

Verifications of growing models for cooperative R&D networks

Hiroyasu Inoue
Osaka Sangyo University, Osaka 574-0013, Japan

Abstract

We investigated a network based on joint patent applications and modeled it to reveal the dynamics of cooperative research and development among organizations. The network uses nodes to represent the offices of organizations and links to represent joint patent applications. We included about five-million Japanese patents issued between 1993 and 2002. The results are summarized as follows. (1) The distributions of degree and node density follow power laws. (2) The distribution of link distance is inversely related to link distance. (3) We found a model that could generate a network consistent with the above results. It is a revised model of preferential attachment that takes into account the distances between nodes.

1 Introduction

Most countries are trying to develop industrial clusters where innovative research and development actively occur because innovation is important for economic growth. Generally, the industrial clusters consist of organizations¹, infrastructures such as roads, and links between organizations (networks²). The last component (networks) includes various networks such as transaction, share-holding, and director-dispatch networks. Networks affiliated with innovation have a strong relation with the concept of open innovation [1]. Cooperative research and development (R&D) is one aspect of open innovation, and a network based on cooperative R&D among organizations is the focus of this paper.

We investigated the structure of a cooperative R&D network among organizations and a growth model for the network. We previously analyzed some aspects of the network [2]. This paper presents other analyses of the network that are necessary to verify the model. In the investigations, we used data from joint applications of Japanese patents because we needed long-term and large-scale data on cooperative R&D. Since there have been no organized Japanese patent data that could have been used for creating networks until recently, there are few studies on networks consisting of joint patent applications in Japan.

It should be noted that joint patent applications are only part of the results of cooperative R&D. However, our focus is not an accurate investigation into cooperative R&D, but an investigation into the structure of networks based on cooperative R&D. Hence, joint patent applications were sufficient for our investigations.

Many previous studies have examined whether organizational networks are closer to small world networks [3] or scale-free networks [4] because they have put forward the hypothesis that these network structures are effective. Indeed, these studies have produced some significant findings, but their approach limits discussion because of their hypothesis. This paper does not discuss how network structures like small world networks or scale-free networks affect the capabilities of organizations. Instead, we analyzed network structures and verified the growth model that can reproduce them.

¹In this paper, an organization means a group that has corporate status.

²Networks consist of nodes and links.

To summarize, the objective of this paper is to present our analysis of cooperative R&D networks and verify their growth model. We assumed that we can speculate on cooperative R&D activities by studying joint patent applications.

This paper is organized as follows. Section 2 explains the patent data and a joint application network. In Section 3, we discuss our analysis of the network, and in Section 4 we propose a growth model. In Section 5, we discuss the results, and we conclude this paper in Section 6.

2 Japanese patent data and joint application networks

The Japanese Patent Office publishes patent gazettes, which are called *Kokai Tokkyo Koho* (Published Unexamined Patent Applications) and *Tokkyo Koho* (Published Examined Patent Applications). These gazettes are digitized, but not organized because they do not trace changes in trade names or firms' addresses. To solve these problems, Tamada et al. organized a database [5], and this paper is based on that database. It includes 4,998,464 patents published from January 1993 to December 2002 in patent gazettes. We used the names and addresses of organizations in these patents, as well as the addresses of the inventors, to create a joint application network.

The network we created has head or branch offices of organizations as nodes. The nodes have identifiers, which are combinations of organizations' names and the addresses of offices. Links are created among nodes when they have at least one joint patent application. The duplication of links has been ignored in this paper.

Applicants are organizations in almost all the patents, and the addresses of their head offices are recorded. However, inventions obviously not only occur at head offices but also at branch offices. Hence, we should use the inventors' addresses to obtain the addresses of branch offices.

Based on the above, the algorithm to create the joint application network can be given as follows.

Algorithm to create joint application network

Execute the following processes for all patents.

1 Create nodes

1.1 Obtain the organizations' names from applicants' names through a filter

The filter deletes a string identifying corporate status from an applicant's name; therefore it also detects whether an applicant is an individual or not. Individuals are omitted. If all applicants are individuals, the patent is skipped without it being reflected in the network. Combinations of organization names and applicant addresses are used for node identifiers, and the identifiers are stored on a list.

1.2 Obtain organizations' other addresses by using inventors' addresses

If an inventor's address includes the name of an organization that is considered to be an address of a branch office, a combination of the matched organization's name and the inventor's address is used as an identifier of the node and it is added to the list. This process is necessary because it is not clear which organization each inventor belongs to in the patents.

1.3 Create new nodes according to the list of node candidates

If there is a node that has the same identifier in the list, a new node for the identifier is not created.

2 Create links

Links are created among nodes included in the list of node candidates so that a complete graph can be created among them. As previously mentioned, duplicated links are ignored.

3 Analyses of joint application network

We will discuss our analyses of the joint application network in this section. Table 1 lists the basic data for the joint application network. Section 2 describes how we used inventors' addresses as office

addresses of organizations. Table 1 summarizes the differences in using inventors' addresses. By using these processes, the number of nodes is increased by 2.2 times, and the number of links is increased by 1.5 times. Since we will discuss the distances among addresses, we cannot neglect these large differences in the increase in addresses.

Figure 1 plots the various distributions for the joint application network. (a) shows the degree distribution. The degree is the number of links a node has. The degree distribution is the number of nodes for each degree. The horizontal axis represents the degree, and the vertical axis shows the rank. They are logarithmic scales. In this paper, rank representations are used for distributions because we can directly understand the forms of a cumulative probability. The degree distribution indicates that the cumulative probability corresponds to a line, and it is clear that it follows a power law. Hence, the joint application network is a scale-free network. The degree exponent obtained by the method of least squares is -1.3 . This value is an approximation, but there is the other method for an accurate value [6]. However, this approximate value is sufficient for our discussion.

If the exponent of the cumulative distribution is -1.3 , the exponent of the degree distribution is -2.3 . A range of degree exponents, more than -3 and less than -2 is observed in most networks [4]. Hence, the network's degree distribution is not unusual.

(b) in Fig. 1 also shows the degree distribution, but there are five different series of plots. These plots represent the degree distributions of networks whose durations are from 1993 to the indicated years. The inclinations of these plots do not change in the accumulation of data.

(c) in Fig. 1 plots the distribution of density, which is calculated by the number of nodes within one square kilometer. The horizontal axis represents the node density, and the vertical axis represents the rank. They are logarithmic scales. We can also see the power law in this figure.

(d) in Fig. 1 plots the distribution for the link distances. The horizontal axis is the link distance, and the vertical axis is the rank. The horizontal axis is a logarithmic scale, and distances of less than one kilometer have been abbreviated. The plot almost forms a line; thus, the probability distribution of the link distance is inversely related to the link distance. In our previous paper [2], we pointed out some peaks in the link distribution. Since the vertical axis of Fig. 1 is the rank, it is ambiguous that some peaks exist. However, we can find a large distribution at around 400 km, and another at around 250 km. These distances correspond with those between Tokyo and Osaka, and between Tokyo and Nagoya. These are large cities in Japan and are connected by efficient public transportation. This implies that link generation may be facilitated by travel time, not distance. Based on this assumption, we may acquire a more smoother line in the graph if we take the travel time as the horizontal axis, not the geographic distance.

These results are similar to Yook et al.'s analytical results on the Internet [7]. Since links on the Internet exist physically and involve various costs, there are intuitive differences between the network in this paper and on the Internet. Therefore, the structural similarities are interesting. We also found that geographic proximity is important in cooperative R&D, although it has only been empirically founded that cooperative R&D is more successful when sites are closer [8].

4 Verification of growth model

This section discusses our verification of a growth model to generate the same structure as the joint application network analyzed in Section 3. The preferential attachment model [9] is well known for generating scale-free networks. This model can be simply explained as follows.

1. A new node is incrementally added.
2. Links are placed between the new node and existing nodes. The probability Π for choosing an existing node is $\Pi(k_i) = k_i / \sum_i k_i$ where k_i is the degree of the node.

The network generated by this model has a degree distribution that is expressed by $p(k) \propto k^{-3}$. This model does not take link distances into account. As a matter of course, the model should include link distance factors to show the structural features of distance.

From the above, we used a minimally modified preferential attachment model that includes distance factors [7, 10, 11]. This model has already been carefully discussed, and we think it is suitable as the first step to investigate the geographic model for generating the joint application network. The algorithm is outlined below.

Algorithm for modified preferential attachment model

1. Start from a complete graph with m_0 nodes.
2. Add a node.
3. Append an address to it that is randomly chosen from the head or branch office addresses extracted using the method outlined in Section 3.
4. Add m links. A link has a terminal of the new added node. A node for the other terminal of the link is chosen from existing nodes using the probability

$$\Pi(k_j, d_{ij}) \propto k_j^\alpha / d_{ij}^\sigma,$$

where i stands for the new node, j stands for the existing node(s), k_j is the degree of node j , and d_{ij} is the distance between nodes i and j . The distance between the nodes is calculated by the distance between two points on an ellipsoid, which is based on the latitude and longitude derived from the addresses of nodes. Finally, α and σ are constants. Duplications of links are prohibited.

5. Repeat steps (2)-(4) a fixed number of times.

In Manna and Sen’s earlier study [10], a model where α was fixed to 1 in the above model was investigated. In that study, the model’s behavior was described during varied σ on the condition that nodes were equally distributed in a square on a plane. Comparisons of the results between those in this paper and those in this earlier study will be discussed later.

There are other models that can generate scale-free networks. The fitness model [12] is representative of these, where each node has a weight, and if $w_i + w_j > \theta$, where w_i and w_j are weights of two arbitrary nodes, and θ is a constant, a link is stretched between them. Although this model has many advantages, a threshold (θ) should be defined in some manner to use it. We do not currently know what the deterministic factor is for cooperative R&D to determine the threshold. On the other hand, the modified preferential attachment model is easier to evaluate because it only involves node degrees and link distances. This is why we chose this model for this paper.

In the rest of this section, we discuss how the model can reproduce the joint application network. To achieve this, we tried various combinations of constants (α and σ), and verified the results with the original network in degree distribution and link distance distribution.

Figure 2 plots the simulation results. In the simulation, $m_0 = 3$, $m = 3$, and the final number of nodes is 3,000. (a) and (b) in Fig. 2 show the results from different α values. In other words, these results show the effects of the exponents of degrees in the probabilities of links. In these simulations, σ is fixed to 1.0.

(a) in Fig. 2 plots the degree distribution. The axes are the same as those in the previous graphs for degree distribution. The original network follows the power law ((a) in Fig. 1), hence $\alpha = 1$ indicates the closest result. Here, $\alpha = 0$ seems to be an exponential distribution, and $\alpha = 2$ has a few large degree nodes.

(a) in Fig. 2 shows the results correspond to the analytical result that there is the relationship between α and the degree distribution [13]. That is “When P_i grows slower than linearly with k ,

the degree distribution decays faster than a power law in k , while for P_i growing faster than linearly in k , a single node emerges which connects to nearly all other nodes". Since spatial effects are not considered in Krapivsky and Rendner's study, the relationship seems to be robust in some σ range, and its analysis should be studied in the future.

(b) in Fig. 2 shows the link distance distribution. The axes are also the same as those in the previous graphs for link distance distribution. The original network is in inverse proportion, hence $\alpha = 0$ and 1 gives a closer result than $\alpha = 2$.

(c) and (d) in Fig. 2 plot the results for different σ values. In other words, these results show the effects of the exponents of link distances in the probabilities of links. In these simulations, α is fixed to 1.0. (c) in Fig. 2 plots the degree distribution. It is clear that $\sigma = 0, 1$, and 2 derive the similar results. The effect of σ seems to be small on the degree distribution. (d) in Fig. 2 plots the link distance distribution. Here, the link distance distribution depends on σ . $\sigma = 1$ achieves the closest results to the original network, $\sigma = 0$ has more long distance links than the original network, and $\sigma = 2$ provides an inverse result. Since $\sigma = 0$ refers to the original preferential attachment model, this value is insufficient for reproducing the link distance.

From these results, we can see that α and σ greatly affect the structure of the network, and $\alpha = 1$ and $\sigma = 1$ is the best combination in this analysis. However, we can gain limited understanding of the relationship between α and σ . This should be studied further in the future.

We already discussed the structure of the joint application network is similar to the structure of the Internet in Section 3. In addition to this, the simulation results are similar. As already mentioned, we used the same model as Yook et al. used in their study [7]. Our best simulation result, $\alpha = 1$ and $\sigma = 1$ is the same as those for the Internet. The data (spatial distribution of nodes) used in the simulation are definitely different because one represents data from joint patent applications, and the other represents data from the Internet in North America. Therefore, this correspondence of the different data is remarkable.

An earlier study reported that the link length distribution exponentially decays on the Internet, i.e., it is proportional to $\exp(-l/l_0)$, where l_0 is a constant [14]. However the results in this paper and Yook et al.'s study are different from that.

In Section 3, we referred to Manna and Sen's study [10]. In this study, α is fixed to 1 and σ is varied. One of their results is the link distance distribution is proportional to $l^{-\alpha+m-1}$ where m is a dimension. In this paper, α is 1 and m is 2, hence the link distance distribution is constant. This conclusion is different from our simulation results where the degree distribution is proportional to l^{-1} . The reason for this difference seems that these simulations are based on different distributions of nodes on a plane.

Another earlier study by Manna and Kabakcioglu [11] points out that the link distance distribution has a stretched exponential tail for a scale free network with the shortest total link length. This result is different from ours as well as Yook's and Waxman's. However, this earlier study assumed that nodes were equally distributed in a square on a plane, and it seems that this assumption can cause differences in the results. Therefore, we cannot immediately say that a joint application network and the Internet have a tendency for redundancy in link length, and we need to investigate this carefully in the future.

5 Discussion

We analyzed the degree distribution, the node density distribution, and the link distance distribution in this paper, and presented the capabilities of the model to reproduce a network that is similar to the original network with these distributions. However, there are other measurements (e.g., clustering coefficients, degree correlations, and betweenness.) for analyzing networks. The model in this paper does not reproduce a network that has the same structure with these other measurements. Indeed, cooperative R&D is based on complex conditions such as correspondence in technical fields; models

that only use degrees and link distances do not necessarily reproduce precise network structures.

There is another limitation to the model. Intuitively, organizations want to increase their links to other organizations. However, this model cannot demonstrate how the number of links can be increased, although it can show where new links are probably created. This should be complemented by other studies.

Even though there are limitations such as these, the findings in this paper can be interpreted differently. Degree distribution is one of the most basic measurements because it reveals the existence of hubs of various sizes. Also, the link distance distribution is also important because policies to support industrial clusters are based on the concept of geographical aggregation. Therefore, the model discussed in this paper can reproduce these two important structures for the network, and it is based on its local information that is possibly acquired by observation. Obviously, global information about the network is generally difficult to obtain.

The rest of this section discusses the implications for business administration derived from the results. We can see that the probability of obtaining links in the growth model increases as the degree of nodes increases. We can consider links as an absorptive capacity (the ability to obtain technological knowledge from the outside), and this ability facilitate nodes obtaining more links. On the other hand, the growth model also reveals that the sites of offices are important. Hence, R&D offices should be located in areas where cooperative R&D is active because it helps these offices to increase their links. As we previously discussed, the requirements to balance degrees and link distances in probability reveal that neither factor has an extreme effect on the other. Since locations can easily be controlled compared to the number of links, a firm that does not have a strong R&D office has an opportunity to complement its shortage of links by taking advantage of the distances between offices. These considerations basically support the Japanese industrial cluster project, which focuses on the geographical aggregation of organizations centered on universities. However, Japanese universities do not have as many links as those in the United States, which has models of industrial clusters. This is because Japanese universities have not officially been encouraged to engage in cooperative R&D with other organizations. A new law to encourage universities to undertake cooperative R&D went into effect recently in Japan. We should take this background into account in discussing Japanese industrial clusters.

6 Conclusion

We analyzed a joint application network in this paper, and verified the growth model for it. We put forward the hypothesis that the structure of the joint application network could be regarded as the structure of a cooperative R&D network.

The degree distribution and node density distribution follow power laws. The link distance distribution is in inverse proportion. Since it has only been empirically stated that cooperative R&D is more successful when sites are close, the importance of geographic proximity for cooperative R&D has been demonstrated for the first time. Also, these structural features of cooperative R&D are similar to those on the Internet.

We verified the growth model for the joint application network. The model determines the new link probability by the degree and distances between nodes. As a result, the network is well reproduced by the model with a new link probability that is proportional to the degree and inversely proportional to the link distance. This result is also similar to that on the Internet.

References

- [1] H.W. Chesbrough. *Open innovation*. Harvard Business School, 2003.

- [2] H. Inoue, W. Souma, and S. Tamada. Spatial characteristics of joint application networks in Japanese patents. *Physica A*, 383:152–157, 2007.
- [3] D.J. Watts and S.H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393:440–442, 1998.
- [4] A.L. Barabási and Z.N. Oltvai. Network biology: Understanding the cell’s functional organization. *Nature Reviews Genetics*, 5:101–113, 2004.
- [5] S. Tamada, Y. Naitou, F. Kodama, K. Gemba, and J. Suzuki. Significant difference of dependence upon scientific knowledge among different technologies. *Scientometrics*, 68(2):289–302, 2006.
- [6] A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *arXiv:0706.1062*, 2007.
- [7] S. Yook, H Jeong, and A.L. Barabási. Modeling the Internet’s large-scale topology. *Proceedings of the National Academy of Sciences*, 99(21):13382–13386, 2002.
- [8] R. Ponds, F. van Oort, and K. Frenken. The geographical and institutional proximity of scientific collaboration networks. *Regional Science*, 86(3):423–444, 2007.
- [9] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
- [10] S.S. Manna and P. Sen. Modulated scale-free network in the Euclidean space. *Physical Review E*, 66, 066114, 2002.
- [11] S.S. Manna and A. Kabakcioglu. Scale-free network on Euclidean space optimized by rewiring of links. *Journal of Physics A: Mathematical and General*, 36(19):L279–L285, 2003.
- [12] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Muñoz. Scale-free networks from varying vertex intrinsic fitness. *Physical Review Letters*, 89(25):258702, 2002.
- [13] P.L. Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, 63, 06123, 2001.
- [14] B.M. Waxman. Routing of multipoint connections. *IEEE Journal on Selected Areas in Communications*, 6(9):1617–1622, 1988.

Table 1: Number of nodes and links in joint application network

Use of Inventors' Addresses	Use (head and branch offices)	No use (head offices only)
Number of nodes	54,197	24,767
Number of links	154,205	105,088

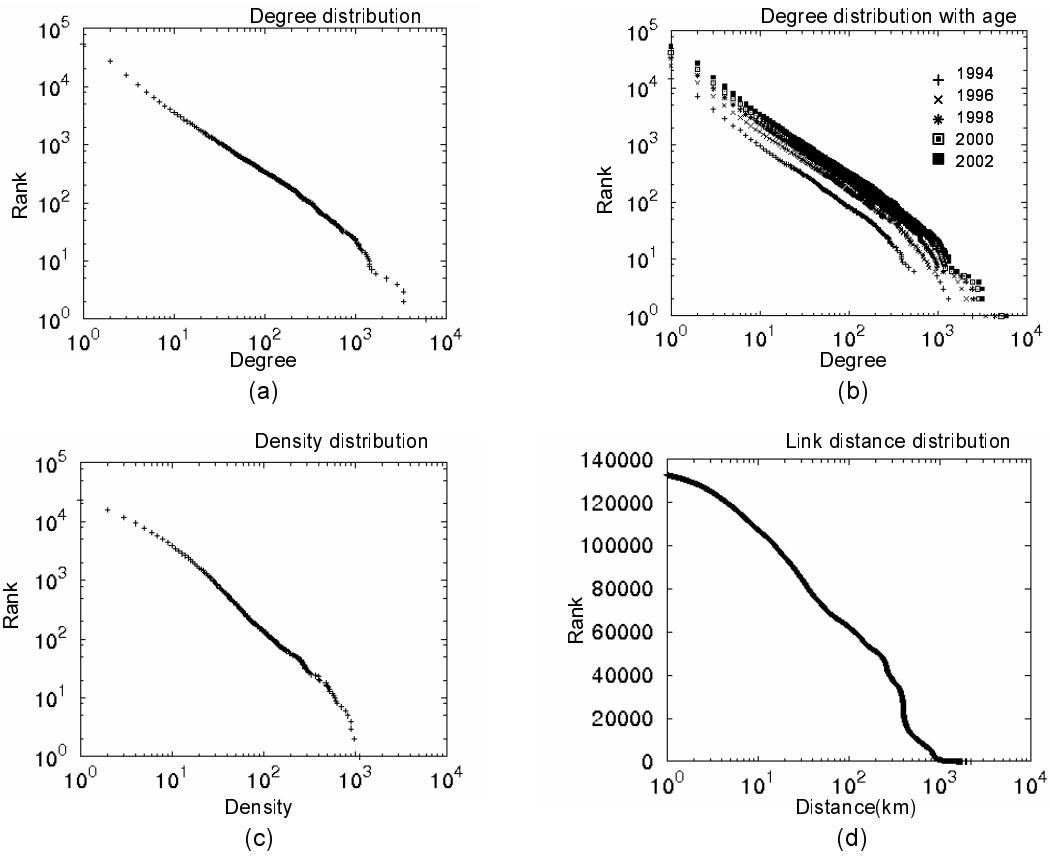


Figure 1: Distribution data. (a): degree distribution from 1993 to 2002. (b): degree distributions from 1993 to indicated years. (c): density distribution. (d): link distance distribution. Vertical axis is linear scale in this graph.

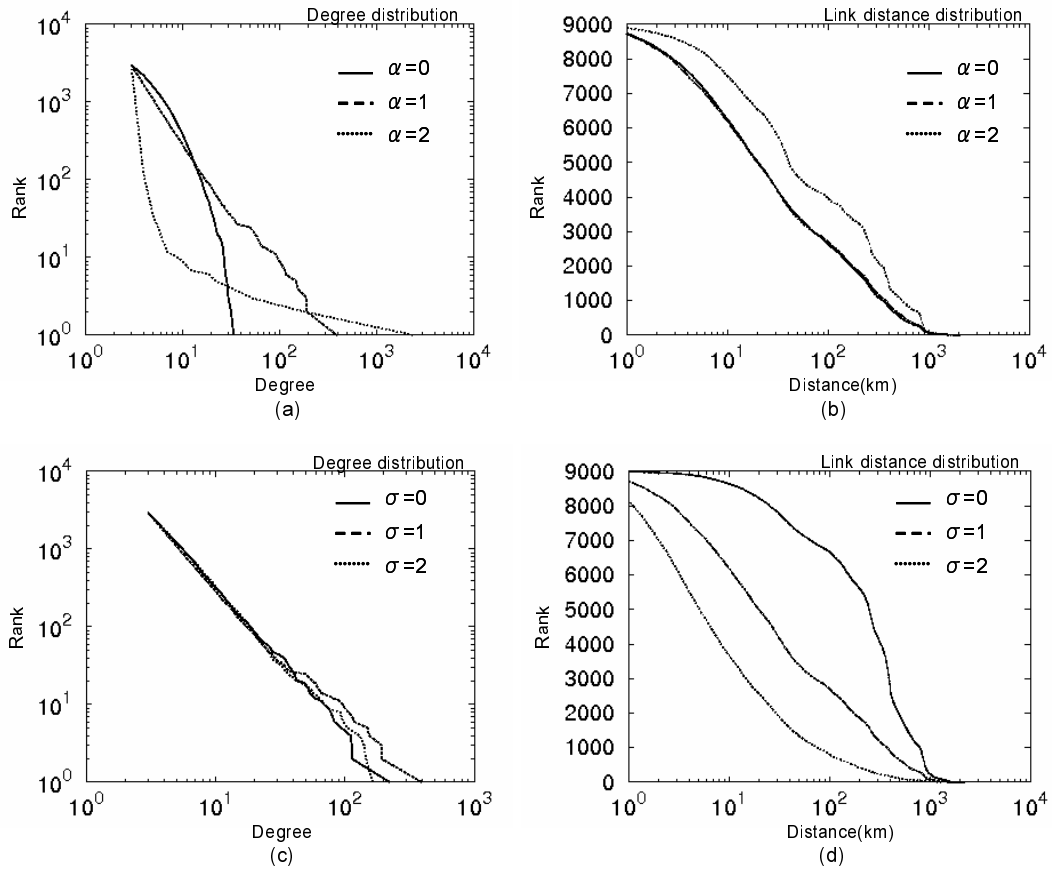


Figure 2: Network data generated by model. (a) is degree distribution and (b) is link distance distribution when $\alpha = 0, 1$ and 2 , and $\sigma = 1$. (c) is degree distribution and (d) is link distance distribution when $\alpha = 1$, and $\sigma = 0, 1$ and 2 .