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

Aditya Ghosh

A joint project with Sayak Chatterjee,
under the supervision of Prof. Antar Bandyopadhyay

June 18, 2020

Introduction

- Consider the Erdős-Rényi binomial random graph model, which we shall denote by $\mathrm{ER}(n, p)$.
- The object of our study is the *typical distance* H_n in $\mathrm{ER}(n, p)$, which is defined as the graph-distance between any two randomly selected vertices.
- It is well-known that a graph $G \sim \mathrm{ER}(n, p)$ is connected with high probability when p is above the connectivity threshold $\log n/n$.
- In the sparse but super-critical regime ($p = c/n$, $c > 1$), 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.

Introduction

- Suppose that $G \sim \mathrm{ER}(n,p)$. The *average distance* of G is defined as the average of all distances $d(u,v)$ for pairs of u and v which belong 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.

Theorem 1. (Chung and Lu, 2002)

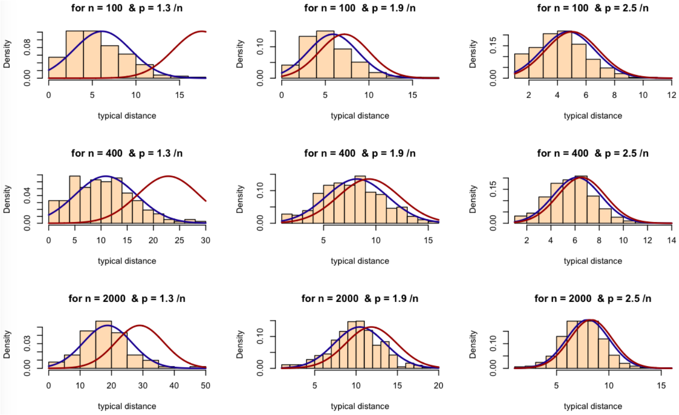
If $G \sim \mathrm{ER}(n,p)$ where $np \geq c > 1$ for some constant c, then almost surely the average distance in G is $(1+o(1))(\log n/\log np)$, provided that $(\log n/\log np)$ goes to infinity as $n \to \infty$.

Introduction

- We simulated $\mathrm{ER}(n, p)$ 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.
- Our major observation from the simulations is that the $o(1)$ term in the theorem is quite ambiguous, both in the sparse regime and the connectivity regime.
- We also studied how the standard deviation of the typical distance varies and performed tests for normality and symmetry of the distribution of the typical distance.

Histograms

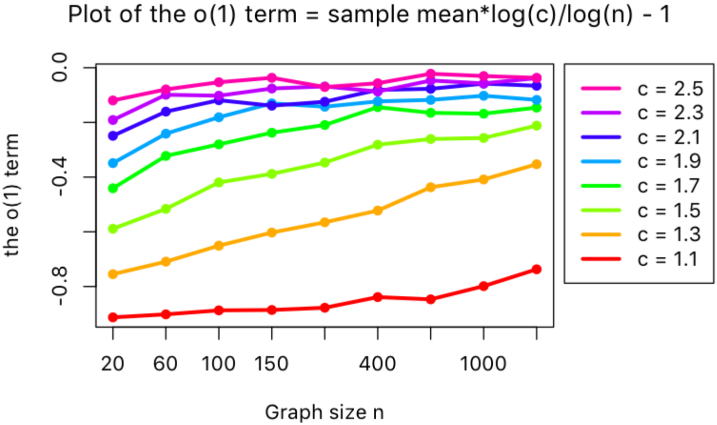
We simulated ER$(n, c/n)$ for each pair of (c, n) where $c = 1.1, 1.3, \ldots, 2.5$ and various values of n. A sampling of the histograms are shown below.



The red and blue curves are normal densities with means $\log n / \log np$ and the sample mean respectively, and s.d. equal to the sample s.d.

The o(1) term

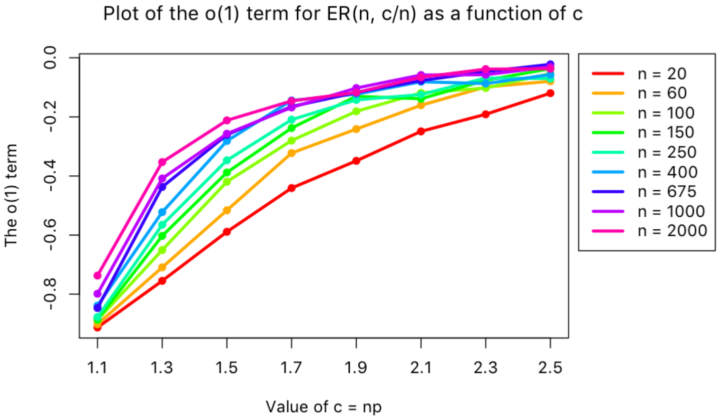
The following diagram shows the o(1) term as a function of *n*.



Observe that the $o(1)$ term falls at a faster rate for higher values of c, especially when $c > 2$.

The o(1) term

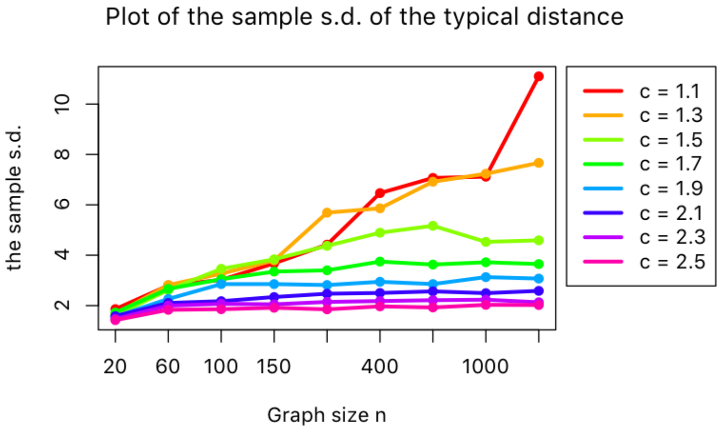
The following diagram shows the o(1) term as a function of c.



We can see that for higher values of c, the o(1) term do not differ much with the graph size n.

The standard deviations

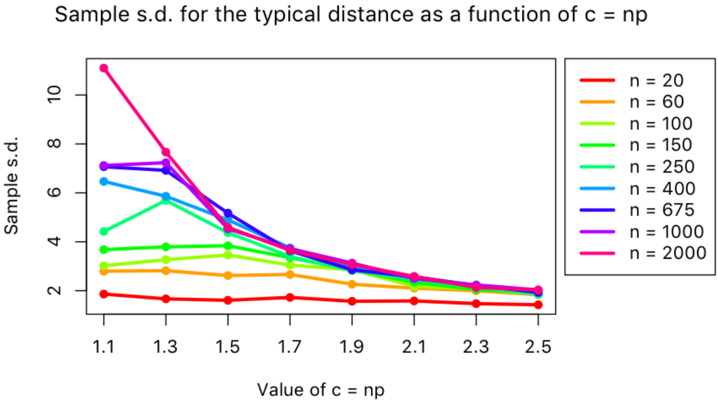
The following diagram shows the standard deviations as a function of *n*.



We can see a steady increase for $c = np > 2$, while for the values of c near 1, the s.d. increases much rapidly.

The standard deviations

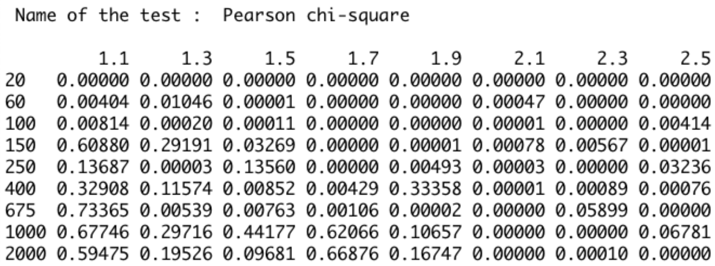
The following diagram shows the standard deviations as a function of c.



Again, we can see that for higher values of c, the standard deviations do not differ much with the graph size n.

Testing Normality

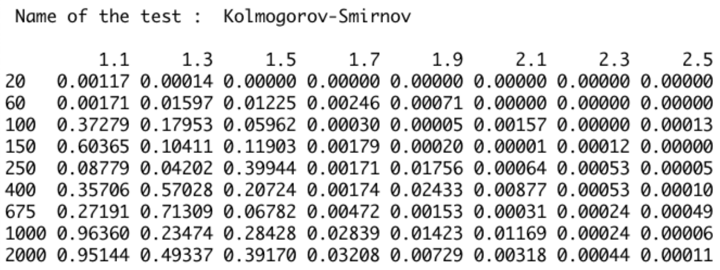
- First we had a look at the histograms for the standardized data and Q-Q plots, for a visual assessment. Then on the standardized samples, we performed Pearson's χ^2 goodness of fit test, Kolmogorv-Smirnov test, and Shapiro-Wilk test for each pair of c and n.
- The table of the p-values for each of these tests showed the following trend: the p-values get bigger towards the lower left corner of the table, which corresponds to small values of c and large values of n.



```
 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
```

Testing Normality



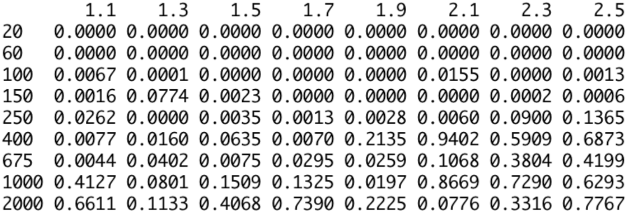
```
 Name of the test :  Kolmogorov-Smirnov

           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
```

- Surprizingly, when n and c both large, despite the standardized histograms being close enough (at least visually) to the standard normal density, 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.

Testing Symmetry

- We performed the Randles-Fligner-Policello-Wolfe test of symmetry on the standardized data. The p-values are shown in the following table.

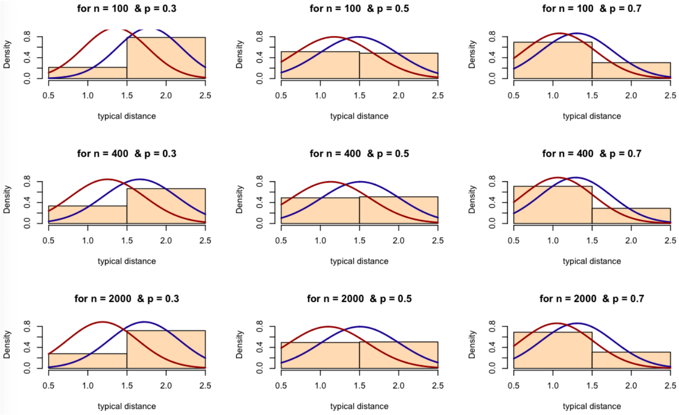


```
        1.1    1.3    1.5    1.7    1.9    2.1    2.3    2.5
20   0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
60   0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
100  0.0067 0.0001 0.0000 0.0000 0.0000 0.0155 0.0000 0.0013
150  0.0016 0.0774 0.0023 0.0000 0.0000 0.0000 0.0002 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
```

- A general pattern here is that the p-values tend to increase as we move towards the bottom. This pattern is also visible from the the standardized histogram plots, the histograms are more symmetric around the mean for larger values of n.
- It is really strange that some of the samples for larger n and smaller c (e.g. $c = 1.1, n = 675$) are accepted with high p-values in the tests of normality but rejected poorly in tests of symmetry.

Histograms

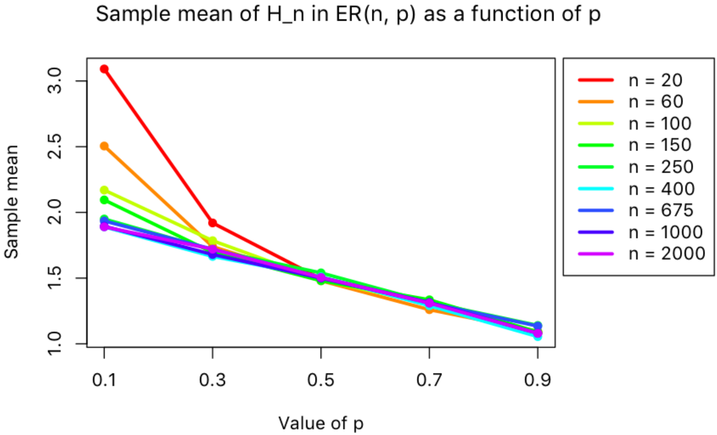
We took $p = 0.1, 0.3, \ldots, 0.9$, and various values of n. A small sampling of the histograms are shown below.



We noted that the typical distance H_n in this case is either 1 or 2 most of the times, which is understandable since $\mathbb{P}(H_n > 2) = (1-p)(1-p^2)^{n-2}$.

The sample mean

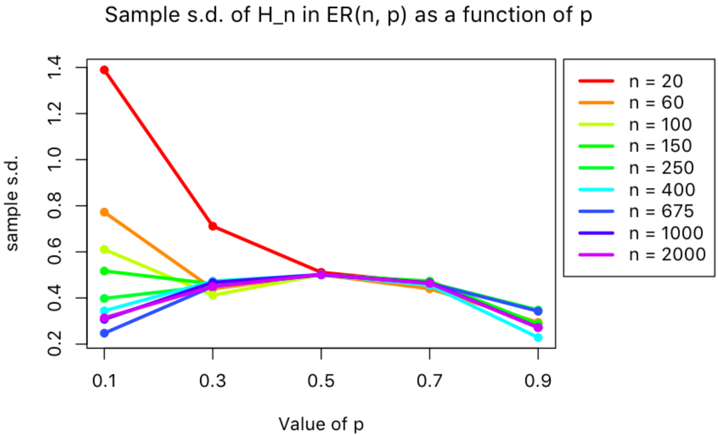
The following diagram shows the sample mean as a function of p.



This curve shows that for a fixed p, and sufficiently large n, the sample mean is close to $2 - p$, which is just $\mathrm{E}[1 + X]$ where $X \sim \mathrm{Ber}(1 - p)$.

The standard deviations

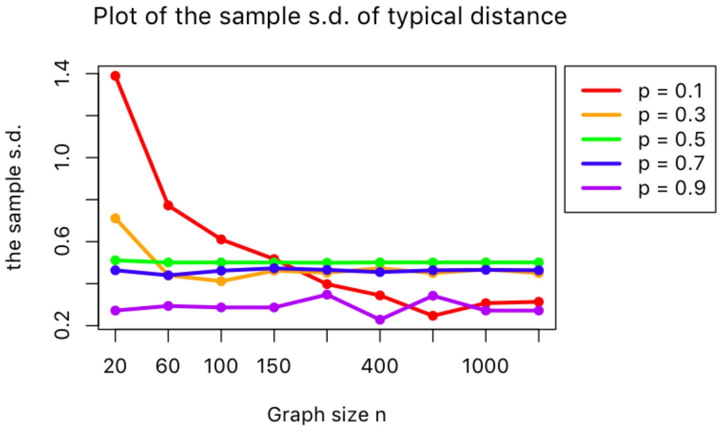
The following diagram shows the standard deviations as a function of p.



Again, for large enough n, the distribution of H_n being close to $1 + X$ where $X \sim \text{Ber}(1-p)$, its standard deviation is close to $\sqrt{p(1-p)}$.

The standard deviations

The following diagram shows the standard deviations as a function of n.



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. The sample s.d. for p close to 0.5 seems to be much steady.

Testing Normality and Symmetry

- It is immediate from the histograms that the distribution of H_n in this case is not anywhere close to normal. Indeed, all the p-values we get for the tests for normality are extremely small.
- The p-values for the RFPW test of symmetry are summarized below.

```
        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
```

- We noted above that even for moderately large n, we have $\mathbb{P}(H_n = 1) = p$, and $\mathbb{P}(H_n = 2) \approx 1 - p$. 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.

Concluding remarks

- The $o(1)$ term in theorem 1 is quite ambiguous. More work needs to be done, specially regarding the rate at which it shrinks.
- The sample s.d. of the typical distance has not been studied in great detail, we suggest a comprehensive study of that.
- Throughout this project we tried to find patterns about how different rates depend on the parameter c. Our simulations suggest that the effects of c on the $o(1)$ term and s.d. needs further study.
- In the sparse but super-critical regime, we suspect from the plots that there is a phase transition at $c = 2$. The rates of decay seem to be different for values of c close to 1 and values of c greater than 2.
- We think that lots of work remain for understanding the limiting distribution of the typical distance as $n \to \infty$.

References

- 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.
- Van Der Hofstad, R. (2009). Random graphs and complex networks. *Available on http://www.win.tue.nl/rhofstad/NotesRGCN.pdf*, 11.
- *Link for Github Repository:* https://github.com/ghoshadi/random-graphs/

Thank you for your attention!