Growth Models of Web Networks

Masahiro Kimura†

Abstract

The World Wide Web is constantly growing through the addition of new pages and hyperlinks created by users according to their particular interests. This article introduces research on a method of statistically modeling the growth dynamics of a network such as the Web.

1. Importance of modeling network growth

The World Wide Web provides a vast repository of information and continues to grow as an important new medium of communication. The pages and hyperlinks of the Web can be viewed as nodes (vertices) and links (edges) of a network (graph). When there is a hyperlink from one Web page to another, we can consider that the administrator of the former Web page recommends the latter Web page or that these two Web pages have some relationship in the real world. Therefore, the network structure of the Web is useful information. For example, like the HITS (hyperlink-induced topic search) algorithm or the PageRank algorithm of Google, this structure is used to improve Web search engines [1].

The Web is a growing network that is constantly evolving as a result of the addition of new pages and hyperlinks. Therefore, modeling the network growth dynamics of the Web is an important research issue. In particular, such a model could be useful to predict the network structure of the Web in the future and to deeply understand the ecology of the Web [2].

2. Scale-free networks

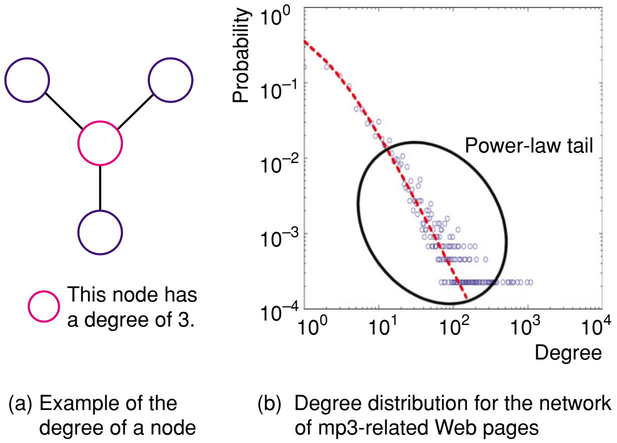
Recently, considerable attention has been devoted to exploring real-world complex networks and there has been progress in research that aims to clarify the statistical regularities of various large-scale networks and model their growth processes [3], [4]. A fundamental characteristic of any network is the degree distribution, which is defined as the distribution of the number of links for every node in the network. **Figure 1(a)** gives an example of the degree of a node in a network. Recent empirical results show that for many large-scale real-world networks, including the Web, the degree distributions do not follow Poisson distributions[*1], which the classical random graph theory[*2] expects, but possess power-law tails [2]-[4]. **Figure 1(b)** shows the degree distribution for the network of mp3-related Web pages[*3] on a double-logarithmic scale. Here, note that the power law is expressed by the linearity of the distribution curve. These observations suggest that for many large-scale real-world networks including the Web, the growth processes cannot be completely random but may obey certain self-organization principles. Moreover, after they have grown sufficiently, their degree distributions must have power-law tails.

† NTT Communication Science Laboratories
Soraku-gun, 619-0237 Japan
E-mail: kimura@cslab.kecl.ntt.co.jp

[*1] Poisson distribution: The probability distribution that arises when counting the number of occurrences of a rare event such as the number of car accidents in one day. If the degree distribution of a network follows a Poisson distribution, the probability that there exists a node with a very high degree called a hub becomes exponentially low. However, if the degree distribution has a power-law tail, this probability cannot become too low, so hubs can exist in such a network.

[*2] Classical random graph theory: This is the mathematical theory for the statistical properties of random graphs. Studies of it were begun by Erdösh and Rény in about 1960. They studied the static graphs generated by randomly adding edges to a fixed set of vertices.

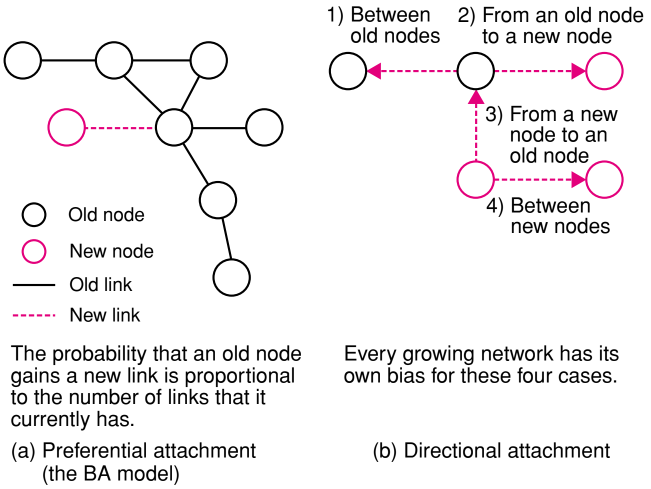
[*3] mp3: MPEG 1 Audio Layer 3. A popular format for compressing digital audio files.



(a) Example of the degree of a node

(b) Degree distribution for the network of mp3-related Web pages

Fig. 1. Degree distributions.


Old node
New node
Old link
New link
The probability that an old node gains a new link is proportional to the number of links that it currently has.
1) Between old nodes
2) From an old node to a new node
3) From a new node to an old node
4) Between new nodes
Every growing network has its own bias for these four cases.

(a) Preferential attachment (the BA model)

(b) Directional attachment

Fig. 2. Mechanism for creating new links.

Albert and Barabási discovered a network growth model satisfying these conditions [4]. The principal ingredient of the Barabási-Albert (BA) model is a mechanism of preferential attachment, in which the probability that an existing node gains a new link is proportional to the number of links that it currently has (**Fig. 2(a)**). However, the degree distribution of the BA model has been proved to follow a power law with index of –3, although large-scale real-world networks can obey various power-law distributions. Therefore, some variants of the BA model have been proposed to construct power-law distributions with indices other than –3 [4].

Since a system with a power law is known to have a scale-free (scale-invariant) nature, the network growth models with power-law degree distributions are generally referred to as scale-free models. With the aim of constructing a more precise statistical-model of a real-world growing network such as the Web, NTT Communication Science Laboratories proposed a new scale-free model for network growth that incorporates directional attachment and community structure [5].

3. Directional attachment

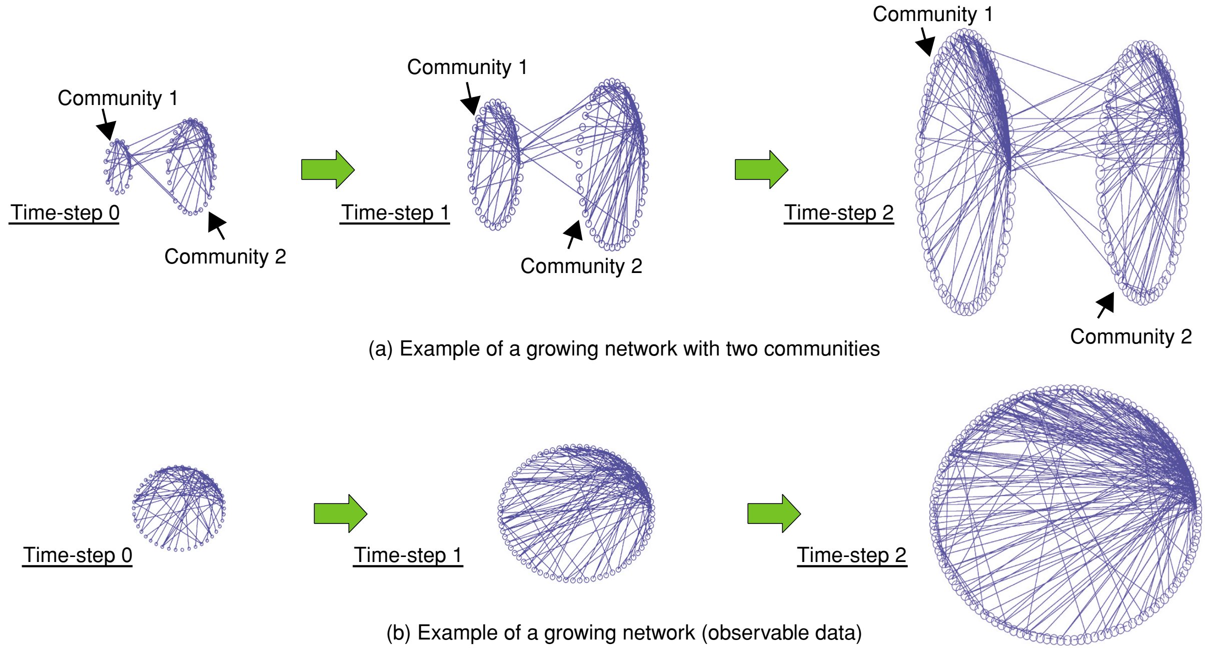
When a new link is created at a time-step, there are four ways it can be attached: 1) from an old node to an old node, 2) from an old node to a new node, 3) from a new node to an old node, or 4) from a new node to a new node (**Fig. 2(b)**). Every growing network has its own bias for these four cases. Namely, the probabilities of each of these four cases occurring can change depending on the growing network to be modeled. The mechanism that appropriately biases these four cases in a new link creation is referred to as directional attachment.

Therefore, given a growing network, for example, a network of Web pages belonging to a certain category, we can model the growth process more precisely by incorporating the directional attachment that is appropriate for it [5].

4. Community structure

We can consider that the community structure of a network is the decomposition of the set of nodes into the clusters that arise from its undirected graph structure. Namely, a community is defined as a collection of nodes in which each member node has more links to nodes within the community than to nodes outside it. One characteristic of the Web is the existence of a community structure, and the Web grows as various clusters are formed. Namely, the following situation can often be observed in its growth process: an increasing number of links are created within each community while the links between communities remain sparse (**Fig. 3(a)**). Incorporating the community structure enables us to model this sort of detailed growth process [5].

In an effort to identify communities, there have been several investigations using graph-theoretic methods. However, those investigations treated only static networks; that is, the numbers of nodes and links were not allowed to increase. Therefore, introducing the community structure into network growth



(a) Example of a growing network with two communities

(b) Example of a growing network (observable data)

Fig. 3. Network growth.

models may be a promising approach.

5. Proposed model

Here, we outline our network growth model that incorporates directional attachment and community structure. The number of communities and the number of links introduced at every time-step are fixed in advance. Now, let us describe the process of creating a new link in a network. First, both the community to which the originating node belongs and the community to which the target node belongs are probabilistically chosen. Next, whether the nodes are new or old is probabilistically decided according to the directional attachment. Next, it is probabilistically decided according to the mixture of preferential and uniform attachment which nodes are chosen as the originating node and the target node. Thus, the growth dynamics is mathematically modeled as a stochastic process. However, there are many parameters that should be adjusted in this model, such as the number of communities.

6. Learning algorithm and evaluation

To acquire a statistical model for the short-term growth process of a growing network such as the Web, we must make the proposed model learn based on the observed data. That is, we must adjust the model parameters. Note that the observed data is time-series data such as that illustrated in **Fig. 3(b)** and that the community structure shown in Fig. 3(a) is not given. At NTT Communication Science Laboratories, we have constructed an efficient learning algorithm for the network growth model with directional attachment and community structure and experimentally confirmed its effectiveness [5].

Using real Web data, we experimentally investigated the effectiveness of incorporating directional attachment and community structure [5]. **Figure 4** shows experimental results for the growing network of mp3-related Web pages for three months. The prediction error is plotted with respect to the number of communities for the models with and without directional attachment. We evaluated the ability of the learned model from the prediction performance of the probability distribution for a new link creation at the next time-step. Figure 4 shows that the prediction performance can be improved by incorporating directional attachment. We also observed that although the prediction performance can be raised by increasing the number of communities, there can be an optimal number of communities (11 in this case). This implies that the prediction performance can be improved by incorporating the community structure. Therefore, incorporating directional attachment and

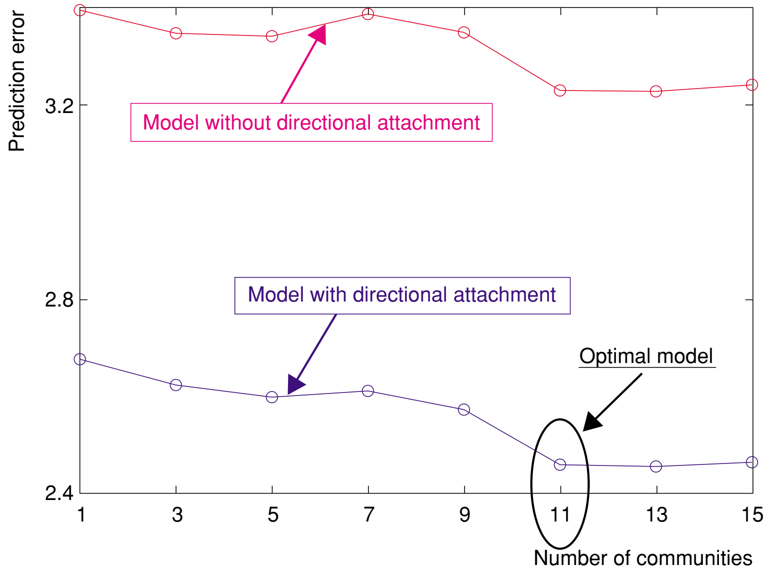


Fig. 4. Prediction performance for the growing network of mp3-related Web pages.

community structure into a scale-free model enables us to acquire a more precise statistical model for the short-term growth process of a network such as the Web.

Moreover, for the growing network of mp3-related Web pages, the inferred community structure with 11 communities can be regarded as the optimal community structure obtained by taking into account the network growth. Namely, the proposed method enables us to identify Web communities by acquiring the statistical model of the network growth dynamics.

7. Future work

We described research on modeling the growth processes of Web networks based on the statistical learning approach. The Web data set is huge and combines many types of features such as text, hyperlinks, images, and user behavior (log data). Moreover, the Web changes over time. From the scientific and technological points of view, mining and modeling this rich collection of data have become important and challenging research issues. We aim to model the Web dynamics as precisely as possible and to extract useful information from the Web based on the acquired models. Currently, in an effort to solve the problem of modeling fluctuations in the number of visitors to Web sites that form a particular market, we are attempting to categorize the sites into groups of competitors and to predict the future shares of the sites based on the observations [6].

References

[1] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, D. Gibson, and J. Kleinberg, “Mining the Web’s link structure,” IEEE Computer, Vol. 32, No. 8, pp. 60-67, 1999.

[2] J. Kleinberg and S. Lawrence, “The structure of the Web,” Science, Vol. 294, pp. 1849-1850, 2001.

[3] S. H. Strogatz, “Exploring complex networks,” Nature, Vol. 410, pp. 268-276, 2001.

[4] R. Albert and A.-L. Barabási, “Statistical mechanics of complex networks,” Reviews of Modern Physics, Vol. 74, pp. 47-97, 2002.

[5] M. Kimura, K. Saito, and N. Ueda, “Modeling of growing networks with directional attachment and communities,” Neural Networks, Vol. 17, pp. 975-988, 2004.

[6] M. Kimura, K. Saito, and N. Ueda, “Modeling share dynamics by extracting competition structure,” Physica D, Vol. 198, pp. 51-73, 2004.

Masahiro Kimura

Research Scientist, Emergent Learning Research Group, Intelligent Communication Laboratory, NTT Communication Science Laboratories.

He received the B.S., M.S., and Ph.D. degrees in mathematics from Osaka University, Toyonaka, Osaka in 1987, 1989, and 2000, respectively. He joined NTT in April 1989. From April 1989 to August 1994, he worked at NTT Human Interface Laboratories, Kanagawa and Tokyo, where he was engaged in computer graphics research. From September 1994 to January 2000, he worked at NTT Communication Science Laboratories, Kyoto, where he was engaged in neural networks research. From February 2000 to September 2001, he worked at NTT-ME Corporation, Tokyo, where he was mainly engaged in developing the 401k system of Japan Investor Solution & Technologies (JIS&T) Co., Ltd. Since October 2001, he has worked at NTT Communication Science Laboratories, Kyoto, where he has been engaged in complex systems research. He is a member of the Mathematical Society of Japan, the Japanese Neural Networks Society, the Japan Society for Industrial and Applied Mathematics, and the Institute of Electronics, Information and Communication Engineers of Japan.