Commun Nonlinear Sci Numer Simulat 44 (2017) 482–494

Contents lists available at ScienceDirect

Commun Nonlinear Sci Numer Simulat

journal homepage: www.elsevier.com/locate/cnsns

Research paper

Effect of homophily on network formation

Kibae Kim[a], Jörn Altmann[b,*]

[a] Moon Soul Graduate School of Future Strategy, Korea Advanced Institute of Science and Technology, Daejeon, South Korea
[b] Technology Management Economics and Policy Program, College of Engineering, Seoul National University, 1 Gwanak-Ro, Gwanak-Gu, Seoul 08826, South Korea

ARTICLE INFO

Article history:
Received 29 November 2014
Revised 24 February 2016
Accepted 15 August 2016
Available online 8 September 2016

Keywords:
Complex social network evolution
Cumulative degree distribution
Preferential attachment rule and homophily
Empirical data and simulation

ABSTRACT

Although there is much research on network formation based on the preferential attachment rule, the research did not come up with a formula that, on the one hand, can reproduce shapes of cumulative degree distributions of empirical complex networks and, on the other hand, can represent intuitively theories on individual behavior. In this paper, we propose a formula that closes this gap by integrating into the formula for the preferential attachment rule (i.e., a node with higher degree is more likely to gain a new link) a representation of the theory of individual behavior with respect to nodes preferring to connect to other nodes with similar attributes (i.e., homophily). Based on this formula, we simulate the shapes of cumulative degree distributions for different levels of homophily and five different seed networks. Our simulation results suggest that homophily and the preferential attachment rule interact for all five types of seed networks. Surprisingly, the resulting cumulative degree distribution in log-log scale always shifts from a concave shape to a convex shape, as the level of homophily gets larger. Therefore, our formula can explain intuitively why some of the empirical complex networks show a linear cumulative degree distribution in log-log scale while others show either a concave or convex shape. Furthermore, another major finding indicates that homophily makes people of a group richer than people outside this group, which is a surprising and significant finding.

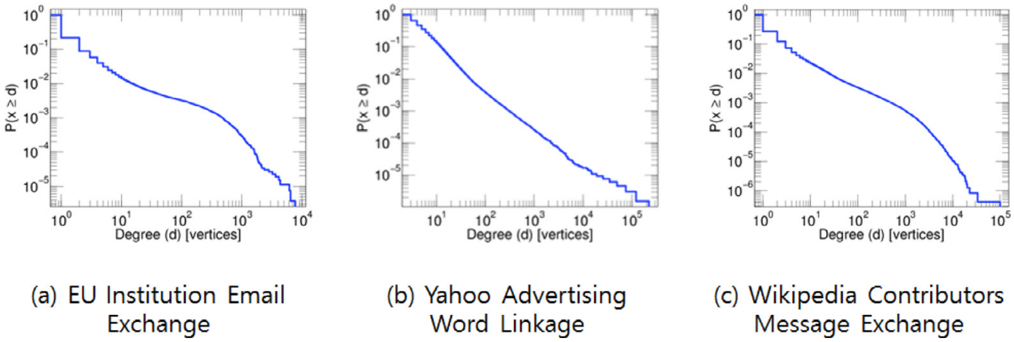
1. Introduction

During the recent decade, empirical research has shown that large complex networks, including the World Wide Web [1], online social networks [2] and Wikipedia [3], can have significantly different network topologies. These topological characteristics are of major interest as they determine the diffusion of information [2,4–6] and the robustness of systems [7,8]. These network topologies have been characterized to have concave, convex, and linear cumulative degree distributions [9,10].

For the formation of these networks, the preferential attachment rule has commonly been accepted. This rule assumes that the probability of adding a new link to a network is proportional to the number of links that a node already has [1]. The resulting cumulative degree distribution has a linear shape, representing a power law function. For reflecting empirical networks more precisely and achieve convex and concave shapes of cumulative degree distributions, prior research has

[*] Corresponding author. Fax: +1 610 956 4404.
E-mail address: jorn.altmann@acm.org (J. Altmann).

http://dx.doi.org/10.1016/j.cnsns.2016.08.011



(a) EU Institution Email Exchange

(b) Yahoo Advertising Word Linkage

(c) Wikipedia Contributors Message Exchange

Fig. 1. Empirical data based cumulative degree distributions of EU institution employees email exchanges, Yahoo advertising word linkages, and Wikipedia contributors message exchanges (taken from http://konect.uni-koblenz.de/).

modified the preferential attachment rule into several versions by introducing links generated among existing nodes [11] and by considering preferences according to the geographical distance between nodes [12,13].

The theoretical relationship between actual behavior of agents and the formulas given only covers linear and concave cumulative degree distributions. It is widely accepted that the basic preferential attachment rule represents the effect that the popularity of a node increases with its node degree. The cumulative degree distribution in log-log scale results in a linear curve. It has also been accepted that the addition of links between existing nodes through friends-of-friends effects results in concavely shaped cumulative degree distributions in log-log scale.

However, the theoretical connection between realistic individual behavior of agents in those networks and the concavely shaped cumulative degree distributions in log-log scale is hard to make with the existing formula. The existing formula introduces an exponent to the degree of a node, which is not intuitive, rather a mathematical tool to generate convexly shape curves [12]. Beside this formula, there is no other mathematical formulation that achieves a convex cumulative degree distribution and is an intuitive representation of a theory on individual behavior. Therefore, the following examples of convex cumulative degree distributions in log-log scale, which are shown in Fig. 1, cannot be explained by a formula based on theory on individual's behavior. The examples include the cumulative degree distribution of words used by advertises in their phrases at Yahoo (Yahoo advertising word linkages), which shows a light convex shape, and the curves of the email exchange between employees of a large EU institution (EU institution employees email exchanges) and the message exchange between Wikipedia users through discussion pages (Wikipedia contributors message exchanges). Those degree distributions depict a convex shape at the beginning of the curve and a concave shape at the tail of the curve [10]. Consequently, it can be stated that a formula is needed that can link the convex shape of a cumulative degree distribution with theory on individual behavior intuitively.

Further empirical studies showed that the network evolution is also impacted by attributes of nodes [14,15]. For example, software services belonging to the same company are likely to be combined to develop new services compared to those software services of different companies [16,17]. People in a virtual community are likely to gather according to their gender, age, and proximity [18,19]. This preference of agents to be connected with other agents that share common attributes is called "homophily" [20]. Theoretical research proved that homophily promotes a group of nodes with common properties to be integrated more densely than groups without that property [21]. However, this effect of homophily on the curvature of cumulative degree distributions has not been investigated yet.

Our research objective is to bring together the research on cumulative degree distributions and the research work on homophily. With respect to this research objective, the following research questions rise: How does the formula look like that combines the preferential attachment rule and homophily? What is the cumulative degree distribution of networks that evolve based on homophily and on the preferential attachment rule? Does a seed network topology influence the evolution of the network?

To answer these research questions, we introduce the modified preferential attachment model that is based on [22]. This model combines the homophily preference with the preferential attachment model [11] within a single formula. For the analysis, we investigate the topology of networks evolved through our modified preferential attachment model for different levels of homophily and five different seed networks (i.e., a single dipole network, a multiple dipole network, a ring network, a star network, and a random network).

The result of the comparison is threefold. First, homophily affects the network evolution based on the preferential attachment rule for all types of seed networks that we used. Surprisingly, the resulting cumulative degree distribution always shifts from a concave shape to a convex shape, as the level of homophily gets larger. Second, there is a level of homophily that can compensate concave-shaping effects, making the cumulative degree distribution linear in log-log scale. Third, the link density of the seed networks can cause local distortions on the network evolution, as our results for high-density, random networks show.

These findings imply that an empirical network might have not only evolved based on the preferential attachment rule but also through homophily and seed networks. In particular, we can explain the convex shapes of cumulative degree distributions of empirical networks (Fig. 1) with a single formula that integrates the homophily property and preferential attachment. This formula also helps explaining the emergence of convex shapes of cumulative degree distributions of empirical networks due to attributes of nodes intuitively. Furthermore, our findings indicate that homophily makes people, which belong to a group, richer than people outside this group, which is a surprising and significant finding. This effect increases the effect of scale-free networks of making the rich richer.

The remainder of this paper is structured as follows: In Section 2, we describe the conventional preferential attachment rule for new nodes and for links between existing nodes. Furthermore, we describe the modified preferential attachment rule, which introduces homophily into the conventional preferential attachment model. In Section 3, our simulation model is introduced. The simulation results of the network evolution under different seed network topologies and different levels of homophily are given in Section 4. Finally, Section 5 concludes the paper with a discussion of the implications of our findings.

2. Scale-free networks and their construction

2.1. Conventional preferential attachment rule

A scale-free network is a network that has a degree distribution following the power law [23]. Empirical analyses showed that many large complex networks in diverse areas (e.g., online social networks, the World Wide Web, and Blog author networks) show the power law between the node degree and its frequency [2,9,24,25]. The power function shows a long tail (or a fat tail) compared to an exponential function and a Gaussian distribution. The long tail indicates that there are a few nodes with a very high node degree. These nodes are called hubs. The hubs play an important role in giving scale-free networks a unique property. For example, a node in a scale-free network is able to reach any other node with a few hops (i.e., a short path length) [26]. In addition to this, a scale-free network is robust against uniformly distributed failures. A few hubs perform the majority of workload while the probability of failure is distributed to all the nodes evenly [7].

Several network evolution models for constructing scale-free networks have been suggested in the past [1,27–30]. These have been applied to overlay networks, Internet topology modeling, and network topology modeling. Among those models, one of the most well-known models is the preferential attachment rule proposed in Barabasi and Albert [1]. It assumes that a node of a network attaches a new node with the probability proportional to the degree of the existing node. Simulations of the preferential attachment rule have shown that the network grows into a scale-free network whose exponent is 3 [1]. That is correct if exactly one new node enters the network with one link at each time step. In particular, the preferential attachment rule defines the probability Π_i that an existing node i gains a new link to a new node entering the network as a linear function of degree k_i of node i.

$$\Pi_i = \frac{k_i}{\sum_{j \in \aleph} k_j} \tag{1}$$

where $\aleph$ is the set of nodes existing in the network before the new node enters.

In many environments (e.g., online social networks, academic co-authorship networks, peer-to-peer networks, and Web2.0 service networks), a new link is not generated only between a new node and an existing node but also between existing nodes [1,2,11,31–34]. For example, in an academic co-authorship network, scholars publish repeatedly articles jointly with different peers throughout their career. This activity of publishing starts at the moment the scholar enters academia and continues until the scholars leaves academia due to retirement [11,32]. In a peer-to-peer network, links between peers are added or replaced, in order to increase the performance of the peer-to-peer network [31,33]. With respect to the Web2.0 service network, users compose new Web services with Web2.0 services that already exist within the Web2.0 service network and that offer open application programming interfaces to their services [34].

In these environments, it is reasonable to consider an additional preferential attachment rule between existing nodes and not only between a new node and an existing node (i.e., Eq. (1)). Existing nodes are linked with each other with a probability that is proportional to the multiplication of the degree of existing nodes [11]. That is, the probability Π_{ij} that node i is connected to node j is defined as a linear function of their node degrees k_i and k_j:

$$\Pi_{ij} = \frac{k_i k_j}{\sum_l \sum_{m>l \in \aleph} k_l k_m} \tag{2}$$

where $\aleph$ is the set of nodes existing in the network.

2.2. Network evolution with homophily

A lot of empirical and theoretical studies show that human beings are homophily in nature. Homophily is the tendency that a person prefers those who share a characteristics to those who do not [20]. It causes from a variety of demographic similarities including geographical proximity, kinship, friendship, and social status. These similarities make “people to associate with similar others for ease communication, shared culture tastes [20].” For example, an observation of the society

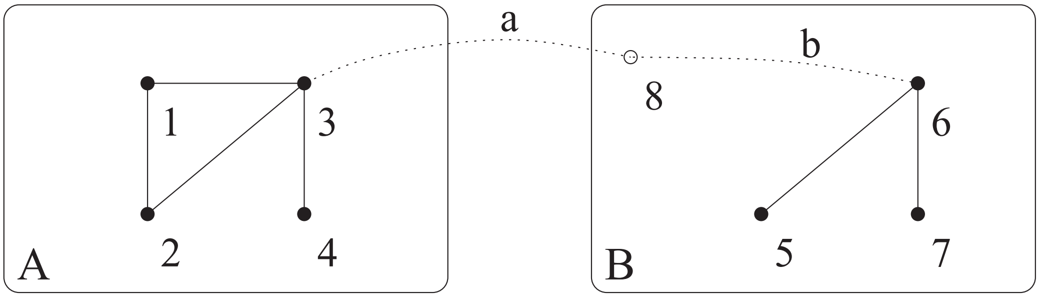


Fig. 2. Example of the modified preferential attachment rule: Selection of the most probable link by a new node, if nodes are not homophilic (link *a*), and the most probable link, if the nodes are completely homophilic (link *b*).

in a university dormitory found that people are likely to gather in a group, when they frequently meet by chance due to proximity [35]. The homophily works in a virtual society as well as in the real world. People joining an online game are likely to make friendship with those who are near by, have similar age and the same gender [18,19]. Theoretical studies show that the homophily can also segregate an ethnic group from other groups [36] and make clusters, in which people are linked more densely with each other than with those outside [21].

Based on the discourse of prior research, we assume that the conflict of two tendencies determines the level of homophily. The first is that a node prefers a link with another node within the same group instead of a link with a node in another group [20]. The second is that a node has also an incentive to link to nodes of other groups for innovation and survivability [37]. To describe it, we use the definition of homophily based on groups, or a set of nodes which share the same property [22]. This definition of homophily indicates the preference of linkages between nodes within the same group to linkages between nodes belonging to different groups. To indicate this likelihood of linkages over groups, the homophily level Λ_s^t between groups s and t is defined as a fraction of the probability of linkages between nodes belonging to different groups to the probability of linkages between nodes within the same group, which is mathematically described as:

$$\Lambda_s^t = \begin{cases} 1 & \text{if } s = t \\ \Lambda_s^t = \Lambda & \text{otherwise} \end{cases} \tag{3}$$

where homophily index Λ is any positive real number between 0 and 1. If $\Lambda = 0$, a node in a group s makes no link with the node in the other group t $(t \neq s)$ but makes a link only with the nodes in the same group. This state describes groups that are completely closed (homophilic). They only prefer groups with the same preference. If $\Lambda = 1$, the group does not affect any linkage between any nodes, independently of whether they belong to the same group or different groups. This state describes nodes that are not homophilic at all.

According to the definition above, the preferential attachment model described in Eqs. (1) and (2) can be extended. That is, the probability Π_i^{pq} that a new node of group q is linked to node i of group p is defined as:

$$\Pi_i^{pq} = \frac{k_i^p \Lambda_p^q}{\sum_{j \in \aleph} k_j^\mu \Lambda_\mu^q} \tag{4}$$

where k_i^p is the degree of node i of group p, and Λ_p^q represents the homophily between the group q of the new node and the group p of node i. The preference of node i by an entrant is normalized by the sum of the preferences of node j of group μ to connect to the entrant in group q. $\aleph$ is the set of existing nodes. If the groups are not homophilic at all (i.e., $\Lambda = 1$), Eq. (4) is identical to Eq. (1). If the groups are completely homophilic (i.e., $\Lambda = 0$), on the other hand, Eq. (4) is reduced to Eq. (1) for each group, and there is no correlation over groups.

Likewise, the probability Π_{ij}^{pq} that an existing node i of group p is linked to another existing node j of group q is defined as:

$$\Pi_{ij}^{pq} = \frac{k_i^p k_j^q \Lambda_p^q}{\sum_{j \in \aleph} \sum_{j > l \in \aleph} k_l^\mu k_m^\nu \Lambda_\mu^\nu} \tag{5}$$

where k_i^p is the degree of existing node i of group p and k_j^q is the degree of another existing node j of group q. Λ_p^q is the homophily level between groups p and q. This probability is normalized by the sum of products of node degrees of node $l \in \aleph$ and node $m \in \aleph$, belonging to groups μ and ν, respectively. $\aleph$ is the set of existing nodes. Eq. (5) is identical to Eq. (2), if nodes are not homophilic at all (i.e., $\Lambda = 1$). If the groups are completely homophilic, like the case of Eq. (4), Eq. (5) goes to Eq. (2) for each group.

The introduction of homophily to the preferential attachment model affects the network evolution significantly. To illustrate this, we consider two examples. The first example is about a new node joining the network, and the second example is about linkage between two existing nodes. Fig. 2 illustrates the first example. It shows the most probable link of a new node, if nodes are not homophilic at all, as well as the link, if the nodes are completely homophilic. In this example, a new

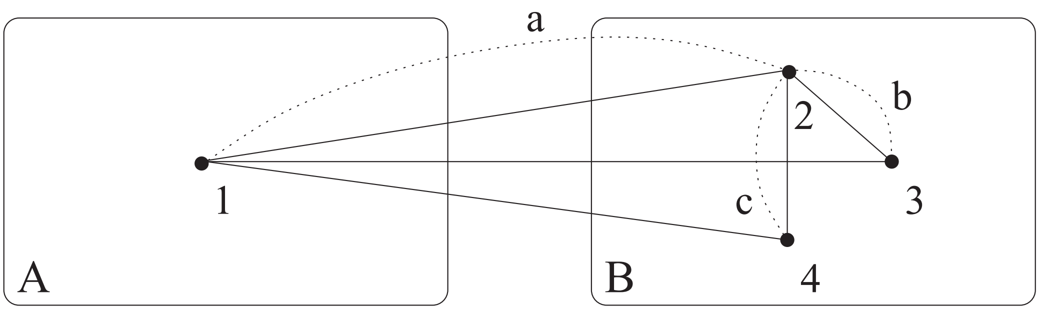


Fig. 3. Example of the most probable link between existing nodes, if nodes are not homophilic at all (link *a*), and the most probable link, if the nodes are completely homophilic (one of links *b* and *c*).

node 8 belong to group B enters the network consisting of seven nodes, among which four nodes are involved in group A, and three nodes in group B. If nodes are not homophilic at all (i.e., $\Lambda = 1$), the preferential attachment rule (Eq. (4)) proposes that node 3 is the most probable node that could gain a link with the entering node 8 (i.e., link *a*). Its degree is the largest in the network. If the nodes are completely homophilic (i.e., $\Lambda = 0$), however, the most probable node gaining a link with node 8 is changed to node 6 (i.e., link *b*). The node with the largest degree in the same group as node 8 (i.e., group B) becomes the most probable node to gain a link with the entering node. The nodes in group A are excluded according to Eq. (4).

Similar to the example above, Fig. 3 shows the effect of homophily on the linkage between existing nodes. In this example, there are four nodes within the network. Among those nodes, one node (i.e., node 1) belongs to group A and the other nodes (i.e., node 2, 3, and 4) belong to group B. If a new link is generated between two existing nodes, and if the nodes are not homophilic (i.e., $\Lambda = 1$), node 1 and node 2 are the most probable nodes that gain the new link according to Eq. (5) (link *a*). Their degrees (i.e., $k_1 = k_2 = 3$) are the largest in the network. If the nodes are completely homophilic ($\Lambda = 0$), however, nodes 2 and 3 or nodes 2 and 4 are selected to gain the new link (i.e., link *b* and link *c*, respectively).

3. Simulation methodology

For investigating the effect of homophily on network evolution, we generate different seed networks for applying the modified preferential attachment rule. The growth of the network follows two rules:

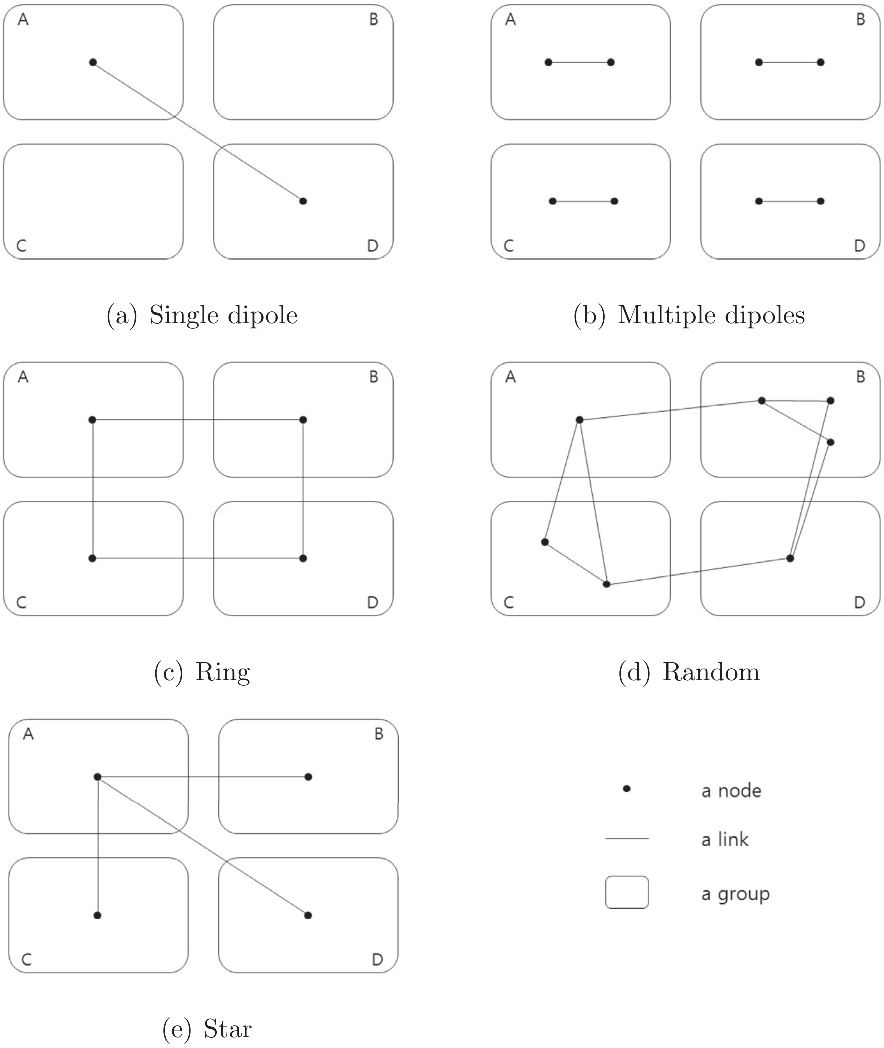
- Rule 1: One node enters the network at each time step and attaches itself with a link to an existing node. The probability that an existing node i of group p gains a link from the entering node that belongs to group q follows the modified preferential attachment rule (Eq. (4)).
- Rule 2: A new link is generated between two existing nodes with probability a at each time step. The probability that the existing node i of group p and node j of group q gain a new link between them follows the modified preferential attachment rule (Eq. (5)).

The probability that a new link appears between existing nodes is set to $a = 0.001$ in all simulations. Our pre-test show that the results, that we present in this paper, do not change for larger a. If the value of $a < 0.001$, then effect of Rule 2 gets less significant and, consequently, homophilic behavior has less impact on the establishing of links.

We use five topologies of seed networks in the simulation to investigate how the effect of homophily changes the network evolution. With these topologies, we investigate whether the topology of seed networks changes the effect of homophily on the network evolution. These five seed networks represent a wide set of network topologies. All other network topologies can be derived from those five networks. In detail, the five seed networks are:

- Single Dipole Network: A single dipole network is a network, in which there are only two inter-connected nodes that are allocated into two groups (Fig. 4a).
- Multiple Dipoles Network: A multiple dipole network is a network, in which there are multiple pairs of inter-connected nodes that are allocated into different groups. Each pair belongs to a same group (Fig. 4b).
- Ring Network: A ring network is a connected network, in which each node is linked with two other nodes and each node belongs to a different group. That is, the degree of all the nodes is 2 and the initial network has 100 links. (Fig. 4c).
- Random Network: A network which consists of N nodes and K links, in which nodes that are linked with each other are randomly chosen according to an equal distribution. The nodes are uniformly allocated to groups (Fig. 4d).
- Star Network: A star network is a network, in which one node (hub) is linked with all other nodes. No other links exists. Each node belongs to a different group (Fig. 4e). Therefore, there are 99 links in the star network with 100 nodes.

With each topology described above, we simulate the evolution of the network under different homophily levels , and analyze the scale-free property of the final networks. In order to avoid the fluctuation in the area of nodes with large node



(a) Single dipole

(b) Multiple dipoles

(c) Ring

(d) Random

(e) Star

Fig. 4. Example of five seed networks with four groups A, B, C and D.

degrees due to small occurrences, we measure the cumulative degree distribution $P(k)$ as proposed in Newman [38]:

$$P(k) = \int_{k_{>}} p(k')k'dk' \tag{6}$$

where $p(k)$ is the frequency of degree k, and $\int_{k_{\geq}}$ means the integration from $k' = k$ to ∞.

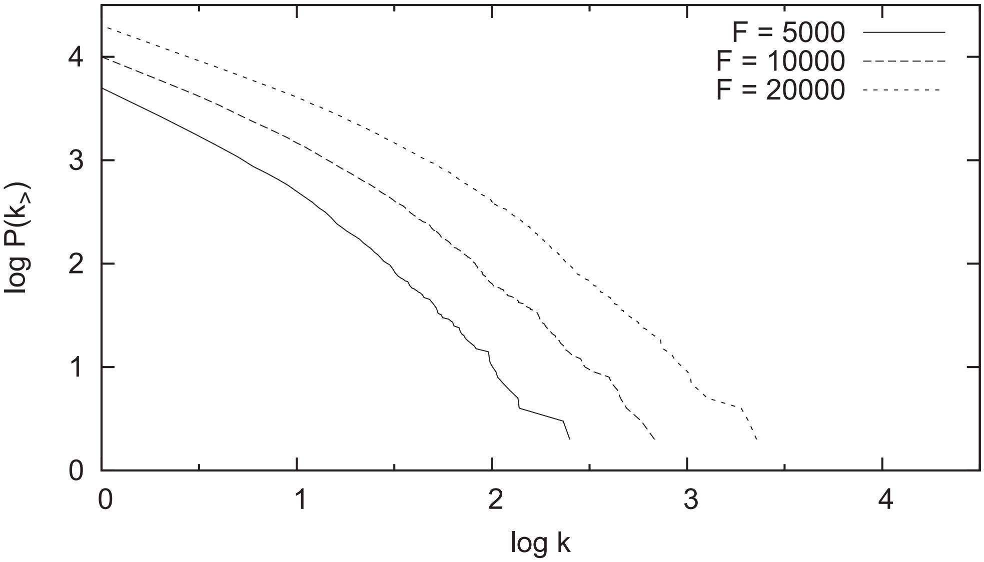


Fig. 5. The cumulative degree distributions of networks that evolve from one dipole without homophily.

By comparing the degree distributions of networks evolved from these five seed network topologies, we can determine whether homophily affects the network evolution and whether the allocation of nodes into groups in a seed network is related with the effect of homophily on the network evolution.

In our simulations, a network evolves from a seed network given at the initial time step until the network reaches a size of 20,000 nodes. We set the number of groups in the network to 100 groups, and the nodes in the network are allocated randomly (i.e., uniformly distributed) into each group, so that each group has in average 200 nodes at the end of the evolution. The size of subgroups is chosen according to anthropological studies showing that a person can manage between 150 and 200 friends at most [39,40].

4. Network evolution from a single dipole without homophily

For stating that our simulations obtain results that are consistent with the one of Barabasi et al. [11], we conducted a network evolution simulation with no homophily impact (i.e., $\Lambda = 1$) and a seed network that contains only a single dipole (Fig. 5). In the figure the size F of evolved networks varies $F = 5000$, $F = 10000$ and $F = 20000$.

As Fig. 5 illustrates, the results of the simulation show curves that have a slight concave shape. The linear regression to the curves say that the data significantly fit to a linear function and the slope are 1.28, 1.24 and 1.20 for network size $F = 5000$, $F = 10000$ and $F = 20000$, respectively. This confirms not only the simulation results of Barabasi et al. [11] but also the results obtained from empirical studies of complex networks [10].

5. Network evolution analysis of homophily for different seed networks

5.1. Network evolution from seed dipole networks

The network evolution is simulated for networks that are initially based on 100 dipoles. Each dipole belongs to a different group (Fig. 4b). Fig. 6 shows the cumulative degree distribution of networks that evolved under five different homophily levels, $\Lambda = 1$, 10^{-1}, 10^{-2}, 10^{-3} and $10^{-\infty}$. The cumulative degree distribution is measured on the basis of Eq. (6) and represented in log-log scale.

The results show that the tail of the cumulative degree distribution moves further to the right as the homophily level goes from $\Lambda = 1$ to $\Lambda = 0$. On the other hand, the head (i.e., the area of low node degrees) does not change for any level of homophily. The midst between head and tail of the cumulative degree distribution shows interesting patterns. The shape of the cumulative degree distribution transforms from concave to convex as the groups are more homophilic. And, it is almost linear when $\Lambda = 10^{-2}$.

In order to understand the impact of the number of initial dipoles on the network evolution, we also analyze networks that evolve from a single dipole network with those from multiple dipoles. Fig. 7 illustrates the cumulative degree distributions of a network that evolves from a single dipole network and six homophily levels, $\Lambda = 1$, 10^{-1}, 3×10^{-2}, 10^{-2}, 10^{-3}, and $10^{-\infty}$. The cumulative degree distribution is also measured on the basis of Eq. (6) and is represented in log-log scale. It is to be noted that the shape of the cumulative degree distribution changes from concave to convex in the range between $10^{-1} < \Lambda < 10^{-2}$ (i.e., if the groups get more homophilic). There is only a slightly difference to networks

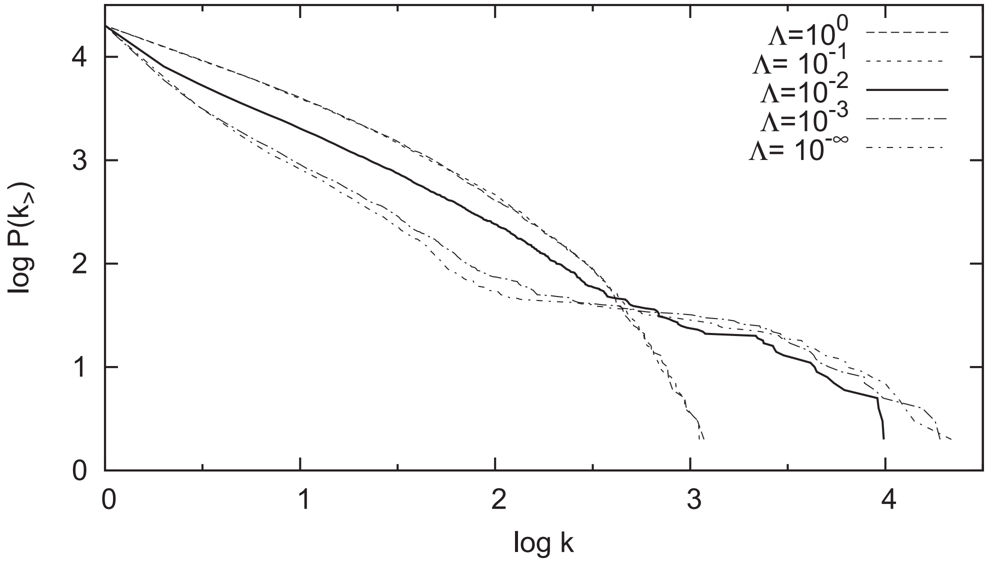


Fig. 6. The cumulative degree distribution of networks that evolve from one dipole per group for different levels of homophily.

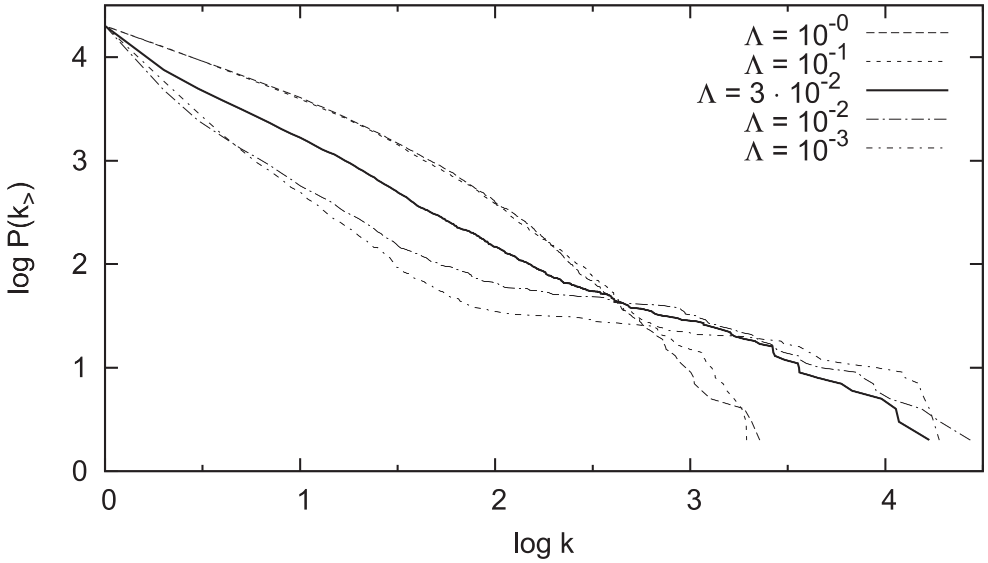


Fig. 7. The cumulative degree distributions of networks that evolve from one dipole for different levels of homophily.

with multiple dipoles. The cumulative degree distribution is almost linear, if $\Lambda = 3 \times 10^{-2}$. Consequently, we can state that the comparison did not reveal any significant differences for the two seed networks.

In order to narrow down the range of homophily levels Λ, in which the degree distribution becomes linear, we generated additional networks with homophily levels between $\Lambda = 10^{-1}$ and $\Lambda = 10^{-2}$. Fig. 8 shows the additional curves for $\Lambda = 5 \times 10^{-2}$, 4×10^{-2}, 2×10^{-2}, illustrating that the cumulative degree distribution curve changes from concave to convex between $\Lambda = 5 \times 10^{-2}$ and $\Lambda = 10^{-2}$. The curves in the middle ($\Lambda = 4 \times 10^{-2}$, 3×10^{-2}, and 2×10^{-2}) can hardly be distinguish. However, it is to be noted that a convex shape can clearly be observed until $\log k < 2.5$. For larger $\log k$ values, the slope is still significantly different to a network without any homophily, showing the impact of homophily on the evolution of the network.

In summary, the results from seed dipole networks suggest that the cumulative degree distribution is transformed from being concave in case of none-homophily to being convex in case of homophily. These results are caused through the combination of the linkage between existing nodes (Rule 2) and homophily. In order to explain this effect, the case of complete homophily (i.e., $\Lambda = 0$) is helpful. If a new link is generated by new nodes only (Rule 1), any group has a chance to get a node and a link due to the design of homogeneous group sizes of about 100 nodes. Furthermore, once a node enters the network, the network in a group evolves independent of the network evolution in other groups due to the complete

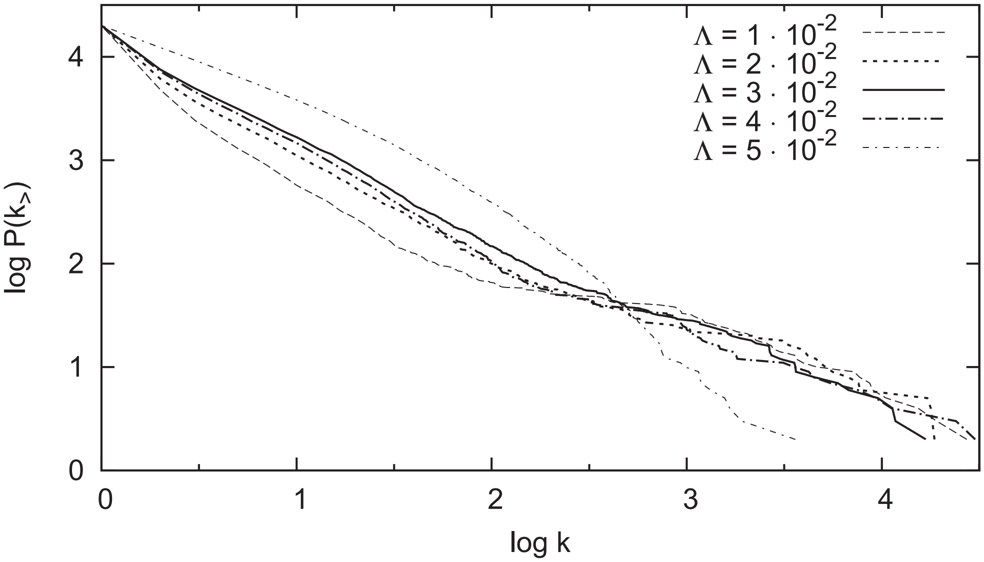


Fig. 8. The cumulative degree distributions of networks that evolve from one dipole for different levels of homophily (scrutinizing the range from $\Lambda = 10^-1 to 10^-2$).

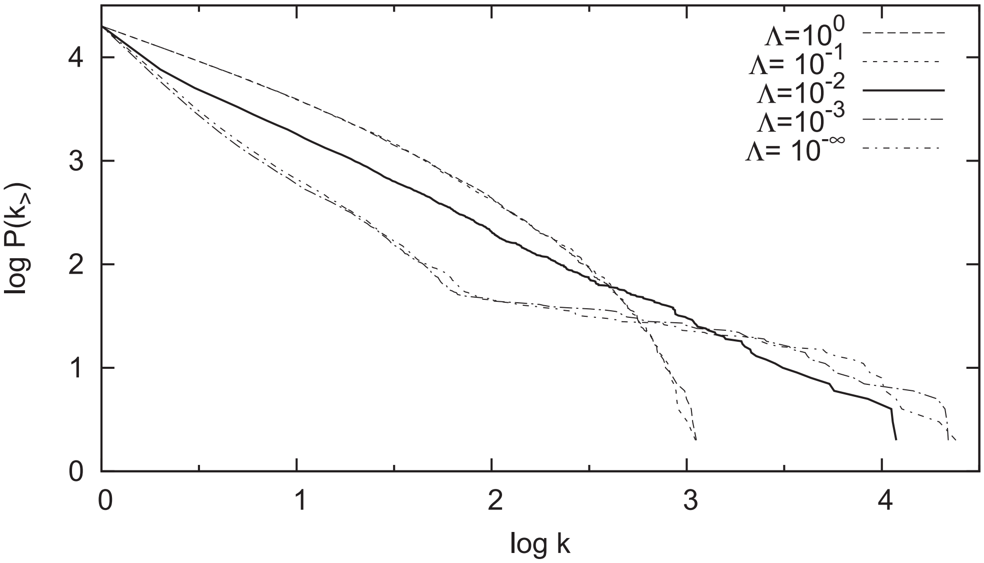


Fig. 9. The cumulative degree distributions of networks that evolve from a seed ring network for different levels of homophily.

homophily. That is, a new node is linked only with an existing node of the same group according to Rule 1. Therefore, the degree distribution does not change even with homophily, if only Rule 1 governs. However, if a new link is also generated between existing nodes and a pair of nodes in a group has slightly larger degrees than the others, a new link according to Rule 2 will only be generated between those nodes (Fig. 3). Thus, the degrees of the nodes in these groups are larger than the degrees of other nodes and the degree distribution of the nodes in this group decays slower than the other nodes. Figs. 6 and 7 show this effect for $\Lambda \leq 10^{-3}$. The slowly decaying parts are to the right-hand side of the kink in those curves.

5.2. Network evolution from seed ring networks and seed star networks

To analyze the findings of the previous sections further, we investigate the network evolution from a seed ring network and a seed star network. It will help determining whether the homogeneity over groups in a seed network affects the network evolution with homophily.

First, the network evolution is simulated from a seed ring topology (100 nodes, one node per group). The left graph of Fig. 9 shows the cumulative degree distribution of networks evolving under five different homophily levels

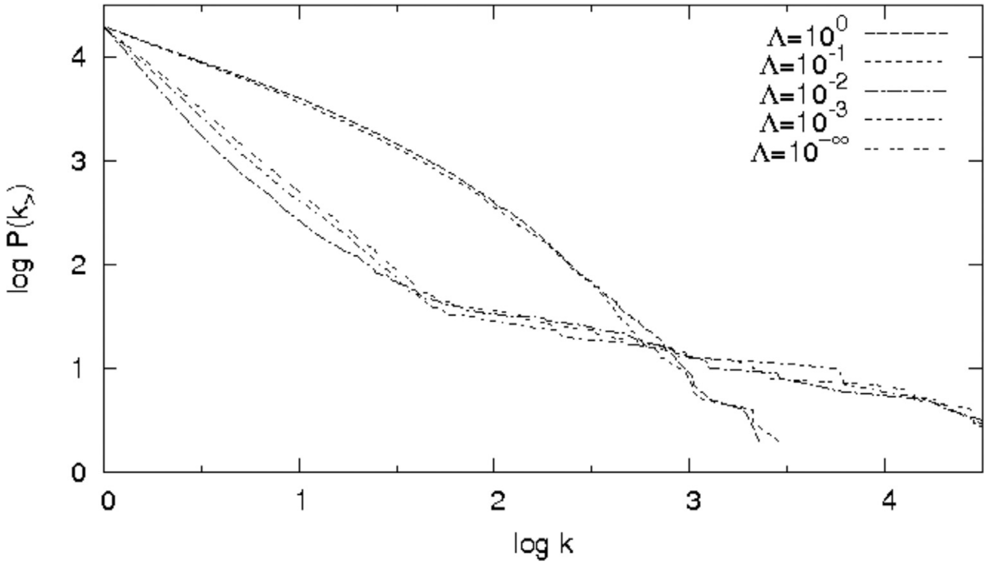


Fig. 10. The cumulative degree distributions of networks that evolve from a seed star network for different levels of homophily.

$\Lambda = 1$, 10^{-1}, 10^{-2}, 10^{-3}, and $10^{-\infty}$. The cumulative degree distribution is shown in log-log scale and is measured on the basis of Eq. (6).

The cumulative frequency of degree $P(k)$ for all homophily level decreases as the degree k increases. But, the curvature changes based on the homophily level Λ. The degree distributions are concave for values $\Lambda = 1$ and 10^{-1}, and they are convex for values $\Lambda = 10^{-3}$, 10^{-4}, and $10^{-\infty}$. Furthermore, the cumulative degree distribution is almost linear, if $\Lambda = 10^{-2}$. In other words, the cumulative degree distribution of a network that evolved under the preferential attachment rule is concave, if the groups are not homophilic, while it is convex if the nodes of the groups are homophilic. The results are similar to the results for dipole networks.

The second simulation uses a star network as an initial network (100 nodes, one node per group). The remaining settings are identical to the ring network simulation. The results shown in Fig. 10 are similar to the one of the ring network. Homophily affects the network evolution under the preferential attachment rules. The cumulative degree distribution is concave, if the nodes are not much homophilic (i.e., $\Lambda \geq 10^{-1}$). It is convex, if the nodes of the groups are homophilic (i.e., $\Lambda \leq 10^{-2}$).

The reason for the course of the cumulative degree distribution of the network evolving from a ring and a star networks is similar to that for the network evolving from a network with dipole topology. In all cases, there are groups involving a single node. In case of completely homophilic nodes and a single node in a group, the nodes in these groups do not gain a new link through Rule 2. Once a group obtained more than two nodes through Rule 1, the nodes in this group can obtain a new link through Rule 2. Since the probability of linkage between existing nodes is dependent on the degree of the nodes, the nodes in groups with two or more nodes gain more links through Rule 2 as time goes on. Once a group involves more one node due to Rule 1 and their degrees are slightly larger than the other nodes, then the nodes gain more links than the others through Rule 2, same as the cases of the dipole networks.

This implies that homophily impacts the network evolution of different seed networks in the same way. The results also suggest that centralization (e.g., star network) or decentralization of nodes (e.g., ring network) does not affect the network evolution through homophily. In conclusion, we obtain the following two propositions. First, homophily impacts the network evolution of a scale-free network. Second, the impact of homophily on the cumulative degree distribution is invariant with respect to the topologies of seed networks.

5.3. Scale-free distortion density of seed random networks

Considering the results of the previous section, we investigate whether the number of links (while keeping the number of nodes constant) impacts the network evolution. In other words, we investigate the effect of density of seed networks on the network evolution, where network density is defined as the ratio of the number of links of a network to the number of nodes of the network.

For the simulation, we generate random networks with 100 nodes. The nodes are allocated randomly (i.e., uniform distribution) into the groups and, then, randomly linked with each other. Furthermore, we set the nodes to be completely homophilic, in order to investigate only the effect of homophily on network evolution.

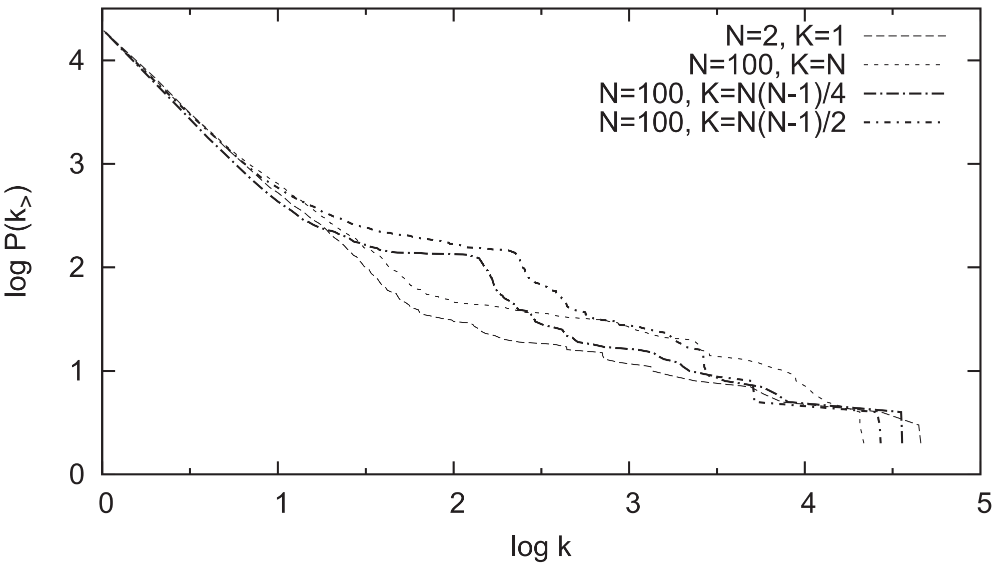


Fig. 11. The cumulative degree distribution of networks evolved from random networks with different numbers of links K and homophily.

Fig. 11 presents the cumulative degree distributions of networks evolving from seed random networks with different numbers of links K. The numbers of links analyzed are $K = N$, $K = N(N-1)/4$, and $K = N(N-1)/2$, where N is the number of nodes. The cumulative degree distribution of the network evolving from a single dipole (i.e., $N = 2$ and $K = 1$) is depicted as a reference in Fig. 11 (dashed line curve).

The simulation results show clearly that the cumulative degree distributions of the three random networks with different densities have a convex shape. However, the results also show that a hump appears, if the network evolves from a seed network with a large number of links. The hump is near the location of the kink, which is generated due to the group homophily that intensifies the link generation in a certain group while keeping the normal evolution in other groups. The kink of cumulative degree distributions can be clearly seen for the cumulative degree distribution of the reference network, the single dipole network ($N = 2$ and $K = 1$).

The hump converts locally the convexity of the overall curve to concavity (Fig. 11). The reason that this hump appears is that the initial 100 nodes have high node degrees, which remain visible until the end of the simulation period. That is, if a random network has a characteristic set of nodes with high degrees, the characteristic node degree will only be blurred slowly. While the network evolves, many of those nodes with this characteristic degree gain more links. As the degree of other nodes increases to this characteristic degree over time as well, the hump becomes indistinguishable for very large networks, disappearing at last.

Concluding, we can state that the network density of a seed network does not affect the distortion of scale-free networks due to homophily in general. It generates irregularities (hump) in a local area of cumulative degree distribution, if the evolution of network has not been sufficiently long. Considering these conclusions, it becomes obvious that the evolution of networks is impacted through homophily and the number of nodes and links in the initial seed networks. These results needs to be considered in information systems research, as it will help network analyses to be more realistic.

6. Discussion and conclusion

In this article, we proposed and implemented a modified preferential attachment rule, which considers homophily, and we investigated whether and how homophily affects the formation of networks for a variety of different seed networks. The simulation results suggest that homophily and the preferential attachment rule affect the network formation reversely, and that this effect is independent of the seed networks (Fig. 4). Furthermore, the network formation is independent of the initial node allocation over groups (e.g., independent of a highly centralized or decentralized allocation among groups). This effect is also not changed for seed networks with a high link density (i.e., a large average number of links per node). In all these cases, the higher the level of homophily is, the more the cumulative degree distribution transforms from a concave shape to a convex shape. As a convex shape shows that the benefit of the network is allocated to a few people (i.e., a few people get a very high benefit, while many people only benefit very little), it becomes clear that homophily is a property that makes the rich even richer than in a pure scale-free network. This effect shows that the impact of homophily increases the effect of the basic preferential attachment rule. This is a surprising result from the analysis of homophily.

The simulation results also show that there is a homophily level Λ that makes the cumulative degree distribution in log-log scale a straight line. The existence of a power function suggests the proposition that a balance between two effects can be achieved. One of the effects is due to the linkages among existing nodes and due to the linkages between new nodes and existing nodes (preferential attachment rule), making the cumulative degree distribution concave in log-log scales. The other effect is due to homophily, making the cumulative degree distribution convex in log-log scales.

Furthermore, previous research has not paid sufficient attention to the significantly different curvatures of cumulative degree distributions of empirical networks. They only considered the concave head [38]. The concave tail has been considered as a local distortion [41,42]. With our modified preferential attachment rule (i.e., with the preferential attachment rule together with factors representing homophily), we suggest that all kinds of curvature of the cumulative degree distributions can be explained.

Many empirical networks (e.g., the social network of Flickr, the message correspondence of UC Irvine, the network of Facebook wall post) show a concave cumulative degree distribution [10]. In this case, we can argue that the level of homophily is not as high as it impacts the network formation strongly. The cumulative degree distributions of some other empirical networks (e.g., Yahoo advertising word linkages, EU institution employees email exchanges, and Wikipedia contributors message exchanges) show convex shapes in log-log scales [10]. This suggests that the revealed topology of these networks might have strongly been impacted by homophily. For example, analyzing the types of these networks, we can assume that the number of emails being send between people in an organization is highly related to the topics (e.g., projects), which they are working on, or the organizational structure, which they belong to. Similar is true for the email exchange between Wikipedia contributors. These people are exchanging emails about specific topics only. In these two examples, the homophily comes from being part of a project or department. Even in the case of word linkages, it can be argued that words are used in certain contexts (i.e., groups), defining the homophily for this example.

An implication of our results is that, once an understanding of the impact of homophily and seed network topologies on the formation of networks based on preferential attachment rules is achieved, appropriate actions for a controlled formation of networks (e.g., for SaaS software service networks) can be taken. For example, when policy makers find a convex cumulative degree distribution in log-log scales, they can assume that there are groups, in which nodes are homophilic. Then, if the network should be open across all groups according to the diagnosis of the policy maker, the policy makers can implement policies, which reduce the homophily, so that the networks evolve to a regular scale-free network. This would spread the benefit of the network in a fairer way. More concrete, from a business perspective, if a company possesses popular software services or first-mover software services, it can build groups that are highly beneficial to them. The company could provide a platform, which supports users to utilize the companies own software services rather than the one of other companies [16]. Therefore, policy makers with the knowledge about the cumulative degree distribution of the software service network can set policies that counteract this tendency, leveling the play field between the companies in the market. It suppresses situations, in which groups are used strategically against competitors.

Although our theoretical research proposes explanations for observations made from empirical networks, it comes with two limitations. First, as the preferential attachment rule and the concept of homophily are simple concepts, they might not reflect reality comprehensively. Therefore, despite our extension of the preferential attachment rule, further factors could also impact the shape of the cumulative degree distributions of network. Second, we showed only the results of a numerical analysis in our paper, and we did not prove the relationship between homophily and the deviation of scale-free topologies through an analytic method. Therefore, further studies could focus on providing an analytic model to prove this relationship.

Acknowledgments

This research was supported by the Basic Science Research Program (2013R1A1A2058665), the Social Sciences Korea Program (NRF-2015-S1A3A-2046742) of the National Research Foundation of Korea, and the project BASMATI, which has received funding from the ICT R&D program of the Korean MSIP/IITP (R0115-16-0001) and from the European Unions Horizon 2020 research and innovation programme under grant agreement no. 723131.

References

[1] Barabási A-L, Albert R. Emergence of scaling in random networks. Science 1999;286(5439):509–12.
[2] Leskovec J, Adamic LA, Huberman BA. The dynamics of viral marketing. ACM Trans Web 2007;1:228–37.
[3] Hendler J, Shadbolt N, Hall W, Berners-Lee T, Weitzner D. Web science: an interdisciplinary approach to understanding the web. Commun ACM 2008;51(7):60–9.
[4] Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. Phys Rev Lett 2001;86(14):3200–3.
[5] Kuandykov L, Sokolov M. Impact of social neighborhood on diffusion of innovation s-curve. Decis Support Syst 2010;48(4):531–5.
[6] Wu F, Huberman BA, Adamic LA, Tyler JR. Information flow in social groups. Phys A: Stat Mech Appl 2004;337(1-2):327–35.
[7] Albert R, Jeong H, Barabási A-L. Error and attack tolerance of complex networks. Nature 2000;406(6794):378–82.
[8] Tu Y. How robust is the internet? Nature 2000;406(6794):353–4.
[9] Albert R, Barabási A-L. Statistical mechanics of complex networks. Rev Mod Phys 2002;74(1):47–97.
[10] Konect-The Koblenz Network Collection. http://konect.uni-koblenz.de/plots/bidd (Accessed 10.09.16).
[11] Barabási AL, Jeong H, Nda Z, Ravasz E, Schubert A, Vicsek T. Evolution of the social network of scientific collaborations. Phys A: Stat Mech Appl 2002;311(3–4):590–614.
[12] Yook S-H, Jeong H, Barabási A-L. Modeling the internet's large-scale topology. Proc Nat Acad Sci 2002;99(21):13382–6.
[13] Liu J, Abbass HA, Zhong W, Green GD. Local-global interaction and the emergence of scale-free networks with community structures. Artif Life 2011;17(4):263–79.
[14] Koohborfardhaghighi S, Altmann J. How variability in individual patterns of behavior changes the structural properties of networks, WIC 2014. In: International Conference on Active Media Technology at Web Intelligence Congress. Poland: Warsaw; 2014.
[15] Koohborfardhaghighi S, Altmann J. How placing limitations on the size of personal networks changes the structural properties of complex networks. In: Proceedings of the 6th International Workshop on Web Intelligence and Communities. Korea: Seoul; 2014.

[16] Kim K, Altmann J, Hwang J. An analysis of the openness of the web2.0 service network based on two sets of indices for measuring the impact of service ownership. In: Proceedings of the 2011 44th Hawaii international conference on system sciences. HICSS '11;. Washington, DC, USA: IEEE Computer Society; 2011. p. 1–11. ISBN 978-0-7695-4282-9.

[17] Koohborfardhaghighi S, Altmann J. A network formation model for social object networks. LISS 2014, International Conference on Logistics, Informatics, and Services Sciences. USA: Berkeley; 2014.

[18] Huang Y, Shen C, Contractor NS. Distance matters: exploring proximity and homophily in virtual world networks. Decis Support Syst 2013;55(4):969–77.

[19] Putzke J, Fischbach K, Schoder D, Gloor PA. The evolution of interaction networks in massively multiplayer online games. J Assoc Inform Syst 2010;11(2):2.

[20] McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: Homophily in social networks. Ann Rev Sociol 2001;27(1):415–44.

[21] Bramoullé Y, Currarini S, Jackson MO, Pin P, Rogers BW. Homophily and long-run integration in social networks. J Econ Theory 2012;147(5):1754–86.

[22] Kim K, Altmann J, Hwang J. The impact of the subgroup structure on the evolution of networks: An economic model of network evolution. In: Proceedings of IEEE Conference on Computer Communications Workshops, 2010; 2010. p. 1–9.

[23] Albert R, Jeong H, Barabási A-L. Internet: Diameter of the world-wide web. Nature 1999;401(6749):130–1.

[24] Newman M. The structure and function of complex networks. SIAM Rev 2003;45(2):167–256.

[25] Valverde S, Solé RV. Self-organization versus hierarchy in open-source social networks. Physical Review E 2007;76(4):046118.

[26] Cohen R, Havlin S. Scale-free networks are ultrasmall. Physical Review Letters 2003;90(5):058701.

[27] Kim BJ, Trusina A, Minnhagen P, Sneppen K. Self organized scale-free networks from merging and regeneration. The European Physical Journal B - Condensed Matter and Complex Systems 2005;43(3):369–72.

[28] Dangalchev C. Generation models for scale-free networks. Phys A: Stat Mech Appl 2004;338(3–4):659–71.

[29] Wang X, Loguinov D. Understanding and modeling the internet topology: economics and evolution perspective. IEEE/ACM Trans Netw 2010;18(1):257–70.

[30] Smaragdakis G, Laoutaris N, Lekakis V, Bestavros A, Byers J, Roussopoulos M. Selfish overlay network creation and maintenance. IEEE/ACM Trans Netw 2011;19(6):1624–37.

[31] Adamic LA, Adar E. Friends and neighbors on the web. Social Netw 2003;25(3):211–30.

[32] Wagner CS, Leydesdorff L. Network structure, self-organization, and the growth of international collaboration in science. Res Policy 2005;34(10):1608–18.

[33] Wang F, Sun Y. Self-organizing peer-to-peer social networks. Comput Intell 2008;24(3):213–33.

[34] Hwang J, Altmann J, Kim K. The structural evolution of the web 2.0 service network. Online Inform Rev 2009;33(6):1040–57.

[35] Festinger L, Back KW, Schachter S. Social pressures in informal groups: a study of human factors in housing. Stanford University Press; 1950. ISBN 9780804701730.

[36] Schelling TC. Micromotives and macrobehavior. Revised edition. New York: W. W. Norton & Company; 2006. ISBN 9780393329469.

[37] Granovetter MS. The strength of weak ties. Am J Sociol 1973;78(6):1360–80.

[38] Newman M. Power laws, Pareto distributions and zipf's law. Contemp Phys 2005;46(5):323–51.

[39] Aiello LC, Dunbar RIM. Neocortex size, group size, and the evolution of language. Current Anthropol 1993;34(2):184–93.

[40] Dunbar RIM. Coevolution of neocortical size, group size and language in humans. Behav Brain Sci 1993;16(04):681–94.

[41] Amaral LAN, Scala A, Barthlmy M, Stanley HE. Classes of small-world networks. Proc Nat Acad Sci 2000;97(21):11149–52.

[42] Fenner T, Levene M, Loizou G. A model for collaboration networks giving rise to a power-law distribution with an exponential cutoff. Social Netw 2007;29(1):70–80.