Kleinberg's small world phenomenon: A modification

Qijia He

SUSTech, Department of Statistics and Data Science
heqj2021@mail.sustech.edu.cn

With the advent of the World Wide Web, we can communicate with anyone anytime and anywhere via the internet. This has brought people closer together, sparking growing interest in the study of social networks: How many friends of my friends do I have? How many friends' friends would it take for me to get acquainted with Bill Gates? In 1960, American sociologist Stanley Milgram conducted a groundbreaking experiment demonstrating that there is a six-degree separation between individuals, leading to the famous Six Degrees of Separation theory. In 2000, Jon Kleinberg further studied the small world phenomenon, modeling social network relationships using a two-dimensional grid. This project is based on Kleinberg's "Small World Phenomenon: An Algorithmic Perspective,"[1]. modifying the assumptions about connection distribution in a two-dimensional network to explore how changes in these assumptions affect both theoretical and experimental results.

1 Introduction

Kleinberg and many social scientists from the late 20th century have pointed out that the world exhibits a small world phenomenon within social networks, which aligns with our experiences: often when talking to strangers, we find that we share some mutual friends. Assuming each of us knows only 50 people, and there is no overlap among our friends, our social network through friends of friends extended six times would reach 50^6, which is 15,625,000,000 people! This number is almost twice the world's population. Our social network world is tightly connected, even with a vast population of over 7 billion people.

In Kleinberg's experiment, he assumed the world is a two-dimensional grid where any individual is a node on the grid, and he simulated Milgram's social network propagation experiment on this grid. Based on a basic empirical assumption that people are familiar with those around them (community) and have a few friends in other places (outside the community), Kleinberg's experiment posited that any node in the two-dimensional space is directly connected to its p neighboring nodes, and has a certain probability ($distance^{-r}$) of connecting with nodes further away. Here, p, q, and r are constant parameters set in the grid simulation, and 'distance' is the Manhattan distance between the current node and the target node. By randomly setting a start and an end point in the grid, Kleinberg explored how many steps it would take for a "letter" to travel from the start to the end point, and which distribution and parameter configuration would optimize the

information transfer within the grid. In the original paper[1], with any fixed p, q, when r reaches 2, the average letter delivery steps reaches the optimal value.

Figure 6 gives a intuitive understanding of how the network looks like: when p=2, the short range connection can reach nodes which has Manhattan distance <= 2, when q=2, it will generates 2 long range connection based on distribution $distance^{-r}$, the point closest to the target point will be selected as the next node.

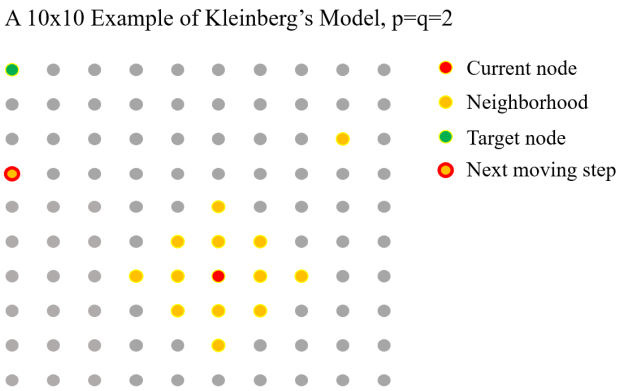

Figure 1: example of Kleinberg's model

2 Problem

However, this model is based solely on empirical assumptions. In the real world, connections between people do not follow this exact pattern. For instance, the distances between communities can be very large, while within a community, almost everyone knows each other. Additionally, long-range connections do not necessarily follow a distance-based power-law distribution. Therefore, we aim to adjust the basic assumptions of this model to explore whether the model's performance remains consistent under different assumptions. What new scenarios might arise after modifying these assumptions? In this project, we focus on one specific case: modifying the criteria for long-range connections. The specific question is:

Problem: What would be the optimal r if long-range connections can only occur in horizontal or vertical directions?

Figure 2 is the pdf of long range connection, while Figure 3 is the modified one. We cannot connect to points that is not on the horizontal or vertical line of the current point, Meanwhile, the probability of total long range connection decreases, since the number of nodes for distance j changed from 4j to 4. Generally speaking, The smaller the value of r, the greater the probability of a long jump (walk through long range contact). Figure 4 shows the distance descending process with regard

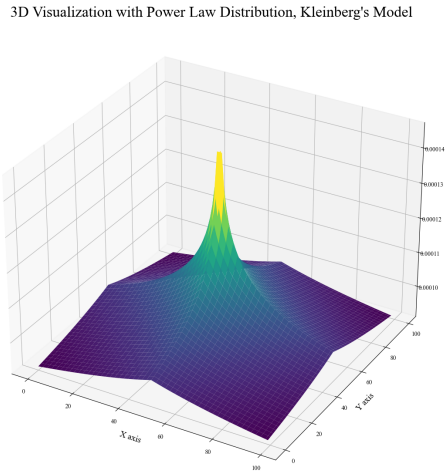

Figure 2: Original, r=0.1

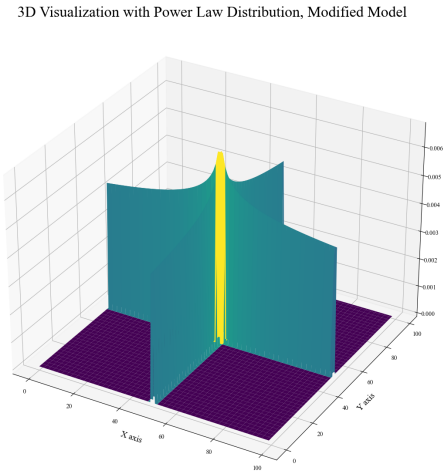

Figure 3: Modified, r=0.1

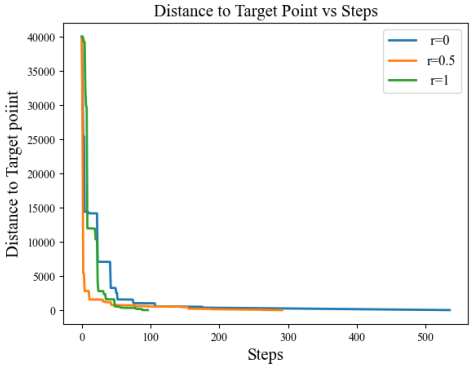

Figure 4: message delivering process, example 1

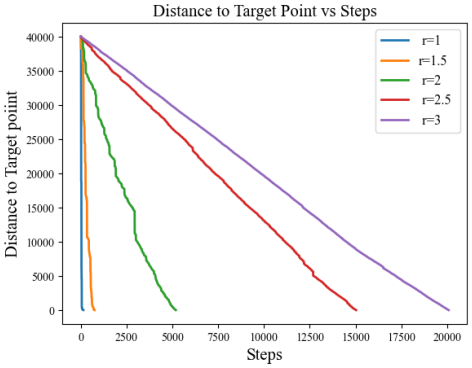

Figure 5: message delivering process, example 2

to steps in 20000x20000 grids. they all converge very quick since all the r are very small. However, as r increases, as Figure 5 shows, when r=3, the long range contact do not make a big difference, and it's like a linear descending process.

In the next 2 sections, I'll theoretically proof $r = 1$ is the optimal value and run several experiments to verify my results.

3 Theoretical Proof

We have few variables to declare before the proof section. Some have been declared in the paragraph before, to make sure its meaning and help readers to find it, they're all summarized on the table below:

parameters	Definition
u	the current point
v	one neighbor of u
t	target node
p	the definition of neighbor: a point is u's neighbor if d(u,v)<=p
q	the number of long range connection of node u
r	power low distribution parameter
n	the size of the grid, which is nxn

Table 1: parameter table

3.1 Distribution's Perspective

According to the paper written by Kleinberg, the best r=2 when the long range distribution is not modified. To reach the optimal in the new distribution, a intuitive understanding is that we should find an r', such that the probability of connecting to nodes at the same distance as in the original network is as close as possible.

Assume the grid is large enough, for the original one, the probability connecting to nodes at distance j is:

$$\mathbb{P}(d(u,v)) = \frac{4j \cdot j^{-2}}{\sum_{i=1}^{n} 4i \cdot i^{-2}} = \frac{j^{-1}}{\sum_{i=1}^{n} i^{-1}} \tag{1}$$

For the modified one, Suppose v and u remain horizontal or vertical, we have:

$$\mathbb{P}(d(u,v)) = \frac{4 \cdot j^{-r'}}{\sum_{i=1}^{n} 4 \cdot i^{-r'}} = \frac{j^{-r'}}{\sum_{i=1}^{n} i^{-r'}} \tag{2}$$

It's clear that equation (1) = equation (2) when r'=1, and in 100*100 dimensional space, the distribution of the comparison is shown in Figure 6. The slightly differences when r>50 is happened due to grid's boundary.

3.2 Time Complexity's Perspective

Now we prove the average number of steps reaches the optimal from the algorithm's time complexity's perspective. In kleinberg's paper [1], He already proved the situation for the original model, only a slightly difference will be made for proving the new situation.

Since the main idea of the proof is exactly the same as Kleinberg's, most formulas from 3$\tilde{3}$2 can be found in his paper[1], the difference is some slightly modification in each formula's parameters. Therefore, I will only give a simplified version of the proof.

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

Proof. : First we define several variables:

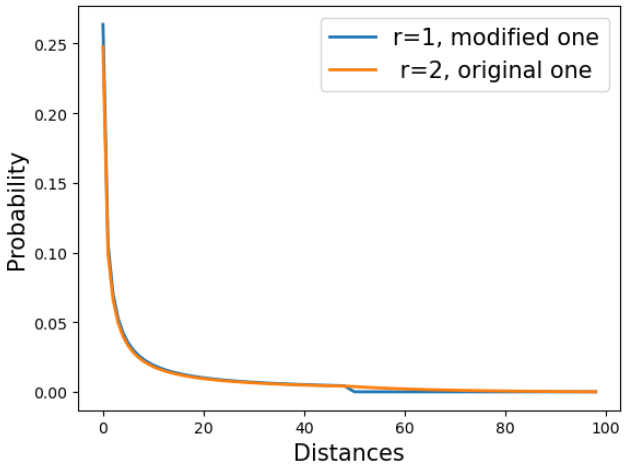

Figure 6: A comparison between $r = 0$ for the original and $r = 1$ for modified model

- phase j: the current node to t is greater than 2^j and at most 2^{j+1}
- B_j: nodes within lattice distance 2^j of t.
- X_j: the total number of steps in phase j

The basic idea of this proof is to use 'phase' defined above. We will then calculate the probability of ending phase j and move to phase j-1 or smaller phases. Since the total amount of phase is fixed, the probability*# phases is the expected delivery time (upper bound).

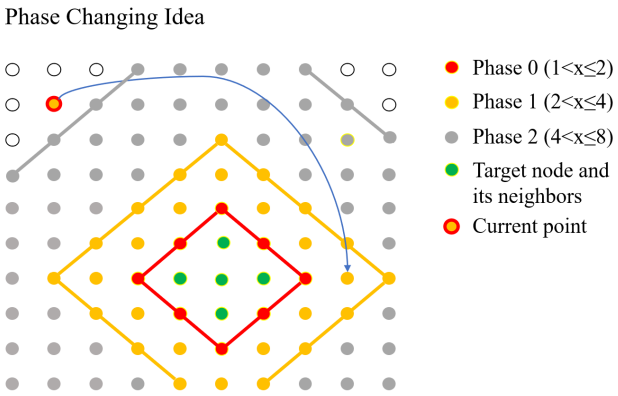

Figure 7: Basic idea for theorem 1

The sum of d(u,v) u!=v is at most:

$$\sum_{v!=u} d(u,v)^{-1} = 4\sum_{j=1}^{n-1} j^{-1} \leq 4 + 4ln(n-1) \leq 4ln(3n) \tag{3}$$

the number of points in B_j is at least:

$$1 + \sum_{i=1}^{2^j} > \frac{1}{2}2^{2j} + \frac{1}{2}2^j + 1 > 2^{2j-1} \tag{4}$$

the probability choosing a node in B_j is at least:

$$\frac{2^{2j-1}}{4ln(3n)2^{j+2}} > \frac{2^{j-1}}{4ln(3n)2^{j+2}} > \frac{1}{32ln(3n)} \tag{5}$$

the expected total number of steps spent in phase j is at most:

$$E(X_j) = \sum_{i=1}^{\inf} P(X_j \geq i) \leq \sum_{i=1}^{\inf}(1 - \frac{1}{32ln(3n)})^{i-1} = 32ln(3n) \tag{6}$$

There's at most logn phase in total, hence average delivery time is at most:

$$X = \sum_{j=0}^{logn} X_j \leq (1 + logn)(32ln(3n)) \leq \alpha_2(logn)^2 \tag{7}$$

as desired. □

Theorem 2. *let o<r<1, there's 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. First we define several variables:

- δ: (1-r)/3; λ: $(2^{6-r}qp^2)^{-1}$
- U: denotes the set of nodes within lattice distance $p^{n\delta}$ of t
- $\mathcal{E}'_i$: in step i, the message reaches a node other than t with a long-range contact in U, $\mathcal{E}' = \bigcup_{i \leq \lambda n^{\delta_i}} \mathcal{E}'_i$
- $\mathcal{E}$: the message reaches target within $p^{n\delta}$ steps
- $\mathcal{F}$: the distance of the initial point and target point $\geq n/4$
- X: the number of steps to reach target

The basic idea of this theorem is to prove the probability of connecting to certain point sets ($\mathcal{U}$) with in certain trials is smaller than a constant (1/4), and then using Bayes' theorem to prove the expectation steps is at least something that can by presented by $cn^{f(r)}$

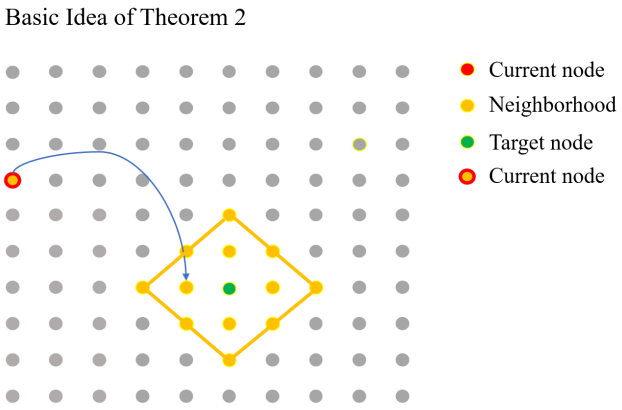

Figure 8: Basic idea for theorem 2

The sum of d(u,v) u!=v is at least:

$$\sum_{v \neq u} d(u,v)^{-r} \geq \sum_{j=1}^{n/2} j^{-r} \tag{8}$$

$$\geq \int_{1}^{n/2} x^{-r} dx \tag{9}$$

$$\geq (1-r)^{-1}((n/2)^{1-r} - 1) \tag{10}$$

$$\geq \frac{1}{(1-r)2^{2-r}} \cdot n^{1-r}, \tag{11}$$

The 11-th inequality holds due to $\frac{n^{1-r}}{2^{2-r}} \geq 1$.

|U| is at least:

$$|U| \leq 1 + \sum_{j=1}^{pn^{\delta}} 4j \leq 4p^2 n^{2\delta}, \tag{12}$$

The message reaches a node other than t with a long range contact in U is at least:

$$\Pr\left[\mathcal{E}_i'\right] \leq \frac{q|U|}{\frac{1}{(1-r)2^{2-r}} \cdot n^{1-r}} \tag{13}$$

$$\leq \frac{(1-r)2^{2-r} q \cdot 4p^2 n^{2\delta}}{n^{1-r}} \tag{14}$$

$$= \frac{(1-r)2^{4-r} q p^2 n^{2\delta}}{n^{1-r}}. \tag{15}$$

the probability that $\mathcal{E}'$ holds is at least $1/4$

$$\Pr\left[\mathcal{E}'\right] \leq \sum_{i \leq \lambda n^{\delta}} \Pr\left[\mathcal{E}_i'\right] \tag{16}$$

$$\leq \frac{(1-r)2^{4-r} \lambda q p^2 n^{3\delta}}{n^{1-r}} \tag{17}$$

$$= (1-r)2^{4-r} \lambda q p^2 \leq \frac{1}{4}. \tag{18}$$

It can be proved that $\Pr[\mathcal{F}] \geq \frac{1}{2}$, since:

$$\Pr\left[\overline{\mathcal{F}} \vee \mathcal{E}'\right] \leq \frac{1}{2} + \frac{1}{4}, \Pr\left[\mathcal{F} \wedge \overline{\mathcal{E}'}\right] \geq \frac{1}{4}. \tag{19}$$

we have:

$$\Pr\left[\mathcal{E} \mid \mathcal{F} \wedge \overline{\mathcal{E}'}\right] = 0 \tag{20}$$

$$E\left[X \mid \mathcal{F} \wedge \overline{\mathcal{E}'}\right] \geq \lambda n^{\delta}. \tag{21}$$

Finally, we get:

$$EX \geq E\left[X \mid \mathcal{F} \wedge \overline{\mathcal{E}'}\right] \cdot \Pr\left[\mathcal{F} \wedge \overline{\mathcal{E}'}\right] \geq \frac{1}{4}\lambda n^{\delta}, \tag{22}$$

as desired. □

Theorem 3. *let r>2, There's 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. First we define several variables:

- ϵ=r-1, β: $\frac{\epsilon}{1+\epsilon}$, γ: $\frac{1}{1+\epsilon}$, λ': $\frac{min(\epsilon,1)}{16q}$
- $\mathcal{E}_i$: in step i, the message reaches a node u!=t that has a long range contact v satisfying $d(u,v) > n^{\lambda}$
- $\mathcal{E}$, $\mathcal{F}$, X are all the same as Theorem 2.

Also, the idea of this proof is similar to theorem 2: First proof given $\mathcal{F}$ the probability of walking through a extremely long range contact within i steps is smaller than a constant (1/4), then use Bayes' theorem to further derived the desired result. all those strange variables (i.e. λ, ϵ, β...) defined above is to simplified the formula derived later and to make the result (constant, 1/4) looks nice.

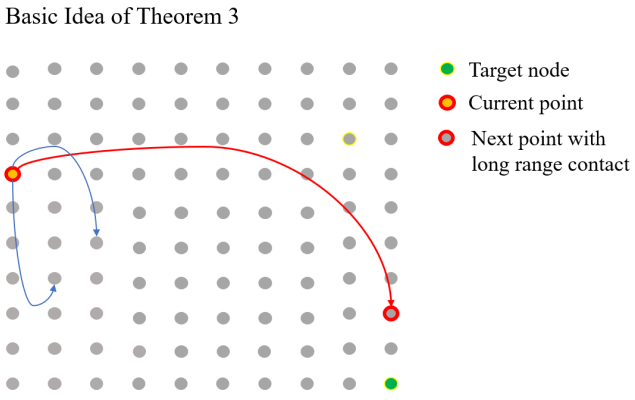

Figure 9: Basic idea for theorem 3

Let m be a random constant(>0), the probability that $d(u,v) > m$ is at most:

$$\Pr[d(u,v) > m] \leq \sum_{j=m+1}^{2n-2} (4)(j^{-r}) \tag{23}$$

$$= 4 \sum_{j=m+1}^{2n-2} j^{-r} \tag{24}$$

$$\leq 4 \int_{m}^{\infty} x^{-r} dx \tag{25}$$

$$\leq 4 \quad (r-1)^{-1} m^{-r} = 4\varepsilon^{-1} m^{-\varepsilon}. \tag{26}$$

The probability that within step i, the message reaches a node u =t that has a has a long range contact $d(u,v) > n^{\lambda}$ is at most

$$\Pr[\mathcal{E}'] \leq \sum_{i \leq \lambda' n^{\beta}} \Pr[\mathcal{E}_i] \tag{27}$$

$$\leq 4\lambda' n^{\beta} \cdot q\varepsilon^{-1} n^{-\varepsilon\gamma} \tag{28}$$

$$= \quad 4\lambda' q\varepsilon^{-1} \leq \frac{1}{4}. \tag{29}$$

Similarly, we can deduce that

$$\Pr[\mathcal{F} \wedge \overline{\mathcal{E}'}] \geq \frac{1}{4}. \tag{30}$$

if this happened, then the message can move a lattice distance of at most n^{γ} in each of its first $\lambda' n^{\beta}$ steps, this is a total lattice distance of at mostn/4 meanwhile the expectation of X given $\mathcal{F} \wedge \overline{\mathcal{E}'}$ gets a lower bound $\lambda' n^{\beta}$:

$$\lambda' n^{\beta+\gamma} = \lambda' n < n/4, E[X \mid \mathcal{F} \wedge \overline{\mathcal{E}'}] \geq \lambda' n^{\beta} \tag{31}$$

by Bayesian formula, we have:

$$EX \geq E[X \mid \mathcal{F} \wedge \overline{\mathcal{E}'}] \cdot \Pr[\mathcal{F} \wedge \overline{\mathcal{E}'}] \geq \frac{1}{4}\lambda' n^{\beta}, \tag{32}$$

as desired. □

By the above 3 theorem, and considering function's continuity, the lower bound for the log number of average steps is roughly like figure 10, It goes down linearly at first, and then it goes up at a rate of 1-1/r, based on the assumption I used when deriving the formulas (the network is very large). The larger the value of n, the closer the actual curve will approach the pattern shown in figure 10 (Since most inequality will become equality when n reaches infinite). and it's for sure that $\alpha = 1$ is the best option when p=q=1 and n is extremely large. (Since $(logn)^2$ always < n^c, $c \in (0, +\infty)$ when $n- > +\infty$). In the next section, few experiments will be done to verify the theorem I derived here.

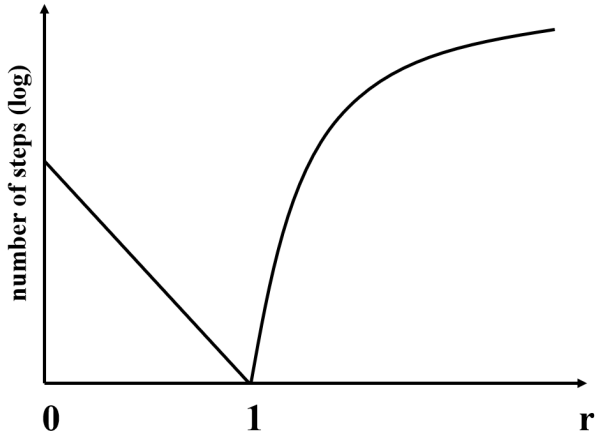

Figure 10: The lower bound of average steps of the modified model

Algorithm 1 Message Delivering Process

Initialize: source (s_1, s_2), target (t_1, t_2)
$current \leftarrow (s_1, s_2)$
while $current \neq (t_1, t_2)$ **do**
    Find neighborhood N of $current$ such that:
        $(x_1, x_2) \in N$ if dist$((x_1, x_2), (s_1, s_2)) \leq p$ **or** (x_1, x_2) has a long range contact based on the probability of the power law distribution
    Select next point $next$ in N such that dist$((next_1, next_2), (t_1, t_2))$ is minimized
    $current \leftarrow next$
end while
return $current$ (which is (t_1, t_2))

4 Experiment

Now we generate a 2D model to simulate the message delivery process, the basic idea is written in Algorithm 1

Algorithm 1 can be implemented in many ways. The most straightforward approach is to model the entire grid by storing a n*n matrix, where each point generates long-range connections based on a distribution, and then performs a walk on the matrix. However, when the matrix is very large (e.g., 20,000 x 20,000), the code may consume a significant amount of memory, and it may even be impossible to initialize the matrix. For simplicity, we will not store the entire matrix but will instead store only the neighborhood of the current point. To address the high time complexity of generating long-range connections probabilistically, we will convert it into a distance sampling problem. The specific solution is as follows(Algorithm 2), this code runs much faster than modeling a whole network, it takes approximately 2 hours to get Figure11 on my laptop.

Algorithm 2 Generating Long-Range Connections

1: Generate a density array arr of length n, where $arr[i] = \frac{f(i)}{\sum_{j=1}^{n} f(j)}$, and $f(x) = x^{-r}$

2: **for** $k = 1$ to q **do**
3: **repeat**
4: Randomly sample an index $dist$ from arr based on its probability distribution
5: Generate candidate points based on the current point (x_1, x_2):

$$(x_1 + dist, x_2), (x_1 - dist, x_2), (x_1, x_2 + dist), (x_1, x_2 - dist)$$

6: **until** at least one candidate point is within the grid bounds
7: Select one point from the valid candidate points as the long-range connection
8: **end for**

In the experiment, I first set p=q=1 and varied r from 0 to 2 with an interval of 0.05. For the grid sizes n=[1000,5000,10000,15000,20000], I conducted 50 simulations for each grid and took the average of these simulations. Figure 11 shows my simulation results.

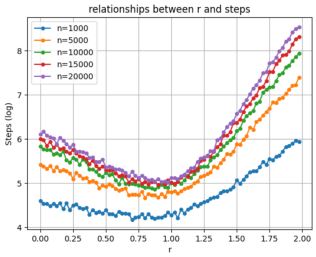

Figure 11: relationships between r and steps

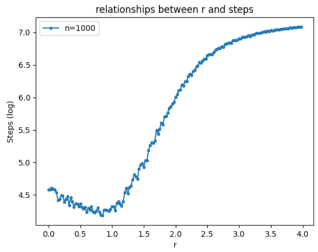

Figure 12: r vs steps, n=1000

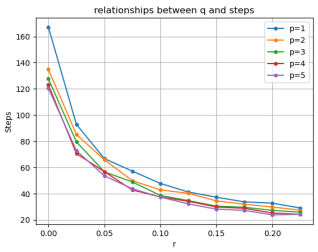

Figure 13: relationships between p and steps

This phenomenon aligns with what we have demonstrated theoretically: the experimental curve also exhibits a process of initially decreasing and then increasing, reaching its minimum point at r=1. Steps will increase when n increases, and the pattern of the function will be more obvious as well.

In figure 12, I increased the range of r (from 0 2 to 0 4), the algorithm's pattern when r>1 is more obvious, just like what we had proved in theorem 3.

when p and q increase, the average steps will decrease. We can quickly derived that the relationship between average steps and p is linear. As for average steps q, I use the Figure13 to illustrate how will the average step change when adjusting q. we set r=1 for this experiment. As figure 13 shows, it decreases very fast at first, and gradually slow down when q is large.

5 Further Discussion

As I mentioned in the introduction section, explaining the small world phenomenon using this method still has its limitations: not all individuals have an equal number of friends, and there are often significant barriers between different communities. Additionally, this model's parameters cannot account for the personalized interactions between individuals. In Assignment 2, we discussed scenarios such as transitioning to one-dimensional dimensions and transforming from a two-dimensional grid to a toroidal lattice, but for practical usage, more real world data and senearios should be included in this model.

For future improvements, we aim to integrate real-world data from social networks into the model through data analysis. This approach will allow us to incorporate a more realistic depiction and further explore the small world phenomenon in the highly information-intensive era.

6 Conclusion

From theoretical proofs and practical simulations, it is evident that when the grid's long-range connections change from being arbitrarily connectable to only generating horizontal and vertical connections, the optimal r parameter shifts from 2 to 1. However, the sampling probability at each distance remains unchanged.

References

[1] J. Kleinberg. "The Small-World Phenomenon: An Algorithmic Perspective". In: *Proceedings of the 32nd ACM Symposium on Theory of Computing* 99.1 (2000), pages 163–170. DOI: 10.1145/335305.335325.