Proposal of Network Generation Model based on Latent Preference Topic

Ikuto Akayama
Graduate School of Engineering Science
Osaka University
Osaka, Japan

Yoshinori Hijikata
School of Business Administration
Kwansei Gakuin University
Hyogo, Japan
contact@soc-research.org

Toshiya Kuramochi
Graduate School of Engineering Science
Osaka University
Osaka, Japan

Nobuchika Sakata
Graduate School of Information Science
Nara Institute of Science and Technology
Nara, Japan

ABSTRACT

People select whom to follow on social networking sites based on the topics that interest them. In this paper, we propose a new generation model for complex networks to mimic people's following behavior. In our proposed model, a node selects a target node to make a directed link based on the latent topic. We examine the features of the networks generated by our model through computer simulation. In the simulations, we calculate the average path length, clustering coefficient, and power exponent, which are representative evaluation indices of the network, and check whether they satisfy the properties of complex networks.

CCS CONCEPTS

• **Networks** → **Network simulations**; *Topology analysis and generation*; *Network dynamics*;

KEYWORDS

Complex network, Topic Model, Simulation, Network generation model

1 INTRODUCTION

The field of complex networks became popular in the late 1990's after the ground-breaking discoveries of structural characteristics of small-world [18] and scale-free networks [3]. Since then, many researchers have been analyzing graph structures such as the World Wide Web, human social networks, and article citation networks. Small-world networks have short average path lengths between nodes and many clusters[18]. In addition, on the scale-free network, the probability distribution of node degree k follows the power law distribution $P(k) \sim k^{-\gamma}$ (where γ is a constant) [2].

Several network generation models have been proposed that artificially generate networks to discover how networks with these properties are created. Major network generation models include the Watts-Strogatz (WS) model [18], the Barabási-Albert (BA) model [2], and Connecting Nearest Neighbor (CNN) model [17]. The WS model can generate small-world networks by randomly swapping edges of grid networks. The BA model can generate scale-free networks by preferentially selecting the link destination node based on the degree of the node (priority attachment)[1, 12]. The CNN model can produce highly clustered networks by randomly selecting a target node from the neighborhood of the source node and creating an edge to it.

In the conventional model, the network grows by using the structural features (such as number of edges and nodes in the vicinity (distant)) with the use of an algorithm, so the entire network is controlled and grown from a unified viewpoint. However, in an actual network in which a person works as a subject, nodes do not necessarily determine the destination of the link under the uniform criteria (e.g. the distance to the node or the degree of the node). It seems that each node (person) determines the destination of an edge based on individual interests and preferences. For example, someone connects to a person having the same hobby and another connects to a person with the same job. In particular, Twitter's social graph is reported to express not only social acquaintance relationships, but also hobbies and concerns [11]. In addition, analysis of the topic has also been performed in the citation relations of research papers[14]. It has been revealed that topics are intervening in citations. In this way, connections are found based on the properties of the individual nodes in actual networks in which people are the subject (especially directed networks). We need to know what type of network will be generated by the network generation model in which a node connects to other nodes with meanings or roles.

We considered that there exist common latent topics between the two nodes of an edge in real directed graphs where

IMCOM '18, January 5–7, 2018, Langkawi, Malaysia

ACM ISBN 978-1-4503-6385-3/18/01…$15.00
https://doi.org/10.1145/3164541.3164580

a person works as a subject (one node corresponds to a person). In this research, we propose a directed network generation model that determines the destination node of an edge based on latent topics. That is, a node has a probability distribution on which topic will be selected, and it selects a topic according to its probability distribution. Furthermore, a topic has a probability distribution on which a node will be selected: it selects a node according to the probability distribution. In this way, our model produces a network with directed edges (oriented edges) from one node to another. This is the same process as document generation realized by Latent Dirichlet Allocation (LDA) [5], which is a popular topic model used for document classification and topic analysis. Thus, we exploit the computational process performed by LDA in our proposed model.

In this research, we conduct a computer simulation using our network generation model to see whether it can produce the features of complex networks (small-world property, scale-free property). We also verify whether the degree of the characteristics of the network to be generated can be changed by changing the parameters of the proposed model.

The paper is organized as follows. Section 2 describes related research and Section 3 explains the proposed model. In Section 4, we evaluate the network generated by the proposed model quantitatively using the evaluation index of a representative network. Section 5 reports the limitations of our model, and Section 6 presents conclusions.

2 RELATED WORK

2.1 Network generation model

There has been a great deal of interest in the study of complex networks. Researchers have reported properties such as the small-world effect, power-law degree distribution, high clustering coefficient, and high degree correlation. Various network generation models that generate artificial networks that have the same property as real-world network have been proposed in the literature. The WS model generates small-world networks by re-connecting edges in a random graph[18]. However, re-connecting edges is an unnatural way of growing of a real-world network. The BA model provides a mechanism that explains the origin of the power-law degree distribution[3]. This model is based on two fundamental properties of real-world networks: natural growth and preferential attachment. Natural growth means that real networks become larger through the addition of nodes and edges. Preferential attachment is the property that new nodes added to the network are attached preferentially to existing high-degree nodes. The BA model cannot generate small-world networks although it explains the power-law degree distribution. To overcome this problem, extended versions of the BA model such as the Holme-Kim model[10], the deactivation model [4], and the CNN model [17] were proposed. These models capture the whole network in a unified viewpoint. In addition, the multi-agent model that assumes nodes to be agents is also proposed[15]. In this model, new edges are generated by the voting results of nodes. The Connecting Nearest Neighbors with Random linkage (CNNR) model is derived by adapting random linkages to the CNN model, which is based on the friends of friends are friends rule [19].

This research provides the network growth models based on the network structure: node degree, path length between nodes, neighbors, and so on. We assume that edges in complex networks are generated based on latent topics that an individual node owns.

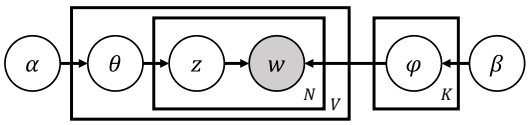


Figure 1: Graphical model of LDA (Reprinted from [5]).

2.2 Application of LDA

Many researchers have attempted to extract latent topics in data and apply them to various applications. LDA is a representative topic extraction method, and is often used in the field of natural language processing. In recent years, some researchers have tried to apply LDA for analyzing network structures, especially for extracting communities. The LDA-G proposed by Henderson and Eliassi-Rad [9] found a community using the LDA that regarded the adjacent node set of the node v as the word set of the document d. Some studies have also applied LDA to Twitter's social graph based on similar ideas. Cha and Cho [6] labeled the user's interest to the directed edge and classified it as a community for each topic. Although there are several studies where LDA is applied to network analysis, to the best of the authors' knowledge no studies where LDA is applied to network generation have been proposed. In this research, we investigate the nature of the generated network when LDA is applied to network generation.

3 PROPOSED METHOD

In this section, we present a network generation model based on latent topics. We propose a new type of network generation model by applying the process of a document-generation model using LDA to network generation. Figure 1 is a graphical model of LDA's document generation process. Blei et al. [5] modeled the selection of words in documents, but we regard documents and words as nodes to be connected and apply them to network generation. In Figure 1, α, β are hyper parameters, w is a node, z is a topic, θ is a probability distribution of the topic selected by a node, and ϕ represents the probability distribution of the node selected by a topic.

The probability distribution θ_v represents the ease of selecting the topic of node v, and

$$\theta_{v,k} = \frac{n_{v,k} + \alpha}{\sum_j (n_{v,j} + \alpha)} \tag{1}$$

represents the probability that the node v selects the topic k. In this equation, $n_{v,k}$ and $n_{v,j}$ are the number of times that node v selected the topic k and j, respectively. In addition, the probability distribution ϕ_z represents the ease of selecting the node of topic z, and

$$\phi_{z,w} = \frac{n_{z,w} + \beta}{\sum_i (n_{z,i} + \beta)} \tag{2}$$

represents the probability that the topic z selects the node w. In this equation, $n_{z,w}$ and $n_{z,i}$ are the number of times that topic z selected the node w and i, respectively.

In the process of network generation in our proposed model, all nodes have a probability distribution θ and all topics have a probability distribution ϕ. In generating the edge from node v to w, node v chooses topic z with preferential selection based on θ_v. In the next step, the topic z selects the node w with preferential selection based on ϕ_z. By following this process, an edge from the node v to the node w is generated with the attribute of the topic z.

In our model, a network is generated by branching the node generation process and the edge generation process with probability. This is a model of a phenomenon common in real networks. For example, after a new web page is created, the author might gradually add links to other websites of their interest or those related to his web page. In addition, in social networking services (SNS), when a new user comes to the service, he or she may gradually follow other users. Furthermore, the edge obtained in the process of Figure 1 has a direction. As the edge from the node v as the starting point to the node v as the end point is generated in the proposed model, the obtained network is a directed graph.

Algorithm 1 shows the details of our model. It generates networks by the iterative addition of new nodes and new edges, as does the BA model. The proposed model starts with a complete graph G_0, which consists of m nodes. Here N is the number of nodes when stopping the algorithm. When edges are generated in the network, the source node (starting point) is selected by preferential selection according to its degree (the number of edges of the node). The out-degree of the node i (source node) is assumed to be k_i, the probability of selecting node i is expressed by

$$p_i \sim \frac{k_i}{\sum_j k_j} \tag{3}$$

4 SIMULATION EXPERIMENT

We conducted a simulation experiment to verify whether the network generated by the proposed model has the features of complex networks and whether the degree of those features of the generated networks can be diversified by changing the model parameters.

Algorithm 1 Proposed model

```
Start with a complete graph G_0(V, E)
# of node n ← 0
repeat
    Generate p ∈ [0, 1] randomly.
    if Probability p ≤ P then
        Add node v_new to V, v_new has no edges.
        n ← n + 1
        for i = 0 to m do
            Select topic z distributed under θ_{v_new}.
            Select node v_old distributed under φ_z.
            Add edge e = (v_new, v_old) to E.
        end for
    else
        Select node x according to out-degree.
        Select topic z distributed under θ_x.
        Select node y distributed under φ_z.
        Add edge e_new = (x, y) to E.
    end if
until n = N
```

4.1 Evaluation indices

We evaluate the characteristics of the network generation model using the following indices.

- Exponent γ_{in}, γ_{out} [3]: The exponent of the degree distribution. Here, γ_{in} corresponds to the exponent of the in-degree distribution; γ_{out} corresponds to the exponent of the out-degree distribution. Larger γ represents that fewer nodes have larger degree.
- Average path length L [12]: The average of shortest path length among all node pairs. Small L means that the network has a strong small-world property.
- Clustering coefficient C [13]: Probability that a node is connected to its neighbors. That is, large C means that the network has the strong cluster property. However, since the clustering coefficient is defined in an undirected graph, we use an extended measure so that the measure can be calculated for a directed graph [8]. The clustering coefficient C used in this study is calculated as follows.

$$d_i^{\leftrightarrow} = \sum_j a_{ij} a_{ji} \tag{4}$$

$$d_i^{total} = \sum_j (a_{ij} + a_{ji}) \tag{5}$$

$$C = \frac{1}{N} \sum_i \frac{\frac{1}{2} \sum_j \sum_h (a_{ij} + a_{ji})(a_{ih} + a_{hi})(a_{jh} + a_{hj})}{[d_i^{total}(d_i^{total} - 1) - 2d_i^{\leftrightarrow}]} \tag{6}$$

Here N indicates the total number of nodes in the network, A corresponds to the adjacency matrix representing whether an edge exists between a node and another node, and a_{ij} is a component of A. The numerator of equation (6) represents the number of clusters generated by the node i and another two nodes j and

Table 1: Standard value of parameters of our proposed model

Valiable parameters	Value
Number of topics K	3
Number of seed nodes m_0	3
Number of edges in edge generation process m	1
Branching probability P	0.5
Hyper-parameter α	0.05
Hyper-parameter β	0.05

h, which are connected to i. In addition, the denominator corresponds the number of clusters that can be generated from node i.

4.2 Environment of simulation

In this experiment, we investigate the degree to which the indices for evaluating the complex network can be changed by each parameter. Values of the indices are calculated by changing only a parameter of interest with the other parameters fixed as standard values.

Parameters in the proposed model are K representing the number of topics, m_0 representing the number of initial seeds, m corresponding to the number of edges generated in the node generation process, P corresponding to the branching probability of the node generation process and edge generation process, and hyper-parameters α, β. We set the standard value when we fix the parameters as listed in Table 1.

The hyper-parameter α affects the probability that a node selects a topic, and the larger the value, the closer the nodes are likely to select a topic randomly. In addition, the hyper-parameter β affects the probability that topics select a node, and the larger the value, the more likely the topics are to select a node randomly.

Among these, the parameters corresponding to the core part of this model are the number of topics K and hyper-parameters α, β. Therefore, we test several values of these parameters in our experiment. In addition, we will change the value of the initial seed number m_0 to see whether the size of initial network will affect the network growth.

We generated 100 networks for one set of parameters and calculated the average of the indices of the network structure. Here, we compare the network generated by the proposed model and baseline network. We use a random graph as a baseline network that is generated in a random manner. We used the Erdős Rényi (ER) model [7] for generating random graphs. The ER model is a network generation algorithm that generates edges with the probability p for all node. Inputs of the ER model are the number of the nodes N and the probability of generating edges p.

Here, the number of nodes N as input is set to 2000, and the probability of edge generation p is set to 0.001 to coincide with the theoretical value of the number of edges generated by the proposed model.

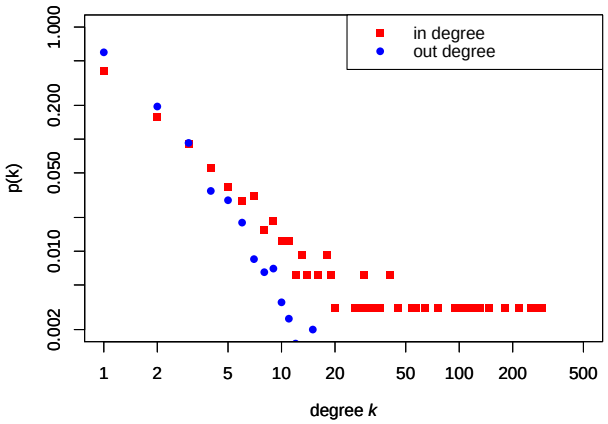


Figure 2: Degree distribution of a network generated by our proposed model.

4.3 Results and discussions

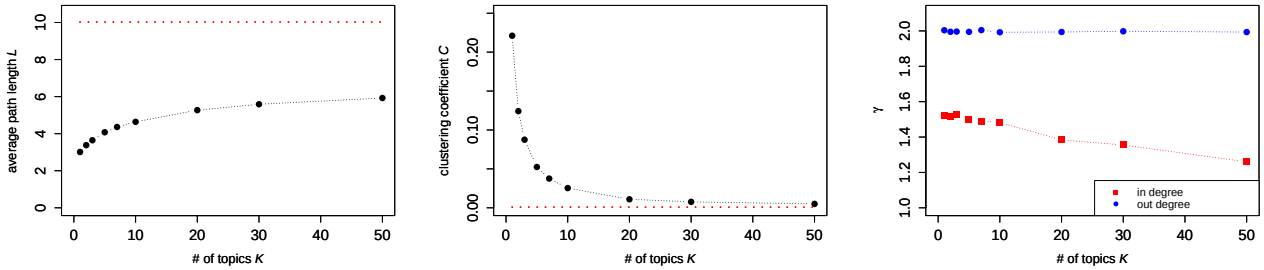
4.3.1 Scale-free property. Figure 2 represents the distribution of in-degree and out-degree when all parameters are set as standard values. The x-axis is the degree of the node and the y-axis is the appearance probability of the node. From the figure, we can see that the in-degree and out-degree follow a power-law distribution. Although it follows a power-law distribution, the distribution spreads horizontally at higher orders. This is because one or two high degree nodes were generated accidentally. Increasing the number of nodes that the network is composed of seems to shift the lateral spread as described previously to the right, that is, to higher degree.

For any combination of parameters, such a distribution was confirmed although the slope was different. From this, it can be said that our proposed model can generate scale-free networks.

4.3.2 Number of topics. Next, we show network structure indices when changing the number of topics K in Figure 3.

In Figure 3-(a), we can see that increasing the number of topics K increases the value of the average path length L. As the number of topics increases, the number of edges per topic decreases, and it seems that the degree of nodes that are likely to be chosen among topics has decreased. This leads to the decrease of nodes with high degree. It is considered that the average path length in the network has become large because we should pass more nodes with lower degree to reach another node from a node.

In addition, from Figure 3-(b), you can see that clustering coefficient C decreases as the number of topics K increases. When the number of topics increases, the nodes with higher degree come to belong to different topics. This leads to a lower probability of connecting these nodes by an edge. If nodes with a large number of edges are connected to each other, a complete graph (clique) can be easily made when another node links to the above two nodes. A complete graph



(a) Average path length **(b) Clustering coefficient** **(c) Exponent γ**

Figure 3: Results when the number of topics K is changed.

of three nodes is essential to make a cluster. However, if such connections between nodes with a large number of edges are reduced, node combinations in which all the edges between three nodes are connected to each other decrease. From this, it is considered that the clustering coefficient is much reduced.

Furthermore, increasing the value of the number of topics K from Figure 3-(c) makes it possible to slightly reduce the exponent γ_{in}. As the number of topics increases, the degree of nodes that are more likely to be selected from a topic decreases. This leads to distributing nodes with higher degree in different topics. From this, it is considered that the exponent that is the slope of the degree distribution has decreased. As described above, it can be seen that by changing the number of topics, the values of these three indices can be altered.

Also from Figure 3 (a) and (b), the average path length L of the proposed model is lower than that of the random network, and the clustering coefficient C is higher. In particular, when the number of topics is small in the proposed model, the average path length L is small and the clustering coefficient C is large. In general, complex networks are known to have smaller average path length, and relatively large clustering coefficients [16, 18]. Therefore, our proposed model can generate small-world networks when the number of topics K is small.

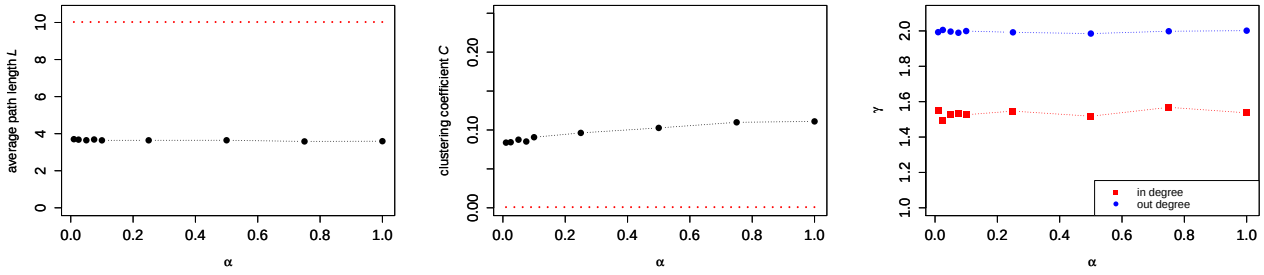
4.3.3 Hyper-parameter α. Figure 4 indicate the three indices when the hyper-parameter α is changed. When changing the value α, the three indices change in little. Thus, the hyper-parameter α cannot be used to make major changes to the characteristics of the network. Increasing the value of the hyper-parameter α smooths the probability distribution that selects the latent topic of each node, and the node is closer to choosing latent topics randomly. Thus, we can conclude that whether each node has a topic is not related to the change of the property of the complex network.

4.3.4 Hyper-parameter β. Next, Figure 5 shows the index of the network structure when the hyper-parameter β is changed.

From Figure 5 (a) and (b), we can see that by increasing the value of hyper-parameter β, the average path length becomes large, the clustering coefficient becomes small, and the power exponent γ_{in} becomes small. This is the same as the change in tendency of each indicator when changing the number of topics K. Increasing the value of the hyper-parameter β smooths the probability distribution, and the latent topic is closer to choosing nodes randomly. These results show that in network generation, when the latent topic chooses a node, stochastic preferential selection leads to reproducing the scale-free and small-world property compared with random selection. From the results that changing the hyper-parameter α did not bring significant changes in the property of the complex network, it can be said that the influential mechanism in network generation is not that each node has a latent topic, but that each latent topic has nodes that are easy to select.

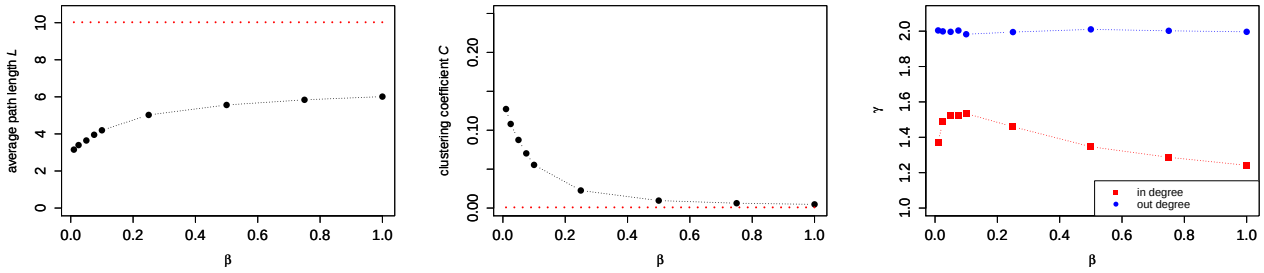
4.3.5 Number of seed nodes. Finally, we conducted an experiment to see whether the index of the network structure changes when the number of initial seeds m_0 change. Figure 6 indicate each indicator. From Figure 6-(a) and (b), we can see that the average path length and power-law exponent did not change much and the clustering coefficient became slightly smaller when we change the initial number m_0. However, the change in the clustering coefficient is small compared with the other parameters. Thus, we can say that the proposed model can generate a network with scale-free property without being influenced by the initial number of nodes.

4.3.6 Summary of network generation. To summarize the results so far, we have found that the network generated by the proposed model has scale-free property. In addition, when the number of topics K and the hyper-parameter β are small, we have found that it has the small-world property. We found that by changing the number of topics K and



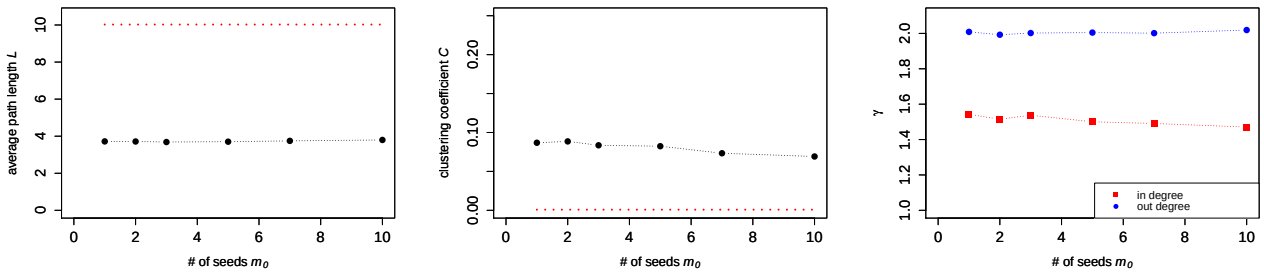
(a) Average path length (b) Clustering coefficient (c) Exponent γ

Figure 4: Results when the hyper-parameter α is changed.



(a) Average path length (b) Clustering coefficient (c) Exponent γ

Figure 5: Results when the hyper-parameter β is changed.



(a) Average path length (b) Clustering coefficient (c) Exponent γ

Figure 6: Results when the number of seed nodes m_0 is changed.

hyper-parameter β, we can change the degree of scale-free and small-world properties. However, the hyper-parameter α is found to have only a minor effect on the above property. We also found that the proposed model can generate a network with the above characteristics regardless of the initial seed number m_0.

5 LIMITATIONS

In this proposed model, we assume that the initial network is a complete graph. However, it is known that the network structure of the seed has an influence on the network growth. Other types of initial networks include a relaxed version of a complete graph by trimming several edges of the complete graph, a small tree-type network, a small star-type network, and a combination of them. It is necessary to confirm how the network grows when these network structures are given as the seed.

Our model considers only the probability distribution of selecting topics by nodes. In practice, however, the topic probability distribution of one node is considered correlated with the probability distribution of selecting nodes by topics. For example, in a SNS, if a certain user i follows another user j on the topic of the economy, the user i itself may be more likely to be selected from economy topics. Therefore, it might be better to make an association between the probability that a node selects a topic and the probability that the topic selects the node.

Finally, we set the hyper-parameter α to a scalar value in this research. However, it seems that bias exists in the ease of selection of topics in the real world. There might be a popular topics as a whole. If we set the hyper-parameter α as a vector and the ease of selecting a topic as according to the distribution of the hyper-parameter α, the above bias can be realized. It is necessary to analyze what kind of network is generated by biasing the ease of selection of topics.

6 CONCLUSION

In this paper, we have proposed a model that generates a directed network based on latent topics. Several networks have been generated by simulation with changing the parameters of the proposed model, and the average path length, clustering coefficient, and power-law exponent were calculated. In the proposed model, it was confirmed that even if any parameter was changed, a scale-free network can be generated. Moreover, by changing the β, which affects the probability distribution of the topic, and K, which is the number of topics, we could diversify the property related to small world and clusters, which is a feature of complex networks. When both values are small, it is clear that the generated network has a small-world property. In future work, we will try vectorizing hyper-parameters to see how different the generated networks are from those generated by our model.

REFERENCES

[1] R. Albert and A.L. Barabási. 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74, 1 (2002), 47–97.

[2] A.L. Barabási and R. Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286 (1999), 509–512.

[3] A.L. Barabási, R. Albert, and H. Jeong. 1999. Mean-field theory for scale-free random networks. *Physica A* 272, 1 (1999), 173–187.

[4] G. Bianconi and A.L. Barabási. 2001. Competition and multiscaling in evolving networks. *Europhysics Letters (EPL)* 54, 4 (2001), 436–442.

[5] D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[6] Y. Cha and J. Cho. 2012. Social-network analysis using topic models. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12)*. ACM Press, New York, NY, USA, 565–574.

[7] P. Erdös and A. Rényi. 1960. On the Evolution of Random Graphs. *PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES* (1960), 17–61.

[8] G. Fagiolo. 2007. Clustering in complex directed networks. *Phys. Rev. E.* 76, 2 (2007), 026107.

[9] K. Henderson and T. Eliassi-Rad. 2009. Applying latent dirichlet allocation to group discovery in large graphs. In *Proceedings of the 2009 ACM symposium on Applied Computing (SAC '09)*. ACM Press, New York, NY, USA, 1456–1461.

[10] P. Holme and B. Kim. 2002. Growing scale-free networks with tunable clustering. *Phys. Rev. E.* 65, 2 (2002), 026107.

[11] H. Kwak, C. Lee, H. Park, and S. Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web (WWW '10)*. ACM Press, New York, NY, USA, 591–600.

[12] M.E.J. Newman. 2003. The Structure and Function of Complex Networks. *SIAM Rev* 45, 2 (2003), 167–256.

[13] M.E.J. Newman and M. Girvan. 2004. Finding and evaluating community structure in networks. *Phys. Rev. E.* 69, 2 (2004), 026113.

[14] H. Okamoto. 2011. Topic-Dependent Document Ranking: Citation Network Analysis by Analogy to Memory Retrieval in the Brain. In *Artificial Neural Networks and Machine Learning ICANN 2011*, T. Honkela, W. Duch, M. Girolami, and S. Kaski (Eds.). Vol. 6791. Springer, Berlin, Heidelberg, 371–378.

[15] K. Shinoda, Y. Matsuo, and H. Nakashima. 2007. Emergence of global network property based on multi-agent voting model. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*. ACM Press, New York, NY, USA, 1–8.

[16] J. Travers and S. Milgram. 1969. An Experimental Study of the Small World Problem. *Sociometry* 32, 4 (1969), 425–443.

[17] A. Vázquez. 2003. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Phys. Rev. E.* 67, 5 (2003), 056104.

[18] Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393, 6684 (1998), 440–442.

[19] K. Yuta, N. Ono, and Y. Fujiwara. 2007. A Gap in the Community-Size Distribution of a Large-Scale Social Networking Site. (2007). arXiv:physics/0701168