CS-E5745 Mathematical Methods for Network Science, Final Project Report

Petri Leskinen, 42638C

August 11, 2019

1 Introduction

In the article *"Simple and accurate analytical calculation of shortest path lengths"*[MG16] Sergey Melnik and James P. Gleeson present an analytical approach for estimating the distribution of shortest path lengths of a random graph. The approach is based on susceptible-infected epidemic model where at a initial step only one node is infected. As the epidemic spreads, the distribution of shortest path lengths can be estimated from the fractional amount of infected nodes.

In the article after the introduction first the analogy between the distribution of shortest path lengths and the infection time is presented in section II. In section III the approach with z-regular random graphs is explained and the involved recurrence relations as well as the analytical solution is presented. In section IV the model is generalized to networks with arbitrary degree distribution. However in this more general case the recurrence relations do not have an analytical solution. Section V describes how the model can be adapted to random networks with joint degree-degree distribution. Most of the derived calculations in this report focuses on the result of section III described in detail in Appendix A, but also the calculations of sections IV and V are discussed. In the appendices of the article also similar models by other researches are explained, but these calculations are not explained in this report.

2 Calculations for z-regular random graphs

As the epidemic spreads in a susceptible-infected model the distribution of vertex distances can be calculated from the difference of consecutive infection ratios according to the Eq. (1):

$$D_n = \rho_n - \rho_{n-1} \tag{1}$$

and the average path length is the expected value $\bar{D} = \sum nD_n$ which could also be written using terms ρ_n: $\bar{D} = n \cdot \rho_n - \sum_{k=0}^{n-1} \rho_k$, where n is the time step when epidemic has reach all nodes in the graph. At the beginning of the process, it is assumed that only one of nodes is infected, so the initial value ρ_0 is:

$$\rho_0 = \frac{1}{N} \tag{2}$$

Similarly, in a z-regular network there are z nodes which have 1 infected neighbor, each with contribution of $1/z$. So the ratio of infected neighbors of a node q_0 is:

$$q_0 = \frac{1}{z} \cdot \frac{z}{N} = \frac{1}{N} \tag{3}$$

The probability that at a time step $n+1$ a node is susceptible if it was originally susceptible which has the probability of $1 - \rho_0$. We only require that all its z neighbors are also susceptible each with probability $1 - q_n$. Therefore the ratio of susceptible nodes is $(1 - \rho_0)(1 - q_n)^z$ and as a complement the ratio of infected nodes is

$$\rho_{n+1} = 1 - (1 - \rho_0)(1 - q_n)^z \tag{4}$$

The formula for the ratio of infected neighbor nodes is similar with the difference of having $z - 1$ in exponent: if a susceptible node A has a susceptible neighbor B it can only have up to $z - 1$ possible infected neighbors.

$$q_n = 1 - (1 - \rho_0)(1 - q_{n-1})^{z-1} \tag{5}$$

Recurrence relations (4)-(5) are solved in Appendix A of [MG16], which we will now derive in detail. When applying the substitutions $y_n = 1 - q_n$ and $y_0 = 1 - p_0 (= 1 - q_0)$, Eq (5) becomes:

$$y_n = y_0 \cdot (y_{n-1})^{z-1} \tag{6}$$

which by applying natural logarithm on both sides and then by making a substitution $u_n = \ln y_n$ becomes:

$$\ln y_n = \ln y_0 + (z - 1) \cdot \ln(y_{n-1}) \qquad \Rightarrow \qquad u_n = (z - 1) \cdot u_{n-1} + u_0 \tag{7}$$

This non-homogeneous linear recurrence relation can be solved using linear algebra:

$$\boldsymbol{x}_n = \boldsymbol{A}\boldsymbol{x}_{n-1} \qquad \Rightarrow \qquad \boldsymbol{x}_n = \boldsymbol{A}^n\boldsymbol{x}_0 \tag{8}$$

Relation is of first order, and we will set the constant term u_0 as a second row of u_n so $\boldsymbol{A}$ is a 2×2 matrix. In a written out format the first part of Eq. (8) becomes:

$$\boldsymbol{x_n} = \boldsymbol{A}\boldsymbol{x_{n-1}} \Rightarrow \begin{bmatrix} u_n \\ u_0 \end{bmatrix} = \begin{bmatrix} z-1 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} u_{n-1} \\ u_0 \end{bmatrix} \tag{9}$$

To solve $\boldsymbol{A}^n$, we derive the eigendecomposition of $\boldsymbol{A}$ by first solving the eigenvalues λ_n of $\boldsymbol{A}$ from the eigenvalue equation $\boldsymbol{A}\lambda = \boldsymbol{v}\lambda$:

$$\boldsymbol{A}\lambda = \boldsymbol{v}\lambda \Rightarrow \det(\boldsymbol{A} - \boldsymbol{I}\lambda) = \det \begin{vmatrix} (z-1)-\lambda & 1 \\ 0 & 1-\lambda \end{vmatrix} = 0 \Rightarrow \tag{10}$$

$$((z-1)-\lambda)(1-\lambda) = 0 \Rightarrow \begin{cases} \lambda_1 = 1 \\ \lambda_2 = z-1 \end{cases}$$

Next the corresponding eigenvectors, first $\boldsymbol{v}_1$ for $\lambda_1 = 1$:

$$\boldsymbol{A}\boldsymbol{v}_1 = \lambda_1 \boldsymbol{v}_1 \Rightarrow \boldsymbol{A} = \boldsymbol{v}_1 \Rightarrow \begin{bmatrix} z-1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \Rightarrow \begin{cases} (z-2)x_1 + y_1 = 0 \\ y_1 = y_1 \end{cases} \tag{11}$$

$$\boldsymbol{v_1} = \begin{bmatrix} \frac{1}{2-z} \\ 1 \end{bmatrix}$$

And $\boldsymbol{v}_2$ for $\lambda_2 = z-1$:

$$\boldsymbol{A}\boldsymbol{v}_2 = \lambda_2 \boldsymbol{v}_2 \Rightarrow \boldsymbol{A}\boldsymbol{v}_2 = (z-1)\boldsymbol{v}_2 \Rightarrow \begin{bmatrix} z-1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} (z-1)x_2 \\ (z-1)y_1 \end{bmatrix} \Rightarrow \begin{cases} y_2 = (z-1)y_2 \\ y_2 = 0 \end{cases} \tag{12}$$

$$\boldsymbol{v_2} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Now the eigendecomposition $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^{-1}$ where $\boldsymbol{C} = [\boldsymbol{v}_1|\boldsymbol{v}_2]$, $\boldsymbol{C}^{-1} = [\boldsymbol{v}_1|\boldsymbol{v}_2]^{-1}$, and $\boldsymbol{D}$ has the eigenvalues on diagonal becomes

$$\begin{bmatrix} z-1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2-z} & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & z-1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & \frac{1}{z-2} \end{bmatrix} \tag{13}$$

Using the eigendecomposition we can convert the matrix power to $\boldsymbol{A}^n = (\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^{-1})^n = \boldsymbol{C}\boldsymbol{D}^n\boldsymbol{C}^{-1}$ and we will write Eq. (9) as

$$\boldsymbol{x}_n = \boldsymbol{A}^n \boldsymbol{x}_0 \Rightarrow \begin{bmatrix} u_n \\ u_0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2-z} & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1^n & 0 \\ 0 & (z-1)^n \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & \frac{1}{z-2} \end{bmatrix} \cdot \begin{bmatrix} u_0 \\ u_0 \end{bmatrix} \tag{14}$$

Simplifying Eq. (14) gives

$$\boldsymbol{x}_n = \boldsymbol{A}^n \boldsymbol{x}_0 \Rightarrow \begin{bmatrix} u_n \\ u_0 \end{bmatrix} = \begin{bmatrix} u_0(z-1)^n + u_0\frac{(z-1)^n - 1}{z-2} \\ u_0 \end{bmatrix} \Rightarrow u_n = \Big((z-1)^n + \frac{(z-1)^n - 1}{z-2}\Big)u_0 \tag{15}$$

The coefficient of u_0 can be simplified:

$$(z-1)^n + \frac{(z-1)^n - 1}{z-2} = \frac{(z-1)^n((z-1)-1)}{z-2} + \frac{(z-1)^n - 1}{z-2} = \tag{16}$$

$$\frac{(z-1)^{n+1}-(z-1)^n+(z-1)^n-1}{z-2}=\frac{(z-1)^{n+1}-1}{z-2}\Rightarrow$$
$$u_n=\frac{(z-1)^{n+1}-1}{z-2}\cdot u_0$$

Notice also that in fact the coefficient is the sum of geometric series $\sum_{k=0}^{n}(z-1)^k$ and we made a simplification $(z-1)^n+\sum_{k=0}^{n-1}(z-1)^k=\sum_{k=0}^{n}(z-1)^k$, e.g. added one term to the series. The same result could have been derived also by induction, e.g. by starting with u_0 and writing the sequence u_1, u_2 and by concluding the general term u_n.

To continue with the equation substituting $u_n=\ln y_n\Rightarrow y_n=e^{u_n}$ gives:

$$y_n=\exp(u_n)=\exp\Big[\frac{(z-1)^{n+1}-1}{z-2}\cdot u_0\Big]=\exp\Big[\frac{(z-1)^{n+1}-1}{z-2}\cdot\ln y_0\Big]\tag{17}$$

Finally replacing $y_0=1-\rho_0$ gives the same result as equation A2 in the article:

$$y_n=\exp\Big[\frac{(z-1)^{n+1}-1}{z-2}\ln(1-\rho_0)\Big].\tag{18}$$

With this result we can get back to Eq. (1) and first write it using Eq. (4):

$$D_n=\rho_n-\rho_{n-1}=\big(1-(1-\rho_0)(1-q_{n-1})^z\big)-\big(1-(1-\rho_0)(1-q_{n-2})^z\big)=$$
$$(1-\rho_0)\big((1-q_{n-2})^z-(1-q_{n-1})^z\big)=(1-\rho_0)\big(y_{n-2}^z-y_{n-1}^z\big)\tag{19}$$

which is equal to equation A3 in [MG16]. Next we will simplify the term $(1-\rho_0)y_n^z$ when the result (18) is plugged in:

$$(1-\rho_0)y_n^z=(1-\rho_0)\exp\Big[z\cdot\frac{(z-1)^{n+1}-1}{z-2}\ln(1-\rho_0)\Big]\Rightarrow$$
$$\exp(\ln(1-\rho_0))\cdot\exp\Big[z\cdot\frac{(z-1)^{n+1}-1}{z-2}\ln(1-\rho_0)\Big]=$$
$$\exp\Big[\Big(\frac{z(z-1)^{n+1}-z}{z-2}+1\Big)\ln(1-\rho_0)\Big]=\exp\Big[\Big(\frac{z(z-1)^{n+1}-z+z-2}{z-2}\Big)\ln(1-\rho_0)\Big]\Rightarrow$$
$$(1-\rho_0)y_n^z=\exp\Big[\Big(\frac{z(z-1)^{n+1}-2}{z-2}\Big)\ln(1-\rho_0)\Big]\tag{20}$$

This result is now applied to Eq.(19):

$$D_n=(1-\rho_0)y_{n-2}^z-(1-\rho_0)y_{n-1}^z\Rightarrow$$
$$D_n=\exp\Big[\Big(\frac{z(z-1)^{n-1}-2}{z-2}\Big)\ln(1-\rho_0)\Big]-\exp\Big[\Big(\frac{z(z-1)^{n}-2}{z-2}\Big)\ln(1-\rho_0)\Big]\tag{21}$$

which is equal to equation A4 in [MG16]. Next in the article an approximation $\ln(1-x) = -x$ is made. This can be justified by looking at the power series of $\ln(1+x)$, which can be derived by integrating the sum for geometric series:

$$\frac{1}{1-t} = \sum_{k=0}^{\infty} t^k = 1 + t + t^2 + t^3 + ..., \text{ for } |t| < 1 \Rightarrow$$

$$\int_0^x \frac{1}{1-t}\,\mathrm{d}t = \int_0^x \sum_{k=0}^{\infty} t^k\,\mathrm{d}t \Rightarrow -\ln(1-t)\Big|_{t=0}^{x} = \sum_{k=0}^{\infty} \frac{t^{k+1}}{k+1}\Bigg|_{t=0}^{x} \Rightarrow$$

$$\ln(1-x) = -x - \frac{1}{2}x^2 - \frac{1}{3}x^3 + ...$$

Now with small values $0 < x \ll 1$, $x^2 \approx 0$, $x^3 \approx 0$, ... and we can approximate

$$\ln(1-x) \approx -x.$$

This approximation is used for the term $\ln(1-\rho_0) = \ln(1-\frac{1}{N}) = -\frac{1}{N}$. Now Eq. (21) becomes

$$D_n^{\mathrm{RRG}} = \exp\Big[-\frac{z(z-1)^{n-1}-2}{(z-2)N}\Big] - \exp\Big[-\frac{z(z-1)^{n}-2}{(z-2)N}\Big]. \tag{22}$$

This results corresponds to equation 4 in the article. The index RRG means Random Regular Graph.

3 Calculations for generalization to networks with arbitrary degree distribution

In the case of random networks with arbitrary degree distribution the authors of the article consider that the epidemic starts from a single degree-k' node. The initial values from excepted fraction of $\rho_0^{k,k'}$ and $q_0^{k,k'}$ are defined with Eq. (23)

$$\rho_0^{k,k'} = q_0^{k,k'} = \frac{\delta_{k,k'}}{Np_{k'}} \tag{23}$$

The Kronecker δ-function in the numerator assures that only nodes with degree equal to k' can have non-zero value. The denominator $Np_{k'}$ is the total number of nodes with degree k'. Next defined variable is $\bar{q}_n^{k,k'}$ which is the excepted value for infected neighbor nodes of a degree-k node at time step n of an epidemic started from a degree-k' node:

$$\bar{q}_n^{k,k'} = \sum_k \frac{kp_k}{z} q_n^{kk'} = \frac{\sum_k kp_k q_n^{kk'}}{\sum_k kp_k} \tag{24}$$

In the last term of Eq. (24) the denominator is the mean degree, and the numerator calculates the mean of infected neighbor nodes. The generalized versions of Eq. (4) and (5) become:

$$\rho_{n+1}^{k,k'} = 1 - (1 - \rho_0^{k,k'})(1 - \bar{q}_n^{k,k'})^k \tag{25}$$

and

$$q_{n+1}^{k,k'} = 1 - (1 - \rho_0^{k,k'})(1 - \bar{q}_n^{k,k'})^{k-1} \tag{26}$$

In Eq. (25) $\rho_{n+1}^{k,k'}$ is the excepted fraction of infected nodes at time step $n+1$, and similarly $q_{n+1}^{k,k'}$ the fraction of infected neighbors. In a case of a z-regular graph Eq. (23)–(26) reduce to Eq. (2)–(5).

The authors assume that the giant connected component (GCC) is so large that a randomly chosen starting node belongs to it with high probability, when they state that when the epidemic spreads in the network, all the nodes in GCC finally become infected. The fraction of degree-k' nodes in GCC can be calculated with the steady state value $\rho_\infty^{k,k'} \leq 1$. Furthermore this ratio does not depend on the degree k' of the starting node, so we can mark that $\rho_\infty^k \equiv \rho_\infty^{k,k'}$. Ratio of infected nodes at a certain time step can be calculated using also the steady state value with formula $\rho_n^{k,k'}/\rho_\infty^{k,k'}$. Using these results we can write a corresponding formula for (1) for the case of random distribution degree network. First the probability that two nodes with degrees k and k' have the path distance of n:

$$D_n^{k,k'} = \frac{\rho_n^{k,k'}}{\rho_\infty^{k,k'}} - \frac{\rho_{n-1}^{k,k'}}{\rho_\infty^{k,k'}} = \frac{\rho_n^{k,k'} - \rho_{n-1}^{k,k'}}{\rho_\infty^{k,k'}} \tag{27}$$

This can be developed further by summing over the product of degree distribution and expected distances for all degrees k':

$$D_n^k = \sum_{k'} p_{k'} D_n^{k,k'} \tag{28}$$

which gives the probability for a degree-k to be at distance n to a random node with any degree. Again, we will sum all the values k to get

$$D_n = \sum_k p_k D_n^k \tag{29}$$

which corresponds to Eq. (1) in the case of a network with random degree distribution giving the probability for any two nodes to be at distance n.

4 Calculations for generalization to networks with joint degree-degree distribution

The authors state that the joint degree-degree distribution contains more information about the structure of the network than mere degree distribution. Degree distribution can be derived from the joint degree-degree distribution using the Eq. 12 in the article, that equation is not be proven in this report. The authors state that the equations Eq. (23)–(26) can be directly applied to a network with known joint degree-degree distributions with the exception that Eq. (24) should be replaced with

$$\bar{q}_n^{k,k'} = \frac{\sum_{k''} P(k,k'') q_n^{k'',k'}}{\sum_{k''} P(k,k'')}. \tag{30}$$

This can be proved by first having the probability that a randomly chosen link starts at a k-degree node:

$$\Pr(k_u = k) = \frac{kp_k}{z}, \text{ where } z = \sum_k kp_k \tag{31}$$

The joint degree-degree distribution is the same as the probability that a random link connects nodes of degrees k and k':

$$P(k,k') = \Pr(k_u = k) \cdot \Pr(k_v = k') = \frac{kp_k \cdot k'p_{k'}}{z^2}. \tag{32}$$

Inserting this into Eq. (30) gives

$$\bar{q}_n^{k,k'} = \frac{\sum_{k''} kp_k \cdot k''p_{k''} z^{-2} \cdot q_n^{k'',k'}}{\sum_{k''} kp_k \cdot k''p_{k''} z^{-2}} = \frac{kp_k \cdot z^{-2} \cdot \sum_{k''} k''p_{k''} \cdot q_n^{k'',k'}}{kp_k \cdot z^{-2} \cdot \sum_{k''} k''p_{k''}} \Rightarrow$$

$$\bar{q}_n^{k,k'} = \frac{\sum_{k''} k''p_{k''} \cdot q_n^{k'',k'}}{\sum_{k''} k''p_{k''}}, \tag{33}$$

which equals Eq. (24).

5 Test with example Networks

The methods were tested with a Python code. The implementation is available in Google Colabs[1]. The results of the test runs are shown in Table 1. Data sets *"500 Airports"* and *"C. Eleg. Neur."* were downloaded from Complex Networks Data Sets [17]. The Erdős-Rényi and z-regular networks were generated with NetworkX[2]. Table has same structure as Table I in the article.

On the left half columns are the network name, number of nodes N, average degree z, and exact average shortest path length $\bar{D}$ are given, and on the right side are the relative errors of estimated values. The error value have a precision of four decimals. The table does not have the estimates FR. and NMN. based on research by other authors[FFH05; NSW01] like in the article. Unlike in the article also the analytical result D^{RRG} (Eq. (22))

[1] `https://colab.research.google.com/drive/19UoafUIfwxlVPNjfOjpNevPMzZc8QaHa`

[2] `https://networkx.github.io/documentation/stable/index.html`

is shown in the table, although it was derived only for z-regular networks. Estimate p_k for degree distribution is calculated with Eq. (24)–(26). Estimate $P(k, k')$ for joint degree-degree distribution is calculated with Eq. (25),(26), and (30). Generally there is not any significant difference in the result values when compared with the original article. Notice that the data set *C. Eleg. Neur.* is a different version from the network tested in the original article, somehow version used in this project gives a larger relative error.

Network	N	z	$\bar{D}$	Relative D^{RRG}	errors in p_k	$\bar{D}$ for $P(k,k')$
500 Airports	500	11.9	2.99	0.0729	0.0631	0.0452
C. Eleg. Neur.	279	16.4	2.43	0.0456	0.0426	0.0585
z-regular	500	4.0	5.00	0.0042	0.0040	0.0040
z-regular	500	10.0	2.95	0.0047	0.0045	0.0045
Erdős-Rényi	500	10.0	2.94	0.0092	0.0056	0.0059
z-regular	10000	4.0	7.73	0.0004	0.0004	0.0004
z-regular	10000	10.0	4.34	0.0002	0.0002	0.0002
Erdős-Rényi	10000	10.0	4.26	0.0197	0.0002	0.0002

Table 1: Test run results

References

[NSW01] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. “Random graphs with arbitrary degree distributions and their applications”. In: *Physical review E* 64.2 (2001), p. 026118.

[FFH05] Agata Fronczak, Piotr Fronczak, and Janusz A Hołyst. “How to calculate the main characteristics of random uncorrelated networks”. In: *AIP Conference Proceedings*. Vol. 776. 1. AIP. 2005, pp. 52–68.

[MG16] Sergey Melnik and James P Gleeson. “Simple and accurate analytical calculation of shortest path lengths”. In: *arXiv preprint arXiv:1604.05521* (2016).

[17] *Complex Networks Data Sets*. `https://www.complex-networks.net/datasets.html`. Accessed: 2019-02-24. 2017.