant probability case

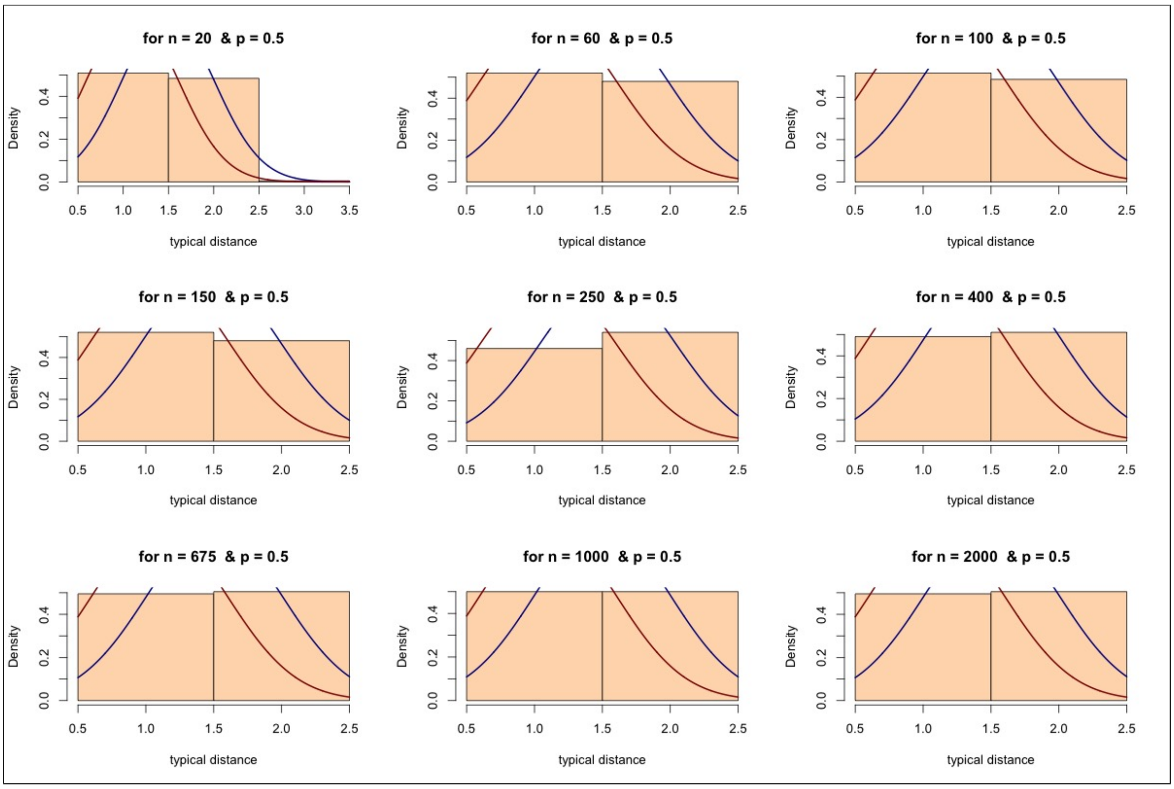
We considered $p = 0.1, 0.3, 0.5, 0.7, 0.9$, and varied the graph size n same as earlier. Since these random graphs are connected with very high probabilities, we simulated these graphs just 200 times for each choice of n and p. Among the observed distances, we threw away the infinite ones (only a few though).

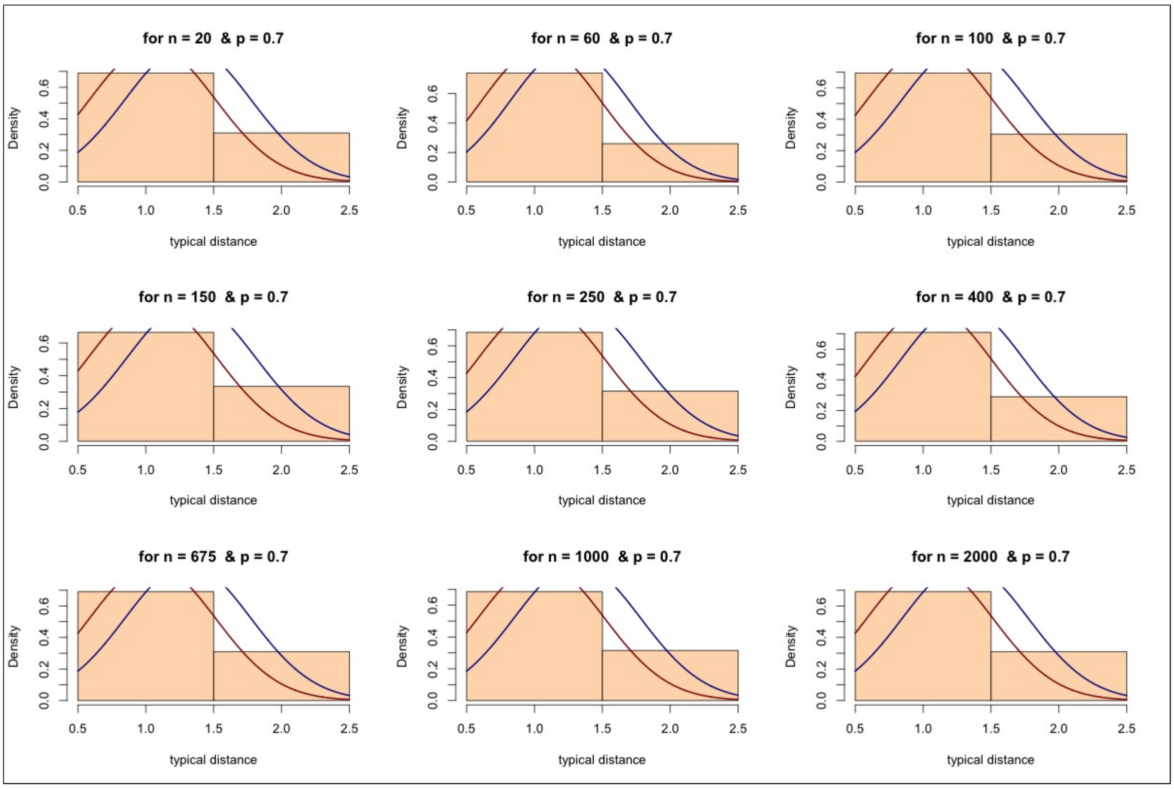
4.1 Histograms

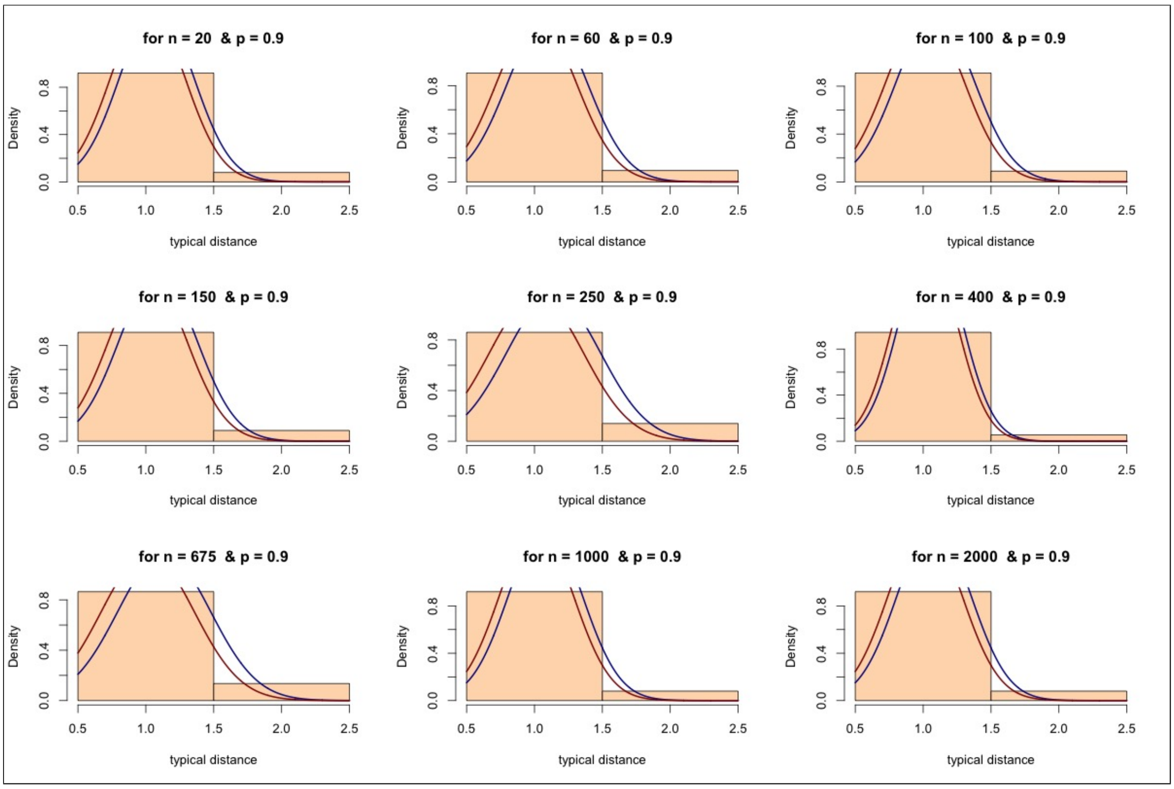
For each of values of p we plot the histogram of the observed (finite) typical distance in ER(n, p) for each choice of the graph size n. The blue line shows the normal density with mean and s.d. estimated by the sample mean and sample s.d. The red line shows the normal density with mean equal to $\log n / \log np$ and s.d. estimated by the sample s.d.




for n = 20 & p = 0.3
for n = 60 & p = 0.3
for n = 100 & p = 0.3
for n = 150 & p = 0.3
for n = 250 & p = 0.3
for n = 400 & p = 0.3
for n = 675 & p = 0.3
for n = 1000 & p = 0.3
for n = 2000 & p = 0.3
Density
typical distance
0.0
0.2
0.4
0.6
0.8
0.5
1.0
1.5
2.0
2.5
1
2
3
4


for n = 20 & p = 0.5
for n = 60 & p = 0.5
for n = 100 & p = 0.5
for n = 150 & p = 0.5
for n = 250 & p = 0.5
for n = 400 & p = 0.5
for n = 675 & p = 0.5
for n = 1000 & p = 0.5
for n = 2000 & p = 0.5
Density
typical distance
0.0
0.2
0.4
0.5
1.0
1.5
2.0
2.5
3.0
3.5





It is notable that the typical distance H_n in this case equals either 1 or 2 most of the times. This is actually not so surprizing if one observes the following.

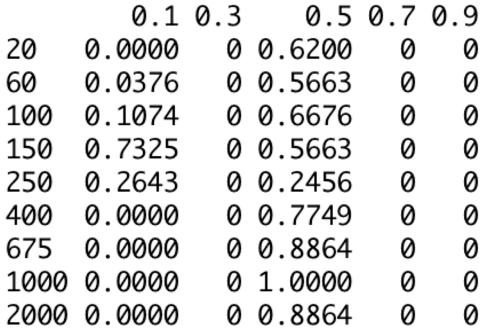
Observations: We have $\Pr(H_n > 2) = (1-p)(1-p^2)^{n-2}$. Clearly, this probability falls rapidly as n increases, for any $0 < p < 1$.

Proof. From the n vertices we are choosing 2 vertices, say X and Y, uniformly at random (without replacement). First let us condition upon the event that the chosen two vertices are (i, j) respectively. To compute the probability $\Pr(H_n > 2|X = i, Y = j)$, note that for this event to occur, the edge $i - j$ must not be present and for any k other than i and j, the path $i - k - j$ must not be present. Probability that the edge $i - j$ is absent is just $(1 - p)$; and the probability that the path $i-k-j$ is absent, is $(1-p^2)$. Note that there are $(n-2)$ many choices for this middle vertex k. Since any these paths and the edge $i - j$ are all formed independently in ER(n, p), it follows that $\Pr(H_n > 2|X = i, Y = j) = (1 - p)(1 - p^2)^{n-2}$. Since this holds for any $1 \leq i \neq j \leq n$, we conclude that $\Pr(H_n > 2) = (1 - p)(1 - p^2)^{n-2}$. □

4.2 Testing normality and symmetry

It is immediate from the histograms that the distribution of H_n for ER(n, p) where p is constant, is far from being close to normal. Indeed, if we perform the tests for normality as we did earlier, all the p-values we get are extremely small.

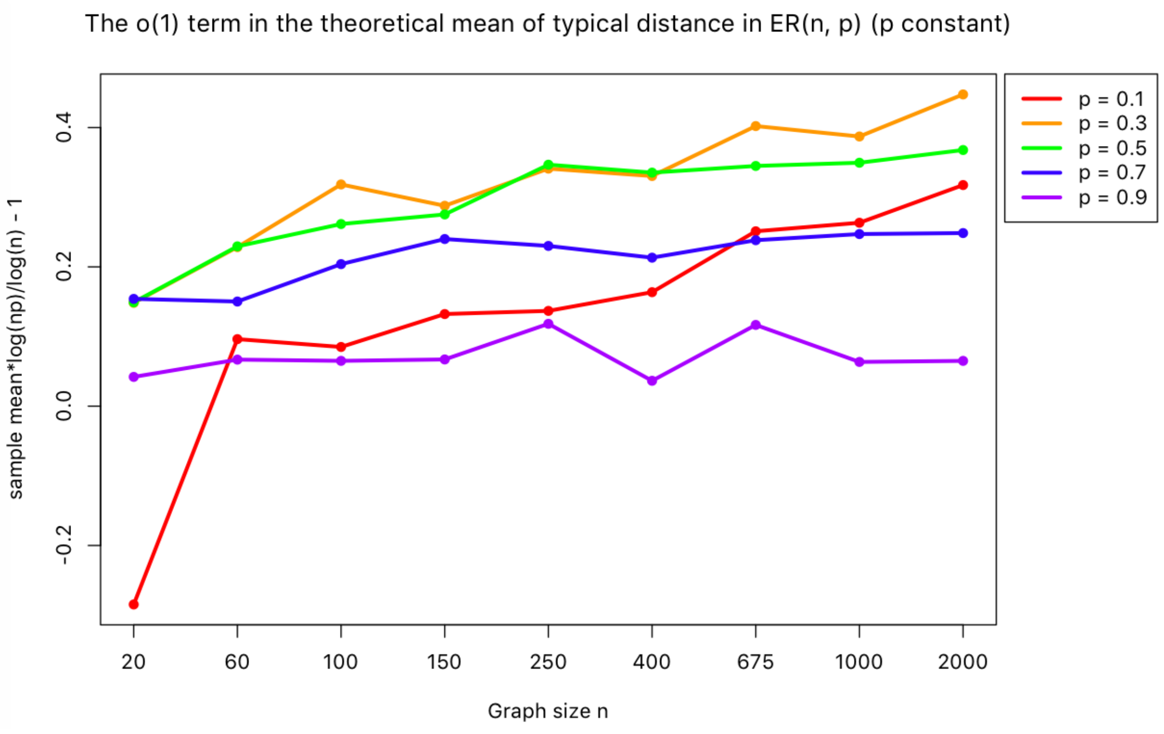
On the other hand, the p-values for the tests of symmetry are quite interesting. They are summarized in the following table.



```
        0.1 0.3    0.5 0.7 0.9
20   0.0000   0 0.6200   0   0
60   0.0376   0 0.5663   0   0
100  0.1074   0 0.6676   0   0
150  0.7325   0 0.5663   0   0
250  0.2643   0 0.2456   0   0
400  0.0000   0 0.7749   0   0
675  0.0000   0 0.8864   0   0
1000 0.0000   0 1.0000   0   0
2000 0.0000   0 0.8864   0   0
```

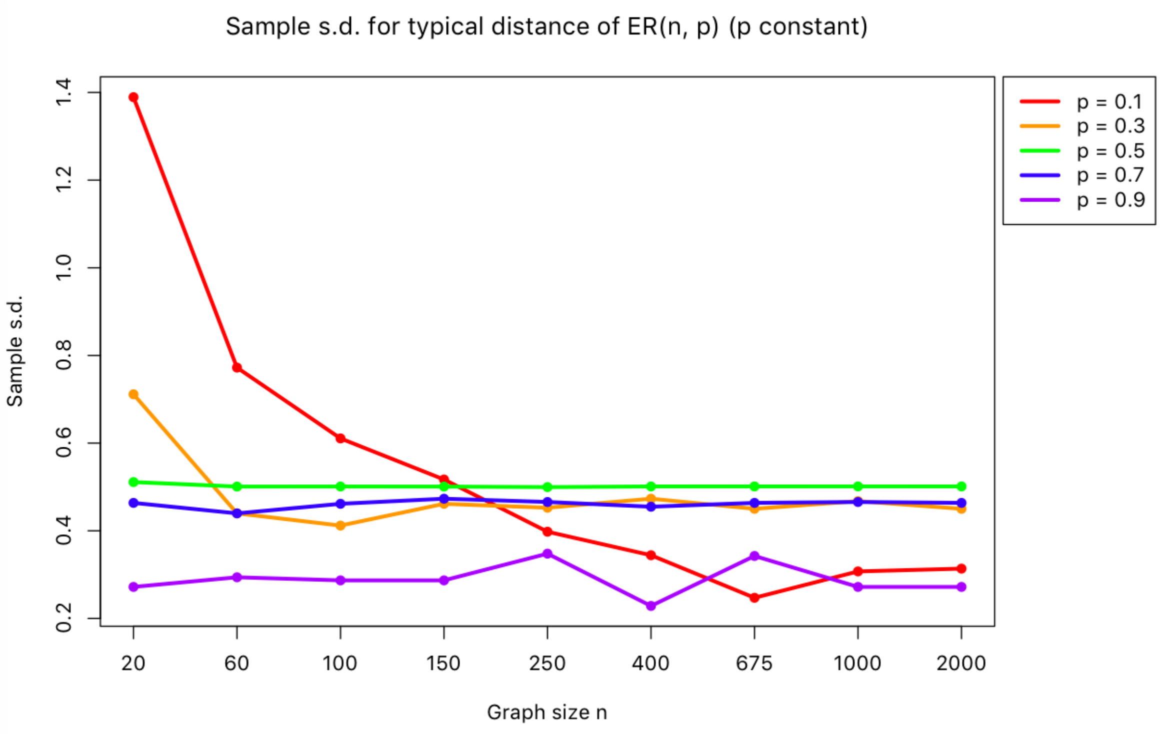
Note that the p-values are much higher for $p = 0.5$ than the others. Following is a plausible explanation for this pattern. We noted above that $\Pr(H_n > 2) = (1-p)(1-p^2)^{n-2}$, which falls off rapidly as n increases. This tells us that $\Pr(H_n = 1 \text{ or } 2) \approx 1$. But $\Pr(H_n = 1) = p$, which gives $\Pr(H_n = 2) \approx 1-p$. When $p = 1/2$, these two become (almost) equal, bringing symmetry in the distribution of H_n. When $p < 0.5$ or $p > 0.5$, one of the two sides gets heavier (compare the histograms for $p = 0.3$ and $p = 0.7$). What about the first column? Look at the histograms for $p = 0.1$. When n is in the middle range $(100 - 200)$, the symmetry in the histogram is due to the classes 1 and 3 having nearly equal frequencies. When n increases further, the class for 3 disappears, and the symmetry is lost.

4.3 Studying the $o(1)$ term


The o(1) term in the theoretical mean of typical distance in ER(n, p) (p constant)

Observations: We observe that the $o(1)$ term takes both positive and negative values (unlike the previous two cases) and the magnitudes are not as small as the connectivity regime. In fact for $p = 0.1$ and 0.3 we can see a slightly increasing pattern, which is strange.

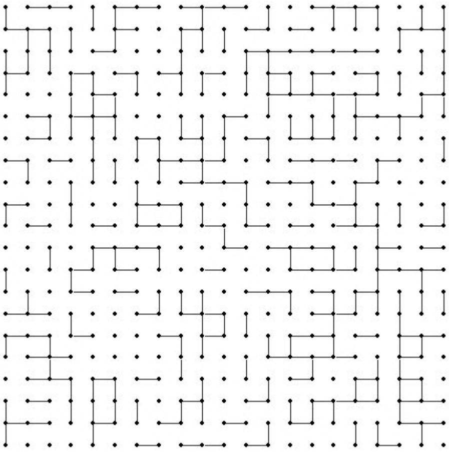
4.4 Studying the standard deviations


Sample s.d. for typical distance of ER(n, p) (p constant)

Observations: We can see that although the sample s.d. is pretty high for small n and $p = 0.1$, it quickly decreases for higher values of n. Surprizingly enough, the lines for p close to 0.5 seem to be much steady.

5 Studying the lattice random graph

For $n \in \mathbb{N}$ consider all the points with integer coordinates lying on the boundary or inside of the square with vertices (n, n), $(-n, n)$, $(-n, -n)$ and $(n, -n)$. In other words we consider all those points with coordinates (x, y) where x and y are integers satisfying $-n \leq x, y \leq n$. Clearly total number of points considered is $(2n + 1)^2$. A graph is created using these points as vertices. We join the vertices which are unit distance apart with an edge. This will form a square grid which we will consider as our complete lattice. Now we consider a subgraph of the complete lattice where each edge of the complete lattice occurs in the subgraph with probability p independently of occurrence of all other edges. We denote this model of random graph by Lat(n, p). A simulation from Lat$(10, 0.4)$ is shown below.

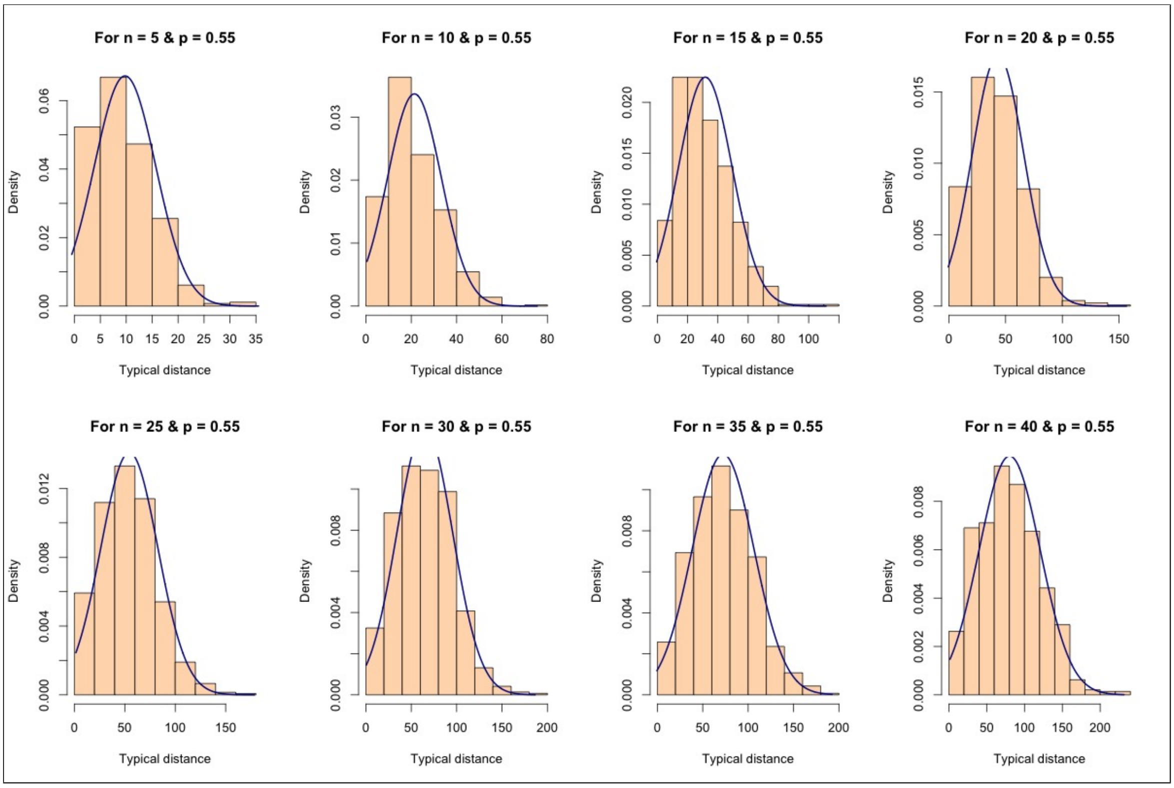


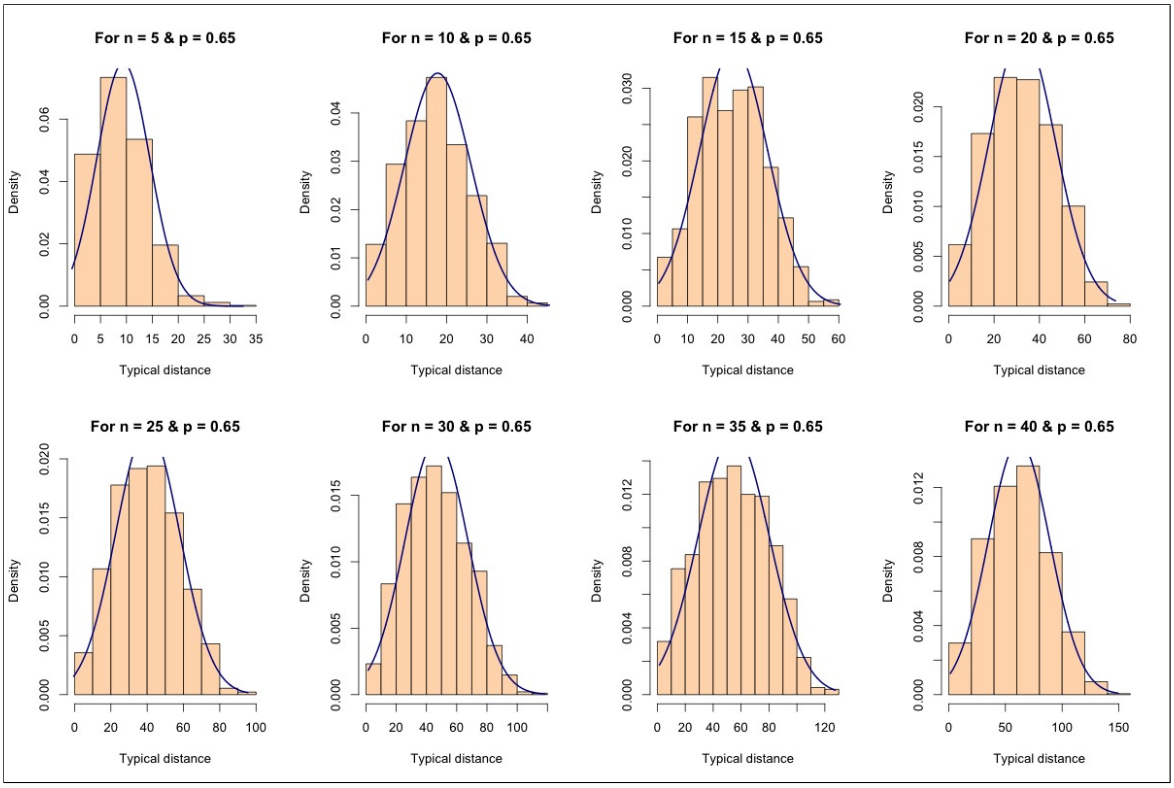
Here we will study the *typical distance* in Lat(n, p) which is same as the one studied for the Erdős-Rényi binomial random graph model, except one thing. We select two vertices randomly *with replacement* and observe their graph distance. If we select the same vertex twice, the distance is defined to be 0. Again we will denote this distance by H_n. Even if the vertices were selected *without replacement*, the results would have been same asymptotically. The connectivity threshold for Lat(n, p) is known to be $p = 1/2$. Here we studied H_n for $p > 1/2$.

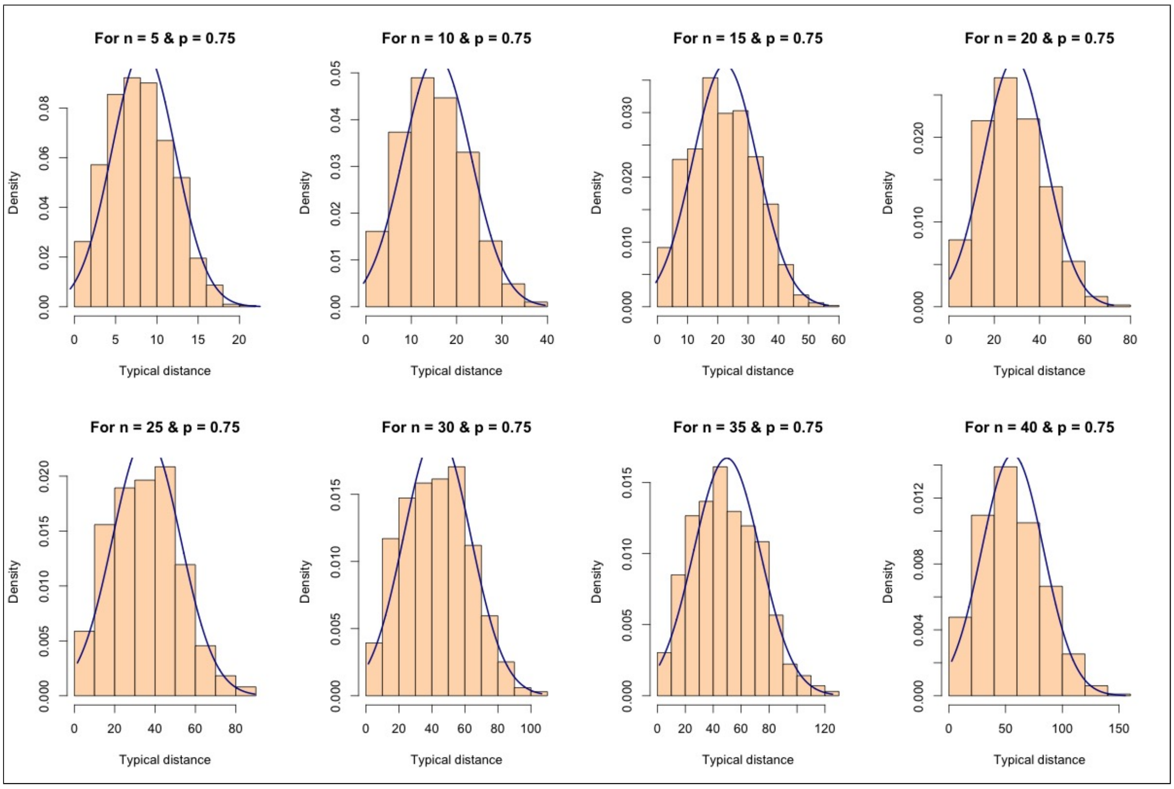
5.1 Histograms

We simulated from the model Lat(n, p) for 1000 times for each pair of (n, p) where $n = 5, 10, 15, \ldots, 40$ and $p = 0.55, 0.6, 0.65, \ldots, 0.95$. So the number of vertices in the graphs varied from $11^2 = 121$ to $81^2 = 6561$. We dropped the observations when the chosen vertices belonged to different connected components.

For each of the values $p = 0.55, 0.65, 0.75, 0.85, 0.95$ the histograms of the observed (finite) typical distances in Lat(n, p) for each choice of the graph size n are given below (histograms for the other values of p show similar pattern; those are omitted here). The blue line shows the normal density with mean and s.d. estimated by the sample mean and sample s.d.







For n = 5 & p = 0.75

For n = 10 & p = 0.75

For n = 15 & p = 0.75

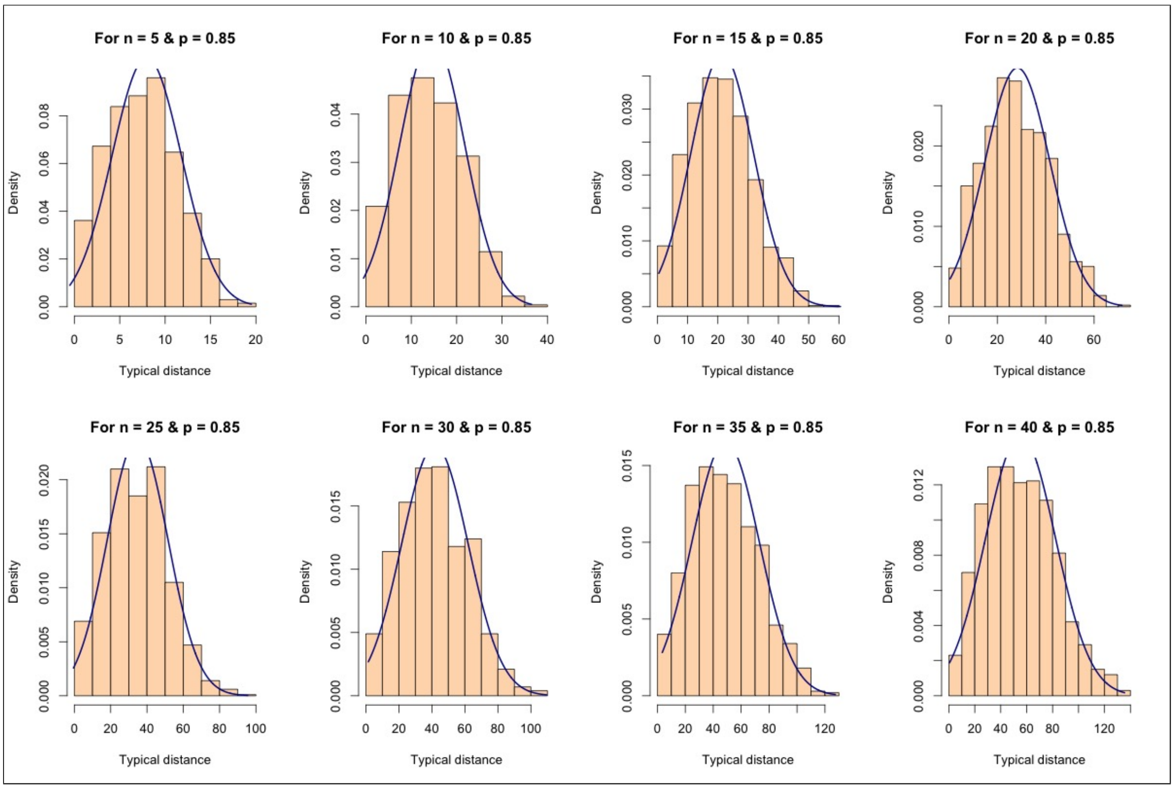
For n = 20 & p = 0.75

For n = 25 & p = 0.75

For n = 30 & p = 0.75

For n = 35 & p = 0.75

For n = 40 & p = 0.75



For n = 5 & p = 0.85

For n = 10 & p = 0.85

For n = 15 & p = 0.85

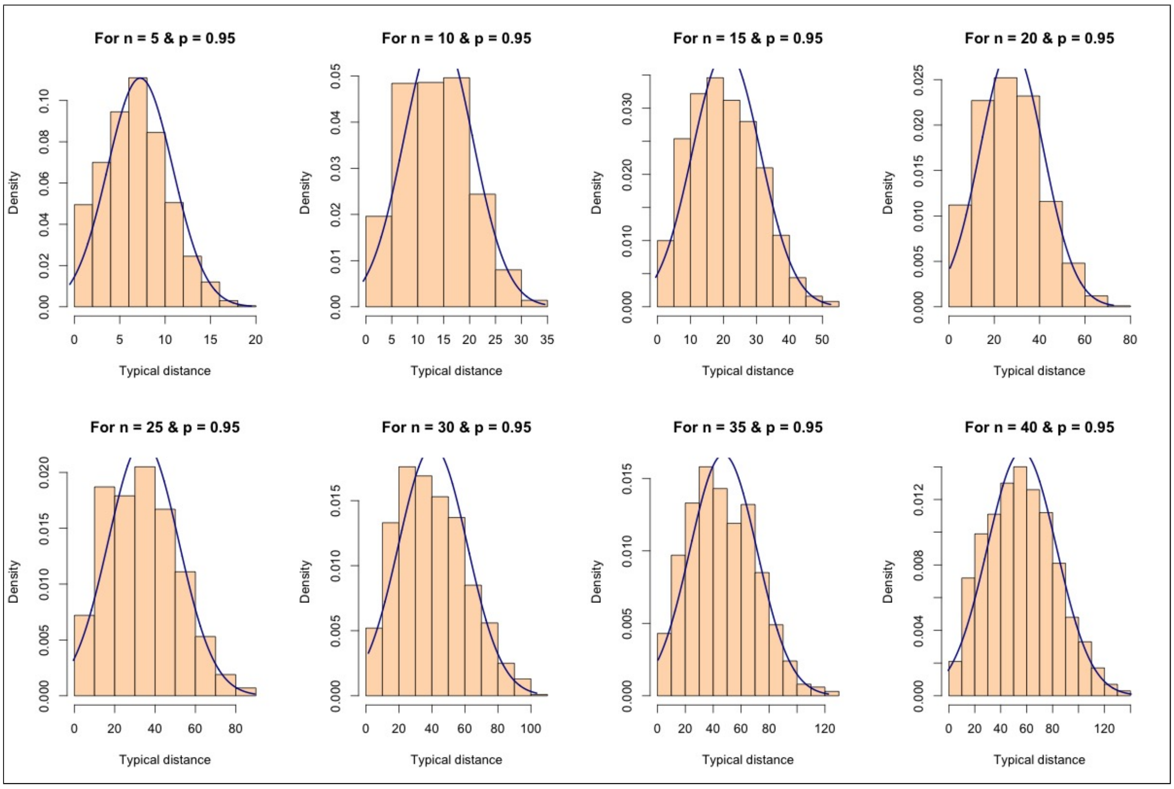
For n = 20 & p = 0.85

For n = 25 & p = 0.85

For n = 30 & p = 0.85

For n = 35 & p = 0.85

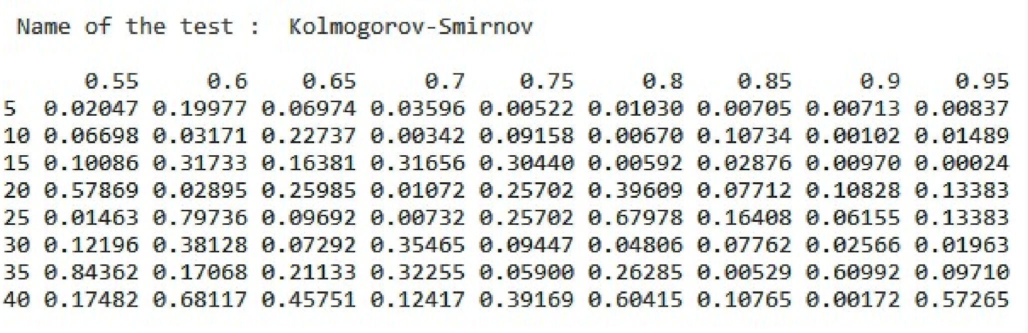
For n = 40 & p = 0.85



Observations: For smaller values of n and p, we notice that the left tail of the distribution is not visible. As n increases, the mode shifts rightward and hence the left tail becomes more and more visible. Hence the test of symmetry is more likely to get accepted for large n. Also, for large n, the histograms of H_n look more or less like the normal density (shown as a blue curve).

5.2 Testing normality

For each choices of (n, p) we performed the Kolmogorov-Smirnov test for Gaussianity on the data standardized by sample mean and sample s.d. To break the ties, we jittered the data by adding random noise to the observations. The noises are drawn from a normal distribution with mean 0 and s.d. 0.001.



```
 Name of the test :  Kolmogorov-Smirnov

      0.55     0.6    0.65     0.7    0.75     0.8    0.85     0.9    0.95
5  0.02047 0.19977 0.06974 0.03596 0.00522 0.01030 0.00705 0.00713 0.00837
10 0.06698 0.03171 0.22737 0.00342 0.09158 0.00670 0.10734 0.00102 0.01489
15 0.10086 0.31733 0.16381 0.31656 0.30440 0.00592 0.02876 0.00970 0.00024
20 0.57869 0.02895 0.25985 0.01072 0.25702 0.39609 0.07712 0.10828 0.13383
25 0.01463 0.79736 0.09692 0.00732 0.25702 0.67978 0.16408 0.06155 0.13383
30 0.12196 0.38128 0.07292 0.35465 0.09447 0.04806 0.07762 0.02566 0.01963
35 0.84362 0.17068 0.21133 0.32255 0.05900 0.26285 0.00529 0.60992 0.09710
40 0.17482 0.68117 0.45751 0.12417 0.39169 0.60415 0.10765 0.00172 0.57265
```

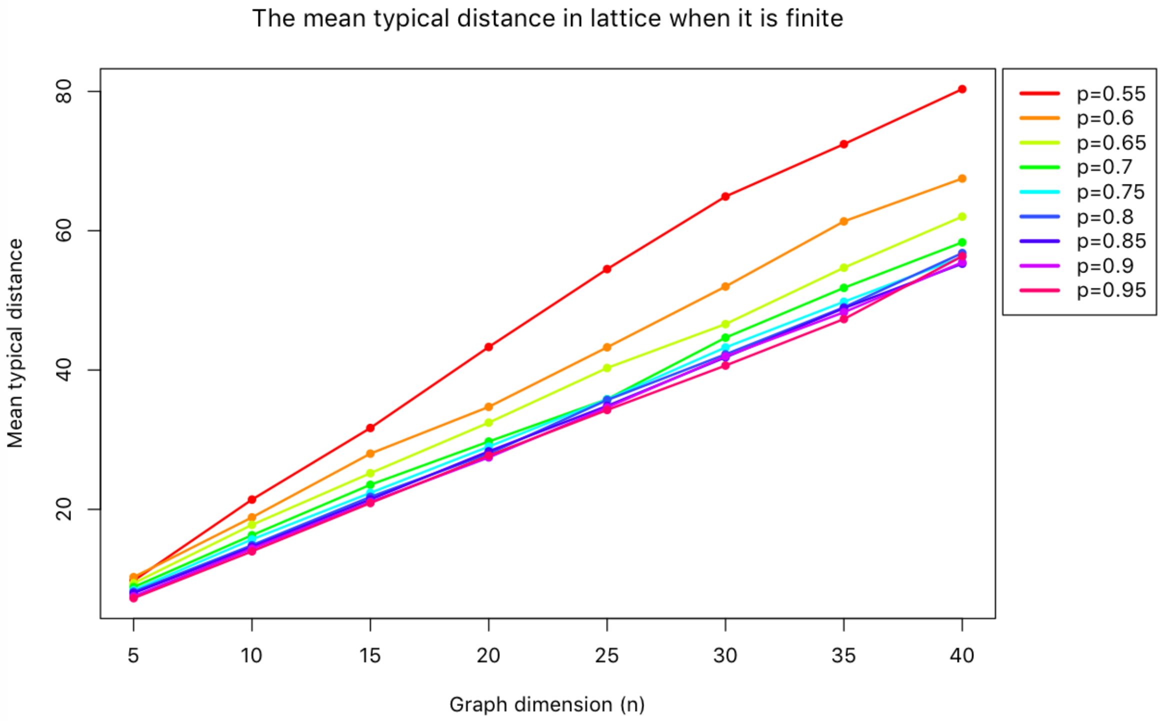
The lower left corner of the above table shows more acceptance of normality than the other corners. So if n increases, it is more likely that normality will get accepted by the test. If we fix n and take p close to 1 we may get deviations from normality due to large number of observations.

We also performed Pearson's chi-square goodness of fit test and Shapiro-Wilk test on standardized data for each pair (n, p). Since these tests are much more sensitive, we got almost all the p-values very close to 0. We are omitting those p-value tables.

Next let us take a closer look at the sample mean and s.d. for different choices of (n, p).

5.3 Studying the sample mean

The sample means of the typical distances are plotted against different values of n. For different choices of p, we have used different colours.



Observations: We notice that as p increases, keeping n fixed, the sample mean decreases in general, which is intuitive. The plot suggests that the sample mean grows more or less linearly with n. If p is taken closer to 1, the slope of the curve of sample mean vs. n approaches a constant which is close to $4/3$. This can be observed in the plot.

To explain this, we consider the case $p = 1$ where we have the complete lattice. We select two points randomly with replacement. Let the coordinates of the selected points are (X_1, Y_1) and (X_2, Y_2) respectively. Then observe that X_1, X_2, Y_1, Y_2 are i.i.d. observations from the uniform distribution on the set $\{-n, -n+1, \ldots, -1, 0, 1, \ldots, n-1, n\}$. Also $H_n = |X_1 - X_2| + |Y_1 - Y_2|$. We have:

$$\Pr(|X_1 - X_2| = k) = \begin{cases} \frac{1}{2n+1} & \text{when } k = 0. \\ \frac{2(2n+1-k)}{(2n+1)^2} & \text{when } k \in \{1, 2, \ldots, 2n\}. \end{cases}$$

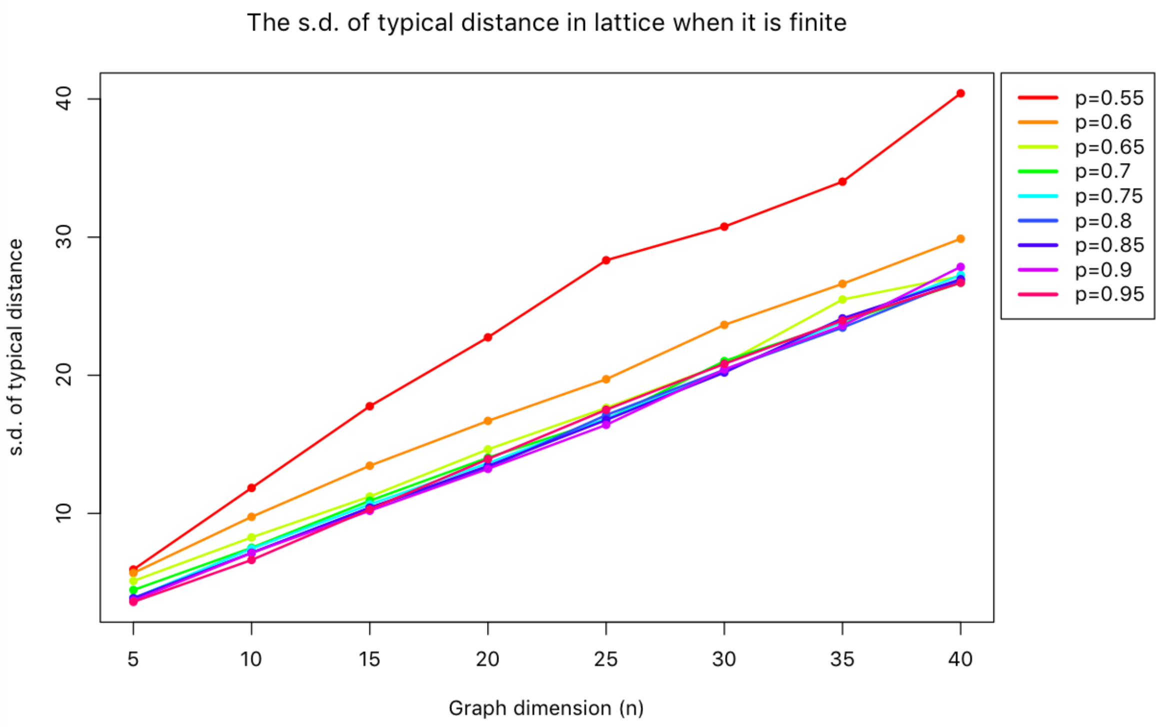
This implies

$$\mathrm{E}(|X_1 - X_2|) = \frac{4n(n+1)}{3(2n+1)} \implies \mathrm{E}(H_n) = 2\mathrm{E}(|X_1 - X_2|) = \frac{8n(n+1)}{3(2n+1)}.$$

Thus $\mathrm{E}(H_n)/n \to 4/3$ as $n \to \infty$. This agrees with what we observed in the above plot.

5.4 Studying the standard deviation

The sample s.d.'s of the typical distances are plotted against different values of n. For different choices of p, we have used different colours.



Observations: We notice that as p increases, keeping n fixed, the sample s.d.'s decreases in general. The plot suggests that the sample s.d. grows more or less linearly with n. If p is taken closer to 1, the slope of the curve of sample mean vs. n approaches a constant which is close to $2/3$. This can be observed in the plot.

Again, to explain this, we take $p = 1$. Define X_1, X_2, Y_1, Y_2 as before. Since X_1, X_2, Y_1, Y_2 are independent, $\mathrm{Var}(H_n) = 2\mathrm{Var}(|X_1 - X_2|)$. From the p.m.f. of $|X_1 - X_2|$, we can show that

$$\mathrm{E}(|X_1 - X_2|^2) = \frac{2n(n+1)}{3} \implies \mathrm{Var}(H_n) = 2\mathrm{Var}(|X_1 - X_2|) = \frac{4n(n+1)(4n^2+4n+3)}{9(2n+1)^2}.$$

Now it is easy to see that $\mathrm{Var}(H_n)/n^2 \to 4/9$, i.e., $\mathrm{sd}(H_n)/n \to 2/3$, as $n \to \infty$.

6 Concluding remarks

Given limited time and resources, there are lots of things that could not be done. However, we think, the following directions might serve as good future scopes of this project.

1. The $o(1)$ term in [1] is quite ambiguous. More work needs to be done, specially regarding the rate at which it shrinks.

2. The sample s.d. of the typical distance has not been studied in great detail, we suggest that it should be studied in a comprehensive manner.

3. Throughout this project we tried to find patterns about how different rates or characteristics

depend on the parameter c, for all the cases ($p = c/n, c \log n/n$, or just c). Our simulation study suggests that the effects of c on the $o(1)$ term and s.d. need further study.

4. In the sparse but super-critical regime, we suspect from the plots that there is a phase transition at $c = np = 2$. The rates of decay seem to be different for values of c close to 1 and values of c greater than 2.

5. The histograms in connectivity regime or constant probability case showed a discrete nature, hence rejecting the hypothesis of normality with very low p-values. We think that lots of work remain for understanding the limiting distribution of the typical distance as $n \to \infty$.

6. For Lat(n, p), our simulations tried to reveal the nature of H_n when $n \to \infty$ and $p \to 1$. The normality tests were rejected most of the times. Nothing could be said from here about H_n if we fix $p \in (1/2, 1)$ and let $n \to \infty$. The limiting distribution of H_n for fixed p is yet to be studied.

7. We have only worked for square lattices. A similar study could be done for typical distances in triangular or hexagonal lattices or even higher dimensional lattice structures.

References

[1] Chung, F., & Lu, L. (2002). The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25), 15879-15882.

[2] Van Der Hofstad, R. (2009). Random graphs and complex networks. *Available on http://www.win.tue.nl/rhofstad/NotesRGCN.pdf*, 11.

[3] *Link for Github Repository:* https://github.com/ghoshadi/Random-Graphs/