Page Rank, Trust Rank and Spam Farm in Graphs

Chenkai Wang

Abstract

In this essay, the author mainly uses the package “networkx”[1] to generate the desired network and uses the Page Rank algorithm to evaluate it. Meanwhile, the spam farm is created to disperse the Page Rank algorithm. Finally, the Trust Rank is applied to defeat the spam farm successfully.

Keywords: Page Rank Algorithm, Trust Rank Algorithm, Spam Farm

SUSTech Southern University of Science and Technology

Contents

1 Network and Page Rank algorithm

1.1 Generate a network

I first design an undirected network with 100 nodes which follows a power law distribution with $\gamma = 2.5$. Moreover, the degree of the nodes is restricted between 2 to 20. The step is shown as follows:

- Step1 Use *powerlaw_sequence* function in the package to generate the samples from power law distribution with$\gamma = 2.5$, say $d_i, i = 1, 2, 3, \cdots$.
- Step2 Generate a sequence p to store the first 100 d_i satisfies $2 \leq d_i \leq 20$, which are denoted by$p_1, p_2, \cdots, p_{100}$.
- Step3 Recall that the degree of an undirected graph should be even. If $\sum_{i=1}^{100} p_i \ mod \ 2 \neq 0$. Repeat the previous step until $\sum_{i=1}^{100} p_i \ mod \ 2 = 0$.

Now we already get the eligible nodes stored in the sequence generated by the desired restriction, which is shown below.

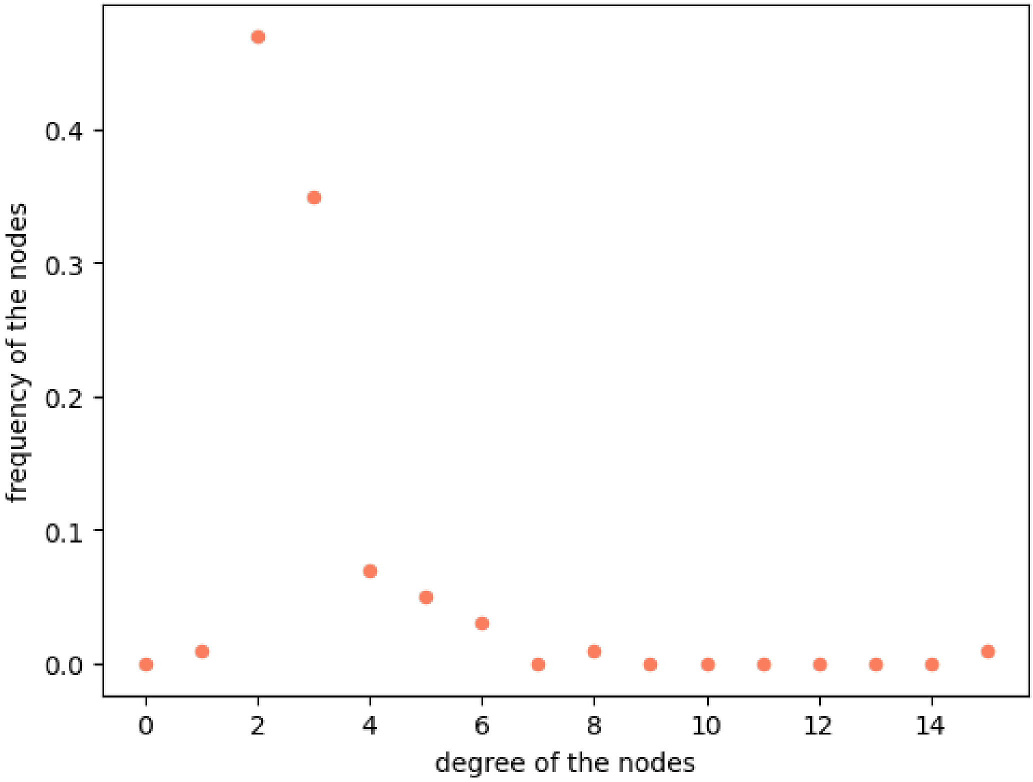

Figure 1: Degree Distribution of the Nodes

Once the desired degree distribution has been constructed, we can input it into the *configuration_model* function to generate the undirected graph. The desired directed graph, as shown in the picture below, can be generated by randomly transforming undirected edges into directed edges.

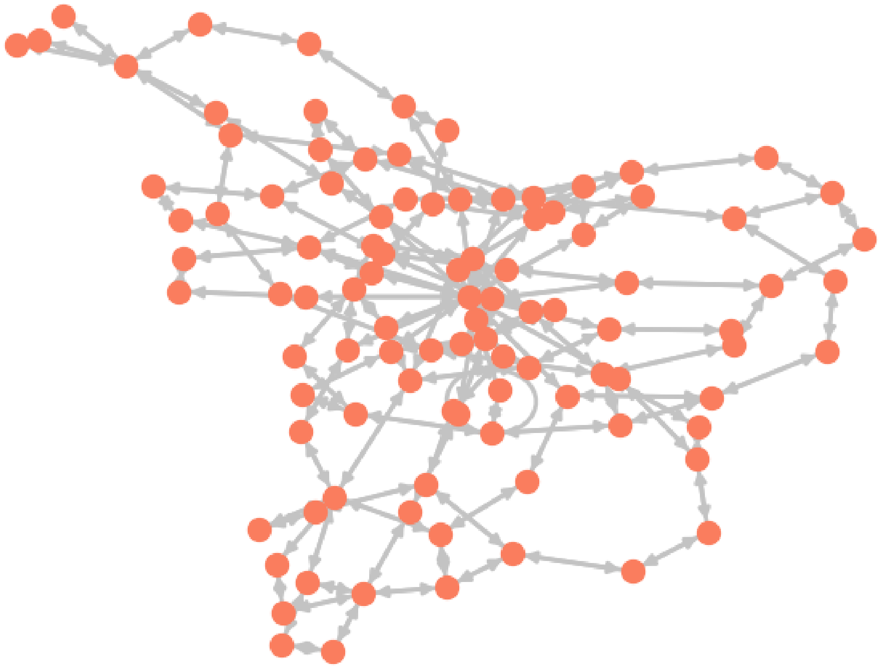

Figure 2: Directed Graph

The graph has 100 nodes and 147 edges, with the maximum and minimum degrees equal to 15 and 2, respectively.

1.2 Page Rank Algorithm

Page Rank[2] is an algorithm used by Google Search to rank web pages in their search engine results. It is named after both the term ”web page” and co-founder Larry Page. Page Rank is a way of measuring the importance of website pages. The following part of this section is details.

Let r be an n-dimensional vector with each component representing the rank of a page. Here we set $n = 100$ to meet the requirements. Define the adjacent matrix M as

$$M_{i,j} = \begin{cases} \frac{1}{d_j} & \text{if} \quad j \to i \\ 0 & \text{if} \quad j \nrightarrow i \end{cases}$$

where d_j denotes the number of out-links of page j. Then we can calculate the rank of the pages by solving $r = Mr$ where M is a stochastic matrix. However, there could exist a spider trap or dead end, which will undermine the reliability of the result or invalidate the algorithm in the worst scenario. Consequently, we will add random teleport in our algorithm to solve the problem. More specifically, at each step, there are two options:

- with probability β, follow a link at random,
- with probability $1 - \beta$, jump to some random page.

The common value of β is around 0.85. In this case, the new adjacent matrix, say A, can be represented as

$$A = \beta \cdot M + (1 - \beta) \cdot \left[\frac{1}{n}\right]_{n \times n}$$

and $n = 100$ by convention.

The algorithm is constructed as follows. Start from the united r, say $r_{\text{old}} = [\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}]^T$ repeat the following steps until $\sum_j \left|r_j^{\text{new}} - r_j^{old}\right| < \epsilon$ where ϵ dominance the accuracy.

- $r^{\text{new}} = \beta M \cdot r^{\text{old}} + (1 - \beta)[\frac{1}{n}]_{n \times n}$
- Standardize r^{new}
- $r^{\text{old}} = r^{\text{new}}$

Using the above algorithm, we can get the importance of each node (page) created in section 1.1. Moreover, the correlation between the page rank value and the degree of nodes is 0.98, which is remarkably close to 1.

2 Spam Farm and Trust Rank Algorithm

2.1 Spam Farm

Randomly Pick a node(page) from the graph generated in section 1.1, say p_{50}. Construct a spam farm with k nodes around it and add create double links between these nodes and the target. We want to find out how the page rank value of a specific node changes with the k, and the result is shown below.

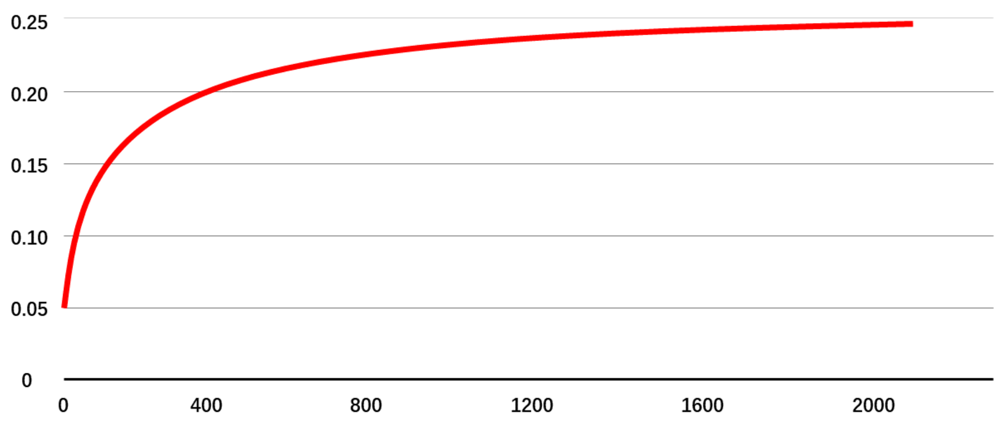

Figure 3: The relationship between Spam Farm size k and page rank value of p_{50}

From the picture, we can see that the page rank value increases extremely fast at the very beginning. However, as n increases, it increases slower and slower. In other words, when n is relatively large, the marginal revenue of improving page rank value through Spam Farm becomes lower and lower, and the marginal cost becomes higher and higher.

2.2 Trust Rank Algorithm and Spam Mass Comparison

2.2.1 Trust Rank Algorithm

Trust Rank is an algorithm that conducts link analysis to separate useful web pages from spam and helps search engines rank pages in the Search Engine Results Page. In other words, Trust Rank Algorithm can help appraise the quality of a web page. Trust Rank comes from a natural assumption that good pages, such as school or government websites, rarely point to bad pages, such as the Spam Farm we have just created. Therefore, to calculate the trust rank value, we need to choose the good pages first and flow the trust to other pages according to the rules. Now back to the graph generated in section 1.1, since generated only with the restriction of the degree and we do not have prior knowledge about the reliability of each node, we can randomly select five nodes from the network to expect p_{50}. Suppose the nodes set $\{18, 32, 45, 63, 72\}$ is chosen randomly. Trust Rank Algorithm is pretty similar to Page Rank. Suppose the trust of page p is t_p and package p has a set of out-links o_p. For each $q \in o_p$, p confers the trust $\frac{\beta t_p}{o_p}$ to q where $0 < \beta < 1$. Note that trust is additive; we can get the trust of the whole graph.

2.2.2 Spam Mass Comparison

Before defining Spam Mass, we first introduce some notation. Let r_p , r_p^+ denote the page rank of page p and the page rank of $p$with teleport into trusted pages only respectively. Then Spam Mass of p can be defined as $\frac{r_p - r_p^+}{r_p}$, which reveals the fraction that page rank comes from Spam Farm. The following pictures show the spam mass of a target node with different spam farm sizes k applied, where the x-axis represents the spam farm k and the y-axis represents the spam mass.

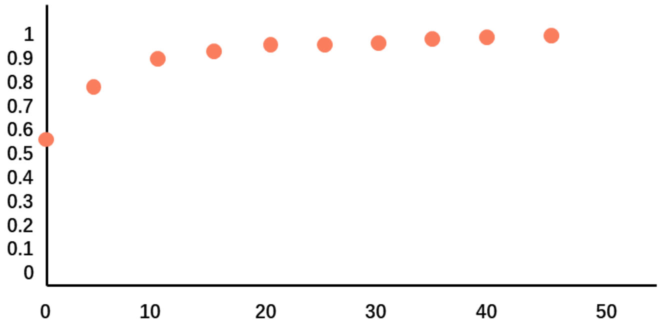

Figure 4: The relationship between Spam Farm size k and spam mass of the target

We can see that the fraction increases rapidly when k increases from 0 to 10. The possible reason could be that the initial graph size is so tiny.

References

[1] Dan Schult et. al Aric Hagberg Pieter Swart. https://networkx.org.

[2] Sergey Brin and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine". In: *Computer networks and ISDN systems* 30.1-7 (1998), pp. 107–117.