Statistica Sinica Preprint No: SS-2022-0136

Title	A New Preferential Model With Homophily for Recommender Systems
Manuscript ID	SS-2022-0136
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0136
Complete List of Authors	Hanyang Tian, Bo Zhang, Ruixue Jiang and Xiao Han
Corresponding Authors	Bo Zhang
E-mails	zhangbo890301@outlook.com

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0136

A NEW PREFERENTIAL MODEL WITH HOMOPHILY FOR RECOMMENDER SYSTEMS

Hanyang Tian, Bo Zhang*, Ruixue Jiang*, and Xiao Han

University of Science and Technology of China

Abstract: "Rich-get-richer" and "homophily" are two essential phenomena in evolving social networks. "Rich-get-richer" means people with greater followings are more likely to attract new followers, and "homophily" means people prefer to bond with others of the same social group, or who have some other attribute in common. To formalize these phenomena simultaneously in the context of an evolving social network, we consider a K-group preferential attachment (KPA) model, which is helpful for social network recommender systems. Our primary contribution is to propose a new evolving social network model that incorporates the mechanisms of rich-get-richer and homophily. We show that the proposed KPA model exhibits a power-law degree distribution for each group, and prove the central limit theorem for the maximum likelihood estimation of the parameters in the model. In addition, we verify the robustness of the proposed parameter estimation methods, and apply them to simulated data and to real-data examples.

Key words and phrases: evolving network, homophily, preferential attachment, recommender system.

1. Introduction

The "rich-get-richer" (or preferential attachment) mechanism has been studied by numerous researchers. In preferential attachment models, new edges in evolving networks are attached to older nodes, chosen according to a probability distribution that is an affine function of the older nodes' degrees. This way, nodes with high degrees are more likely to attract edge connections and attain higher degrees, which explains why they are called rich-get-richer models. A basic introduction to the preferential attachment model can be found in Easley et al. (2010) and Barabási and Albert (1999). To further understand the statistical properties of this model, Chung and Lu (2006), Durrett (2007), and Van Der Hofstad (2024) show the limit theories and asymptotic characteristics.

The "homophily" mechanism has a profound effect on individuals and groups in society (Lazarsfeld et al. (1954)). This well-documented phenomenon appears in social networks, and describes how people often prefer to connect with others who have similar characteristics (McPherson et al. (2001)). For example, people are more likely to build social relations, such as marriage, friendship, and colleagues, with someone of the same age or education, or who has similar hobbies. In other words, homophily influences the connection structure in human society. Furthermore, many studies show that homophily affects not only society's static structures, but also its dynamic operations. Jackson et al. (2008) and Jackson and López-Pintado (2013) show the effect of homophily on the welfare of individuals and diffusion patterns of information in social networks.

On social network platforms (e.g., Twitter, TikTok, and Sina Weibo), rich-get-richer and homophily often co-occur. This study supposes that it takes two steps for a person to connect with another in a social network: (1) become aware of someone through a referral or a social media feed, and (2) decide whether to follow or connect with that person. Internet celebrities are more likely to be introduced to others (the first step), implying that the rich-get-richer phenomenon affects the celebrity's followers. In the second step, a person prefers to follow somebody with, for example, the same hobby, which means homophily is also involved. Moreover, homophily can work contrary to the influence of rich-get-richer. For example, Lebron James is a basketball superstar with a huge following. His popularity makes it easy for him to get more followers, but someone with no interest in basketball will not follow him, even if friends recommend him. Thus, it is meaningful to study evolving networks, while considering interactions of rich-get-richer and homophily. Unfortunately, most previous studies typically consider rich-get-richer and homophily separately.

Some recent papers (Lee et al. (2019), Avin et al. (2020), and Hajek and Sankagiri (2019)) have tried to combine these ideas. However, they do not focus on statistical problems, such as estimators and central limit theorems. Hajek and Sankagiri (2019) concentrated on the community recovery problem of the preferential attachment model without edge-steps, which greatly inspired our research. Compared with their work, ours estimates the homophily parameter and our model considers edge-steps, which are common in evolving social networks.

In this paper, we propose the K-group preferential attachment (KPA) model, based on the Barab*á*si–Albert model (Barabási and Albert (1999) and the work of Albert and Barabási (2002)). The unit time point of the dynamic process is divided into two parts: (1) [rich-get-richer] The evolving network tries to connect a chosen older node to the new node. The higher the degree of the older node, the higher the probability that it will be chosen. (2) [homophily] The new node will accept the connection with a probability dependent on the similarity of the two nodes.

We divide all nodes into K groups according to a specific feature. Homophily states that nodes in the same group are more easily connected. We introduce a parameter θ to the classic Barab*á*si–Albert model to represent the influence of homophily on the generation of evolving networks.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0136

Using the KPA model, we obtain theoretical results about degrees. Then, we propose the estimators of the homophily and the other parameters in an evolving network featuring both rich-get-richer and homophily. We also give the joint asymptotic distribution of these estimators. It is commonly acknowledged that recommender systems play a vital role in big data (see Jannach et al. (2010) and Ricci et al. (2015)). Accurate estimation of the effects of homophily helps improve the recommender system of any social network platform. If the homophily is strong, recommending connections from other groups (i.e., groups to which recommended node does not belong) is inefficient. In contrast, when the homophily is not strong, recommending nodes with high degrees from different groups is meaningful.

The remainder of the paper is organized as follows. In Section 2, we introduce the specific construction process of the KPA model and the meaning of each random variable. The main asymptotic results are presented in Section 3. The parameter estimation methods are given in Section 4. Section 5 focuses on the change point, and Section 6 discusses the robustness of the estimations. Section 7 contains simulations to illustrate the theoretical results. In Section 8, we apply our methods to real-life data. Simulations showing the robustness of the estimators and the proofs of the lemmas and theorems are provided in the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0136

Notation

In this paper, for a set $\mathcal{M}$, $|\mathcal{M}|$ means the number of elements in $\mathcal{M}$, and for a constant $x \in \mathbb{R}$, $|x|$ means the absolute value of x. We use $[n] := \{1, \cdots, n\}$, for some $n \in \mathbb{Z}^+$, and $\mathbb{1}\{\cdot\}$ denotes the indicator function.

We use the graph $G(t) = (\mathcal{V}(t), \mathcal{E}(t), \mathcal{G}(t))$ to record the state of an evolving network at time $t \in \mathbb{Z}$, where $\mathcal{V}(t)$ is the set of nodes, $\mathcal{E}(t)$ is the set of edges, and $\mathcal{G}(t)$ is the set of node group labels. Furthermore, t is discrete, and if the network structure changes, t changes ($t \to t+1$, $G(t) \to G(t+1)$). An evolving network on the time range $[0, T]$ means the discrete process $\{G(t)\}_{t=0}^{T}$, where $T \in \mathbb{Z}^+$. The graph $G(0)$ is the initial state of the dynamic process, and the graph $G(T)$ is the final stage.

2. Model

According to the classic Barab*á*si–Albert model, an initial graph $G(0)$ has an isolated node with one loop, and its degree is one (Chung and Lu (2006)). There are two operations on the evolving network:

- Vertex-step: A new node w is added to the network, and connects to node u with the edge (w, u).

- Edge-step: No new node arrives, but a new edge (w, u) is added to the network. Nodes w, u are pre-existing in the network.

Here, $\mathcal{V}(t)$ is the node set in graph $G(t)$. Let $v(t) := |\mathcal{V}(t)| - |\mathcal{V}(t-1)|$. For $t > 0$, $G(t)$ is formed from $G(t-1)$ by performing a vertex-step when $v(t) = 1$, or an edge-step when $v(t) = 0$. Assume there are K groups of nodes in the network, where K is a fixed known constant. Let $g_i \in [K]$ be the group label of node i. We make the following assumptions:

Assumption 1. *Group label g_i is known (or observed) for each node i.*

Assumption 2. *The new node comes from group k with unknown probability p_k at the vertex-step, for $k \in [K]$, where $p_k \in [0,1]$ and $\sum_{k=1}^{K} p_k = 1$.*

Assumption 3. *$\{v(t)\}_{t=1}^{T}$ are independent and identically distributed (i.i.d.) random variables with a Bernoulli distribution $B(1, q)$, where $q \in (0, 1]$ is an unknown constant.*

Assumption 3 implies that qt is an approximation for $|\mathcal{V}(t)|$.

Assumption 4. *In the initial graph $G(0)$, the number of nodes from group k is $p_k n_0$. We assume that n_0 is a constant large enough for $p_k n_0$ to be an integer, for $k \in [K]$. Each node is isolated and has a loop (the node's degree is one).*

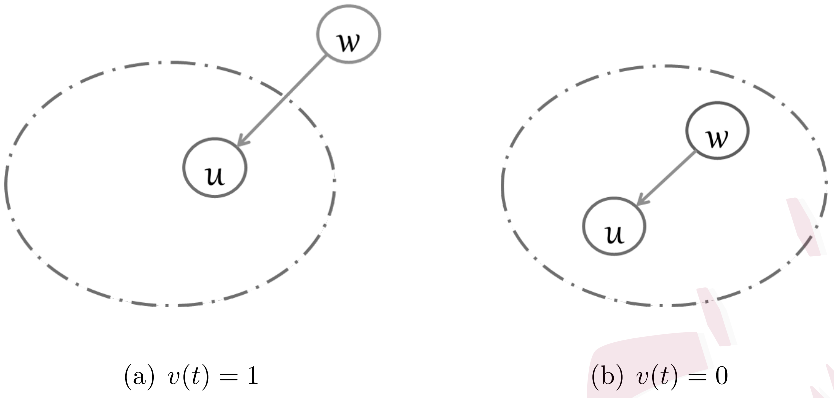


Figure 1: The dotted circle shows the graph $G(t-1)$, the node outside the circle represent the new one and the arrow represent the new edge at time t.

Next, we discuss the process of generating network data by using a KPA model in detail. Beginning with the initial graph $G(0)$, with n_0 ($n_0 \geq K$) isolated nodes (each node has a loop) from K groups, $v(t)$ is generated randomly at time t. Let $d_i(t)$ be the degree of node i in $G(t)$. Next, the two operations are as follows:

- Vertex-step:

Step 1. A new node w comes to the network at time $t+1$.

Step 2. An older node u_1 in $G(t)$ is chosen to connect to w with probability $d_{u_1}(t)/[\sum_{i \in \mathcal{V}(t)} d_i(t)]$. If u_1 and w are from the same group, w accepts the connection without hesitation. Otherwise, w rejects the connection with probability $1-\gamma$, $\gamma \in (0,1]$.

Step 3. If w successfully connects to u_1 in Step 2, the vertex-step process stops. Otherwise, the network chooses another node in $G(t)$ to connect with w in Step 4, after w rejects u_1.

Step 4. Another node u_2 in $G(t)$ is chosen with probability $\alpha d_{u_2}(t)/[\sum_{i\in\mathcal{V}(t),g_i=g_w} d_i(t)]$, for $\alpha \in (0,1]$, if u_2 and w are from the same group. Otherwise, the chosen probability is $(1-\alpha)d_{u_2}(t)/[\sum_{i\in\mathcal{V}(t),g_i\neq g_w} d_i(t)]$. Furthermore, if u_2 and w are from the same group, w accepts the connection without hesitation. Otherwise, w rejects u_2 with probability $1-\gamma$.

Step 5. If w successfully connects to u_2 in Step 4, the vertex-step process stops. Otherwise, the network chooses another node to connect to w, and goes back to Step 4.

- Edge-step: This process is identical to the vertex-step, except for Step 1.

Step 1. No new node arrives at time t. Randomly select a node w in the network with probability $d_w(t)/[\sum_{i\in\mathcal{V}(t)} d_i(t)]$.

Remark 1. We allow multiple edges between any two nodes in the edge-step. The first edge between the two nodes denotes following to become a friend, and subsequent edges denote liking, commenting, and other interactions after becoming friends. Both following and interactions increase the degrees of the nodes.

Remark 2. In the vertex-step, both the number of nodes and the number of edges increase by one. In contrast, only the number of edges increases by one in the edge-step. When q is away from zero and t tends to infinity, the number of nodes and the number of edges have the same order. From Page 128 of Newman (2018), the *connectance* or *density* ρ of the present network is $|\mathcal{E}|/[(|\mathcal{V}|-1)|\mathcal{V}|]$, where $|\mathcal{E}|$ is the number of edges, and $|\mathcal{V}|$ is the number of nodes. Thus, ρ tends to zero, and the network is sparse.

Remark 3. The case $\gamma = 1$ means connections from all groups are accepted without hesitation, which implies there is no homophily effect on the network. If $\gamma < 1$, then connections from other groups are rejected with a certain probability, which implies that nodes of the same group are more easily connected, and homophily exists in the network. To illustrate the meaning and role of the parameter γ in dynamic networks, we present three examples.

1. Groups with different political orientations from different communities on Twitter. After observation, people with different political

orientations follow each other very little. At this point, the dynamic network parameter γ is close to zero.

2. On TikTok, users can follow each other. Users who like different ball games come from different communities. Users who like basketball tend to follow other users who like basketball, or who post about basketball. However, they sometimes also follow or know about other popular sports-related users, such as football or tennis players. Here, the dynamic network parameter γ is away from zero.

3. Some groups in social networks overlap significantly. For example, Chinese young people who are interested in an animation also pay close attention to the game. In this case, γ may be close to one.

If $\gamma = 1$, the recommender should recommend older nodes with high degrees from different groups to the new node to construct a large network with centralized nodes. If γ is small, recommending nodes from the same group is safer. As a result, many sub-networks of different groups are generated.

We suppose that people do not refuse to make friends with people who share their interests, so nodes must accept connections from the same group. Furthermore, we assume that if a person refuses to be friends with

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0136

a celebrity because their interests do not match, he or she will pay more attention in future to people with the same interests, rather than simply because the person is famous. Thus, if the new node rejects connections from other groups the first time, the older nodes in the same group will be chosen with probability α next time, where α is large. If we know information about a connection rejection, we can infer and estimate γ. However, general network data shows only information about successful connections. Thus, we introduce the following discussion and a new parameter θ.

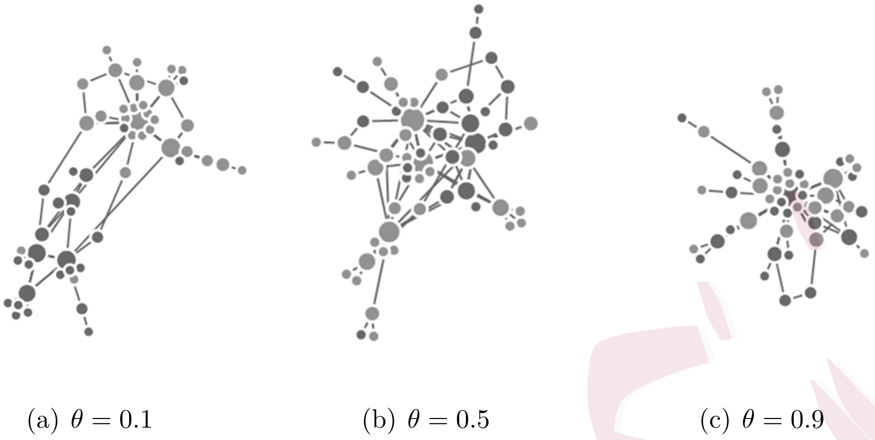
Using the above details for the KPA model, we can calculate the conditional probability of the edge $e(t+1) := (e_1(t+1), e_2(t+1))$ connected at time $t+1$, where $e_1(t+1), e_2(t+1)$ are the two nodes of the edge $e(t+1)$:

$$
\begin{aligned}
&P(e(t+1) = (w,u) | g_w, g_u, G(t)) \\
&= \begin{cases}
\frac{d_u(t)}{\sum_{i\in\mathcal{V}(t)} d_i(t)} + \frac{\alpha(1-\gamma)}{\gamma+\alpha(1-\gamma)} \frac{\sum_{i,\, g_i\neq g_w} d_i(t)}{\sum_{i\in\mathcal{V}(t)} d_i(t)} \frac{d_u(t)}{\sum_{i,g_i=g_w} d_i(t)}, & \text{case } 1; \\
\frac{\gamma}{\gamma+\alpha(1-\gamma)} \frac{d_u(t)}{\sum_{i\in\mathcal{V}(t)} d_i(t)}, & \text{case } 2; \\
\frac{d_w(t)}{\sum_i d_i(t)} \left[\frac{d_u(t)}{\sum_i d_i(t)} + \frac{\alpha(1-\gamma)}{\gamma+\alpha(1-\gamma)} \frac{\sum_{i,g_i\neq g_w} d_i(t)}{\sum_i d_i(t)} \frac{d_u(t)}{\sum_{i,g_i=g_w} d_i(t)} \right], & \text{case } 3; \\
\frac{d_w(t)}{\sum_{i\in\mathcal{V}(t)} d_i(t)} \frac{\gamma}{\gamma+\alpha(1-\gamma)} \frac{d_u(t)}{\sum_{i\in\mathcal{V}(t)} d_i(t)}, & \text{case } 4.
\end{cases}
\end{aligned}
\tag{2.1}
$$

where case 1: $g_w = g_u$, $v(t+1) = 1$; case 2: $g_w \neq g_u$, $v(t+1) = 1$; case 3: $g_w = g_u$, $v(t+1) = 0$; case 4: $g_w \neq g_u$, $v(t+1) = 0$.

Let $\theta := \gamma/[\gamma + \alpha(1-\gamma)] \in (0,1]$. We find that θ is the parameter that ultimately determines the influence of homophily on the network structure. Figure 2 shows the influence of θ on the network structure. Although γ is unobservable, we can infer the homophily by estimating θ and testing whether $\theta = 1$, using the methods described in Section 4. We can see that $\theta = 1$ infers $\gamma = 1$, and $\theta \geq \gamma$ implies that $\theta < 1$ can infer $\gamma < 1$. Then, we can recommend older users to new users using a strategy based on θ, as follows. If the result of statistical inference is $\theta = 1$ and there is no homophily, we recommend older users with high degrees in different groups. This process contributes to constructing a large network. Otherwise, when the result of the statistical inference is $\theta < 1$ and there is homophily, we recommend older users from the same group to ensure the new users can connect to the network quickly. Thus, in this study, we foucus on how to obtain information about the parameter θ.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0136



(a) $\theta = 0.1$ (b) $\theta = 0.5$ (c) $\theta = 0.9$

Figure 2: The network generated by the KPA model with the time range $[0, 100]$ and parameters $K = 2$, $q = 0.5$, $p_1 = p_2 = 0.5$.

In the following sections, we focus on θ instead of γ and α.

3. Asymptotic results

Theorem 1. *Under Assumptions 1–3, let $d_i(t)$ be the degree of node i in graph $G(t)$. Let $D_k(t) = \sum_{i \in \mathcal{V}(t)} d_i(t) \mathbb{1}\{g_i = k\}$ be the total degrees from group k in $G(t)$, for $k \in [K]$. When t tends to infinity,*

$$\frac{D_k(t)}{2t} \xrightarrow{a.s.} p_k.$$

Here, $D_k(t)/2t$ is the ratio of the degrees from group k, the sum of the nodes' degrees from group k divided by the total degrees at time t.

Consider the edge added at time t, $e(t) = (e_1(t), e_2(t))$. Let $X(t) =$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0136

$\mathbb{1}\{g_{e_1(t)} = g_{e_2(t)}\}$ and $S(t) = \sum_{i=1}^{t} X(i)$. Here, $S(t)$ is the number of edges when two nodes are from the same group in graph $G(t)$.

Corollary 1. *Under the conditions of Theorem 1, for any node i in graph $G(t)$, we have when t tends to infinity,*

$$\frac{d_i(t)}{2t} \xrightarrow{i.p.} 0.$$

Corollary 1 coincides with the sparsity in Remark 2.

Corollary 2. *Under the conditions of Theorem 1, we have that when t tends to infinity,*

$$\frac{S(t)}{t} \xrightarrow{a.s.} 1 + \theta \left(\sum_{k=1}^{K} p_k^2 - 1 \right).$$

Theorem 1 implies a limit of the ratio of nodes' degrees from group k. However, the limit might differ from what we get at a particular time t. We give a probabilistic estimate of the difference in the following theorem.

Theorem 2. *Under Assumptions 1–4, for some time point t, we have:*

$$\begin{cases} P\left(|D_k(t) - p_k(2t+n_0)| \geq 2c_1(t)t^{1/2}\right) \leq C_1 e^{-[c_1(t)]^2}, & \text{if } \frac{1}{2-\theta} < q \leq 1, \\ P\left(|D_k(t) - p_k(2t+n_0)| \geq 2c_2(t)\log^{1/2}(t)\right) \leq C_2 e^{-\frac{[c_2(t)]^2}{t}}, & \text{if } q = \frac{1}{2-\theta}, \\ P\left(|D_k(t) - p_k(2t+n_0)| \geq 2c_3(t)\right) \leq C_3 e^{-\frac{[c_3(t)]^2}{t^{2-q(2-\theta)}}}, & \text{if } 0 < q < \frac{1}{2-\theta}, \end{cases}$$

where $c_1(t)$, $c_2(t)$, and $c_3(t)$ are strictly monotonically increasing positive functions of t, and C_1, C_2, and C_3 are constants greater than zero.

Remark 4. The phase transition phenomenon in Theorem 2 is because of $\sum_{j=1}^{t} j^{q(2-\theta)-2}$ with different q, which is described in detail in Section S5 of the Supplementary Material.

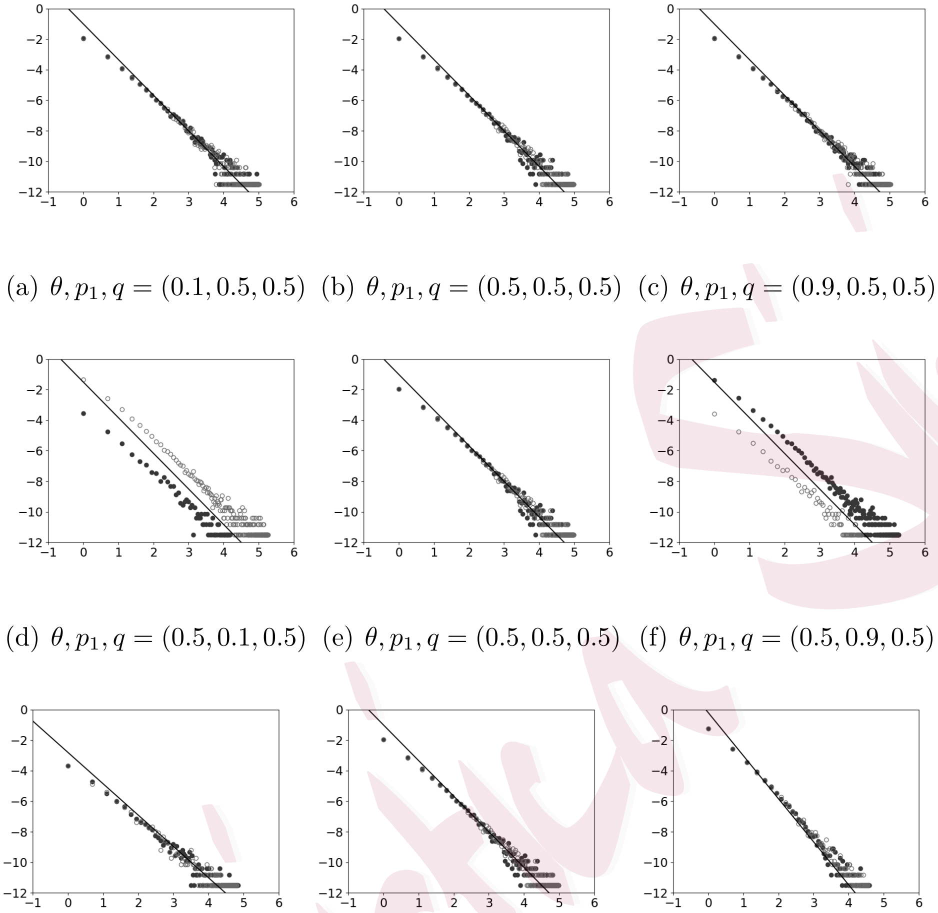
Remark 5. To better understand the convergence rate of the group degree, let $c_1(t) = \log^{1/2}(t)$, $c_2(t) = (t\log(t))^{1/2}$, and $c_3(t) = (t^{2-q(2-\theta)}\log(t))^{1/2}$. Then,

$$\begin{cases} P\left(|D_k(t) - p_k(2t+n_0)| \geq 2t^{1/2}\log^{1/2}(t)\right) \leq C_1 t^{-1}, & \text{if } \frac{1}{2-\theta} < q \leq 1, \\ P\left(|D_k(t) - p_k(2t+n_0)| \geq 2t^{1/2}\log(t)\right) \leq C_2 t^{-1}, & \text{if } q = \frac{1}{2-\theta}, \\ P\left(|D_k(t) - p_k(2t+n_0)| \geq 2t^{1-q(2-\theta)/2}\log^{1/2}(t)\right) \leq C_3 t^{-1}, & \text{if } 0 < q < \frac{1}{2-\theta}. \end{cases}$$

The degree distribution obeying the power law is an attractive property of the classic preferential attachment model. For the KPA model, the nodes are from K groups. We now have the power-law degree distribution for each group.

Theorem 3. *Under the conditions of Theorem 2, let $m_{k,d}(t)$ denote the number of nodes with degree d from group k in graph $G(t)$. Note that $m_{k,1}(0) = p_k n_0$ and $m_{k,0}(t) = 0$. For $k \in [K]$, letting $M_{k,d} = \lim_{t\to\infty} E(m_{k,d}(t))/t$, we have*

$$M_{k,d} = \frac{2qp_k}{4-q}\prod_{j=2}^{d}\frac{(j-1)(2-q)}{2+j(2-q)} \propto d^{-[1+2/(2-q)]}.$$



(a) $\theta, p_1, q = (0.1, 0.5, 0.5)$ (b) $\theta, p_1, q = (0.5, 0.5, 0.5)$ (c) $\theta, p_1, q = (0.9, 0.5, 0.5)$

(d) $\theta, p_1, q = (0.5, 0.1, 0.5)$ (e) $\theta, p_1, q = (0.5, 0.5, 0.5)$ (f) $\theta, p_1, q = (0.5, 0.9, 0.5)$

(g) $\theta, p_1, q = (0.5, 0.5, 0.1)$ (h) $\theta, p_1, q = (0.5, 0.5, 0.5)$ (i) $\theta, p_1, q = (0.5, 0.5, 0.9)$

Figure 3: The x-axis is $\log(d)$ and the y-axis is $\log(m_{k,d}(T)/T)$.

Figure 3 shows the power-law degree distribution of a simulated network data set where $T = 100000$, and $K = 2$. Hollow nodes and solid nodes come from two groups, and the solid line is $y = -[1 + 2/(2 - q)]x + C$.

Theorem 3 implies that the rich-get-richer mechanism leads to a degree distribution following the power law. For network data with a power-law degree distribution, stochastic block models (SBMs) cannot explain the significant difference between the nodes' degrees, and degree correction block models (DCBMs) cannot explain the existence of nodes with enormous or tiny degrees. In contrast, our KPA model is suitable for interpreting network data with a power-law degree distribution. We show the significant difference in the degree distributions between the SBM model and the rich-get-rich model in Section S1 of the Supplementary Material.

4. Parameter estimation

4.1 With historical information

The parameters involved in the KPA model are $(\theta,\ \{p_k\}_{k=1}^K,\ q)$, where $\sum_{k=1}^K p_k = 1$. Let $\psi = (\theta, \{p_k\}_{k=1}^{K-1}, q)^\top$. Then, $\{G(t)\}_{t=0}^T$ is an evolving network process generated by a KPA model with the time range $[0, T]$. Based on $\{G(t)\}_{t=0}^T$, we can obtain the MLE of the parameters:

$$\text{For } k \in [K],\ \hat{p}_k = \frac{\sum_{t=1}^T \mathbb{1}\{v(t) = 1, g_{e_1(t)} = k\}}{\sum_{t=1}^T v(t)}; \quad \hat{q} = \frac{\sum_{t=1}^T v(t)}{T};$$
$$\hat{\theta} = \arg \max_{\theta \in (0, 1+\epsilon)} \log L_2(\theta | \{G(t)\}_{t=0}^T),$$

where

$$
\begin{aligned}
& \log L_2(\theta|\{G(t)\}_{t=0}^{T}) \\
= \quad & \sum_{t=1}^{T}\sum_{k=1}^{K} \mathbb{1}\{v(t)=1, g_{e_1(t)}=g_{e_2(t)}=k\}\log\left[P_k(t)+(1-\theta)\left(1-P_k(t)\right)\right] \\
& +\sum_{t=1}^{T}\sum_{k=1}^{K} \mathbb{1}\{v(t)=1, g_{e_1(t)}=k, g_{e_2(t)}\neq k\}\log\left[\left(1-P_k(t)\right)\theta\right] \\
& +\sum_{t=1}^{T}\sum_{k=1}^{K} \mathbb{1}\{v(t)=0, g_{e_1(t)}=g_{e_2(t)}=k\}\log\left[P_k(t)\left(P_k(t)+(1-\theta)\left(1-P_k(t)\right)\right)\right] \\
& +\sum_{t=1}^{T}\sum_{k=1}^{K} \mathbb{1}\{v(t)=0, g_{e_1(t)}=k, g_{e_2(t)}\neq k\}\log\left[P_k(t)\left(1-P_k(t)\right)\theta\right]. \qquad (4.1)
\end{aligned}
$$

$e(t)=(e_1(t),e_2(t))$, $\epsilon=\min_{k\in[K]}\{\min\{P_k(t)/(1-P_k(t)) : t\in[1,T], g_{e_1(t)}= g_{e_2(t)}=k\}\}$, $P_k(t)=D_k(t-1)/[2(t-1)+n_0]$.

Definition 1. To avoid confusion of symbols, let $\psi^*=(\theta^*,\{p_k^*\}_{k=1}^{K-1},q^*)^\top$ be the vector of true parameters of the KPA model. Let $\hat{\psi}=(\hat{\theta},\{\hat{p}_k\}_{k=1}^{K-1},\hat{q})^\top$ be the MLE based on $\{G(t)\}_{t=0}^{T}$.

Theorem 4. *Assume the evolving network $\{G(t)\}_{t=0}^{T}$ is generated by a KPA model under Assumptions 1–3. Then, when T tends to infinity,*

$$\hat{\psi}\xrightarrow{a.s.}\psi^*.$$

Theorem 4 guarantees the convergence of the MLE, and we can obtain the asymptotic normality.

Theorem 5. *Under the conditions of Theorem 4, when T tends to infinity,*

$$T^{1/2}\left(\hat{\psi}-\psi^*\right)\xrightarrow{d}N(0,\mathbf{\Sigma}^{-1}),$$

$$\mathbf{\Sigma}=\begin{bmatrix}\sum_{k=1}^{K}\left[\frac{p_k^*(1-p_k^*)}{\theta^*}+\frac{p_k^*(1-p_k^*)^2}{p_k^*+(1-p_k^*)(1-\theta^*)}\right] & 0 & 0\\ 0 & \mathbf{\Sigma_{22}} & 0\\ 0 & 0 & \frac{1}{q^*(1-q^*)}\end{bmatrix},$$

where $p_K^*=1-\sum_{k=1}^{K-1}p_k^*$ and $\mathbf{\Sigma_{22}}$ is a $(K-1)\times(K-1)$ symmetric matrix satisfying

$$\Sigma_{22}(ij)=\begin{cases}q^*/p_K^*, & i\neq j;\\ q^*(p_l^*+p_K^*)/(p_l^*p_K^*), & i=j=l.\end{cases}$$

Corollary 3. *Under the conditions of Theorem 4, when T tends to infinity,*

$$T^{1/2}\hat{\mathbf{\Sigma}}^{\mathbf{1/2}}(\hat{\psi}-\psi^*)\xrightarrow{d}N(0,I_{K+1}),$$

where I_{K+1} is the $(K+1)\times(K+1)$ identity matrix, and $\hat{\mathbf{\Sigma}}$ is the estimator of $\mathbf{\Sigma}$:

$$\hat{\mathbf{\Sigma}}=\begin{bmatrix}\sum_{k=1}^{K}\left[\frac{\hat{p}_k(1-\hat{p}_k)}{\hat{\theta}}+\frac{\hat{p}_k(1-\hat{p}_k)^2}{\hat{p}_k+(1-\hat{p}_k)(1-\hat{\theta})}\right] & 0 & 0\\ 0 & \hat{\mathbf{\Sigma}}_{\mathbf{22}} & 0\\ 0 & 0 & \frac{1}{\hat{q}(1-\hat{q})}\end{bmatrix},$$

where $\hat{p}_K=1-\sum_{k=1}^{K-1}\hat{p}_k$ and

$$\hat{\Sigma}_{22}(ij)=\begin{cases}\hat{q}/\hat{p}_K, & i\neq j;\\ \hat{q}(\hat{p}_l+\hat{p}_K)/(\hat{p}_l\hat{p}_K), & i=j=l.\end{cases}$$

Theorems 4–5 exhibit the excellent asymptotic properties of the MLE. We can construct a confidence interval for $\hat{\theta}$ with level α,

$$\mathrm{CI} = [\hat{\theta} - \mu_{\alpha/2}\Delta, \hat{\theta} + \mu_{\alpha/2}\Delta], \tag{4.2}$$

where $\Delta = \left[T \sum_{k=1}^{K} \left\{\hat{p}_k(1-\hat{p}_k)/\hat{\theta} + \hat{p}_k(1-\hat{p}_k)^2/\{\hat{p}_k + (1-\hat{p}_k)(1-\hat{\theta})\}\right\}\right]^{-1/2}$, and $\mu_{\alpha/2}$ is the $\alpha/2$ upper-quantile of the standard normal distribution. Corollary 3 implies $\lim_{T\to\infty} P(\theta^* \in \mathrm{CI}) = 1-\alpha$ if the null hypothesis is valid.

For the case $\theta^* = 1$, there is no homophily in the evolving network. Whether an evolving network has homophily is a significant issue for researchers. Therefore, we construct a hypothesis test as follows:

$$H_0 : \theta^* = 1 \longleftrightarrow H_1 : \theta^* < 1. \tag{4.3}$$

The validity of the null hypothesis indicates that there is no homophily in the network. Note that if $\theta^* = 1$, then $\sum_{k=1}^{K} [p_k^*(1-p_k^*)/\theta^* + p_k^*(1-p_k^*)^2/\{p_k^* + (1-p_k^*)(1-\theta^*)\}] = K-1$. Construct the rejection region:

$$\begin{aligned} &\mathcal{D} = \{\theta : \theta \in (0, \theta_\alpha)\}, \\ &\theta_\alpha = 1 - \mu_\alpha [T(K-1)]^{-1/2}. \end{aligned} \tag{4.4}$$

Theorem 5 implies that $\lim_{T\to\infty} P(\hat{\theta} \in \mathcal{D}) = \alpha$ under the null hypothesis $H_0 : \theta^* = 1$. Thus, when we have $\hat{\theta}$ by the MLE, the criterion for homophily existing is as follows:

$$\begin{cases} \text{if } \hat{\theta} \geq \theta_\alpha, & \text{there is no homophily effect in the evolving network;} \\ \text{if } \hat{\theta} < \theta_\alpha, & \text{the evolving network has a homophily structure.} \end{cases} \tag{4.5}$$

Corollary 3 also allows us to test the hypothesis:

$$H_0 : \theta^* = \theta_0 \longleftrightarrow H_1 : \theta^* \neq \theta_0, \tag{4.6}$$

where $\theta_0 \in (0, 1)$, and

$$\begin{aligned} &\mathcal{D} = \{\theta : |\theta - \theta_0| > c_\alpha\}, \\ &c_\alpha = \left[T \sum_{k=1}^{K} \left\{ \frac{\hat{p}_k(1-\hat{p}_k)}{\theta_0} + \frac{\hat{p}_k(1-\hat{p}_k)^2}{\hat{p}_k + (1-\hat{p}_k)(1-\theta_0)} \right\} \right]^{-1/2} . \end{aligned} \tag{4.7}$$

Theorem 5 implies that $\lim_{T\to\infty} P(\hat{\theta} \in \mathcal{D}) = \alpha$ under the null hypothesis $H_0 : \theta^* = \theta_0$.

4.2 Snapshot

A snapshot of an evolving network refers to the present state of the network without historical information. In our work, the snapshot is the graph $G(T)$ of an evolving process $\{G(t)\}_{t=0}^{T}$.

We propose a parameter estimation procedure based on $G(T)$.

$$\begin{aligned} &\text{For } k \in [K],\ \tilde{p}_k = \frac{|\mathcal{V}_k(T)|}{|\mathcal{V}(T)|}; \quad \tilde{q} = \frac{|\mathcal{V}(T)|}{|\mathcal{E}(T)|}; \\ &\tilde{\theta} = \arg \max_{\theta \in (0, 1+\tilde{\epsilon})} L_T(\theta | G(T)), \end{aligned} \tag{4.8}$$

where $\mathcal{V}_k(T)$ is the set of nodes from group k in $G(T)$, and $\mathcal{E}_{k,1}(T) = \{(e_1, e_2) \in \mathcal{E}(T) : g_{e_1} = g_{e_2} = k\}$, $\mathcal{E}_{k,0}(T) = \{(e_1, e_2) \in \mathcal{E}(T) : g_{e_1} = k, g_{e_2} \neq k\}$. And $L_T(\theta|G(T))$ satisfies

$$\begin{aligned} & L_T(\theta|G(T)) \\ = \quad & \sum_{k=1}^{K} |\mathcal{E}_{k,1}(T)| \log \left[\frac{D_k(T)}{2|\mathcal{E}(T)|} \left\{ \frac{D_k(T)}{2|\mathcal{E}(T)|} + (1-\theta)(1 - \frac{D_k(T)}{2|\mathcal{E}(T)|}) \right\} \right] \\ & + \sum_{k=1}^{K} |\mathcal{E}_{k,0}(T)| \log \left[\frac{D_k(T)}{2|\mathcal{E}(T)|} \theta (1 - \frac{D_k(T)}{2|\mathcal{E}(T)|}) \right], \end{aligned} \tag{4.9}$$

where $\tilde{\epsilon} = \min_{k \in [K]} \left[\frac{D_k(T)}{2|\mathcal{E}(T)|} / (1 - \frac{D_k(T)}{2|\mathcal{E}(T)|}) \right]$.

Theorem 6. *Let the evolving network $\{G(t)\}_{t=0}^{T}$ be generated by a KPA model under Assumptions 1–3. Let $\psi^* = (\theta^*, \{p_k^*\}_{k=1}^{K-1}, q^*)^\top$ be the vector of true parameters of the KPA model. The parameter estimator $\tilde{\psi} = (\tilde{\theta}, \{\tilde{p}_k\}_{k=1}^{K-1}, \tilde{q})^\top$ based on $G(T)$ satisfies the following: when T tends to infinity,*

$$\tilde{\psi} \xrightarrow{a.s.} \psi^*.$$

Theorems 4–6 provide consistent estimations of the homophily parameter – θ^*. The absence of homophily can be viewed as the case $\theta^* = 1$.

In the real world, most evolving network data sets are in the form of a snapshot. Thus, Theorem 6 alone cannot infer whether an evolving network has homophily. However, when we have obtained $\tilde{\theta}$ from the snapshot

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0136

estimation method, we can use the following algorithm to test whether the evolving network has a homophily structure:

Algorithm 1 Homophily structure test on the snapshot.

```
Input: A snapshot graph $G(T)$; the number of randomized trials $R$; the
    statistical significance level $\alpha$.
Output: 1 or 0.
        "1": this evolving network has a homophily structure;
        "0": there is no homophily effect in this evolving network.
Estimate $\psi$ by the Equation (4.8) to get $\tilde{\psi} = (\tilde{\theta}, \{\tilde{p}_k\}_{k=1}^{K-1}, \tilde{q})^\top$ and $\mathcal{E}(T)$;
Assuming $\psi^* = (1, \{\tilde{p}_k\}_{k=1}^{K-1}, \tilde{q})^\top$, generate an evolving network with the
    time range $[0, |\mathcal{E}(T)|]$ by $G(0)$ with $n_0$ nodes satisfying Assumption 4
    and the KPA model in $R$ trials;
Estimate $\theta$ using the snapshot estimation method in the $r$th trial, and
    record the estimator as $\tilde{\theta}_r$, for $r \in [R]$;
Let $\theta_{\alpha,R}$ be the $\alpha$-quantile of $\{\tilde{\theta}_r\}_{r=1}^{R}$;
return $\mathbb{1}\{\tilde{\theta} < \theta_{\alpha,R}\}$;
```

The effectiveness of Algorithm 1 is demonstrated in Simulation 7.4.

5. Location of the change point

A change point $\tau^* \in [2, T]$ means the parameter θ^* of a KPA model changes at time τ^*, which implies that the influence of homophily on the network structure has changed. Assume that the network follows a KPA model with parameter θ_1^* in the time range $[0, \tau^*]$, and follows a KPA model with parameter θ_2^* in the time range $[\tau^* + 1, T]$, such that $\theta_2^* \neq \theta_1^*$.

Assumption 5. *$\tau^* \in [t_0, T - t_0]$, where $t_0/T \equiv c_0$, $c_0 \in (0, 1)$ is a constant.*

Assumption 6. *There is no difference in the parameters $(\{p_k^*\}_{k=1}^{K-1}, q^*)$ before and after the change point τ^*.*

Assumption 5 guarantees that we have enough information before and after the change to locate the point τ^*. Assumption 6 excludes the influence of other factors.

We estimate the change point τ^* by using the maximum likelihood method:

$$\begin{aligned}
\hat{\tau} &= \arg \max_{t_0 \leq \tau \leq T - t_0} [\max_{\theta_1 \in (0, 1+\epsilon_1)} \log L_2(\theta_1 | \{G(t)\}_{t=0}^{\tau}) + \max_{\theta_2 \in (0, 1+\epsilon_2)} \log L_2(\theta_2 | \{G(t)\}_{t=\tau}^{T}];\\
\hat{\theta}_1 &= \arg \max_{\theta_1 \in (0, 1+\epsilon_1)} \log L_2(\theta_1 | \{G(t)\}_{t=0}^{\hat{\tau}});\\
\hat{\theta}_2 &= \arg \max_{\theta_2 \in (0, 1+\epsilon_2)} \log L_2(\theta_2 | \{G(t)\}_{t=\hat{\tau}}^{T}), \qquad (5.1)
\end{aligned}$$

where $\epsilon_1 = \min_{k \in [K]}[\min_t \{P_k(t)/(1 - P_k(t)) : t \in [1, \tau], g_{e_1(t)} = g_{e_2(t)} = k\}]$, and $\epsilon_2 = \min_{k \in [K]}[\min_t \{P_k(t)/(1 - P_k(t)) : t \in [\tau + 1, T], g_{e_1(t)} = g_{e_2(t)} = k\}]$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0136

Here, $\log L_2(\theta|\{G(t)\})$ is defined by Equation (4.1).

Theorem 7. *Under Assumptions 1–6, we have when T tends to infinity,*

$$\frac{|\hat{\tau} - \tau^*|}{T} \xrightarrow{a.s.} 0.$$

6. Robustness of the estimation and the group label recovery

This section discusses the robustness of the proposed estimators from various perspectives. The KPA model's estimation accuracy is influenced by several factors: the incorrect or missing assignment of group labels to nodes; unobserved edge connections in the network generation process; and the instability of the distribution parameter q of the vertex-steps.

Based on our research and simulations, we conclude the following about the robustness of our estimators:

- The absence of group labels for nodes can be considered a special case of erroneous group labels, because a random label can be assigned to each node without a label. When $\theta^* = 1$, the CLT of $T^{1/2}(\hat{\theta} - 1)$ still works, even though the nodes are labeled as wrong groups with a probability away from one. When $\theta^* < 1$, the effect on the estimation is slight when nodes are assigned to incorrect groups with a small probability. However, if the nodes with erroneous group labels have high degrees, they can severely affect the parameter estimation, even in small and finite numbers. To address this issue, we propose a method for recovering group labels, following Hajek and Sankagiri (2019). Note the following:

- Our estimators still performance well, even if edges added to the network are unobserved with a certain probability.
- The convergence of our estimation method is proved both theoretically and experimentally, even if the parameter q has undergone a finite number of changes.

Therefore, our estimators are robust. Details of the methods, theoretical results, and proofs and simulations related to the robustness can be found in Sections S2–S3 of the Supplementary Material.

7. Simulations

This section verifies the theorems in Sections 3–5 by randomly generating an evolving network from the KPA model in B trials. We design the simulations as follows:

- The evolving network's time range is $[0, T]$, and the number of groups is K.

- The initial graph has n_0 isolated nodes with loops, and $n_0 \times p_k$ nodes are from group k, for $k \in [K]$.

- For each time t, a vertex-step occurs with probability q. In a vertex-step, the node from group k arrives with probability p_k.

- Record $\{v(t), e(t) = (e_1(t), e_2(t))\}$ and $\{D_k(t)\}_{k=1}^K$ at each time $t \in [1, T]$ in each trial.

Set

$$p(K) = \begin{cases} (0.5, 0.3, 0.2), & K = 3; \\ (0.4, 0.2, 0.2, 0.1, 0.1), & K = 5; \\ (0.2, 0.2, 0.1, 0.1, 0.1, \underbrace{0.06, 0.06, \cdots}_{5}), & K = 10. \end{cases} \tag{7.1}$$

7.1 Performance of $D_k(T)$

This subsection verifies Theorems 1–2. Set $B = 500$, $T = 10000$, $K \in \{3, 5, 10\}$, $n_0 = 100$, $\theta \in \{0.8, 0.5, 0.2\}$, $q \in \{0.9, 1/(1-\theta), 0.1\}$, and $p = p(K)$ in Equation (7.1).

Table 1 shows the convergence of $D_k(T)/2T$ and the effect of q on the convergence rate. “Bias” records the absolute sum of the bias from B trials: $\sum_{k=1}^K \left| \sum_{b=1}^B [(D_{k,b}(T)/2T) - p_k]/B \right|$. “MSE” records the sum of the mean squared error from B trials: $\sum_{k=1}^K \sum_{b=1}^B [(D_{k,b}(T)/2T) - p_k]^2/B$.

Table 1: Behavior of $D_k(T)/2T$ with $T = 10000$.

		Bias			MSE		
		$q = 0.9$	$q = 1/(2-\theta)$	$q = 0.1$	$q = 0.9$	$q = 1/(2-\theta)$	$q = 0.1$
$K = 3$	$\theta = 0.8$	$1.1564e-03$	$1.4024e-03$	$1.844e-03$	$1.5215e-04$	$1.8449e-04$	$4.3004e-03$
	$\theta = 0.5$	$3.5264e-04$	$4.6249e-04$	$3.7849e-03$	$1.0701e-04$	$2.2359e-04$	$4.4340e-03$
	$\theta = 0.2$	$1.3015e-04$	$9.8866e-04$	$1.9522e-04$	$8.6749e-05$	$2.9715e-04$	$5.6653e-03$
$K = 5$	$\theta = 0.8$	$1.4314e-03$	$1.6414e-03$	$7.5471e-03$	$1.9349e-04$	$2.3103e-04$	$5.5345e-03$
	$\theta = 0.5$	$9.2040e-04$	$1.7393e-03$	$7.9041e-03$	$1.2686e-04$	$2.8392e-04$	$5.4508e-03$
	$\theta = 0.2$	$6.3920e-04$	$1.6032e-03$	$3.4957e-03$	$1.0529e-04$	$3.7104e-04$	$6.2910e-03$
$K = 10$	$\theta = 0.8$	$1.7843e-03$	$2.7916e-03$	0.0102	$2.2084e-04$	$2.8741e-04$	$5.9235e-03$
	$\theta = 0.5$	$1.2434e-03$	$1.3576e-03$	$7.7375e-03$	$1.5407e-04$	$3.4378e-04$	$6.7235e-03$
	$\theta = 0.2$	$1.6965e-03$	$3.8501e-03$	$8.6197e-03$	$1.2460e-04$	$4.1959e-04$	$7.1033e-03$

7.2 Estimators of parameters with historical information

This subsection verifies Theorems 4–5 and Corollary 3. We test the convergence of $\hat{\theta}$ in B simulation trials. Set $B = 500$, $T = 10000$, $K \in \{3, 5, 10\}$, $n_0 = 100$, $\theta \in \{0.8, 0.5, 0.2\}$, $q \in \{0.9, 0.1\}$, and $p = p(K)$ in Equation (7.1).

$\{\{D_k(t)\}_{k=1}^K, v(t), g_{e_1(t)}, g_{e_2(t)}\}_{t=1}^T$ are used to construct the maximum likelihood equation and record the estimators of θ and Σ_{11} in the bth trial:- $\hat{\theta}_b$, $\hat{\Sigma}_{11,b}$, $b \in [B]$.

For $\hat{\theta}$, we have the following. "Bias" records the absolute sum of the bias from B trials: $\sum_{b=1}^B (\hat{\theta}_b - \theta)/B$. "MSE" records the sum of the mean square error from B trials: $\sum_{b=1}^B (\hat{\theta}_b - \theta)^2/B$. "Cover rate" records the percentage of B trials that θ fall in the confidence interval: $\sum_{b=1}^B \mathbb{1}\{\theta \in \text{CI(b)}\}/B$, where $\mathrm{CI(b)} = [\hat{\theta}_b - \mu_{\alpha/2}\Delta, \hat{\theta}_b + \mu_{\alpha/2}\Delta]$, $\alpha = 0.05$,

$$\Delta = \left[T \sum_{k=1}^{K} \{p_k(1-p_k)/\theta + p_k(1-p_k)^2/\{p_k + (1-p_k)(1-\theta)\}\}\right]^{-1/2}.$$

For $(\hat{\theta}-\theta)/\hat{\Sigma}_{11}^{1/2}$, "Bias" records $\sum_{b=1}^{B}(\hat{\theta}_b - \theta)/(\hat{\Sigma}_{11,b}^{1/2}B)$, "MSE" records $\sum_{b=1}^{B}(\hat{\theta}_b - \theta)^2/(\hat{\Sigma}_{11,b}B)$, and "Cover rate" records $\sum_{b=1}^{B} \mathbb{1}\{(\theta - \hat{\theta}_b)/\hat{\Sigma}_{11,b}^{1/2} \in \mathrm{CI}\}/B$, where $\mathrm{CI} = [-\mu_{\alpha/2}, \mu_{\alpha/2}]$, $\alpha = 0.05$.

The results of $\hat{\theta}_b$ and $(\hat{\theta}_b - \theta)/\hat{\Sigma}_{11,b}^{1/2}$ are recorded in Table 2.

Table 2: Performance of $\hat{\theta}$ and $(\hat{\theta}-\theta)/\hat{\Sigma}_{11}^{1/2}$.

$\hat{\theta}$		$q = 0.9$			$q = 0.1$		
		Bias	MSE	Cover rate	Bias	MSE	Cover rate
$K = 3$	$\theta = 0.8$	$2.9744e-05$	$5.6869e-05$	0.956	$3.8720e-04$	$5.9814e-05$	0.958
	$\theta = 0.5$	$-2.9054-04$	$5.2692e-05$	0.95	$-3.0766e-05$	$5.6951e-05$	0.94
	$\theta = 0.2$	$2.3003e-05$	$3.0974e-05$	0.952	$1.5615e-04$	$2.9925e-05$	0.946
$K = 5$	$\theta = 0.8$	$-1.6917e-04$	$3.9761e-05$	0.96	$2.7934e-04$	$3.6486e-05$	0.962
	$\theta = 0.5$	$-8.6336e-05$	$4.0488e-05$	0.95	$1.5401e-04$	$3.9474e-05$	0.96
	$\theta = 0.2$	$-1.6008e-04$	$2.2379e-05$	0.956	$-2.0367e-04$	$2.1806e-05$	0.96
$K = 10$	$\theta = 0.8$	$1.8203e-05$	$2.6514e-05$	0.954	$6.8822e-05$	$2.5242e-05$	0.956
	$\theta = 0.5$	$-1.6593e-04$	$3.3950e-05$	0.96	$2.3263e-05$	$3.5618e-05$	0.944
	$\theta = 0.2$	$-1.2789e-04$	$1.8692e-05$	0.958	$-2.3689e-04$	$2.0406e-05$	0.942
$(\hat{\theta}-\theta)/\hat{\Sigma}_{11}^{1/2}$		Bias	MSE	Cover rate	Bias	MSE	Cover rate
$K = 3$	$\theta = 0.8$	-0.0306	1.0564	0.946	0.0366	1.0159	0.944
	$\theta = 0.5$	0.0274	0.9302	0.96	-0.0248	0.9675	0.956
	$\theta = 0.2$	0.0237	0.9094	0.964	-0.0457	0.9996	0.958
$K = 5$	$\theta = 0.8$	-0.0478	1.0184	0.948	-0.0236	0.9684	0.962
	$\theta = 0.5$	$-9.3054e-04$	1.0922	0.94	0.0407	0.9865	0.952
	$\theta = 0.2$	-0.0229	0.9492	0.96	0.0563	0.9652	0.958
$K = 10$	$\theta = 0.8$	0.0262	0.9633	0.966	-0.0834	0.9840	0.954
	$\theta = 0.5$	-0.0724	0.9287	0.958	-0.0288	1.0255	0.942
	$\theta = 0.2$	-0.0923	1.0773	0.95	0.0378	0.9487	0.958

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0136

7.3 Snapshot

This subsection verifies Theorem 6, the convergence of the snapshot estimation based on graph $G(T)$. Set $B = 500$, $T = 10000$, $K \in \{3, 5, 10\}$, $n_0 = 100$, $q = 0.5$, and $p = p(K)$ in Equation (7.1).

$\{\{D_k(t)\}_{k=1}^K, v(t), g_{e_1(t)}, g_{e_2(t)}\}_{t=1}^T$ are used to construct the maximum likelihood equation (Equation (4.1)) and record the estimator of θ as $\hat{\theta}_{\text{mle}}(b)$ in the bth trial, $b \in [B]$. $\{D_k(T), \mathcal{E}_{k,1}(T), \mathcal{E}_{k,0}(T)\}_{k=1}^K$ are used to construct the snapshot estimation (Equation (4.9)) and record the estimator of θ as $\hat{\theta}_{\text{snap}}(b)$ in the bth trial, for $b \in [B]$.

We calculate the mean absolute error (MAE) and the mean squared error (MSE) for $\hat{\theta}_{\text{mle}} = \sum_{b=1}^B \hat{\theta}_{\text{mle}}(b)/B$ and $\hat{\theta}_{\text{snap}} = \sum_{b=1}^B \hat{\theta}_{\text{snap}}(b)/B$, for $b \in [B]$.

Table 3 compares the results of the two methods.

Table 3: Comparison of the MLE and snapshot estimations.

		Bias		MAE		MSE	
		$\hat{\theta}_{\text{mle}} - \theta$	$\hat{\theta}_{\text{snap}} - \theta$	$\sum_{b=1}^{B} \lvert\hat{\theta}_{\text{snap}}(b) - \theta\rvert/B$	$\sum_{b=1}^{B} \lvert\hat{\theta}_{\text{mle}}(b) - \theta\rvert/B$	$\sum_{b=1}^{B} (\hat{\theta}_{\text{snap}}(b) - \theta)^2/B$	$\sum_{b=1}^{B} (\hat{\theta}_{\text{mle}}(b) - \theta)^2/B$
$K = 3$	$\theta = 0.8$	$-1.2620e - 04$	$-3.9170e - 05$	$6.1264e - 03$	$6.1511e - 03$	$5.9312e - 05$	$5.9918e - 05$
	$\theta = 0.5$	$-4.4441e - 04$	$-5.7596e - 04$	$6.0634e - 03$	$6.1599e - 03$	$5.7998e - 05$	$5.7998e - 05$
	$\theta = 0.2$	$-4.6652e - 04$	$-4.7906e - 04$	$4.1808e - 03$	$4.2455 - e03$	$2.8312e - 05$	$2.8898e - 05$
$K = 5$	$\theta = 0.8$	$-1.2516e - 04$	$-1.0208e - 04$	$4.9228e - 03$	$4.9908e - 03$	$3.8822e - 05$	$3.9828e - 05$
	$\theta = 0.5$	$-8.7880e - 05$	$-1.6979e - 04$	$5.0875e - 03$	$5.1120e - 03$	$4.0902e - 05$	$4.1537e - 05$
	$\theta = 0.2$	$2.4947e - 04$	$2.4115e - 4$	$3.6752e - 03$	$3.6703e - 03$	$2.1145e - 05$	$2.1027e - 05$
$K = 10$	$\theta = 0.8$	$-3.8862e - 05$	$-4.8347e - 05$	$4.185e - 03$	$4.1786e - 03$	$2.6769e - 05$	$2.6822e - 05$
	$\theta = 0.5$	$6.6961e - 05$	$5.3035e - 05$	$4.3469e - 03$	$4.3455e - 03$	$2.9281e - 05$	$2.9428e - 05$
	$\theta = 0.2$	$-1.4646e - 04$	$-1.8334e - 04$	$3.7049e - 03$	$3.7002e - 03$	$2.1952e - 05$	$2.1979e - 05$

7.4 Homophily structure test on the snapshot

This subsection verifies Algorithm 1. Set $B = 500$, $n_0 = 10$, $T \in \{200, 500, 1000\}$, $K \in \{3, 5, 10\}$, $\theta \in \{1, 0.9, 0.95\}$, $q \in \{0.9, 0.5\}$, and $p = p(K)$ in Equation (7.1) if $K \in \{3, 5\}$, and $p = (\underbrace{0.1, 0.1, \cdots}_{10})$ if $K = 10$.

$\{D_k(T), \mathcal{E}_{k,0}(T), \mathcal{E}_{k,1}(T)\}_{k=1}^{K}$ are used to construct the snapshot estimation and record the estimator of $(\theta, \{p_k\}_{k=1}^{K-1}, q)$ in the bth trial: $(\hat{\theta}_b, \{\hat{p}_{k,b}\}_{k=1}^{K-1}, \hat{q}_b)$, $b \in [B]$.

Using Algorithm 1, we obtain $\theta_{\alpha,R,b}$ based on the parameters $(1, \{\hat{p}_{k,b}\}_{k=1}^{K-1}, \hat{q}_b)$ in the bth trial, where $\alpha = 0.05$ and $R = 500$. Table 4 records the percentage of B trials that accept the null hypothesis — $\sum_{b=1}^{B} \mathbb{1}\{\hat{\theta}_b \geq \theta_{\alpha,R,b}\}/B$ —

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0136

when $\theta = 1$, and the percentage of B trials that reject the null hypothesis — $\sum_{b=1}^{B} \mathbb{1}\{\hat{\theta}_b < \theta_{\alpha,R,b}\}/B$ — when $\theta < 1$.

Table 4: The effect of Algorithm 1.

		$\theta = 1$		$\theta = 0.95$		$\theta = 0.9$	
		$\sum_{b=1}^{B} \mathbb{1}\{\hat{\theta}_b \geq \theta_{\alpha,R,b}\}/B$		$\sum_{b=1}^{B} \mathbb{1}\{\hat{\theta}_b < \theta_{\alpha,R,b}\}/B$		$\sum_{b=1}^{B} \mathbb{1}\{\hat{\theta}_b < \theta_{\alpha,R,b}\}/B$	
		$q = 0.9$	$q = 0.5$	$q = 0.9$	$q = 0.5$	$q = 0.9$	$q = 0.5$
$K = 3$	$T = 200$	0.946	0.956	0.256	0.286	0.622	0.616
	$T = 500$	0.956	0.942	0.452	0.506	0.922	0.932
	$T = 1000$	0.958	0.94	0.722	0.736	1	0.996
$K = 5$	$T = 200$	0.956	0.954	0.428	0.396	0.872	0.858
	$T = 500$	0.938	0.942	0.724	0.736	0.99	0.994
	$T = 1000$	0.944	0.95	0.894	0.926	1	1
$K = 10$	$T = 200$	0.946	0.948	0.712	0.652	0.986	0.976
	$T = 500$	0.95	0.94	0.93	0.922	1	1
	$T = 1000$	0.948	0.948	0.998	1	1	1

7.5 Change point

This subsection verifies Theorem 7, the method of locating the change point τ. Set $B = 500$, $T \in \{10000, 20000\}$, $K = 10$, $n_0 = 100$, $q = 0.5$, and $p = (0.2, 0.2, 0.1, 0.1, 0.1, \underbrace{0.06, 0.06, \cdots}_{5})$.

Set $c_0 = t_0/T = 0.1$, $\tau/T \in \{0.25, 0.5, 0.75\}$. The homophily parameter is θ_1 before the change point, and θ_2 after the change point, with $(\theta_1, \theta_2) \in \{(0.4, 0.6), (0.1, 0.9)\}$.

$\{\{D_k(t)\}_{k=1}^{K}, v(t), g_{e_1(t)}, g_{e_2(t)}\}_{t=1}^{T}$ are used to construct the maximum likelihood equation (Equation (4.1)), and to obtain the estimator $\hat{\tau}(b)$, $\hat{\theta}_1(b)$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0136

and $\hat{\theta}_2(b)$ from Equation (5.1) in the bth trial, $b \in [B]$. Table 5 shows the results for $\hat{\tau} = \sum_{b=1}^{B} \hat{\tau}(b)/B$ and $\hat{\theta}_1 = \sum_{b=1}^{B} \hat{\theta}_1(b)/B$, $\hat{\theta}_2 = \sum_{b=1}^{B} \hat{\theta}_2(b)/B$.

Table 5: Accuracy of change point location.

		Estimator				MSE		
$T = 1000$		$\hat{\tau}$	$(\hat{\tau}-\tau)/T$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\sum_{b=1}^{B}[(\hat{\tau}(b)-\tau)/T]^2/B$	$\sum_{b=1}^{B}(\hat{\theta}_1(b)-\theta_1)^2/B$	$\sum_{b=1}^{B}(\hat{\theta}_2(b)-\theta_2)^2/B$
$\tau = 250$	$(\theta_1,\theta_2) = (0.1, 0.9)$	250.092	$9.2e-05$	0.0989	0.9008	$2.132e-06$	$3.7703e-04$	$2.6951e-04$
	$(\theta_1,\theta_2) = (0.4, 0.6)$	255.1	0.0051	0.3928	0.6051	$2.0025e-03$	$1.3744e-03$	$5.0610e-04$
$\tau = 500$	$(\theta_1,\theta_2) = (0.1, 0.9)$	500.22	$2.2e-04$	0.1008	0.9028	$2.548e-06$	$1.8752e-04$	$4.6358e-04$
	$(\theta_1,\theta_2) = (0.4, 0.6)$	500.4	$4e-04$	0.3942	0.6056	$2.5436e-03$	$6.036e-04$	$8.1849e-04$
$\tau = 750$	$(\theta_1,\theta_2) = (0.1, 0.9)$	750.264	$2.64e-04$	0.0997	0.9002	$2.328e-06$	$1.4643e-04$	$7.7421e-04$
	$(\theta_1,\theta_2) = (0.4, 0.6)$	749.58	$-4.2e-04$	0.3967	0.6098	$2.5096e-03$	$3.897e-04$	$1.7278e-03$
$T = 2000$		$\hat{\tau}$	$(\hat{\tau}-\tau)/T$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\sum_{b=1}^{B}[(\hat{\tau}(b)-\tau)/T]^2/B$	$\sum_{b=1}^{B}(\hat{\theta}_1(b)-\theta_1)^2/B$	$\sum_{b=1}^{B}(\hat{\theta}_2(b)-\theta_2)^2/B$
$\tau = 500$	$(\theta_1,\theta_2) = (0.1, 0.9)$	500.028	$1.4e-05$	0.0996	0.9001	$6.53e-07$	$1.8623e-04$	$1.3056e-04$
	$(\theta_1,\theta_2) = (0.4, 0.6)$	494.156	$-2.922e-03$	0.3931	0.5993	$5.5524e-04$	$7.4999e-04$	$2.2043e-04$
$\tau = 1000$	$(\theta_1,\theta_2) = (0.1, 0.9)$	999.996	$-2e-06$	0.0996	0.9016	$9.19e-07$	$1.0328e-04$	$2.1915e-04$
	$(\theta_1,\theta_2) = (0.4, 0.6)$	998.892	$-5.54e-04$	0.3961	0.6007	$4.6902e-04$	$3.5954e-04$	$3.5254e-04$
$\tau = 1500$	$(\theta_1,\theta_2) = (0.1, 0.9)$	1500.04	$2e-05$	0.0993	0.9026	$4.6e-07$	$7.2221e-05$	$4.3321e-04$
	$(\theta_1,\theta_2) = (0.4, 0.6)$	1499.3	$-3.5e-04$	0.3980	0.6042	$3.8961e-04$	$1.8925e-04$	$8.2859e-04$

8. Data application

We selected two real network data sets to test our KPA model and the estimation methods. Both have group labels for nodes, and one has timestamp information about edges. The data sets are as follows:

- CL-10K-1d8-L5 is a network data set with group information, but without a timestamp from the Network Repository (Rossi and Ahmed (2015)) available at

  https://networkrepository.com/CL-10K-1d8-L5.php.

- Soc-political-retweet is a network data set with group information and timestamp information from the Network Repository (Rossi and Ahmed (2015)) available at

  https://networkrepository.com/soc-political-retweet.php.

The basic information about these network data sets is in Table S10 of the Supplementary Material.

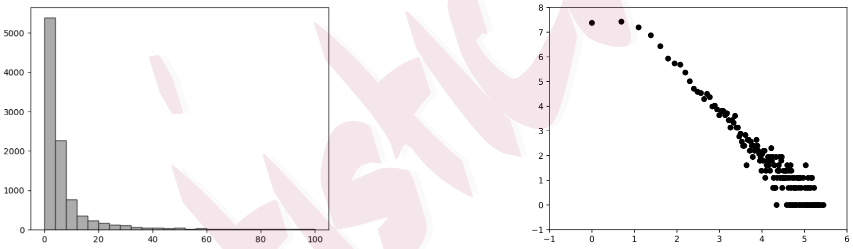
The $\hat{\theta}_{\text{snap}}$ of CL-10K-1d8-L5 by the snapshot estimation is 0.9999, and the $\hat{\theta}_{\text{mle}}$ of soc-political-retweet is 0.0413.

For the snapshot estimation of CL-10K-1d8-L5, the estimated parameters are $(\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5, \hat{q}) = (0.2, 0.2, 0.2, 0.2, 0.2, 0.2227)$. Based on $(\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5, \hat{q})$ and $\theta^* = 1$, with the time range $[0, 44896]$, we can test the homophily by using Algorithm 1. Letting the statistical significance level be $\alpha = 0.05$ and the number of randomized trials be $R = 500$, we have $\theta_{\alpha,R} = 0.9963$ and $\mathbb{1}\{0.9999 < 0.9963\} = 0$. These results imply there is no homophily in CL-10K-1d8-L5.

For the MLE of soc-political-retweet, the estimated parameters are $(\hat{p}_1, \hat{p}_2, \hat{q}) = (0.3852, 0.6148, 0.3020)$, $T = 61157$, and $K = 2$. By Equation (4.4), we get $\theta_\alpha = 0.9933$, where $\alpha = 0.05$ and $\hat{\theta} < \theta_\alpha$, which means the evolving network has a homophily structure.

Tables S1–S2 in the Supplementary Material show that our estimations are robust when the nodes are mislabeled with a small probability. Furthermore, consider the mislabeling that arose with probability, as specified in Assumption S.1. Theorem S.1 implies that the rejection region is the same, regardless of any mislabeling. Thus, Equation (4.5) is still valid. We can still infer that soc-political-retweet has a homophily structure, because it rejects the null hypothesis.

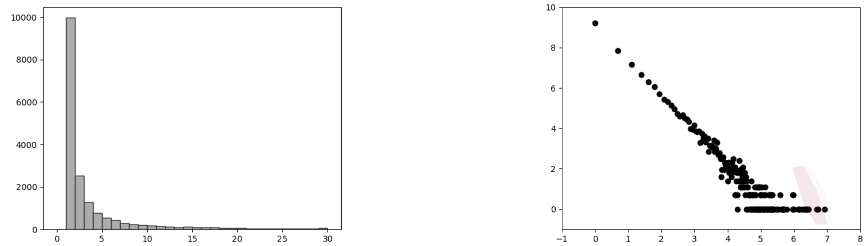
Rich-get-richer is another essential mechanism of the KPA model. Figures 4–5 show the power-law degree distribution of the data sets. Each group's power-law degree distribution is provided in the Supplementary Material.



(a) Histogram of degree distribution (b) log-log scale of degree distribution

Figure 4: Degree distribution of CL-10K-1d8-L5.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0136



(a) Histogram of degree distribution (b) log-log scale of degree distribution

Figure 5: Degree distribution of soc-political-retweet.

Supplementary Materials

In the supplementary material, we discuss the applicability of the KPA model. We also give proofs of the theoretical results presented in Sections 3–6 and the basic information of datasets in Section 8. More simulations for the estimations' robustness and the group label recovery in Section 6 are also listed.

Acknowledgments

The research of Bo Zhang was partially supported by the National Natural Science Foundation of China No.12001517 & 72091212, the Double First-Class Initiative Research Fund YD2040002005, and the Fundamental Research Funds for the Central Universities WK2040000026 & WK2040000027.

The research of Xiao Han was supported by the National Natural Science Foundation of China No.12001518.

References

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

Avin, C., H. Daltrophe, B. Keller, Z. Lotker, C. Mathieu, D. Peleg, et al. (2020). Mixed preferential attachment model: Homophily and minorities in social networks. *Physica A: Statistical Mechanics and its Applications 555*, 124723.

Barabási, A.-L. and R. Albert (1999). Emergence of scaling in random networks. *science 286*(5439), 509–512.

Chung, F. and L. Lu (2006). *Complex graphs and networks.* Number 107. American Mathematical Soc.

Durrett, R. (2007). *Random graph dynamics*, Volume 200. Cambridge university press.

Easley, D., J. Kleinberg, et al. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*, Volume 1. Cambridge university press.

Hajek, B. and S. Sankagiri (2019). Community recovery in a preferential attachment graph. *IEEE Transactions on Information Theory 65*(11), 6853–6874.

Jackson, M. O. et al. (2008). *Social and economic networks*, Volume 3. Princeton university press Princeton.

Jackson, M. O. and D. López-Pintado (2013). Diffusion and contagion in networks with heterogeneous agents and homophily. *Network Science 1*(1), 49–67.

Jannach, D., M. Zanker, A. Felfernig, and G. Friedrich (2010). *Recommender systems: an introduction*. Cambridge University Press.

Lazarsfeld, P. F., R. K. Merton, et al. (1954). Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society 18*(1), 18–66.

Lee, E., F. Karimi, C. Wagner, H.-H. Jo, M. Strohmaier, and M. Galesic (2019). Homophily and minority-group size explain perception biases in social networks. *Nature human behaviour 3*(10), 1078–1087.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology 27*(1), 415–444.

Newman, M. (2018). *Networks*. Oxford university press.

Ricci, F., L. Rokach, and B. Shapira (2015). Recommender systems: introduction and challenges. In *Recommender systems handbook*, pp. 1–34. Springer.

Rossi, R. and N. Ahmed (2015, Mar.). The network data repository with interactive graph analytics and visualization. *Proceedings of the AAAI Conference on Artificial Intelligence 29*(1).

Van Der Hofstad, R. (2024). *Random graphs and complex networks*. Cambridge university press.

Hanyang Tian, E-mail: thy@mail.ustc.edu.cn

Department of Statistics & Finance, School of Management,

University of Science and Technology of China.

Bo Zhang: Corresponding author, E-mail: wbchpmp@ustc.edu.cn

Department of Statistics & Finance, International Institute of Finance, School of Management,

University of Science and Technology of China.

Ruixue Jiang: Co-corresponding author, E-mail: ruixue1@ustc.edu.cn

International Institute of Finance, School of Management, University of Science and Technology of China.

Xiao Hanm E-mail: xhan011@ustc.edu.cn

Department of Statistics & Finance, International Institute of Finance, School of Management,

School of Management, University of Science and Technology of China.