Selected Topics in Frontiers of Statistics Project

Yixuan Ding[*]

June 25, 2024

Abstract

Modelling "small world phenomenon" - the principle that we are all linked by short chains of acquaintance is very popular and important topic in social science and complex network system. For simplicity, we consider two-dimensional grid lattice. In the paper "The Small-World Phenomenon: An Algorithmic Perspective" by Jon Kleinberg, it considers the long-range contact happens between current node and all others nodes except local nodes in the network. In this project, we consider a special case where long-range contact happens horizontally and vertically. By theoretical proofing and simulation, we draw the conclusion that the best clustering exponent locates in 1.

1. Introduction

A social network exhibits "small world phenomenon" if roughly speaking, any two individuals in the network are very likely to be linked with each other by a relatively short sequence of intermediate acquaintance. The problem was originated from a "message passing" story: The experiment is conducted between a "source" person and "target" person and they don't know each other. The source is told to deliver the letter to the target by knowing only his/hers location and occupation, and the source can only transfer the message to the person who his/her knows on the "first-name basis". Then the length of sequence is measured to indicate the "six degrees of separation" principle. In the paper of Kleinberg [Kle00], the efficiency of clustering exponent is discussed and the value of 2 is found to be the most efficient in the situation of two-dimensional lattice and the long-range contacts are constructed between the current node and all other nodes except local connection. More specifically, there are several values we need to set:

- **p**: lattice distance for local contact
- **q**: number of long-contact nodes

[*] Department of Statistics and Data Science; Southern University of Science and Technology; 12111620@mail.sustech.edu.cn;

- **r**: clustering exponent, used to calculate the probability when connecting long-contact nodes.

In the discussion of Kleinberg, p and q are set as 1, and r is found to be 2 as the best value. The illustration via visualization is present in figure 1 (from the paper of Kleinberg).

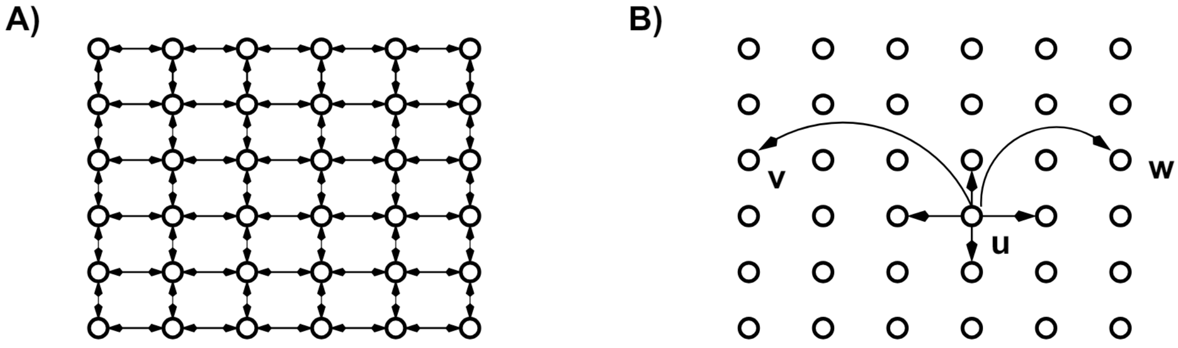

Figure 1.1: Original Network Setting in Kleinberg's Paper

In the paper of Kleinberg, there are three main theorems mentioned to indicate the efficiency of different clustering exponent:

Theorem 1.1. *There is a constant α_0, depending on p and q but independent of n, so that when $r = 0$, the expected delivery time of any decentralized algorithm is at least $\alpha_0 n^{\frac{2}{3}}$. (Hence exponential in the expected minimum path length.)*

Theorem 1.2. *There is a decentralized algorithm $\mathcal{A}$ and a constant α_2, independent of n, so that when $r = 2$ and $p = q = 1$, the expected delivery time of $\mathcal{A}$ is at most $\alpha_2(\log n)^2$*

Theorem 1.3. *(a) Let $0 \leq r < 2$. There is a constant α_r, depending on p, q, r, but independent of n, so that the expected delivery time of any decentralized algorithm is at least $\alpha_r n^{(2-r)/3}$.*

(b) Let $r > 2$. There is a constant α_r, depending on p, q, r, but independent of n, so that the expected delivery time of any decentralized algorithm is at least $\alpha_r n^{(r-2)/(r-1)}$.

However, in this project, we are going to discuss another unique setting, where the long-range contact can only be construct horizontally and vertically (See Figure 1.2). Under this setting, we are going to find the best (the average delivery steps are the shortest) clustering exponent. First we are going to theoretically investigate the location of clustering exponent when it attains the maximum efficiency, and then verify out result via simulation.

2. Theoretical Investigation

This section aims to investigate how the construction process of the network affect the shortest path length between nodes via decentralized algorithm. When $r = 0$, long-range

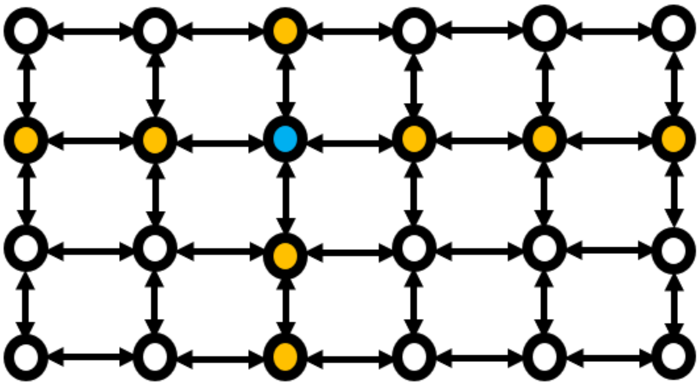

Figure 1.2: New Setting Investigated in This Project

contact between nodes is uniform distributed. Same as *Theorem 1* in the paper of Kleinberg, there is no way for a decentralized algorithm that the expected delivery time between any two nodes is bounded by a polynomial of $\log n$, exponentially smaller than the total number of nodes.

However, as the clustering exponent r becoming larger, we can further make better use of the "geographical structure" of the network. As r growing, the long-range contact becomes less possible between two nodes with large distance. Hence we can foresee that there exists a particular r where the trade-off is being best exploited.

Theorem 2.1. *There is a decentralized $\mathcal{A}$ and a constant α_2, independent of n, so that when $r = 1$, and $p = q = 1$, the expected delivery time of $\mathcal{A}$ is at most $\alpha_2(\log n)^2$.*

Proof: We consider a two-dimensional lattice and $p = q = 1$ for simplicity, for a source node u, the probability that u having a long-range connection of node v can be viewed as: $d(u,v)^{-2}/\sum_{v\neq u} d(u,v)^{-2}$. Then we have:

$$\sum_{v\neq u} d(u,v)^{-1} \leq 4\sum_{i=1}^{2n-2} = 4\sum_{j=1}^{2n-2} j^{-1} \leq 4 + 4\ln(2n-2) \leq 4\ln(6n)$$

Thus, the probability that v is chosen is at least $[4\ln(6n)d(u,v)^2]^{-1}$. Now, we denote for $j > 0$, the execution of *mathcalA* is in phase j when the lattice distance from the current node to the target node is greater than 2^j and at most 2^{j+1}. Thus the initial value of j is at most $\log n$. Suppose we are in phase j, where $\log(\log n) \leq j < \log n$, we denote there are B_j of nodes within the lattice distance 2^j of target node. Then the number of nodes in B_j is

$$1 + \sum_{i=1}^{2^j} i = \frac{1}{2}2^{2j} + \frac{1}{2}2^j + 1 > 2^{2j-1}$$

For nodes in phase j, each has a probability of at least $(4\ln(6n)2^{2j+4})^{-1}$ of being the long-range contact of u. Then the message enters B_j with probability at least:

$$\frac{2^{2j-1}}{4(\ln(6n)2^{2j+4})} = \frac{1}{128\ln(6n)}$$

. Let X_j denote total number of steps spent in phase j, then we have:

$$\begin{aligned} EX_j &= \sum_{i=1}^{\infty} \Pr[X_j \geq i] \\ &\leq \sum_{i=1}^{\infty} \left(1 - \frac{1}{128\ln(6n)}\right)^{i-1} \\ &= 128\ln(6n) \end{aligned}$$

Now, if X denotes the total number of steps spent by the algorithm, we have:

$$X = \sum_{j=0}^{\log n} X_j$$

so by linearity of expectation we have $EX \leq (1+\log n)(128\ln(6n)) \leq \alpha_2(\log n)^2$ for a suitable choice of α_2 ■

Theorem 2.2. *Let $0 \leq r < 1$. There is a constant α_r, depending on p, q, r, but independent of n, so that the expected delivery time of any decentralized algorithm is at least $\alpha_r n^{(1-r)/3}$.*

Proof:

$$\begin{aligned} \sum_{v\neq u} d(u,v)^{-r} &\geq \sum_{j=1}^{n/2} j^{-r} \\ &\geq \int_1^{n/2} x^{-r} dx \\ &\geq (1-r)^{-1}((n/2)^{1-r} - 1) \\ &= \frac{n^{1-r}}{(1-r)2^{1-r}} - \frac{1}{1-r} \\ &\geq \frac{1}{(1-r)2^{2-r}} \cdot n^{1-r} \end{aligned}$$

the last line follows if we assume $n^{1-r} \geq 2^{2-r}$.
Let $\delta = (1-r)/3$, and U denote the set of nodes within lattice distance pn^δ of t, then:

$$|U| \leq 1 + \sum_{j=1}^{pn^\delta} 4j \leq 4p^2 n^{2\delta}$$

where we assume n is large enough so that $pn^{\delta} \geq 2$. Define $\lambda = (2^{7-r}qp^2)^{-1}$, let ϵ' be the event that within λn^{δ} steps, the message reaches a node other than t with a long-range contact in U, also denote ϵ'_i be the event that in step i, the message reaches a node other than t with a long-range contact in U. Thus $\epsilon' = \bigcup_{i \leq \lambda n^{\delta}} \epsilon'_i$. For the node reached at step i having q long-range contact, we can derive:

$$\begin{aligned}
Pr[\epsilon'_i] &\leq \frac{q|U|}{\frac{1}{(1-r)2^{2-r}} \cdot n^{1-r}} \\
&\leq \frac{(1-r)2^{2-r}q \cdot 4p^2n^{2\delta}}{n^{1-r}} \\
&= \frac{(1-r)2^{4-r}qp^2n^{2\delta}}{n^{1-r}}
\end{aligned}$$

The probability of a union event is bounded by the sum of their elements' probabilities:

$$\begin{aligned}
Pr[\epsilon'] &\leq \sum_{i \leq \lambda n^{\delta}} Pr[\epsilon'_i] \\
&\leq \frac{(1-r)2^{4-r}\lambda qp^2n^{3\delta}}{n^{1-r}} \\
&= (1-r)2^{4-r}\lambda qp^2 \leq \frac{1}{4}
\end{aligned} \quad \blacksquare$$

We now denote $\mathcal{F}$ the event that the chosen source node and target node are separated by a lattice distance of at least $n/4$. We can verify that $Pr[\mathcal{F}] \geq \frac{1}{2}$.

Finally, we let X the random variable equal to the number of steps taken for the message to reach t, and let ϵ denote the event that the message reaches t within λn^{δ} steps. We can know that if $\mathcal{F}$ occurs and ϵ' does not occur, then ϵ cannot occur. Thus $Pr[\epsilon|\mathcal{F} \wedge \bar{\epsilon}'] = 0$, then $E[X|\mathcal{F} \wedge \bar{\epsilon'}] \geq \lambda n^{\delta}$. Then

$$EX \geq E[X|\mathcal{F} \wedge \bar{\epsilon'}] \cdot Pr[\mathcal{F} \wedge \bar{\epsilon'}] \geq \frac{1}{4}\lambda n^{\delta}$$

Theorem 2.3. *Let $r > 1$. There is a constant α_r, depending on p, q, r, but independent of n, so that the expected delivery time of any decentralized algorithm is at least $\alpha_r n^{(r-1)/r}$*

Proof: Proof of this theorem has almost the same procedure as the proof of *Theorem 2.2* $\blacksquare$

Therefore, ideally we want to confirm that the relationship between clustering exponent and lower bound of the average delivery time is behaving like in Figure 2.2.

3. Simulation

For calculation efficiency, we do not construct the whole network at the beginning, we use python tuple to initiate each node in the network. For each run, we sample two nodes in

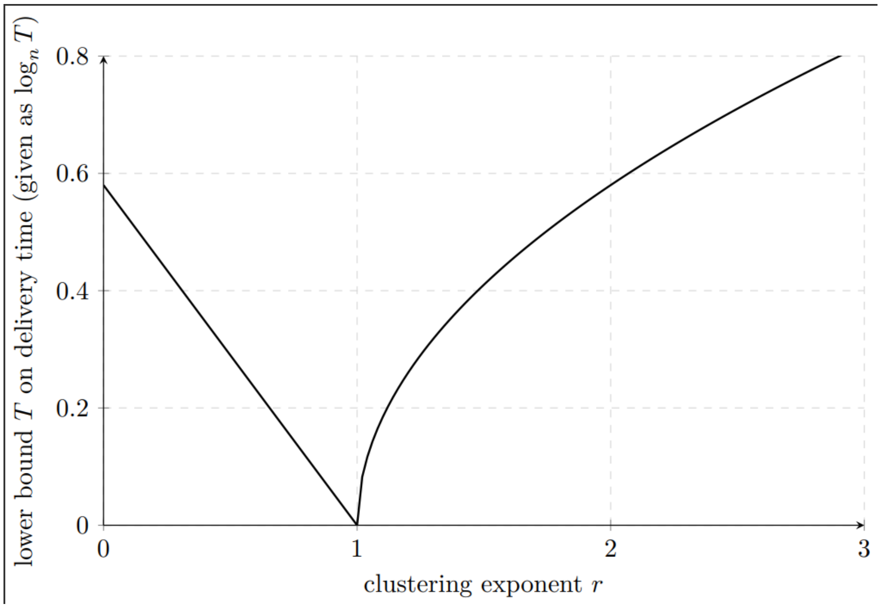

Figure 2.1: Theoretical Relationship between Clustering Exponent and Lower Bound Log Average Delivery Time

the network via random trials, denoting them as source node and target node. Then we can obtain local connection of source node, and sample long-range contact for the source node and operate message passing process. We then denote the next message holder as source node and iterate the process until the message is transferred to target node. Therefore we accomplish one message passing, and we can leverage each process to calculate the overall delivery efficiency. More specifically, the simulation involves the following functions:

- **cal_prob**: Given a network size and clustering exponent, we can calculate cumulative logits $\sum_{u \neq v} d(u,v)^{-r}$ and the corresponding probability for different distance. (Note that the set of probability is same for different clustering exponent, hence we just need to generate once per simulation).

- **assign_available_node**: For each message passing step, we need to derive next candidate message holder. For calculation efficiency, candidate nodes are derived when the message is sent to the current holder. This function calculate the local contact and long-range contact at the same time. For local contact, we need to consider the boundary situation. As for long-range contact, we need to make use of the probability calculated above, and assign the contact via random trial.

- **cal_distance**: A simple function return the Manhattan distance between two nodes.

- **find_path**: Given a source node and a target node, we can do an iterative loop: assign the candidate next message holder for source node → find the nearest candidate node to the target node → denote the new message holder as source node → continue the process.

In our simulation process, network size is set as 5000. For each r, we conduct 1000 runs to obtain the average delivery steps. The simulation result are as follows:

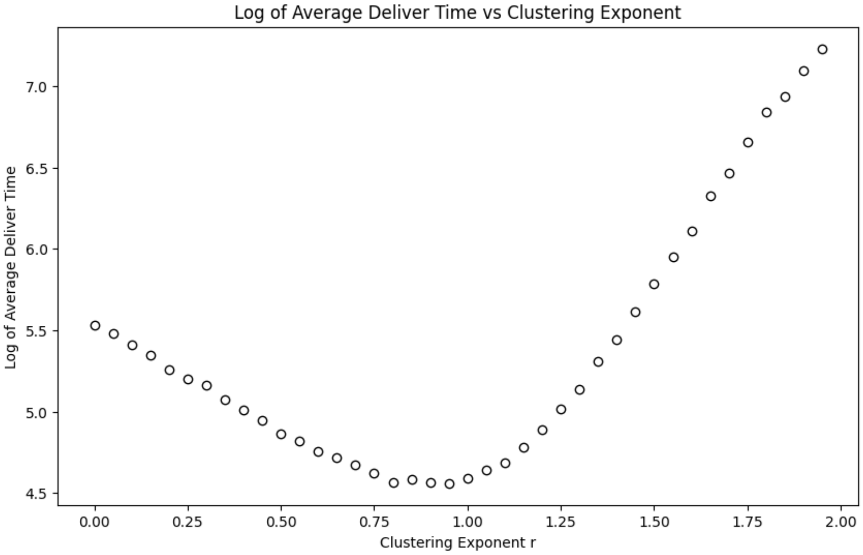

Figure 3.1: New Setting Investigated in This Project

From the simulation, we can observe that the average delivery time is first decreases then increases as the growing of clustering exponent. Also, the simulation result is consistent with our theoretical proofing that the best clustering exponent lies around 1.

4. Conclusion

In this project, we review the paper of Kleinberg and discuss a new setting that long-range contact exists horizontally or vertically for current message holder. Through theoretical investigation and synthesis data simulation, we evaluate our thinking that the most efficient clustering exponent under the setting is 1.

References

[Kle00] J. M. Kleinberg. The small-world phenomenon: an algorithmic perspective.. *STOC*, 163–170, 2000.