2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

Proposal for a Growth Model of Social Network Service

Ken ISHIDA
Graduate School of Information Science,
Nagoya University
Nagoya, Japan
ishida@kishii.ss.is.nagoya-u.ac.jp

Fujio TORIUMI
tori@is.nagoya-u.ac.jp

Kenichiro ISHII
kishii@is.nagoya-u.ac.jp

Abstract

In this paper, we analyze the network structure of two SNSs, Academic Community System (ACS) and Amippy. From the viewpoint of network topology, the major characteristics of these data sets can be summarized as follows: low average shortest-path length, high clustering coefficient, presence of a power law degree distribution and negative assortativity. Based on our analysis, we propose a growth model of SNS networks. We conducted numerical simulations to compare actual data sets with networks generated by the proposed model. Results of simulations indicated that the processes of the CNN model and Fitness model are needed to reproduce the networks of SNSs.

1. Introduction

As part of the steady growth of new network communication tools, the expansion of Social Network Services (SNSs) such as MySpace, orcut, Cyworld, and mixi, is becoming a social phenomenon impacting societies all over the world. With expansion of these services, many kinds of SNSs can also be found, such as campus, company, and local area SNSs. SNSs provide individuals with an online space for communication on the Internet. SNSs help people to find others with common interests, exchange opinions, establish a forum for communication, and so on.

Studies of social networks are fundamental in the mining and analysis of social phenomena. A number of studies have focused on the structure of social networks. In traditional studies, since it is difficult to obtain large data sets, the focus has been on small-size networks. In the 21st century, on the other hand, networks of thousands or millions of vertices are not unusual and we can obtain data sets through online communication tools, such as SNSs and blogs.

There are some studies about social networks on the Internet as communication tools. Adamic *et al.* researched the university SNS called *Nexus*. They analyzed the structure of the network, along with attributes and personalities of the users [4]. Yuta *et al.* [16] investigated the network structure of the SNS, mixi. They found the existence of a gap in the distribution of community size, which was not present in real-world social networks. In addition, they presented a simple model, accounting for this feature. Yong-Yeol Ahn *et al.* [5] compared the structures of three online SNSs: Cyworld, MySpace, and orcut, each with more than 10 million users, respectively. Moreover, they analyzed the historical evolution of the topological characteristics of Cyworld. Backstorm *et al.* [8] investigated group formation in an SNS, LiveJournal, and co-authorship and conference publications in DBLP. They studied how the evolution of these communities relates to properties such as the structure of the underlying social networks.

In this paper, we analyze the data sets of the two SNSs, ACS [1] and Amippy [2]. In each SNS, users can make a link of friendship and establish their friendship network. Our concerns here are statistics and dynamics of this network. Through the analysis, we investigate features of SNSs and propose a the growth model of SNS networks. This model is based on simple stochastic processes and does not attempt to capture the microscopic details. However, a number of intuitively reasonable results emerge from this model.

This paper is organized as follows. In section 2, we describe details of two SNSs. In section 3, we analyze the structure of these SNS networks. In section 4, we propose a growth model of SNSs. In section 5, by conducting numerical simulations, we demonstrate the validity of the proposed model, and consider the characteristics of the model in detail. In section 6, we conclude this paper.

2. Description of data sets

2.1. Academic Community System (ACS)

ACS is a social network service at Nagoya University that is designed for communities consisting of various hu-

978-0-7695-3496-1/08 $25.00 © 2008 IEEE
DOI 10.1109/WIIAT.2008.194

man relationships. This service helps users establish and maintain an online network with other users. ACS accounts are given to anyone who belongs to Nagoya University.

ACS began its service in January 2006. We received the data set of the user friendship network in May 2008, in which the number of users and links of friendship are 709 and 2,222, respectively. The network is separated into some connected components: the biggest connected component consists of 307 users. In this paper, we investigate this biggest component, which includes 307 users and 1,892 links.

2.2. Amippy

Amippy is a social network service managed by TRY-WARP [3], an incorporated nonprofit organization in West Chiba city in Japan. People living in West Chiba or who are related to the city participate in this SNS. One purpose of Amippy is to revitalize local communities through online communications. In order to build an SNS network with a sense of security and healthiness, Amippy employs a closed invitation policy. That is, Amippy's accounts are given only to people invited by an existing user, which is different from ACS.

Amippy began its service in January 2006, just about the same time that ACS began its operation. From "TRY-WARP," the provider of the Amippy service, we received an anonymized data set of the user friendship network in May 2008. The data set contains 2,610 users and 22,434 links of friendship. The biggest connected component consists of 2,459 users and 21,258 links. We investigate this component in this paper.

3. Analysis of Social Network Services

In this section, we analyze the network topology of two SNSs (ACS and Amippy) from their average shortest-path length, clustering coefficient, degree distribution, and assortativity.

3.1. Average Shortest-path Length

Shortest-path length is defined as the shortest distance between node pairs in a network [6]. Therefore, average shortest-path length is defined as the following equation:

$$L = \frac{1}{\frac{1}{2}N(N-1)} \sum_{i \geq j} l_{ij} \tag{1}$$

where N is the number of nodes, and l_{ij} is the shortest-path length between node i and j.

3.2. Clustering Coefficient

The clustering coefficient is the average probability that two neighbors of a node i are connected [6]. For a node i, the clustering coefficient C_i is given by the ratio of existing links between its neighbors to the possible number of such connections. Thus, the clustering coefficient C_i is defined as the following equation:

$$C_i = \frac{2E_i}{k_i(k_i-1)} \tag{2}$$

where E_i is the number of links between node i's neighbors, and k_i is the degree of node i, meaning the number of links connected to node i.

Averaging C_i over all nodes of a network yields the clustering coefficient of the network C. It provides a measure of how well the neighbors of a node are locally interconnected.

3.3. Assortativity

Assortativity r is the standard Pearson correlation coefficient of the degrees at either ends of a link, and lies in the range $-1 \leq r \leq 1$. It shows whether or not the nodes in the network that have many connections tend to be connected to other nodes with many connections [13, 14]. Assortativity is defined as follows:

$$r = \frac{M^{-1}\sum_i j_i k_i - [M^{-1}\sum_i \frac{1}{2}(j_i+k_i)]^2}{M^{-1}\sum_i (j_i^2+k_i^2) - [M^{-1}\sum_i \frac{1}{2}(j_i+k_i)]^2} \tag{3}$$

where M is the number of links, and j_i,k_i are the degrees of the nodes at the ends of the ith link, with $i = 1, ..., M$.

When $r > 0$, there is a preference for high-degree nodes to attach to other high-degree nodes. The network is said to show assortative mixing. On the other hand, $r < 0$, there is a preference for high-degree nodes to attach to low-degree ones. The network is said to show disassortative mixing. It is said that the positive value of the assortativity is considered as a unique property of real-world social networks, while technological and biological networks have the negative value.

3.4. Degree Distribution

The degree distribution is defined by $p(k)$, the fraction of nodes in the network that have degree k. In other words, $p(k)$ is the probability that a node chosen at random has degree k.

Figure 1 shows the probability $p(k)$ obtained from two data sets, indicating that the degree distribution follows a power law $p(k) \propto k^{-\gamma}$.

In ACS the maximum degree user has $k_{max} = 93$, which means that the user is connected with roughly 35% of all

users. In contrast, 31.6% of all users have only one link. In Amippy, maximum degree user has $k_{max} = 628$, it corresponds to 25.5% of all users, and 38.4% of all users have only one link.

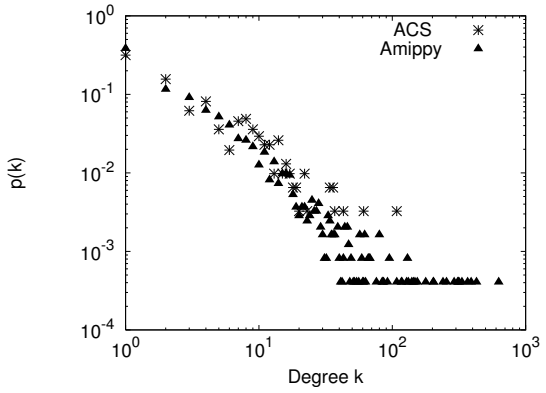

Figure 1. Degree distribution of two SNSs

3.5. Summary of Analysis

Table 1 shows the summary of the network topology analysis. The major characteristics of two data sets can be summarized as follows:

- Low average shortest-path length and high clustering coefficient (Small world network),
- Presence of a power law degree distribution (Scale free network),
- Negative assortativity.

Table 1. Properties of two SNSs

	ACS	Amippy
Users N	307	2459
Links M	1892	21258
Average shortest-path length L	3.28	3.10
Clustering coefficient C	0.479	0.395
Scaling exponent γ	1.134	1.169
Maximum degree k_{max}	108	628
Mean degree $\langle k \rangle$	6.16	8.64
Assortativity r	-0.152	-0.255

4. Modeling of Social Network Services

4.1. Characteristics of Social Network Service

Let us begin modeling by considering the following characteristics of SNSs. In general, it is said that SNSs can be characterized by the following features:

- Preferential Attachment,
- Friends of one's friends,
- Influence of Special Interest Group.

4.1.1 Preferential Attachment

The degree distribution of many real-world social networks does not follow a power-law distribution, while the degree distribution of the Web does. Instead, the distribution appears to be strongly peaked around a certain mean degree [7], meaning that many people have approximately the same degree as the mean degree. The typical explanation for this distribution is that there is recurring cost in terms of time and effort to maintain a friendship. In cases of SNS networks, however, there is only a one-time cost to increasing one's degree, and there is less cost to maintain a friendship than in a real-world social network. Thus, as with the Web, it is reasonable to consider the degree distribution of SNS networks follow a power-law distribution. The Barabási-Albert model [9] is a network growth model that leads to a network with a power-law degree distribution. In the Barabási-Albert model, both links and nodes are added, and one end of each link is added with linear *preferential attachment*, meaning that links are more likely to connect to nodes of high degree than to ones of low degree. Starting with a small number of nodes, at every timestep a new node i is added with m links (that will be connected to the nodes already present in the network). When choosing the nodes to which the new node connects, the probability Π_i is defined as a new node that will be connected to node i depending on the degree of that node, such that

$$\Pi_i = \frac{k_i}{\sum_j k_j} \tag{4}$$

where k_i is the degree of the node i. This equation incorporates the fact that new nodes link preferentially to nodes with a higher degree.

It is likely that the rate at which nodes in a network increase their connectivity depends on their ability to compete for links. For example, in social networks some individuals acquire more social links than others, or on the WWW some web pages attract considerably more links than others. That is to say that nodes in networks have various capabilities, such as the social skills of an individual and the content of a web page. Also in SNSs, it is reasonable to suppose that users have various abilities to make links of friendship. For example, gregarious users and users who have high motivation for using SNSs are supposed to acquire more links than others. The *Fitness model* is the network growth model that accounts for such difference in the ability of nodes to

compete for links [11, 10]. This model is a simple extension model of the BA model. In this model, each node has a parameter called *fitness*, describing its ability to compete for links. A node will increase its degree at the rate that is proportional to its degree and fitness. Starting with a small number of nodes, at every timestep a new node i is added with fitness η_i, where η_i is chosen from the distribution $\rho(\eta)$. Each new node i has m links that are connected to the nodes already present in the network. It is assumed that the probability Π_i that a new node will connect to a node i already present in the network depends on the degree and on the fitness of that node, such that

$$\Pi_i = \frac{\eta_i k_i}{\sum_j \eta_j k_j} \tag{5}$$

where k_i is the degree of the node i, and η_i is the fitness of the node i. This equation incorporates the fact that fitness and the number of links jointly determine the attractiveness and evolution of a node.

4.1.2 Friends of One's Friends

In a real-world social network, it is more probable that two people with a common friend get to know each other than two people without a common friend. In almost all SNSs, one's friends are listed on one's top page. Through this list, people can easily see the "friends of their friends in the SNS." It helps users to make a link of friendship with friends of their friends. Thus, as well as in a real-world social network, a relation of "friends of friends" plays an important role in an SNS.

The connecting nearest-neighbor (CNN) model [15] is a network growth model that incorporates a process of *connecting nearest neighbors* [12]. The basic assumption of this model is that the evolution of connections is mainly determined by the creation of new relations between pairs of individuals with a common friend. This model proposed the concept of a *potential link*, in which a pair of nodes is connected by a *potential link* if (1) they are not connected by a link and (2) they have at least one common neighbor. This model, starting with a single node and an empty set of links, iteratively performs the following rules.

1. With probability $1-u$, add a new node in the network, and create a $real$ link from the new node to a randomly selected node i. At the same time, create $potential$ links from the new node to all the neighbors of node i.

2. With probability u, convert one $potential$ link selected at random into a $real$ link.

In this process, a new node participates in the network at the probability $1-u$, and a new link is either generated by converting a potential link or created by random linkage at the rate u. Therefore, the rate u is determined by the number of nodes and links.

4.1.3 Influence of Special Interest Group

In many networks including SNSs, individuals are related not only with links, but also with characteristics attributed to them. Examples include participation in particular groups with specific interests, living in the same region, and relations of families. In some SNSs, users can participate in special interest groups called *Community*. Through this *Community*, users can communicate with other users who have similar interests, and may have a chance to get acquainted with other people, even if they are at a great distance on a network. It means that people can build their friendship network beyond the range of each neighborhood. We can consider this process to be a random linkage that does not depend on the current structure of the network.

4.2. Generalized Growth Model of Social Network Service

4.2.1 Overview

Taking into account our consideration in the previous section, we propose a model based on a combination scheme of the Fitness model [11], the CNN model [16] and the process of random linkage. Here, we deal with six models, which are the Fitness model, the CNN model, and four combination models. Combination models are defined as follows: (1) CNN and random linkage model (CR model [16]), (2) Fitness, CNN and random linkage model (FCR model), (3) Fitness and CNN model (FC model), (4) Fitness and random linkage model (FR model).

Each model's properties and corresponding name are summarized in Table 2. The CR model was originally proposed as the CNNR model [16]. In this section, however, we call this model the CR model for abbreviation.

4.2.2 Description of the Model

By introducing the parameters u, p_f, and p_r, we describe these models as a single procedure. The model, starting with a single node and an empty set of links, iteratively performs the following rules.

1. With probability $1-u$, one of the following two processes is performed.

   (a) With probability p_f, add a new node in the network, and create a $real$ link from the new node to a node i, which is selected by the probability Π_i.

$$\Pi_i = \frac{\eta_i k_i}{\sum_j \eta_j k_j} \quad (6)$$

where η_i is the fitness of the node i, and k_i is the degree of the node i. At the same time, create $potential$ links from the new node to all the neighbors of node i.

(b) With probability $1 - p_f$, add a new node in the network, and create a link from the new node to a randomly selected node i. At the same time, create $potential$ links from the new node to all the neighbors of node i.

2. With probability u, one of the following two processes is performed.

   (a) With probability p_r, connect one pair of nodes selected randomly with a $real$ link.

   (b) With probability $1 - p_r$, convert one $potential$ link selected randomly into a $real$ link.

In this process, a new node participates in the network at the probability $1 - u$, and a new link is either generated by converting a potential link or created by random linkage at the rate u. Therefore, the rate u is determined by the number of nodes and links. The rate p_f is the relative frequency of selection by the probability Π_i with that of random selection. The rate p_r is the relative frequency of random linkage compared with that of converting a potential link. The value of node's fitness η is chosen from uniform distribution in the range $0 \leq \eta \leq 1$, and it is unchanged in iteration of the rules.

5. Numerical Simulation

5.1. Condition

In order to verify the validity of the proposed model, we compare networks generated by the Fitness model, CNN model, CR model, FR model, FC model, and FCR model with the two data sets.

By adjusting parameter u, the numbers of nodes and links in these simulations are precisely equal to the numbers of the ACS or Amippy. In the Fitness model, the parameter $u = 1$. In other models, based on the mean degree of the ACS and Amippy, set the parameter $u = 0.66$ and $u = 0.77$, respectively. As for p_f, set the parameter $p_f = 1$ in the Fitness model, FCR model, FC model, and FR model. In other models, we set the parameter $p_f = 0$. As for p_r, in CR and FCR, we follow a study by Yuta $et\ al.$ [16] and assume $p_r = 0.04$. In the FR model, the process of connecting nearest neighbors is not performed at all, thus we set the parameter $p_r = 1$. In the CNN model and FC model, the process of random linkage is not performed at all, thus we set the parameter $p_r = 0$.

5.2. Results of ACS Simulation

Table 3 demonstrates the results of the numerical simulations. In the Fitness model, both the average shortest-path length and clustering coefficient are less than the value of ACS. In the CNN model and CR model, average shortest-path length is greater than the value of ACS, and assortativity indicates positive value. In the FR model, the clustering coefficient is much lower than the value of ACS, and it shows no assortative mixing. In the FC model and FCR model, the clustering coefficient is slightly lower than the value of ACS, however, assortativity shows negative value and the average shortest-path length is similar to that of ACS. For more detailed investigation, the degree distributions of ACS and that of one of the networks generated by the FC model are shown in Fig. 2. We see from Fig. 2 that the degree distribution of these networks seems to be similar. Indeed, the scaling exponent γ of each network's degree distributions are $\gamma_{acs} = 1.184$ and $\gamma_{fc} = 1.227$.

From the viewpoint of these properties, the results of the simulations show that the FC model and FCR model seem to be able to reproduce the network structure of ACS. In particular, the FC model is the most similar to them.

5.3. Results of Amippy Simulation

Table 4 shows the results of the numerical simulations. We see from Table 4 that the results of the Fitness model,

Table 2. Characteristic feature of each model ("+" means use the property "-" means do not use the property.)

	Fitness	CNN	CR	FR	FC	FCR
Fitness	+	-	-	+	+	+
Connecting nearest neighbors	-	+	+	-	+	+
Random linkage	-	-	+	+	-	+

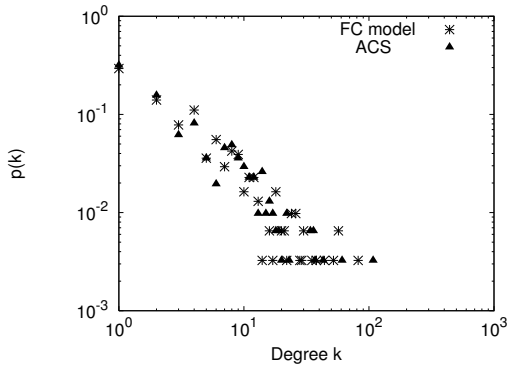

Figure 2. Degree distribution(ACS and FC)

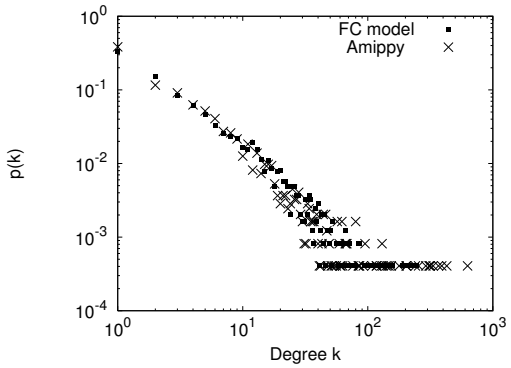

Figure 3. Degree distribution(Amippy and FC)

CNN model, CR model and FR model are similar to the results in section 5.2. That is, the Fitness model and FR model are different from the actual data set by the clustering coefficient. The CNN model and CR model are different from the actual data set by its assortativity. On the other hand, the results of the FC model and FCR model are different from the results in section 5.2. First of all, the FC model and FCR model show week negative assortativity ($r = -0.079$), although the actual data set shows strong negative assortativity ($r = -0.225$). Secondly, the scaling exponents γ differ from one another. In order to explain these differences, we show the degree distributions of Amippy and one of the networks generated by the FC model in Fig. 3. Although, these degree distributions seem to be similar, the scaling exponent γ are $\gamma_{amippy} = 1.169$ and $\gamma_{fc} = 1.441$. The main reason for this is the difference of high degree domains: the maximum degree node has only 245 links in the FC model, while in Amippy it has 628 links. This means that the FC model fails to reproduce the presence of a "super-hub" node. Since the presence of a "super-hub" node leads to a negative value of assortativity, it is reasonable to suppose that the difference of high degree domains also causes the divergence of the value of assortativity.

5.4. Discussion

As noted above, the results of the network generated by the FC model are the most similar to those of the two SNS networks. In this section, we discuss why the FC model is able to reproduce networks similar to the two SNS networks.

First, from the Tables 3 and 4, it is reasonable to suppose that the concept of the CNN model leads to a high clustering coefficient. Since the networks of the two SNSs show a high clustering coefficient, the concept of the CNN model is needed to reproduce the networks of the SNSs. Secondly, as noted in section 3.4, the user of maximum degree in ACS is connected with roughly 35% of all users, even though 31.6% of all users have only one link. In the same way, maximum degree user in Amippy has a link of friendship with 25.5% of all users, although 38.4% of all users have only one link. As Bianconi *et al.* demonstrated, the concept of the Fitness model leads the phenomenon called "fit get rich(FGR)" in which a single node captures a macroscopic fraction of links [10]. It is clear that the phenomenon is conducive to the value of negative assortativity from its definition. Since the networks of the two SNSs show the FGR feature and the negative value of assortativity, we concluded that the concept of the Fitness model is necessary to reproduce the networks of the SNSs. Lastly, we can see from Tables 3 and 4, that the process of random linkage doesn't have much effect on the results. Since ACS and Amippy are used in the university and local community respectively, their range of friendship is limited. It is likely that users in each SNS seldom make a relationship beyond the range of each neighborhood. These results suggest that the process of random linkage is not certainly indispensable for reproducing the networks of the two SNSs. For these reasons,

Table 3. Network topology of ACS network and results of simulations with the models

	ACS	Fitness	CNN	CR	FR	FC	FCR
Average shortest-path length L	3.28	2.80	4.42	4.19	3.31	3.22	3.24
Clustering coefficient C	0.479	0.153	0.378	0.356	0.047	0.375	0.354
Assortativity r	-0.152	-0.219	0.165	0.170	0.007	-0.115	-0.103
Scaling exponent γ	1.134	1.387	1.302	1.334	1.389	1.237	1.253

Table 4. Network topology of Amippy network and results of simulations with the models

	Amippy	Fitness	CNN	CR	FR	FC	FCR
Average shortest-path length L	3.10	3.08	5.28	4.75	3.75	3.51	3.54
Clustering coefficient C	0.394	0.068	0.393	0.360	0.010	0.341	0.300
Assortativity r	-0.255	-0.132	0.144	0.136	0.201	-0.079	-0.079
Scaling exponent γ	1.169	1.425	1.493	1.573	1.872	1.441	1.430

it is concluded that the FC model is able to reproduce the most similar network to SNSs. On the other hand, from the results of Amippy simulations, we also found deficiencies of the proposed model. For example, the model failed to reproduce the presence of "super-hub" nodes and a network of strong negative assortativity. This means that the model is not always able to reproduce all characteristics of SNSs. In order to reproduce the SNS network with higher accuracy, it is necessary to incorporate some other characteristics into our model.

6. Conclusion

In this study, we analyzed the data sets of the ACS and Amippy. The major characteristics of the two data sets can be summarized as follows: (1) Low average shortest-path length, and high clustering coefficient, (2) Presence of a power law degree distribution, (3) Negative assortativity. Based on our analysis, we proposed a growth model of SNS networks. We confirmed the model's validity by conducting numerical simulations with the proposed model. Results of simulations indicated that the processes of the CNN model and Fitness model are needed to reproduce the networks of SNSs, and the proposed model is able to reproduce the network structure of SNSs. However, it was found that the model cannot always reproduce all characteristics of SNSs.

For future work, in order to reproduce the SNS network with higher accuracy, we need to incorporate other characteristics into our model. To take an examples it may be presumed that the fitness η obey other distributions, although we chose fitness η from uniform distribution for simplicity. In addition, it is an interesting to define the fitness η as a variable parameter, which varies with communication in SNSs.

Acknowledgments

We would like to thank Associate Professors N. Kawaguchi, I. Ide, and Mr. K. Takai for providing the data set of the ACS, and Mr. M. Toraiwa, Mr. H. Kato and Mr. T. Yoshino for providing the data set of Amippy.

References

[1] ACS. `http://acs.is.nagoya-u.ac.jp/`.
[2] Amippy. `http://amippy.jp/`.
[3] TRYWARP. `http://trywarp.com/`.
[4] L. Adamic, O. Buyukkokten, and E. Adar. A social network caught in the Web. *First Monday*, 8(6):29, 2003.
[5] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. *Proceedings of the 16th international conference on World Wide Web*, pages 835–844, 2007.
[6] R. Albert and A. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.
[7] L. Amaral, A. Scala, M. Barthélémy, and H. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 102(30):10421–10426, 2005.
[8] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, 2006.
[9] A. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509, 1999.
[10] G. Bianconi and A. Barabási. Bose-Einstein Condensation in Complex Networks. *Physical Review Letters*, 86(24):5632–5635, 2001.
[11] G. Bianconi and A. Barabási. Competition and multiscaling in evolving networks. *Europhysics Letters*, 54(4):436–442, 2001.
[12] J. Davidsen, H. Ebel, and S. Bornholdt. Emergence of a Small World from Local Interactions: Modeling Acquaintance Networks. *Physical Review Letters*, 88(12):128701, 2002.
[13] M. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):26126, 2003.
[14] M. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):36122, 2003.
[15] A. Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, 67(5):56104, 2003.
[16] K. Yuta, N. Ono, and Y. Fujiwara. A Gap in the Community-Size Distribution of a Large-Scale Social Networking Site. *Arxiv preprint physics/0701168*, 2007.