Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: Interdisciplinary statistical mechanics

On the analysis of fitness change: fitness-popularity dynamic network model with varying fitness

Hohyun Jung[1], Jae-Gil Lee[2] and Sung-Ho Kim[1]

[1] Department of Mathematical Sciences, KAIST, Daejeon, Republic of Korea
[2] Graduate School of Knowledge Service Engineering, KAIST, Daejeon, Republic of Korea
E-mail: hhjung@kaist.ac.kr, jaegil@kaist.ac.kr and sung-ho.kim@kaist.edu

Received 16 April 2019
Accepted for publication 29 January 2020
Published 16 April 2020

Online at stacks.iop.org/JSTAT/2020/043407
https://doi.org/10.1088/1742-5468/ab7754

Abstract. There are various dynamic networks around us. Many researchers have investigated the fitness, i.e. ability to get edges from other nodes, and popularity effects on network growth. The fitness-popularity dynamic network (FPDN) model was introduced recently. In the FPDN model, the fitness of a node is assumed invariant for a given period of time. In many real networks, however, the fitness may change over time in various ways. Herein, we propose a varying fitness-popularity dynamic network (V-FPDN) model by allowing variable fitness. Through the V-FPDN model, we can estimate the strength of fitness and popularity effects and show how the fitness of the nodes changes. The magnitude of these effects and fitness values are estimated simultaneously using the expectation-maximization (EM) algorithm combined with the Markov chain Monte Carlo (MCMC) method. We apply the FPDN and V-FPDN model to the Facebook wallpost network and compare the results. The YouTube subscription network is investigated using the V-FPDN model in various categories. We explain the superiority of the proposed model with remarkable interpretations.

Keywords: network dynamics, stochastic processes

1742-5468/20/043407+21$33.00

Contents

1. Introduction **2**
1.1. Background and motivation 2
1.2. Key contributions and outline 3

2. Related works **4**

3. Fitness-popularity dynamic network model with varying fitness **5**
3.1. Model 5
3.2. Popularity measures 6

4. Estimation **7**
4.1. Likelihood functions 7
4.2. Algorithm 8
4.3. Inference on the model parameter and fitness 12

5. Synthetic data analysis **14**

6. Real data analysis **15**
6.1. Comparison with existing model 15
6.2. YouTube subscription network 17

7. Concluding remarks **20**
Acknowledgment **20**

References **20**

1. Introduction

1.1. Background and motivation

A network is a way of representing relationships among members. There are a variety of networks around us, including social networks, citation networks, biological networks, and semantic networks. These have been actively studied in various fields. Many researchers have investigated snapshots of network data and found important characteristics, such as community structures [1–4] and influential nodes [5–7]. However, many real networks do not stay in a fixed state. They usually change over time in a complicated way. It is essential to understand how this change occurs, and in fact there are many studies on the mechanisms supporting network growth [8–15].

Among the growth mechanisms, the rich-get-richer and fit-get-richer phenomena are commonplace. In the *rich-get-richer* phenomenon, popular nodes become more popular. This phenomenon has been observed in many real networks [16–20]. In the YouTube subscription network, popular videos get more exposure to users, and the popular channels are likely to receive more subscribers and become more popular. Conversely, if a popular node becomes stagnant, it is called a rich-get-poorer phenomenon. By contrast,

https://doi.org/10.1088/1742-5468/ab7754

J. Stat. Mech. (2019) 043407

in the *fit-get-richer* phenomenon, capable nodes become more popular. The ability to sing and dance, as well as the ability to edit videos, will have a major impact on the improvement in the channel's popularity. We can explain it through the concept of fitness. In the complex network theory, *fitness* is defined as the intrinsic ability to establish connections with other nodes.

Researches abound in literature on the popularity effect in the fields of statistical physics and social networks. Barabasi and Albert (1999) [21] proposed the BA model that explains the popularity effect using the preferential attachment rule. The degree of the node is considered as a popularity measure, and a node with a higher degree is assumed to be more likely to be connected to a new node. The following process supports the network growth model.

- Growth: At each time point, a node enters the network. Then the node tries to connect with m_{BA} nodes in the network.
- Preferential attachment: The newly entering node connects with node i with a probability that is proportional to the popularity of node i.

This mechanism explains the existence of hub nodes, which have high connectivity in real networks. Many variant models followed the BA model [22–36].

Among them, Bianconi and Barabasi [37] proposed the fitness model, i.e. the BB model, leveraging both fitness and popularity effects on network growth. They assign an inherent fitness value to each node and assume that the connection probability is related to fitness as well as popularity. Many associated models have been developed where the growth mechanism is affected by the fitness and popularity of nodes [15, 38–41]. However, these models are not suitable for comparing the strength of fitness and popularity effects. Besides, they are essentially based on the growth mechanism of the BA model whereby an edge is only generated when a new node enters the system. To address these problems, Jung *et al* [42] proposed the fitness-popularity dynamic network (FPDN) model that can estimate the two effects on an equal footing in terms of magnitude and allows flexibility on node deletion and addition, edge formations among the existing nodes, and so on.

In the FPDN model, popularity changes over time, while fitness remains constant. In many real networks however, it is appropriate to assume that fitness changes over time. For example, the singing, dancing, video editing skills, and activeness of channel operators vary over time in YouTube. In Facebook, fitness is linked to the extroversion and sociability, and it is not proper to be assumed as constant. The FPDN model cannot capture the change in node fitness and applying this model to these types of networks can yield inaccurate estimates of fitness and strengths of the two effects. As a remedy for this weakness, the varying fitness model is proposed as a possible solution.

1.2. Key contributions and outline

In this paper, we propose a varying fitness-popularity dynamic network (V-FPDN) model that allows fitness change and enables a more flexible inference on fitness and popularity effects. The estimation algorithm presented can estimate fitness change and the strength of the two effects simultaneously using the EM (expectation maximization)

J. Stat. Mech. (2020) 043407

algorithm combined with the Markov chain Monte Carlo (MCMC) method. We demonstrate the validity of our model by carrying out experiments on synthetic networks in which the number of nodes increases over time. The Facebook data is analyzed by the proposed model, whose performance is compared favorably with the FPDN model. Moreover, we observe the prevalence of the rich-get-richer phenomenon in YouTube, and the fitness change in channels is investigated with noteworthy interpretations.

The remainder of this paper is organized as follows. We provide a brief review of the preceding works in section 2. In section 3, we propose the V-FPDN model and explain popularity measures. The estimation procedure for model parameters and varying fitness values are described in section 4. In section 5, we apply the V-FPDN model to a synthetic network and investigate the validity of the estimation procedure. In section 6, we compare the proposed model with the FPDN model using the Facebook wallpost network data. Moreover, we analyze the strength of the fitness and popularity effects on the YouTube subscription network with several case studies. Finally, we close the paper with some concluding remarks in section 7.

2. Related works

Many network scientists have worked on the popularity effect through preferential attachment. Barabasi and Albert [21] proposed the BA model and explained the scale-free nature of the network. Preferential attachment is the notion that the probability of the connection between a new node and an existing node i at time t is proportional to the degree $k_i(t)$ of node i. They explained the existence of hub nodes in real networks and derived a power-law degree distribution.

Many successive models based on the growth mechanism of the BA model have been developed. There were several works using *attachment exponents* α where a node i is connected with a new node with a probability proportional to $k_i(t)^\alpha$ rather than directly proportional to the degree [43, 44]. Many generalized versions of the BA model have been proposed in the presence of degree correlation [26], accelerating growth [27–29], aging of nodes [30, 31], internal link formation [24, 25], and node deletion [32–36].

The fitness effect has been investigated in the context of network growth. The node fitness can affect various properties such as the degree distribution of the network [45]. The BB model [37] assumes that the connection probability of node i is proportional to the product of fitness and degree. Pham *et al* developed the model PAFit [40] assuming that the connection probability is proportional to the product of the degree-related and fitness-related terms, similarly to the BB model. They estimate both preferential attachment and fitness effects simultaneously with no assumption on the specific forms of fitness and popularity. Wang *et al* [15] analyzed the paper citation dynamics by adding a time-dependent aging term. They employ fitness and preferential attachment together with the assumption of a decreasing number of citations over time. Similarly, the aging phenomenon on fitness has been investigated [38, 39]. Ghoshal *et al* [41] considered combining edge and node deletion with fitness and popularity. Recently, Jung *et al* [42] proposed the FPDN model to compare fitness and popularity on an equal footing.

J. Stat. Mech. (2019) 043407

Many dynamic network models have been developed to explain the change of the network topology. They attempt to explain the characteristics of social networks by taking into account various features of nodes and networks. These dynamic network models can be classified into non-latent variable models and latent variable models.

Concerning the non-latent variable models, the temporal exponential random graph model (TERGM) [46] is a very general model for the analysis of the network and node-related features on the growth of a network. An actor-oriented model [47] was developed to measure the influence of factors, such as activity, reciprocity, and transitivity on network growth. Recently, the temporal extension of stochastic block model (SBM) [48] was developed to consider the community structure on network dynamics.

However, the non-latent variable models are not suitable for explaining hidden features on network dynamics, such as the fitness of nodes. It is often useful to investigate network changes using static or time-varying latent features, and a review on the latent variable models can be found in [49]. Sarkar and Moore [50] proposed the dynamic social network in latent space (DSNL) model, which is the generalization of [51]. The hidden Markov [52] and the mixed effect models [53] are statistical models involving latent variables, and several related studies have been conducted on network dynamics [54, 55]. Mazzarisi *et al* [56] recently proposed the model that concerns the link persistence and node-specific latent variables. Finally, several generalizations of non-latent variable models, TERGM and temporal SBM, were introduced in [57, 58].

3. Fitness-popularity dynamic network model with varying fitness

3.1. Model

Let us assume that we have time series network data $G^0, G^1, \cdots, G^T$. The directed graph $G^t = (V^t, E^t)$, $t = 0, 1, \cdots, T$ consists of a node set V^t and an edge set E^t. Let $V = \cup_{t=0}^{T} V^t$ and express it as $V = \{1, \cdots, N\}$, unless any confusion arises. Let A^t be the adjacency matrix of G^t. The in-degree and out-degree of node i in G^t are expressed as $D^t_{\text{in},i} = \sum_j A^t_{ji}$ and $D^t_{\text{out},i} = \sum_j A^t_{ij}$. In addition, the fitness and popularity of node i at time t are expressed as f^t_i and u^t_i, respectively.

Let θ^t_{ij} be the connection probability from node i to node j. It is defined in the V-FPDN model as

$$\theta^t_{ij} = g\left(\beta_0 + \beta_1 f^{t-1}_j + \beta_2 u^{t-1}_j\right), \qquad t = 1, 2, \cdots, T, \tag{1}$$

where $g(x) = 1/(1 + e^{-x})$. As mentioned before, the fitness of node j, f^t_j, changes over time.

Specifically, if node j enters and leaves the system at time $t_{0,j}$ and $t_{1,j}$, i.e. $j \in V^t$ if and only if $t = t_{0,j}, t_{0,j} + 1, \cdots, t_{1,j}$, then the distribution of fitness f^t_j is assumed as follows:

$$\begin{aligned} f^{t_{0,j}}_j &\sim N(0, 1), \\ f^t_j &= f^{t-1}_j + \epsilon^t_j, \quad \epsilon^t_j \sim N(0, \sigma^2_j), \quad t = t_{0,j} + 1, \cdots, t_{1,j} - 1. \end{aligned} \tag{2}$$

J. Stat. Mech. (2020) 043407

Throughout this paper, we assume that the fitness variability σ_j^2 of node j has an exponential distribution with rate $\lambda = 100$. When node enters the system, it may link with existing nodes. Note that $t_{1,j} = T$ if node j does not leave the system.

As for the edge, we assume

$$A_{ij}^t \sim \text{Bernoulli}(\theta_{ij}^t).$$

We assign 0 to θ_{ij}^t for the pair of nodes, i and j, for which connection is impossible.

The β_0 is an *intercept* parameter related to the average number of edges created over time. The parameters β_1 and β_2 represent the magnitude of the fitness and popularity effects, respectively. A large β_1 implies that the fitness of a node has great influence on network growth. A positive β_2 indicates that the rich-get-richer effect is in order, and a negative β_2 to the other direction, namely the rich-get-poorer effect.

In this model, we assume that the connection probability of a receiver node is affected by the fitness and popularity of the node as it is in the FPDN model. We also consider the normal distribution for the node fitness.

3.2. Popularity measures

The *popularity* is an indicator of fame and preference among network members. There are several popularity measures including centrality measures, and the measure depends on the characteristics of a network [59]. In this paper, we consider the following two measures:

- In-degree: It is obvious that popular nodes have a large in-degree. We use the logarithm transformation of the in-degree given as

$$u_j^t = \ln(1 + D_{\text{in},j}^t). \tag{3}$$

  The logarithmic value is used because the in-degree is usually right-skewed [60]. Pham *et al* [40] used similar measures.
- Betweenness centrality: The betweenness centrality of a node can be interpreted as the ratio of the short cuts between nodes that pass through the node. The betweenness centrality of node j is defined as

$$u_j^t = \sum_{i \neq j, k \neq j} \frac{\eta_{ik}^t(j)}{\eta_{ik}^t}, \tag{4}$$

  where η_{ik}^t is the number of shortest paths from node i to k at time t and $\eta_{ik}^t(j)$ is the number of shortest paths from node i to k that pass through j at time t. A node with a high betweenness centrality usually has many nodes adjacent to it.

J. Stat. Mech. (2019) 043407

4. Estimation

4.1. Likelihood functions

In the proposed model, the fitness and popularity of the sender node are assumed not to affect the connection probability. We also assume that there are no edge deletions to focus on the fit-get-richer and rich-get-richer phenomena. Suppose that node j has y_j^t incoming edges from nodes in V^{t-1} at time t. Then, we have

$$y_j^t = D_{\text{in},j}^t - D_{\text{in},j}^{t-1} - w_j^t + v_j^t, \qquad t = t_{0,j}+1, \cdots, t_{1,j},$$

where $w_j^t = \sum_{i \in V^t \setminus V^{t-1}} A_{ij}^t$ is the number of incoming edges from the newly entered nodes at time t and $v_j^t = \sum_{i \in V^{t-1} \setminus V^t} A_{ij}^t$ is the number of dropped-out incoming edges by the nodes that left the system at time $t-1$. For two sets A and B, $A \setminus B$ is the set of all elements of A that are not included in B.

Let n_j^t be the number of possible connections to receiver node j at time t. As mentioned above, the subscript i is a dummy variable in θ_{ij}^t, and we will write θ_j^t instead. We regard a connection as the random behavior between nodes, and the random variable y_j^t follows a binomial distribution,

$$y_j^t \quad \sim \quad \text{Binomial}(n_j^t, \theta_j^t),$$

with the probability mass function (pmf)

$$p(y_j^t|\theta_j^t) = \binom{n_j^t}{y_j^t} \left(\theta_j^t\right)^{y_j^t} \left(1-\theta_j^t\right)^{n_j^t - y_j^t}.$$

Using equation (1), we can write the pmf as a function of the fitness, popularity, and the model parameters as

$$\begin{aligned} p(y_j^t|f_j^{t-1}, u_j^{t-1}, \beta) = &\binom{n_j^t}{y_j^t} \left(g(\beta_0 + \beta_1 f_j^{t-1} + \beta_2 u_j^{t-1})\right)^{y_j^t} \\ &\cdot \left(1 - g(\beta_0 + \beta_1 f_j^{t-1} + \beta_2 u_j^{t-1})\right)^{n_j^t - y_j^t}, \end{aligned}$$

where $\beta = (\beta_0, \beta_1, \beta_2)$ and $g(x) = 1/(1+e^{-x})$. The pmf can be reexpressed as

$$\begin{aligned} p(y_j^t|f_j^{t-1}, u_j^{t-1}, \beta) = &\binom{n_j^t}{y_j^t} \exp\left(y_j^t(\beta_0 + \beta_1 f_j^{t-1} + \beta_2 u_j^{t-1})\right) \\ &\cdot \left(1 + \exp\left(\beta_0 + \beta_1 f_j^{t-1} + \beta_2 u_j^{t-1}\right)\right)^{-n_j^t}. \end{aligned} \tag{5}$$

For simplicity, we denote the vector of variables by omitting the superscript t as follows:

- $y_j = (y_j^{t_{0,j}+1}, y_j^{t_{0,j}+2}, \cdots, y_j^{t_{1,j}})$
- $f_j = (f_j^{t_{0,j}}, f_j^{t_{0,j}+1}, \cdots, f_j^{t_{1,j}-1})$
- $u_j = (u_j^{t_{0,j}}, u_j^{t_{0,j}+1}, \cdots, u_j^{t_{1,j}-1})$

https://doi.org/10.1088/1742-5468/ab7754

Similarly, we express the data for all the nodes by omitting the subscript j:

- $y = (y_1^{t_{0,1}+1}, \cdots, y_1^{t_{1,1}}, \cdots, y_N^{t_{0,N}+1}, \cdots, y_N^{t_{1,N}})$
- $f = (f_1^{t_{0,1}}, \cdots, f_1^{t_{1,1}-1}, \cdots, f_N^{t_{0,N}}, \cdots, f_N^{t_{1,N}-1})$
- $u = (u_1^{t_{0,1}}, \cdots, u_1^{t_{1,1}-1}, \cdots, u_N^{t_{0,N}}, \cdots, u_N^{t_{1,N}-1})$
- $\sigma^2 = (\sigma_1^2, \sigma_2^2, \cdots, \sigma_N^2)$

We write the probability distribution of y_j given f_j, u_j, β as

$$p(y_j|f_j, u_j, \beta) = \prod_{t=t_{0,j}+1}^{t_{1,j}} p(y_j^t|f_j^{t-1}, u_j^{t-1}, \beta).$$

The probability distribution of fitness f_j given σ_j^2 is given by

$$p(f_j|\sigma_j^2) = p(f_j^{t_{0,j}}) \prod_{t=t_{0,j}+1}^{t_{1,j}-1} p(f_j^t|f_j^{t-1}, \sigma_j^2),$$

where $p(f_j^{t_{0,j}})$ is the probability density function (pdf) of $N(0,1)$ and $p(f_j^t|f_j^{t-1}, \sigma_j^2)$ is the pdf of $N(f_j^{t-1}, \sigma_j^2)$ as specified in (2). Then we have the probability distribution of y_j, f_j, σ_j^2 given u_j, β for node j,

$$p(y_j, f_j, \sigma_j^2|u_j, \beta) = p(\sigma_j^2)\, p(f_j|\sigma_j^2)\, p(y_j|f_j, u_j, \beta).$$

The total probability distribution of the model can be written by

$$p(y, f, \sigma^2|u, \beta) = \prod_{j \in V} p(y_j, f_j, \sigma_j^2|u_j, \beta).$$

Our goal is to estimate the latent variables f, σ^2 and the model parameter β. We use the EM algorithm for the estimation by regarding the latent variables f and σ^2 as missing. We need a complete data likelihood function for the EM,

$$L(\beta|f, \sigma^2, u, y) = p(y, f, \sigma^2|u, \beta),$$

and the complete data log-likelihood function is thus given by

$$l(\beta|f, \sigma^2, u, y) = \ln L(\beta|f, \sigma^2, p, y).$$

4.2. Algorithm

The EM algorithm consists of the expectation step (E-step) and the maximization step (M-step). In the E-step, we compute the expected value $Q(\beta|\hat{\beta}^{(s)})$ of the complete data log-likelihood function for the given parameter value $\hat{\beta}^{(s)} = (\hat{\beta}_0^{(s)}, \hat{\beta}_1^{(s)}, \hat{\beta}_2^{(s)})$. Formally, we have

J. Stat. Mech. (2019) 043407

$$Q\left(\beta|\hat{\beta}^{(s)}\right) = E\left[l(\beta|f,\sigma^2,u,y)|u,y,\hat{\beta}^{(s)}\right] = \int l(\beta|f,\sigma^2,u,y)\, p(\, f,\sigma^2|u,y,\hat{\beta}^{(s)})d(\, f,\sigma^2), \tag{6}$$

where the expectation is over $p(\, f,\sigma^2|u,y,\hat{\beta}^{(s)})$. The superscript s is the iteration count in the EM process. This log-posterior of the static parameter β is to be maximized in the M-step.

The integral is hard to handle analytically due to a complicated formula for $p(\, f,\sigma^2|u,y,\hat{\beta}^{(s)})$. An alternative is an approximation method through sampling. We are interested in the problem of inferring f and σ^2 when the observations y are given for all time points. In this paper, we use the MCMC method with a Gibbs sampler to approximate the joint distribution of fitness f and fitness variability σ^2 given u, y, $\hat{\beta}^{(s)}$.

For $j \in V$ and $t = t_{0,j}+1, t_{0,j}+2, \cdots, t_{1,j}-1$, we have, using the Markov structure of the network and the Bayes theorem,

$$\begin{aligned} &p(\, f_j^{t-1}|f \setminus \{\, f_j^{t-1}\}, \sigma^2, u, y, \hat{\beta}^{(s)}) \\ &\quad = p(\, f_j^{t-1}|f_j^{t-2}, f_j^t, \sigma_j^2, u_j^{t-1}, y_j^t, \hat{\beta}^{(s)}) \\ &\quad \propto p(\, f_j^t|f_j^{t-1}, \sigma_j^2)\, p(y_j^t|f_j^{t-1}, u_j^{t-1}, \hat{\beta}^{(s)})\, p(\, f_j^{t-1}|f_j^{t-2}, \sigma_j^2), \end{aligned} \tag{7}$$

where $p(\, f_j^{t-1}|f_j^{t-2}, \sigma_j^2)$ is replaced by $p(\, f_j^{t_{0,j}})$ when $t = t_{0,j}+1$.

When $t = t_{1,j}$, we have

$$\begin{aligned} &p(\, f_j^{t_{1,j}-1}|f \setminus \{\, f_j^{t_{1,j}-1}\}, \sigma^2, u, y, \hat{\beta}^{(s)}) \\ &\quad = p(\, f_j^{t_{1,j}-1}|f_j^{t_{1,j}-2}, \sigma_j^2, u_j^{t_{1,j}-1}, y_j^{t_{1,j}}, \hat{\beta}^{(s)}) \\ &\quad \propto p(y_j^{t_{1,j}}|f_j^{t_{1,j}-1}, u_j^{t_{1,j}-1}, \hat{\beta}^{(s)})\, p(\, f_j^{t_{1,j}-1}|f_j^{t_{1,j}-2}, \sigma_j^2). \end{aligned} \tag{8}$$

We can readily check that the product of two normal distributions is also normal under the following condition.

Proposition 1. *Suppose that $X|Y \sim N(Y, \sigma_{X|Y}^2)$ and $Z|X \sim N(X, \sigma_{Z|X}^2)$ for random variables X, Y and Z. Then the distribution of X given Y and Z follows the normal distribution with mean $\mu_{X|Y,Z}$ and variance $\sigma_{X|Y,Z}^2$, i.e. $X|Y,Z \sim N(\mu_{X|Y,Z}, \sigma_{X|Y,Z}^2)$, where*

$$\mu_{X|Y,Z} = \frac{Y/\sigma_{X|Y}^2 + Z/\sigma_{Z|X}^2}{1/\sigma_{X|Y}^2 + 1/\sigma_{Z|X}^2}, \quad \sigma_{X|Y,Z}^2 = \frac{1}{1/\sigma_{X|Y}^2 + 1/\sigma_{Z|X}^2}.$$

From proposition 1, the function in (7), $p(\, f_j^t|f_j^{t-1}, \sigma_j^2)\, p(\, f_j^{t-1}|f_j^{t-2}, \sigma_j^2)$ is proportional to the pdf of $N\left((\, f_j^{t-2} + f_j^t)/2,\ \sigma_j^2/2\right)$ for $t = t_{0,j}+2, t_{0,j}+3, \cdots, t_{1,j}-1$, and $p(\, f_j^{t_{0,j}+1}|f_j^{t_{0,j}}, \sigma_j^2)\, p(\, f_j^{t_{0,j}})$ is proportional to the pdf of $N\left((1/\sigma_0^2 + 1/\sigma_1^2)^{-1} f_j^1/\sigma_1^2,\ (1/\sigma_0^2 + 1/\sigma_1^2)^{-1}\right)$ for $t = t_{0,j}+1$.

For the σ^2 part, we have

$$\begin{aligned} & p(\sigma_j^2|f,\sigma^2-\{\sigma_j^2\},u,y,\hat{\beta}^{(s)}) \\ & \quad = p(\sigma_j^2|f_j,u_j,y_j,\hat{\beta}^{(s)}) \\ & \quad \propto p(\sigma_j^2)\,p(\,f_j|\sigma_j^2). \end{aligned} \tag{9}$$

Now we have obtained the required conditional distributions, and we are ready to run the Gibbs sampling algorithm. The procedure is summarized in algorithm 1.

Algorithm 1. Gibbs sampling.

input: Initial values $f_{j(0)}=\left(f_{j(0)}^{t_{0,j}},f_{j(0)}^{t_{0,j}+1},\cdots,f_{j(0)}^{t_{1,j}-1}\right)$ and $\sigma_{j(0)}^2$, $j\in V$, and parameter estimate $\hat{\beta}^{(s)}$.

for $b=1,2,\cdots,B$ **do**
  for $j\in V$ **do**
    for $t=t_{0,j}+1,2,\cdots,t_{1,j}-1$ **do**
      Sample $f_{j(b)}^{t-1}$ from $p(\,f_j^{t-1}|f-\{\,f_j^{t-1}\},\sigma^2,u,y,\hat{\beta}^{(s)})$ according to (7) employing $f_j^{t'-1}=f_{j(b)}^{t'-1}$ for $t'=t_{0,j}+1,t_{0,j}+2,\cdots,t-1$, $f_j^{t'-1}=f_{j(b-1)}^{t'-1}$ for $t'=t+1,t+2,\cdots,t_{1,j}$, and $\sigma_j^2=\sigma_{j(b-1)}^2$.
    end
    Sample $f_{j(b)}^{t_{1,j}-1}$ from $p(\,f_j^{t_{1,j}-1}|f-\{\,f_j^{t_{1,j}-1}\},\sigma^2,u,y,\hat{\beta}^{(s)})$ according to (8) employing $f_j^{t'-1}=f_{j(b)}^{t'-1}$ for $t'=t_{0,j}+1,2,\cdots,t_{1,j}-1$ and $\sigma_j^2=\sigma_{j(b-1)}^2$.
    Sample $\sigma_{j(b)}^2$ from $p(\sigma_j^2|f,\sigma^2-\{\sigma_j^2\},u,y,\hat{\beta}^{(s)})$ according to (9) employing $f_j=f_{j(b)}$.
  end
end

output: Gibbs samples $f_{j(b)}=\left(f_{j(b)}^{t_{0,j}},f_{j(b)}^{t_{0,j}+1},\cdots,f_{j(b)}^{t_{1,j}-1}\right)$ and $\sigma_{j(b)}^2$, $j\in V$, $b=1,2,\cdots,B$.

The B samples of fitness $f_{(b)}$ and fitness variability $\sigma_{(b)}^2$, $b=1,\cdots,B$ are extracted from the conditional distribution $p(\,f,\sigma^2|u,y,\hat{\beta}^{(s)})$ through algorithm 1. The sampling distributions are described in (7)–(9). They are products of several distributions, and we use adaptive rejection sampling (ARS) [61] to sample ad-hoc distributions in a relatively stable and effective manner. The requirement is that the target distribution is log-concave.

Proposition 2. *The probability density function in equation (5), $p(y_j^t|f_j^{t-1},u_j^{t-1},\beta)$ is a log-concave function of the variable f_j^{t-1}.*

Proof. By taking the logarithm of $p(y_j^t|f_j^{t-1},u_j^{t-1},\beta)$, we obtain

$$\begin{aligned} & \ln\,p(y_j^t|f_j^{t-1},u_j^{t-1},\beta) \\ & \quad = \ln\binom{n_j^t}{y_j^t}+y_j^t(\beta_0+\beta_1 f_j^{t-1}+\beta_2 u_j^{t-1}) \\ & \quad -n_j^t\ln\left(1+\exp\left(\beta_0+\beta_1 f_j^{t-1}+\beta_2 u_j^{t-1}\right)\right). \end{aligned} \tag{10}$$

On the right-hand side of equation (10), the first term does not depend on f_j^{t-1}

J. Stat. Mech. (2019) 043407

Table 1. Parameter estimation results for the synthetic networks.

Synthetic network	β_0 (intercept)		β_1 (fitness)		β_2 (popularity)	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
SD1	−7.8632	0.1542	1.0692	0.1471	0.4765	0.1456
SD2	−7.9277	0.1159	1.0748	0.1496	19.1191	5.2388

and the second term is linear in f_j^{t-1}. The second-order derivative of the last term $-n_j^t \ln\left(1+\exp\left(\beta_0+\beta_1 f_j^{t-1}+\beta_2 u_j^{t-1}\right)\right)$ is given by

$$\frac{-n_j^t\beta_1^2\exp\left(\beta_0+\beta_1 f_j^{t-1}+\beta_2 u_j^{t-1}\right)}{\left(1+\exp\left(\beta_0+\beta_1 f_j^{t-1}+\beta_2 u_j^{t-1}\right)\right)^2},$$

which is negative for any f_j^{t-1}. This means that the last term is concave in f_j^{t-1}. Therefore, the log-concavity of $p(y_j^t|f_j^{t-1},u_j^{t-1},\beta)$ with respect to f_j^{t-1} holds true. □

By the log-concavity of a normal distribution and proposition 2, the probability density functions in the right-hand-side of (7) and (8) are log-concave. Since the product of log-concave functions is log-concave, the target distribution on fitness $p(\,f_j^{t-1}|f-\{\,f_j^{t-1}\},\sigma^2,u,y,\hat{\beta}^{(s)})$, $t=t_{0,j}+1,t_{0,j}+2,\cdots,t_{1,j}$ are log-concave in f_j^{t-1}.

Unfortunately, $p(\sigma_j^2|f,\sigma^2-\{\sigma_j^2\},u,y,\hat{\beta}^{(s)})$ in (9) is not log-concave in σ_j^2. Hence, we make a change of variable $\eta_j=\ln(\sigma_j^2)$. The idea is that we sample η_j instead of σ_j^2, from

$$\begin{aligned}&p(\eta_j|f,\sigma^2-\{\sigma_j^2\},u,y,\hat{\beta}^{(s)})\\&\quad=p(\sigma_j^2=\mathrm{e}^{\eta_j}|f,\sigma^2-\{\sigma_j^2\},u,y,\hat{\beta}^{(s)})\left|\frac{\mathrm{d}\sigma_j^2}{\mathrm{d}\eta_j}\right|,\end{aligned}$$

which can be easily shown to be a log-concave function in η_j using the convexity of the exponential function. Then we can obtain a sample $\sigma_j^2=e^{\eta_j}$ from $p(\sigma_j^2|f,\sigma^2-\{\sigma_j^2\},u,y,\hat{\beta}^{(s)})$.

We set the initial value $f_{j(0)}=(0,0,\cdots,0)$ and $\sigma_{j(0)}^2=0.1^2$ for the first step of the EM algorithm. Then, we use the values $f_{j(B)}$ and $\sigma_{j(B)}^2$ sampled from the Gibbs sampling using the parameter estimate $\hat{\beta}^{(s-1)}$ which is obtained at the previous step. It is well known that the early samples from the Gibbs sampler are susceptible to the initial fitness values. This is why the first $B_0(<B)$ samples are not used for making inferences. We set $B=250$ and $B_0=50$ for data analysis.

For sufficiently large B_0 and $(B-B_0)$, we have

$$\begin{aligned}Q\left(\beta|\hat{\beta}^{(s)}\right)&\approx\frac{1}{B-B_0}\sum_{b=B_0+1}^{B}l(\beta|f_{(b)},\sigma_{(b)}^2,u,y)\\&=\frac{1}{B-B_0}\sum_{b=B_0+1}^{B}\sum_{j\in V}\ln p(y_j,f_{j(b)},\sigma_{j(b)}^2|u_j,\beta).\end{aligned}$$

J. Stat. Mech. (2020) 043407

In a nutshell, we begin the EM process with the initial values $\hat{\beta}^{(0)}$. At the $(s+1)$th iteration, $s = 0, 1, \cdots$, the E-step consists of computing the function $Q\left(\beta|\hat{\beta}^{(s)}\right)$ and the M-step consists of finding $\beta^{(s+1)}$ as

$$\hat{\beta}^{(s+1)} = \underset{\beta}{\operatorname{argmax}}\, Q\left(\beta|\hat{\beta}^{(s)}\right).$$

We repeat the E- and M-steps until $\hat{\beta}^{(s)}$ converges. Let $\hat{\beta}$ be the final estimate of β.

4.3. Inference on the model parameter and fitness

Throughout this subsection, let $f_{(b)}$ and $\sigma^2_{(b)}$, $b = 1, 2, \cdots, B$ be samples from algorithm 1 using the converged parameter $\hat{\beta}$. The observed information matrix of β is given by [62]

$$\begin{aligned} I(\hat{\beta}) \approx & -\nabla^2 Q\left(\hat{\beta}|\hat{\beta}\right) + \left[\nabla Q\left(\hat{\beta}|\hat{\beta}\right)\right]\left[\nabla Q\left(\hat{\beta}|\hat{\beta}\right)\right]' \\ & - \frac{1}{B - B_0}\sum_{b=B_0+1}^{B}\left[\nabla l(\hat{\beta}|f_{(b)}, \sigma^2_{(b)}, u, y)\right]\left[\nabla l(\hat{\beta}|f_{(b)}, \sigma^2_{(b)}, u, y)\right]' \end{aligned}$$

where $\nabla = (\partial/\partial\beta_0, \partial/\partial\beta_1, \partial/\partial\beta_2)'$. An estimate of the asymptotic covariance matrix of β is $\left[I(\hat{\beta})\right]^{-1}$. The standard error of $\hat{\beta}$ is approximated by the square root of the diagonal elements of $\left[I(\hat{\beta})\right]^{-1}$.

Let $\hat{\sigma}^2_j$ be a point estimate of the fitness variability σ^2_j of node j, given by

$$\hat{\sigma}^2_j = \frac{1}{B - B_0}\sum_{b=B_0+1}^{B}\sigma^2_{j(b)}.$$

Similarly, let $\hat{f}^{t-1}_j$ be a point estimate of the fitness of node j at time $t-1$, given by

$$\hat{f}^{t-1}_j = \frac{1}{B - B_0}\sum_{b=B_0+1}^{B} f^{t-1}_{j(b)}. \tag{11}$$

We will now discuss how the inference on fitness depends on the amount of data y. To simplify, we omit the popularity u_j, the fitness variability $\hat{\sigma}^2_j$, and the parameter estimate $\hat{\beta}$ hereafter. The in-degree increments of node j, y^t_j, up to time t, are denoted by $\mathbf{y}^t_j = \{y_j^{t_{0,j}+1}, y_j^{t_{0,j}+2}, \cdots, y^t_j\}$ and we define $\mathbf{y}^t_j$ as the empty set for $t \leqslant t_{0,j}$ and $\{y_j^{t_{0,j}+1}, y_j^{t_{0,j}+2}, \cdots, y_j^{t_{1,j}}\}$ for $t \geqslant t_{1,j}$. Let $f^{t|t'}_j = E\left(f^t_j|\mathbf{y}^{t'}_j\right)$ and $\xi^{t|t'}_j = Var\left(f^t_j|\mathbf{y}^{t'}_j\right)$ be the conditional mean and variance of f^t_j given $\mathbf{y}^{t'}_j$.

The operation, *smoothing*, is the process of collecting information about f^{t-1}_j, given all the data up to T, $\mathbf{y}^T_j$. We already discussed how the fitness samples are obtained. The sample mean and variance of $\{f^{t-1}_{j(b)}\}_{b=B_0+1,\cdots,B}$ are approximations of $f^{t-1|T}_j$ and $\xi^{t-1|T}_j$, respectively.

It is often important to directly estimate the fitness value at time $t-1$ when the observed data is given at time t. In other words, we can get information on the fitness

J. Stat. Mech. (2019) 043407

Table 2. Parameter estimates for the V-FPDN model based on YouTube subscription network data. 15 categories of YouTube are analyzed. The rich-get-richer phenomenon is observed in every category.

		β_0 (intercept)		β_1 (fitness)		β_2 (popularity)	
Category	N[a]	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Animals	237	−15.1684	0.0299	2.2689	0.0043	0.7086	0.0021
Autos	242	−17.0107	0.0255	2.5127	0.0026	0.8650	0.0018
Comedy	246	−19.9789	0.0197	5.2138	0.0016	1.0461	0.0013
Education	243	−17.2405	0.0221	2.0177	0.0033	0.8968	0.0015
Entertainment	240	−16.2190	0.0153	3.0139	0.0014	0.8105	0.0009
Film	247	−17.9354	0.0256	2.7234	0.0054	0.9284	0.0018
Games	241	−15.5797	0.0237	3.5885	0.0052	0.7615	0.0015
Howto	247	−14.1884	0.0162	3.5489	0.0017	0.6888	0.0010
Music	241	−21.2754	0.0167	7.0531	0.0048	1.0706	0.0010
News	244	−10.9966	0.0184	2.3082	0.0021	0.4752	0.0013
Nonprofit	242	−18.4771	0.0284	2.9152	0.0036	0.9533	0.0021
People	244	−16.8054	0.0265	3.1533	0.0030	0.8444	0.0017
Sports	243	−18.9790	0.0158	2.6886	0.0028	1.0080	0.0010
Tech	246	−12.3753	0.0176	2.8151	0.0026	0.5525	0.0011
Travel	248	−13.5024	0.0198	2.7528	0.0053	0.6063	0.0015
Average	243.4	−16.3821	0.0214	3.2383	0.0033	0.8144	0.0014

[a] N is the total number of channels.

in real time, and this is called *filtering*. In this case, the quantities like $f_j^{t-1|t}, \xi_j^{t-1|t}, f_j^{t|t}$ and $\xi_j^{t|t}$ are important measures of fitness. Again, due to the non-Gaussian nature of the observation and the non-linearity, distributions are not explicit. Instead, estimates can be obtained by algorithm 2.

Algorithm 2. Getting estimations of $f_j^{t-1|t}, \xi_j^{t-1|t}, f_j^{t|t}$, and $\xi_j^{t|t}$.

1 Initialize: Let $\hat{f}_j^{t_{0,j}|t_{0,j}} = 0$ and $\hat{\xi}_j^{t_{0,j}|t_{0,j}} = 1$.
2 **for** $t = t_{0,j} + 1, 2, \cdots, t_{1,j}$ **do**
3 Sample $z_{j1}^{t-1}, z_{j2}^{t-1}, \cdots, z_{jm}^{t-1}$ from $p(f_j^{t-1}|\mathbf{y}_j^t)$.
4 Let $\hat{f}_j^{t-1|t}$ and $\hat{\xi}_j^{t-1|t}$ be the sample mean and variance of $\{z_{j1}^{t-1}, z_{j2}^{t-1}, \cdots, z_{jm}^{t-1}\}$.
5 Let $\hat{f}_j^{t|t} = \hat{f}_j^{t-1|t}$ and $\hat{\xi}_j^{t|t} = \hat{\xi}_j^{t-1|t} + \hat{\sigma}_j^2$.
6 **end**
output: $\hat{f}_j^{t-1|t}, \hat{\xi}_j^{t-1|t}, \hat{f}_j^{t|t}$, and $\hat{\xi}_j^{t|t}$, $t = t_{0,j} + 1, t_{0,j} + 2, \cdots, t_{1,j}$.

The proposed model assumes that the distribution of $f_j^{t_{0,j}}$ follows a normal distribution with mean 0 and variance 1 as described in (2). Then we have $f_j^{t_{0,j}|t_{0,j}} = E(f_j^{t_{0,j}}|\mathbf{y}_j^{t_{0,j}}) = E(f_j^{t_{0,j}}) = 0$ and $\xi_j^{t_{0,j}|t_{0,j}} = \mathrm{Var}(f_j^{t_{0,j}}|\mathbf{y}_j^{t_{0,j}}) = Var(f_j^{t_{0,j}}) = 1$. With these starting values, we can obtain $\hat{f}_j^{t-1|t}, \hat{\xi}_j^{t-1|t}, \hat{f}_j^{t|t}$, and $\hat{\xi}_j^{t|t}$, for $t = t_{0,j} + 1, \cdots, t_{1,j}$ through an iterative process. For ease of computation, we approximate $f_j^{t-1}|\mathbf{y}_j^{t-1}$ by the normal distribution with mean $f_j^{t-1|t-1}$ and variance $\xi_j^{t-1|t-1}$. Note that these distributions

J. Stat. Mech. (2020) 043407

are Gaussian when $t = t_{0,j} + 1$. In practice, they are close to a normal distribution due to the similarity of the normal and binomial distributions [63]. Since the relation $p(f_j^{t-1}|\mathbf{y}_j^t) \propto p(y_j^t|f_j^{t-1})\, p(f_j^{t-1}|\mathbf{y}_j^{t-1})$ holds, we can obtain samples $z_{j1}^{t-1}, z_{j2}^{t-1}, \cdots, z_{jm}^{t-1}$ from $p(f_j^{t-1}|\mathbf{y}_j^t)$. Again, we employ the ARS method because the target distribution $p(f_j^{t-1}|\mathbf{y}_j^t)$ is log-concave according to proposition 2. We approximate $f_j^{t-1|t}$ and $\xi_j^{t-1|t}$ using the sample mean and variance of m $(= 200)$ ARS samples $\{z_{j1}^{t-1}, z_{j2}^{t-1}, \cdots, z_{jm}^{t-1}\}$. Finally, we compute $\hat{f}_j^{t|t}$ and $\hat{\xi}_j^{t|t}$ by using the equations,

$$f_j^{t|t} = E(f_j^t|\mathbf{y}_j^t) = E(f_j^{t-1} + \epsilon_j^t|\mathbf{y}_j^t) = f_j^{t-1|t},$$
$$\xi_j^{t|t} = \mathrm{Var}(f_j^t|\mathbf{y}_j^t) = \mathrm{Var}(f_j^{t-1} + \epsilon_j^t|\mathbf{y}_j^t) = \xi_j^{t-1|t} + \hat{\sigma}_j^2.$$

They are derived by applying independence between ϵ_j^t and $\mathbf{y}_j^t$.

5. Synthetic data analysis

We set up a true model using various popularity measures discussed earlier. In this section, we generate time series networks according to (1) and (2), and we apply the V-FPDN model and check whether it can satisfactorily estimate parameters and fitness values. In addition, we apply the filtering and smoothing techniques to detect changes in the fitness values.

We generate two synthetic network datasets by using in-degree and betweenness centrality as popularity measures. For each dataset, there are 200 nodes in total and 11 time points, i.e. $N = 200$ and $T = 10$. Let an initial network G^0 be composed of 110 nodes, $V^0 = \{1, 2, \cdots, 110\}$, where it consists of a random network with 100 nodes $\{1, 2, \cdots, 100\}$ (the edge connection probability is 0.01) plus 10 isolated nodes $\{101, 102, \cdots, 110\}$. Ten new nodes $\{10t + 101, 10t + 102, \cdots, 10t + 110\}$ enter the system as isolated nodes at each time $t = 1, 2, \cdots, 9$. The fitness levels of nodes are generated according to (2) with $\sigma_j^2 = 0.1^2$, $j = 1, 2, \cdots, 200$. We assume that every connection is possible, i.e. $n_j^t = |V^{t-1}| - D_{\mathrm{in},j}^{t-1} - 1$. We set the β_2 values differently considering the scale of each measure and two dynamic network datasets, SD1 and SD2, are generated.

- SD1: We use the in-degree popularity measure given in equation (3) and the parameter is set as $\beta = (-8.0, 1.0, 0.5)$.
- SD2: We use the betweenness centrality given in equation (4) and the parameter is set as $\beta = (-8.0, 1.0, 20.0)$.

We apply the V-FPDN model to the two datasets and estimate the parameters and fitness values. Table 1 shows the parameter estimation results. The algorithm of the V-FPDN model is successful in the estimation of the model parameters. We have the average estimated fitness variability $\frac{1}{N}\sum_j \hat{\sigma}_j^2 = 0.0366$ and 0.0365 for SD1 and SD2, respectively.

Next, we discuss the inference of fitness. The smoothed, filtered, and true fitness values are plotted in figure 1, for which three nodes are chosen corresponding to the first, second, and third quartiles of the true fitness values at the initial time. In addition, the

J. Stat. Mech. (2019) 043407

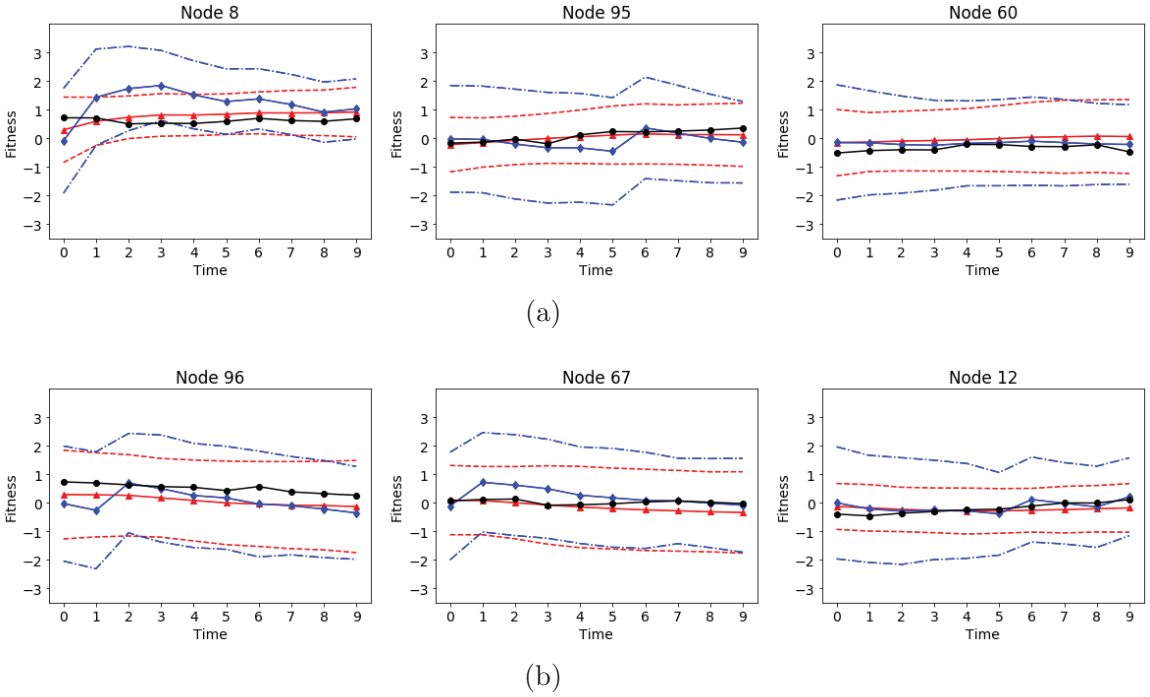

Figure 1. The estimated and true fitness values based on the network data, SD1 and SD2. The nodes corresponding to the first, second, and third quartiles of the true fitness values at the initial time are shown in order. The true fitness value f_j^{t-1}, the smoothed estimate $\hat{f}_j^{t-1|T}$, and the filtered estimate $\hat{f}_j^{t-1|t}$ are expressed in black, red, and blue, respectively. In addition, the approximated confidence intervals are shown in dotted lines. The formulae of the intervals are in (12).

approximated confidence intervals for the true fitness are displayed according to the formulae,

$$\hat{f}_j^{t-1|T} \pm 2\sqrt{\hat{\xi}_j^{t-1|T}} \quad \text{and} \quad \hat{f}_j^{t-1|t} \pm 2\sqrt{\hat{\xi}_j^{t-1|t}}, \tag{12}$$

corresponding to the smoothing and filtering, respectively.

We can see in figure 1 that almost every true fitness value lies inside the confidence intervals. The smoother contributes to a smaller variability over time than the filter. On the other hand, the filtering tends to yield a relatively large variability over time. The standard deviation $\sqrt{\hat{\xi}_j^{t-1|T}}$ of the smoothed estimate is smaller than that of the filtered estimate because the smoother employs a larger amount of observed data than the filtering.

6. Real data analysis

6.1. Comparison with existing model

The strength of the fitness and popularity effects in the Facebook wallpost network is analyzed in Jung *et al* [42], where the FPDN model was used. In this section, we apply

J. Stat. Mech. (2020) 043407

both models, i.e. FPDN and V-FPDN, and analyze the difference between the models. In the Facebook wallpost network [64, 65], the nodes are users, and the directed edge from i to j is created when user i posts a message on user j's wall. In this network, the flow of posting is important. Therefore, we use the betweenness centrality as the popularity measure and standardize it for comparison.

We use data from September 14, 2006 to November 22, 2008. We construct the initial network G^0 based on the edges created up to $T_0 = 400$ days. The time interval is $\Delta T = 40$ days and $T = 10$ time points after the initial network is considered. As in Jung *et al* [42], $N = 1000$ nodes are sampled and applied to the model with $n_j^t = 300$. The result of the parameter estimation for the FPDN model with $m_{\text{FPDN}} = 200$ [42] is given by

$$\theta_{ij}^t = g\left(-7.3336_{(0.0518)} + 1.2270_{(0.0332)} f_{j,\text{FPDN}} + 0.1260_{(0.0075)} u_j^{t-1}\right), \quad t = 1, 2, \cdots, T,$$

where the values in parentheses are the standard errors. We can see that the impact of fitness on network growth is larger than that of popularity. Next, the result of the V-FPDN model is given by

$$\theta_{ij}^t = g\left(-7.5843_{(0.0498)} + 1.6314_{(0.0954)} f_{j,\text{V-FPDN}}^{t-1} + 0.1071_{(0.0232)} u_j^{t-1}\right), \quad t = 1, 2, \cdots, T,$$

with the average estimated fitness variability $\frac{1}{N}\sum_j \hat{\sigma}_j^2 = 0.0490$. Compared with the FPDN model, β_1 increases from 1.2270 to 1.6314, and the popularity effect parameter β_2 decrease slightly, from 0.1260 to 0.1071. We can conclude that the fitness effect dominates the popularity effect in the V-FPDN model, which is consistent with the findings of Pham *et al* [40] and Jung *et al* [42]. As mentioned before, it is appropriate to assume that the node fitness changes in the Facebook wallpost network, and we can explain the network growth with higher accuracy by allowing for fitness change.

Figure 2 is a summary of the fitness estimation for three users 4769, 8610, and 13 477. The black line represents the fitness levels estimated under the FPDN model, and the red line the fitness levels obtained by equation (11). The dotted line represents the estimated fitness plus or minus two times the standard deviation. In the V-FPDN model, fitness estimates for user 4769 are similar to the FPDN model and show little variation over time. By contrast, the fitness estimates for user 13 477 exhibit a significant difference and change considerably over time. The fitness of user 8610 tends to decrease over time. We can detect the change in the fitness levels of users through the proposed V-FPDN model.

The fitness, in-degree, and standardized betweenness centrality of three users are shown in figure 3, exposing the factors causing the fitness change in the V-FPDN model. The in-degree of user 8610 tends to increase slightly at the beginning and then stagnate afterwards. In other words, user 8610 is active at first, but then does not get any more messages from time $t = 4$ on and this tendency is reflected in the fact that the fitness estimates decrease gradually. We observe that user 4769s increments of in-degree are almost constant as shown in figure 3(b), and fitness levels show little variability. In case of user 13 477, fitness level tends to rise when the in-degree increment is

J. Stat. Mech. (2019) 043407

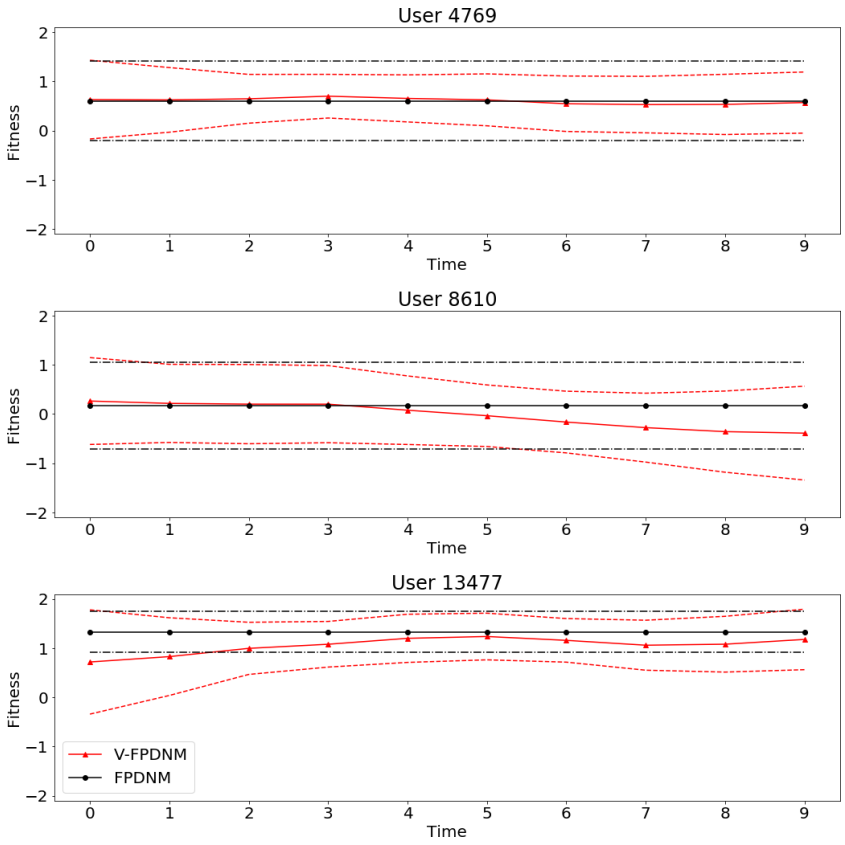

Figure 2. Fitness estimates of three users 4769 (top), 8610 (middle), and 13 477 (bottom). The black line represents the fitness estimates according to the FPDN model, and the red line the fitness estimates according to the V-FPDN model. The dotted line represents the estimated fitness values plus or minus two times the standard deviation.

large. Changes in fitness are sensitive to the number of incoming edges. We also present the popularity measure in figure 3(c). The standardized betweenness centrality tends to increase as the in-degree increases, but this is not always true as shown in figure 3.

6.2. YouTube subscription network

The YouTube subscription network consists of channels with directed subscription relations. We obtain data from Socialblade[3], a website that collects and provides publicly available information from various social networks. According to the Socialblade, YouTube channels are divided into 16 categories such as 'music' and 'sports'. In this paper, we use the top 250 channels based on the number of subscribers on November 13, 2018. We use August 2018 data, and the network on August 1 is used as an initial network[4]. We then observe at every three days until August 31, 2018. In other words, we set $\Delta T = 3$ days and $T = 10$. We use the in-degree popularity measure in equation (3) and set $n_j^t = 1000\,000$ for analysis. We exclude channels with missing data, i.e.

[3] http://socialblade.com
[4] The composition of the top 250 channels changes very little over time.

J. Stat. Mech. (2020) 043407

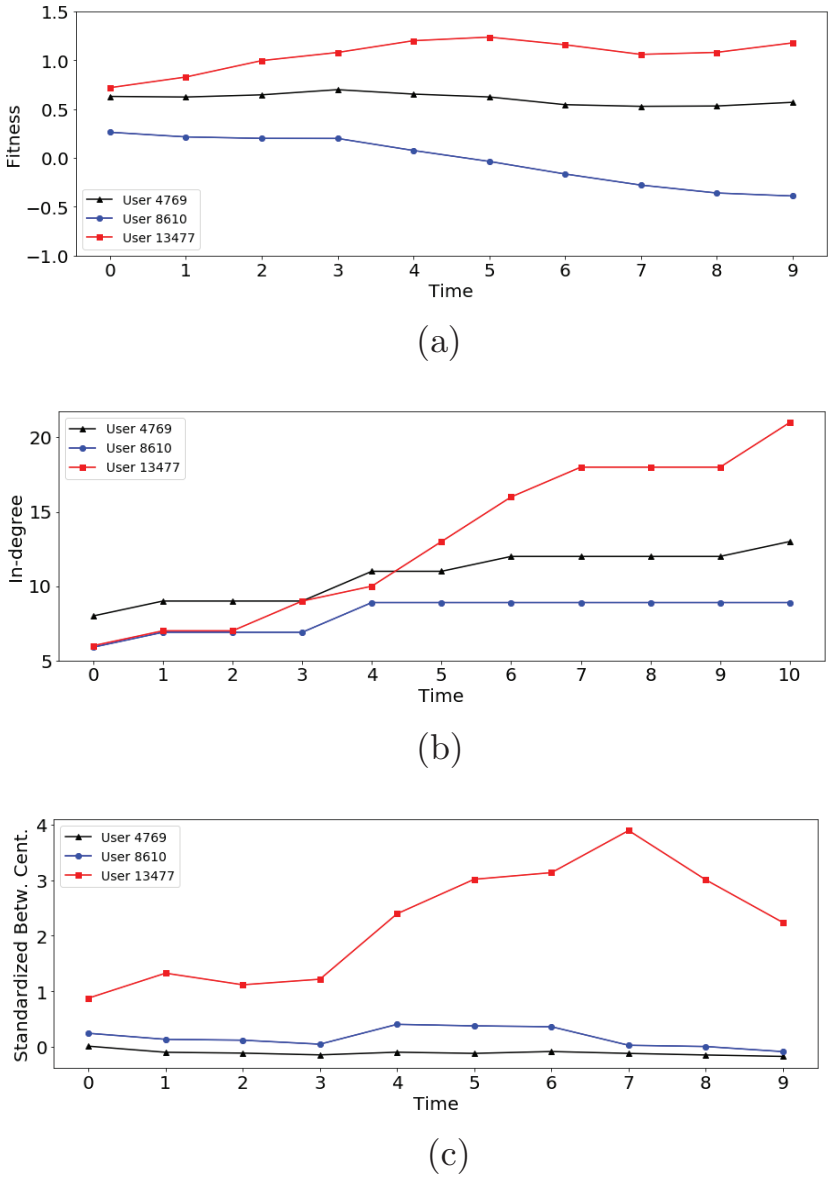

Figure 3. The fitness (top), in-degree (middle), and standardized betweenness centrality (bottom) of three users 4769 (black), 8610 (blue), and 13 477 (red). (a) Fitness. (b) In-degree. (c) Standardized betweenness centrality.

with no pages, no subscription information on at least one of the observation dates, or no subscriber updates in at least seven time points. The total number of the channels in the data are in table 2. There are a small number of negative in-degree increments, which are replaced with zeros for analysis. We exclude the category 'shows' as it has many missing values. Since we use the channels with many subscribers, the results are valid for the popular channels. The results of parameter estimation under the V-FPDN model are summarized in table 2. All categories have the average estimated fitness variability $\frac{1}{N}\sum_j \hat{\sigma}_j^2$ between 0.0154 and 0.0309.

According to the result, the fitness and popularity effect parameters are 3.2383 and 0.8144 on average. The popularity effects are found significant in all the categories.

https://doi.org/10.1088/1742-5468/ab7754

J. Stat. Mech. (2019) 043407

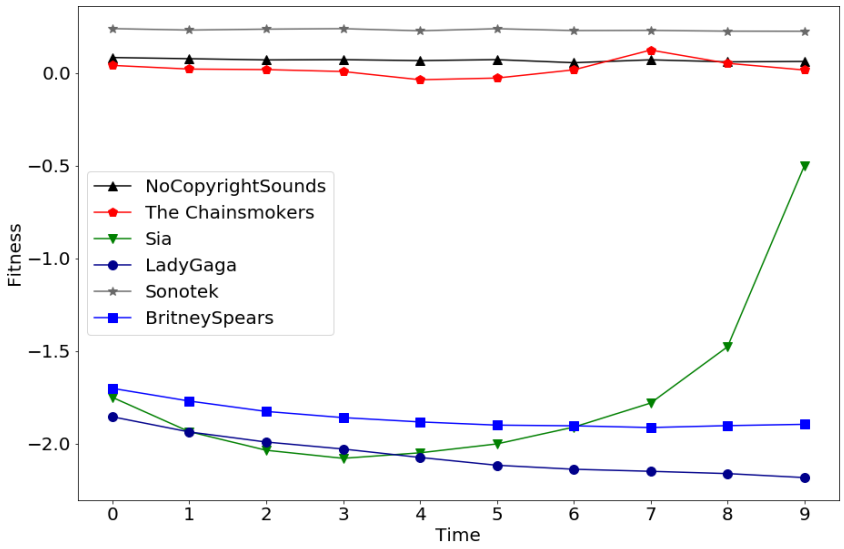

Figure 4. Fitness values of the six channels in the 'music' category of YouTube: 'NoCopyrightSounds', 'The Chainsmokers', 'Sia', 'Lady Gaga', 'Sonotek', and 'Britney Spears'.

In other words, the rich-get-richer effect is working on the YouTube subscription networks in every category. It may be explained by easy accessibility to popular channels through YouTube search, recommendation, and other functions. Significant popularity effects are observed in all categories, especially in 'comedy'. There are considerable fitness effects in active areas such as 'comedy' and 'music', and relatively benign fitness effects in static areas such as 'education', 'animals', and 'news'.

Next, we examine the changes in fitness in some YouTube channels in the 'music' category, which is linked to the most significant fitness effect. Figure 4 shows the fitness values of the six channels, 'NoCopyrightSounds', 'The Chainsmokers', 'Sia', 'Lady Gaga', 'Sonotek', and 'Britney Spears'. American music production duo 'The Chainsmokers' released the song 'Side Effects' around the investigated time, and we can observe that the fitness increases when the official video[5] is released some time between $t = 6$ and $t = 7$. A singer-songwriter 'Sia' announced the song 'Thunderclouds' on August 9, 2018. The audio version and the official video were uploaded in August, and the fitness increased rapidly. We observe that activities such as the song presentation have a great influence on fitness in the 'music' category, and the V-FPDN model captures this well.

The musicians, 'Lady Gaga' and 'Britney Spears', had little activity in or before August 2018. As a result, fitness continued to decline. Some channels upload videos periodically with a certain theme on music-related channels such as, 'NoCopyrightSounds' and 'Sonotek'. 'NoCopyrightSounds' is a music organization and one of the open-source labels that release unlicensed musics, and 'Sonotek' is an Indian music company. We observe that the fitness levels are almost constant over time, which may be because they regularly upload similar videos related to a specific theme.

[5] https://youtu.be/nuckTcoZG4Q

J. Stat. Mech. (2020) 043407

7. Concluding remarks

In this paper, we proposed a novel model, V-FPDN model, allowing changes in fitness over time in an effort to generalize the FPDN model. We presented an estimation procedure using the EM algorithm and MCMC, which worked well in the data analysis of networks. The inference on the change in fitness was substantiated via the smoothing and filtering. We analyzed various data using in-degree and betweenness centrality as popularity measures. Depending on the characteristics of the network, various popularity measures can be employed, and the relationship between popularity and network growth can be inferred.

We investigated two real datasets, Facebook and YouTube. In Facebook, we estimated the size of the fitness and popularity effects in the framework of varying fitness and compared the results with those of the FPDN model. We also investigated the change in fitness of the nodes and made interpretations in the context of data. We observed a significant rich-get-richer phenomenon in all categories of YouTube. We found that fitness effects are strong in active areas such as 'music' and 'comedy' and mild in static areas such as 'news', 'education', and 'animals'. The proposed model is shown capable of analyzing changes in fitness and has successfully detected changes that show various patterns in the YouTube channels.

Acknowledgment

Sung-Ho Kim was supported for this work by the NRF Grant (No. 2016R1D1A1B03936155) of the Republic of Korea and Jae-Gil Lee by the MOLIT (The Ministry of Land, Infrastructure and Transport), Korea, under the national spatial information research program supervised by the KAIA (Korea Agency for Infrastructure Technology Advancement) (19NSIP-B081011-06).

References

[1] Kudělka M, Horák Z, Snášel V, Krömer P, Platoš J and Abraham A 2012 *Log. J. IGPL* **20** 634–43
[2] Lim S and Lee J G 2016 *J. Stat. Mech.* 123401
[3] Fortunato S 2010 *Phys. Rep.* **486** 75–174
[4] Coscia M, Giannotti F and Pedreschi D 2011 *Stat. Anal. Data Min.* **4** 512–46
[5] Chen D, Lü L, Shang M S, Zhang Y C and Zhou T 2012 *Physica* A **391** 1777–87
[6] Zamora-López G, Zhou C and Kurths J 2010 *Front. Neuroinform.* **4** 1
[7] Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E and Makse H A 2010 *Nat. Phys.* **6** 888
[8] Braha D and Bar-Yam Y 2006 *Complexity* **12** 59–63
[9] Hill S A and Braha D 2010 *Phys. Rev.* E **82** 046105
[10] Bringmann B, Berlingerio M, Bonchi F and Gionis A 2010 *IEEE Intell. Syst.* **25** 26–35
[11] Juszczyszyn K, Budka M and Musial K 2011 The dynamic structural patterns of social networks based on triad transitions *Proc. IEEE ACM Int. Conf. on Advances in Social Network Analysis and Mining* pp 581–6
[12] Blonder B, Wey T W, Dornhaus A, James R and Sih A 2012 *Methods Ecol. Evol.* **3** 958–72
[13] Aggarwal C and Subbian K 2014 *ACM Comput. Surv.* **47** 10
[14] Leskovec J, Kleinberg J and Faloutsos C 2005 Graphs over time: densification laws, shrinking diameters and possible explanations *Proc. of ACM SIGKDD* pp 177–87
[15] Wang D, Song C and Barabási A L 2013 *Science* **342** 127–32
[16] Merton R K 1968 *Science* **159** 56–63

J. Stat. Mech. (2019) 043407

[17] Tufekci Z 2010 Who acquires friends through social media and why? 'get richer' versus 'seek and ye shall find' *4th Int. AAAI Conf. on Weblogs and Social Media*
[18] Kondor D, Pósfai M, Csabai I and Vattay G 2014 *PLoS One* **9** e86197
[19] Van de Rijt A, Kang S M, Restivo M and Patil A 2014 *Proc. Natl Acad. Sci.* **111** 6934–9
[20] Perc M 2014 *J. R. Soc. Interface* **11** 20140378
[21] Barabási A L and Albert R 1999 *Science* **286** 509–12
[22] Dorogovtsev S N, Mendes J F F and Samukhin A N 2000 *Phys. Rev. Lett.* **85** 4633
[23] Gabel A and Redner S 2013 *J. Stat. Mech.* P02043
[24] Dorogovtsev S N and Mendes J F F 2000 *Europhys. Lett.* **52** 33
[25] Albert R and Barabási A L 2000 *Phys. Rev. Lett.* **85** 5234
[26] Fotouhi B and Rabbat M G 2013 *Eur. Phys. J.* B **86** 510
[27] Yu X, Li Z, Zhang D, Liang F, Wang X Y and Wu X 2006 *J. Phys. A: Math. Gen.* **39** 14343
[28] Jung S, Kim S and Kahng B 2002 *Phys. Rev.* E **65** 056101
[29] Dorogovtsev S N and Mendes J F 2002 (arXiv:cond-mat/0204102)
[30] Dorogovtsev S N and Mendes J F F 2000 *Phys. Rev.* E **62** 1842
[31] Zhu H, Wang X and Zhu J Y 2003 *Phys. Rev.* E **68** 056121
[32] Saavedra S, Reed-Tsochas F and Uzzi B 2008 *Proc. Natl Acad. Sci. USA* **105** 16466–71
[33] Chung F and Lu L 2004 *Internet Math.* **1** 409–61
[34] Cooper C, Frieze A and Vera J 2004 *Internet Math.* **1** 463–83
[35] Moore C, Ghoshal G and Newman M E 2006 *Phys. Rev.* E **74** 036121
[36] Bauke H, Moore C, Rouquier J B and Sherrington D 2011 *Eur. Phys. J.* B **83** 519–24
[37] Bianconi G and Barabási A L 2001 *Europhys. Lett.* **54** 436
[38] Wang M, Yu G and Yu D 2008 *Physica* A **387** 4692–8
[39] Medo M, Cimini G and Gualdi S 2011 *Phys. Rev. Lett.* **107** 238701
[40] Pham T, Sheridan P and Shimodaira H 2016 *Sci. Rep.* **6** 32558
[41] Ghoshal G, Chi L and Barabási A L 2013 *Sci. Rep.* **3** 2920
[42] Jung H, Lee J G, Lee N and Kim S H 2018 *J. Stat. Mech.* 123403
[43] Newman M E 2003 *SIAM Rev.* **45** 167–256
[44] Krapivsky P L and Redner S 2001 *Phys. Rev.* E **63** 066123
[45] Caldarelli G, Capocci A, De Los Rios P and Munoz M A 2002 *Phys. Rev. Lett.* **89** 258702
[46] Hanneke S *et al* 2010 *Electron. J. Stat.* **4** 585–605
[47] Snijders T A, Van de Bunt G G and Steglich C E 2010 *Soc. Netw.* **32** 44–60
[48] Peixoto T P and Rosvall M 2017 *Nat. Commun.* **8** 582
[49] Kim B *et al* 2018 *Stat. Surv.* **12** 105–35
[50] Sarkar P and Moore A W 2006 Dynamic social network analysis using latent space models *Advances in Neural Information Processing Systems* pp 1145–52
[51] Hoff P D, Raftery A E and Handcock M S 2002 *J. Am. Stat. Assoc.* **97** 1090–8
[52] Baum L E and Petrie T 1966 *Ann. Math. Stat.* **37** 1554–63
[53] Hoff P D 2005 *J. Am. Stat. Assoc.* **100** 286–95
[54] Holland P W, Laskey K B and Leinhardt S 1983 *Soc. Netw.* **5** 109–37
[55] Xing E P, Fu W and Song L 2010 *Ann. Appl. Stat.* **4** 535–66
[56] Mazzarisi P, Barucca P, Lillo F and Tantari D 2019 *Eur. J. Oper. Res.* **281** 50–65
[57] Lee J, Li G and Wilson J D 2017 (arXiv:1702.03632)
[58] Xu K S and Hero A O 2014 *IEEE J. Sel. Top. Signal. Process.* **8** 552–62
[59] Landherr A, Friedl B and Heidemann J 2010 *Bus. Inf. Syst. Eng.* **2** 371–85
[60] Clauset A, Shalizi C R and Newman M E 2009 *SIAM Rev.* **51** 661–703
[61] Gilks W R and Wild P 1992 *Appl. Stat.* **41** 337–48
[62] Louis T A 1982 *J. R. Stat. Soc.* B **44** 226–33
[63] Morelande M R and Garcia-Fernandez A F 2013 *IEEE Trans. Signal Process.* **61** 5477–84
[64] Rossi R and Ahmed N 2015 The network data repository with interactive graph analytics and visualization *Proc. Conf. AAAI Artificial Intelligence* pp 4292–3
[65] Viswanath B, Mislove A, Cha M and Gummadi K P 2009 On the evolution of user interaction in facebook *Proc. Online Social Networks* pp 37–42

J. Stat. Mech. (2020) 043407