PAPER *Special Section on Data Engineering and Information Management*

Modeling N-th Order Derivative Creation Based on Content Attractiveness and Time-Dependent Popularity

Kosetsu TSUKUDA[†a)], **Masahiro HAMASAKI**[†b)], ***Nonmembers*, *and* Masataka GOTO**[†c)], ***Member***

SUMMARY For amateur creators, it has been becoming popular to create new content based on existing original work: such new content is called derivative work. We know that derivative creation is popular, but *why* are individual derivative works created? Although there are several factors that inspire the creation of derivative works, such factors cannot usually be observed on the Web. In this paper, we propose a model for inferring latent factors from sequences of derivative work posting events. We assume a sequence to be a stochastic process incorporating the following three factors: (1) the original work's attractiveness, (2) the original work's popularity, and (3) the derivative work's popularity. To characterize content popularity, we use content ranking data and incorporate rank-biased popularity based on the creators' browsing behaviors. Our main contributions are three-fold. First, to the best of our knowledge, this is the first study modeling derivative creation activity. Second, by using real-world datasets of music-related derivative work creation, we conducted quantitative experiments and showed the effectiveness of adopting all three factors to model derivative creation activity and considering creators' browsing behaviors in terms of the negative logarithm of the likelihood for test data. Third, we carried out qualitative experiments and showed that our model is useful in analyzing following aspects: (1) derivative creation activity in terms of category characteristics, (2) temporal development of factors that trigger derivative work posting events, (3) creator characteristics, (4) N-th order derivative creation process, and (5) original work ranking.

key words: *user-generated content, derivative creation, latent variable model, music content*

1. Introduction

These days not only professional creators but also amateur creators who used to be just consumers can easily create content, which is known as user-generated content (UGC), and make them accessible via the Web. Since not all amateur creators can create new content from scratch, it is popular to use existing original (1st generation) work as the basis for new content: such content is called *derivative work* [1] or 2nd generation work. For example, on YouTube*, there are many videos in which amateur creators dance to an existing song, or perform a cover of it [2], [3]. To be more specific, there have been various cases where an original work gains more popularity with increasing its derivative works such as "Nyan Cat**" and "PPAP***" phenomena. Although original content creators could identify their derivative works by using the Content ID function and ask YouTube to delete them, they do not bother to do that. This is because gaining the popularity of an original work benefits its creator. Thingiverse**** is a Web service that facilitates derivative work creation, where amateur creators can share 3D model data intended for a 3D printer. On Thingiverse, it is popular for creators to download original 3D model data created by others, modify it, and upload their new version [4]. In this kind of derivative work creation activity, a creator influenced by 2nd generation content can create 3rd generation content. Similarly, N-th generation content can be transformed into N+1-th generation content. Such derivative work creation activity is called "N-th order derivative creation [5]."

When a creator creates a derivative work and uploads it to the Web, there are various factors that inspire the creation of the derivative work. However, since the factors that trigger derivative creation cannot usually be observed on the Web, they are difficult to detect. To get around this problem, we assume that when a creator creates a derivative work, triggering factors can be divided into two categories. The first one includes factors related to original work features. For example, a creator may cover an original song because he/she likes the melody, even if the song is not popular among consumers. The second one includes social factors. For example, a creator may dance to an original song only because the song is popular. Based on this assumption, we propose a probabilistic model to estimate the factors that triggered derivative work creation. More specifically, our model incorporates three factors: (1) an original work's attractiveness, (2) an original work's popularity, and (3) a derivative work's popularity. The details of each factor are given in Sect. 3.2. Since the relative influence of the three factors varies among creators (*e.g.*, one creator may put a high priority on factor (1), while another creator may put a high priority on factor (2)), our model also incorporates the latent relationships between creators and each of the three factors. Moreover, our model uses content ranking information to take into account the popularity of original and derivative works. By referring to the examination model of a Web search result [6], [7], we model popularity based on the hypothesis that higher ranked content has a larger influence because such content is, with high probability, viewed by many creators. By using efficient Bayesian inference based on the stochastic expectation-maximization (EM) al-

Manuscript received June 14, 2019.
Manuscript revised November 1, 2019.
Manuscript publicized February 5, 2020.

[†]The authors are with National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba-shi, 305–8568 Japan.
a) E-mail: k.tsukuda@aist.go.jp
b) E-mail: masahiro.hamasaki@aist.go.jp
c) E-mail: m.goto@aist.go.jp

DOI: 10.1587/transinf.2019DAP0008

*http://www.youtube.com

**http://www.youtube.com/watch?v=QH2-TGUlwu4

***http://www.youtube.com/watch?v=0E00Zuayv9Q

****http://www.thingiverse.com

gorithm [8], we can obtain the latent triggers for derivative work posts.

Modeling derivative creation activity is worth studying from various viewpoints.

- Our model can find original work that has a significant influence on derivative creation activity. This enables us to generate original work ranking based on popularity among creators, even though it is common to rank original works based on popularity among consumers (*e.g.*, rank original works based on the view count). Such a ranking enables consumers to search for original works from a new viewpoint.
- Given a category (*e.g.*, "3D models of chairs" or "music videos covering songs"), our model can show the characteristics of the category in derivative creation activity (*e.g.*, most creators put a high priority on original work attractiveness in a category). Understanding such characteristics in a category is important from the social scientific point of view.
- Our model can also show creator characteristics (*e.g.*, a creator puts a high priority on derivative work popularity). There are potentially many applications using this data such as ads and recommendation. For example, if a creator puts a high priority on original work attractiveness, it would be useful to recommend original works similar to the original works the creator used in the past to encourage more derivative work creation.

In Sect. 6, we discuss the application of our model to a real-world dataset and show that our model can be used to obtain this kind of information.

Our main contributions in this paper are summarized as follows [†].

- To the best of our knowledge, this is the first study modeling derivative creation activity. Our model can simultaneously take into account the influences of three factors: (1) original work attractiveness, (2) original work popularity, and (3) derivative work popularity.
- We describe the details of inference of model parameters based on the stochastic EM algorithm.
- We quantitatively evaluated our model by using derivative creation data of the music content. Our experimental results show that the model adopting all three factors achieves the best result in terms of the log likelihood computed by using test data. We also show that when we consider the content popularity based on popularity ranking, the method reflecting creators' browsing behaviors is the most effective to model derivative creation activity.
- We carried out qualitative experiments in terms of (1) category characteristics, (2) temporal development of factors that trigger the derivative work posting events, (3) creator characteristics, (4) N-th order derivative creation process, and (5) original work ranking, and showed that our model can be used to analyze derivative work creation activity.

The remainder of this paper is organized as follows. Section 2 describes related work in two areas: (1) analysis of derivative creation activity and (2) modeling influences in social communities. Section 3 describes the model that adopts the creator influence factor, which is used in related work, in addition to the aforementioned three factors. Section 4 presents a Bayesian inference procedure to infer the latent triggers for derivative work posts. Sections 5 and 6 report on our quantitative and qualitative experiments, respectively. Finally, Sect. 7 concludes this paper.

2. Related Work

2.1 Analysis of Derivative Creation Activity

A limited number of studies have investigated derivative creation activity. Eto *et al.* [10] developed a 3D modeling application and a model sharing Web service called Modulobe, which allows users to create 3D models from scratch or based on the work of other creators. They reported that 10.4% of models were parents of other models and the chains of creation reached four generations. Cheliotic and Yew [11] examined the remixing activity in the ccMixter online music community[††]. They reported that derivative creation greatly boosted the output of a community as well as increased the diversity of the output. Hamasaki *et al.* [1] analyzed derivative creation activity on Niconico[†††], which is one of the most popular video sharing Web services in Japan. They used explicit citation information between an original work and its derivative works and discussed certain statistics (*e.g.*, the number of derivative works of an original work). Hamasaki *et al.* [12] also developed a Web service called Songrium[††††] that helps a user browse original songs and their derivative works by visualizing their relations.

All the studies mentioned above analyzed *how* derivative works had been created by using a network based on the relationships between the original content and derivative works. In this work, we focus on *why* derivative works were created and propose a model to estimate the factors and their influences.

2.2 Modeling Influences in Social Communities

Since estimating influences among users in social activities is useful for various applications, such as influential user detection [13] and personalized recommendation [14],

[†] This paper is an extended version of a conference paper [9]. We extend the paper by describing the details of inference of model parameters and by carrying out the qualitative evaluations in terms of creator characteristics and original work ranking.

[††] http://ccmixter.org

[†††] http://www.nicovideo.jp

[††††] http://songrium.jp

many methods for estimating such influences have been proposed. One major approach is to use an information diffusion model such as the independent cascade model [15]. Although discrete time is assumed with this model, Saito *et al.* [16] proposed a model based on Poisson processes that allows for continuous time modeling. However, their model requires a network of users in which a node corresponds to a user and an edge between users represents the existence of influence. To overcome this limitation, Iwata *et al.* [8] proposed a model that discovers latent influences between users without a network. Although the cascade Poisson process [17] models a sequence of cascading events, the model proposed by Iwata *et al.* [8], which is called the Shared Cascade Poisson Process (SCPP), can handle multiple sequences of adoption events for multiple items by sharing parameters. Iwata *et al.* [8] used a Bayesian approach to discourage overfitting during parameter inference. They evaluated the model by using social bookmark data, where adopting an item corresponds to bookmarking a Web page. Tanaka *et al.* [18] extended the SCPP to estimate the factors that trigger item purchase events. They considered the users' view histories for TV advertisements in addition to influences between users and showed that the SCPP is also effective in modeling purchase events.

Our model extends the SCPP and the model proposed by Tanaka *et al.* [18], differing from them in the following two respects. First, in the other models, there is no need to consider the effect of adopted items such as bookmarked Web pages and purchased items. However, in derivative creation activity, adopted items (*i.e.*, derivative works) also influence other creators' creation activity. Therefore, we extended the SCPP so that we can handle the effect of both original works and derivative works. Second, although the other models assume that the popularity of items is constant regardless of time, we assume that content popularity depends on time. Hence, our model incorporates the time-dependent popularity of both original works and derivative works by considering content ranking data and the creators' ranking browsing behavior.

In our previous work [19], we implemented a public Web service for browsing the derivation factors called Songrium Derivation Factor Analysis, which was developed by applying our proposed model to the original works and derivative works uploaded to a video sharing service. Songrium Derivation Factor Analysis has several functions that could enable users to browse and watch videos from a new viewpoint and decide which content they want to use to create a new derivative work. In our previous paper [19], we focused on application interfaces realized by the proposed model, but did not evaluate the model itself. In this paper, we describe the details of the proposed model and conduct quantitative and qualitative evaluations to show the effectiveness of the model.

3. Model

In an online social activity model, it is common to consider user preference for content (we refer to the factor as original work attractiveness) and influences among users [8], [18]. However, in derivative creation activity, the existence of user influence is unlikely because no obvious influences among creators (users) have been observed in derivative creation activity analysis [1], [10], [11]. Instead, it seems that the rich-get-richer phenomenon [20] exists in the activity [1]. Hence, we assume that the popularity of the original and derivative works represents their exposure to creators and that it is an important factor in modeling derivative creation activity. Note that although we describe the *complete* model as incorporating four factors (original work attractiveness, creator influence, original work popularity, and derivative work popularity) in this section, our proposed model incorporates three of these (setting aside the creator influence factor).

3.1 Notations

In this section, we summarize the notations used in our model. Given a category (*e.g.*, "3D models of chairs" or "music videos covering songs") and observation time period T, let $\mathcal{I}$ be a set of original works posted to a Web service (*e.g.*, Thingiverse or YouTube) between time 0 and time T. Let $(t^{\mathrm{p}}_{ij}, u^{\mathrm{p}}_{ij})$ denote the jth derivative work posting event of original work i. More specifically, creator $u^{\mathrm{p}}_{ij} \in \mathcal{U}$ posts i's derivative work at time t^{p}_{ij}. Here, $\mathcal{U}$ is the set of creators. Without loss of generality, we assume that derivative work posting events are sorted in ascending order of their timestamps: $t^{\mathrm{p}}_{ij} \leq t^{\mathrm{p}}_{ij'}$ for $j < j'$. When J_i represents the total number of i's derivative works posted during the observation time period, a set of derivative work posting events of i is given by $\mathcal{D}_i = \{(t^{\mathrm{p}}_{ij}, u^{\mathrm{p}}_{ij})\}_{j=1}^{J_i}$. Hence, a set of derivative work posting events of all original works is given by $\mathcal{D} = \{\mathcal{D}_i\}_{i\in\mathcal{I}}$.

Suppose creators can see the ranking of original works on the Web service, where original works are ranked based on the popularity computed using statistics such as view count. Let (t^o_{ik}, r^o_{ik}) denote the kth ranked event of $i \in \mathcal{I}$. That is, i is ranked at the r^o_{ik}th place at time t^o_{ik}. We also assume that the events are sorted in ascending order of their timestamps without loss of generality: $t^o_{ik} \leq t^o_{ik'}$ for $k < k'$. Let K^o_i be the total number of i's ranked events between time 0 and time T, then a set of ranked events of i is given by $O_i = \{(t^o_{ik}, r^o_{ik})\}_{k=1}^{K^o_i}$. Therefore, a set of ranked events of all original works is given by $O = \{O_i\}_{i\in\mathcal{I}}$.

Similarly, suppose creators can also see the ranking of derivative works. In the same manner as with the ranked event of the original work, let (t^c_{ik}, r^c_{ik}) denote the kth ranked event of i's derivative work. Let K^c_i be the total number of ranked events of i's derivative works between time 0 and time T; then a set of ranked events of i's derivative works is given by $C_i = \{(t^c_{ik}, r^c_{ik})\}_{k=1}^{K^c_i}$. Note that C_i includes ranked events of all i's derivative works: (t^c_{ik}, r^c_{ik}) and $(t^c_{ik'}, r^c_{ik'})$ can be ranked events of different derivative works. Finally, a set of ranked events of all derivative works of all original works

Table 1 Notations used in our model

Symbol	Description
$\mathcal{I}$	set of original works
$\mathcal{U}$	set of creators
i	original work, $i \in \mathcal{I}$
u^{p}_{ij}	creator of jth derivative work posting event of original work i, $u^{\mathrm{p}}_{ij} \in \mathcal{U}$
t^{p}_{ij}	time of jth derivative work posting event of i
r^{o}_{ik}	rank of kth ranked event of i
t^{o}_{ik}	time of kth ranked event of i
r^{c}_{ik}	rank of kth ranked event of i's derivative work
t^{c}_{ik}	time of kth ranked event of i's derivative work
$\mathcal{D}$	set of derivative work posting events
$\mathcal{O}$	set of ranked events of original works
$\mathcal{C}$	set of ranked events of derivative works
T	observation period
J_i	number of derivative work posting events for i in $t \in [0, T]$
K^{o}_{i}	number of ranked events for i in $t \in [0, T]$
K^{c}_{i}	number of ranked events for derivative works of i in $t \in [0, T]$

is given by $\mathcal{C} = \{\mathcal{C}_i\}_{i \in \mathcal{I}}$.

Table 1 summarizes the notations used in this paper.

3.2 Factors

3.2.1 Original Work Attractiveness

A creator may create original work i's derivative work because he/she thinks that i is attractive even if it is not popular. The attractiveness of i can be due to i's various features; in the case of a song, the features can be the melody, beat, lyrics, etc. We assume that each creator has a different preference for original content attractiveness. For example, a creator may put a high priority on original work attractiveness when he/she decides whether or not to create a derivative work of i, while another creator may put a low priority on it. We also assume that the post rate based on original work attractiveness is constant in the time period from 0 to T as described in Fig. 1(a). Here, the rate at time t represents the instantaneous probability of a creator posting i's derivative work at t. This kind of constant rate is known as the "background rate" in the point process framework [21]. Based on these assumptions, we model the rate at which creator u posts i's derivative work triggered by i's attractiveness as follows:

$$f_i(u) = \alpha_i \theta_{0u}, \tag{1}$$

where $\alpha_i \geq 0$ is the original work attractiveness. In other words, α_i is the rate at which i's derivative work is posted without being triggered by the preceding events, and $\theta_{0u} \geq 0$ represents the probability that u is influenced by original work attractiveness when he/she creates a derivative work, and $\sum_{u \in \mathcal{U}} \theta_{0u} = 1$. If u puts a higher priority on original work attractiveness than other factors, θ_{0u} becomes large. In Fig. 1(a), the height of the blue line corresponds to $\alpha_i \theta_{0u}$.

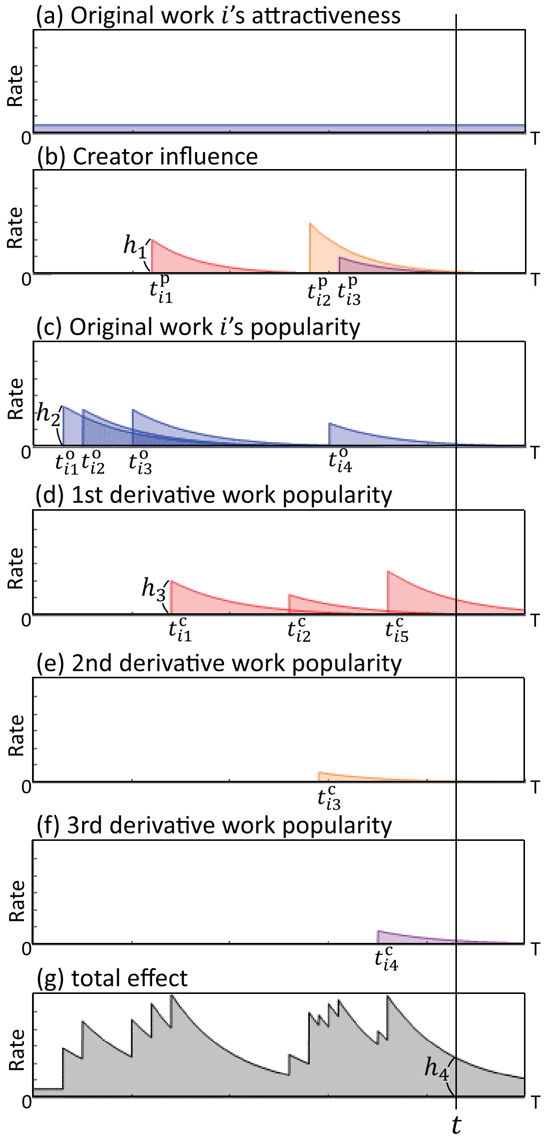

Fig. 1 Rate at which creator u posts original work i's derivative work at time t.

3.2.2 Creator Influence

Creator u may create original work i's derivative work because creator u' posted i's derivative work; in other words, u is influenced by u'. We assume that the influences of u' on other creators are different from one creator to another. For example, if u is a fan of u', u' has a larger influence on u than on other creators. We also assume that a creator's influence on another creator decays over time. This assumption is often used to model information diffusion processes between users [22], [23]. Based on these assumptions, we model the rate at which u posts i's derivative work at time t based on the influence of u' who posted i's derivative work at time t' as follows:

$$g_{(i,t',u')}(t, u) = \begin{cases} \alpha_{u'} \theta_{u'u} \mathrm{e}^{-\gamma_{\mathrm{p}}(t-t')} & \text{if} \quad t' < t \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\alpha_{u'} \geq 0$ is the influence of u' on other creators, $\theta_{u'u} \geq 0$ represents the strength of the relation between u' and u, and $\sum_{u \in \mathcal{U} \backslash u'} \theta_{u'u} = 1$, where $\mathcal{U} \backslash u'$ is the set of creators excluding u'. Hence, $\alpha_{u'}\theta_{u'u}$ means the influence of u' on u. Finally, $\mathrm{e}^{-\gamma_{\mathrm{p}}(t-t')}$ models the decay of influence over time with decay parameter $\gamma_{\mathrm{p}} \geq 0$. Note that if u' posts i's derivative work after u, u' does not influence u: $g_{(i,t',u')}(t,u) = 0$ if $t' \geq t$.

In Fig. 1(b), three creators post original work i's derivative works. Let the first creator (shown in red) be u'. The influence of u' is $\alpha_{u'}\theta_{u'u}$, which corresponds to h_1 in the figure, when u' posts the derivative work. The influence decreases as time proceeds.

In derivative creation activity, the derivative work's attractiveness may also have an influence. We assume that the derivative work's attractiveness is determined by the creator of the derivative work. For example, when the creator creates a derivative work by covering an original work's song, we think the derivative work's attractiveness depends on the creator's singing voice in the derivative work. Therefore, by considering the creator's influence, we can also consider the derivative work's attractiveness.

3.2.3 Original Work Popularity

If original work i is popular among consumers, creator u may create i's derivative work because his/her derivative work might also become popular. As mentioned in Sect. 3.1, we assume creators can see the popularity ranking of original works. When two original works are ranked, we hypothesize that the higher ranked one has a larger influence than the lower ranked one. This hypothesis comes from the position bias in the Web search: it has been proved that higher ranked results receive more user attention and have larger probabilities of being examined during search sessions [6], [7]. In addition, we assume that each creator has a different preference for original work popularity: one creator may be susceptible to popularity and put a high priority on original work popularity when he/she decides whether to create a derivative work of the original work, while another creator may not. As is the case with creator influences, we also assume that the influence of original work popularity on a creator decays over time. Based on these assumptions, we model the rate at which u posts i's derivative work at time t based on the influence of i's popularity as follows:

$$h_{o(i,t',r')}(t,u) = \begin{cases} rb(r')\omega_i\theta_{-1u}\mathrm{e}^{-\gamma_{\mathrm{o}}(t-t')} & \text{if} \quad t' < t \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where r' represents the rank of i at time t', and function rb computes the rank bias. As reported in studies on behavior analysis of search result examination, the probability that each ranked item is viewed dramatically decreases as the rank drops [6], [7]. Based on the examination behavior, we compute the rank bias as $rb(r') = \frac{1}{r'}$. In Sect. 5.3, we evaluate the usefulness of rank bias. The term $\omega_i \geq 0$ represents the influence of i's popularity, $\theta_{-1u} \geq 0$ represents the probability that u is influenced by original work popularity when he/she creates a derivative work, and $\sum_{u \in \mathcal{U}} \theta_{-1u} = 1$. Finally, $\mathrm{e}^{-\gamma_{\mathrm{o}}(t-t')}$ models the decay of influence over time with decay parameter $\gamma_{\mathrm{o}} \geq 0$.

In Fig. 1(c), the original work i appears four times in the popularity ranking. Let r' be the rank of the first ranked event. The influence of the event is $rb(r')\omega_i\theta_{-1u}$, which corresponds to h_2 in Fig. 1(c) at t^o_{i1}. Then, the influence decreases as time proceeds.

3.2.4 Derivative Work Popularity

If original work i's derivative work created by creator u' is popular among consumers, creator u may also create i's derivative work because his/her derivative work might also become popular even if u is not a fan of u'. As mentioned in Sect. 3.1, we assume creators can see the popularity ranking of derivative works. Based on similar assumptions and the hypothesis described in Sect. 3.2.3, when i's derivative work was ranked r'th at time t', we model the rate at which u posts i's derivative work at time t based on the influence of i's derivative work popularity as follows:

$$h_{d(i,t',r')}(t,u) = \begin{cases} rb(r')\sigma_i\theta_{-2u}\mathrm{e}^{-\gamma_{\mathrm{d}}(t-t')} & \text{if} \quad t' < t \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $\sigma_i \geq 0$ represents the influence of the popularity of i's derivative work, $\theta_{-2u} \geq 0$ represents the probability that u is influenced by derivative work popularity when he/she creates a derivative work, and $\sum_{u \in \mathcal{U}} \theta_{-2u} = 1$. Finally, $\mathrm{e}^{-\gamma_{\mathrm{d}}(t-t')}$ models the decay of influence over time with decay parameter $\gamma_{\mathrm{d}} \geq 0$.

Figure 1(d), (e), and (f) show the influences of i's first, second, and third derivative work popularity, respectively: the first derivative work appears three times in the ranking, while the second and third ones appear one time. Let r' be the rank of the first ranked event in Fig. 1(d). The influence of the first ranked event is $rb(r')\sigma_i\theta_{-2u}$, which corresponds to h_3 in Fig. 1(d) at t^c_{i1}. Then, the influence decreases as time proceeds.

3.3 Derivative Work Post Rate

Based on the factors described in Sects. 3.2.1 to 3.2.4, the rate at which u posts i's derivative work at t is given by:

$$\begin{aligned} \lambda_i(t,u) = f_i(u) &+ \sum_{(t',u') \in \mathcal{D}_{it\backslash u}} g_{(i,t',u')}(t,u) \\ &+ \sum_{(t',r') \in O_{it}} h_{o(i,t',r')}(t,u) + \sum_{(t',r') \in C_{it}} h_{d(i,t',r')}(t,u), \end{aligned} \tag{5}$$

where $\mathcal{D}_{it\backslash u} = \{(t',u') | (t',u') \in \mathcal{D}_i \text{ and } t' < t \wedge u' \neq u\}$ is the set of derivative work posting events before t excluding u's one; $O_{it} = \{(t',r') | (t',r') \in O_i \text{ and } t' < t\}$ is the set of ranked events of i before t; and $C_{it} = \{(t',r') | (t',r') \in C_i \text{ and } t' < t\}$ is the set of ranked events of i's derivative works before t. Here, $\lambda_i(t,u)$ corresponds to h_4 in Fig. 1(g).

4. Inference

Given derivative work posting events $\mathcal{D}$, original works ranked events $\mathcal{O}$, and derivative works ranked events $\mathcal{C}$, we infer the model parameters in Table 2 by using the stochastic EM algorithm. Following Iwata *et al.* [8], we assume that a set of i's derivative work posting events $\mathcal{D}_i$ is generated from a marked point process [24] at a rate of $\lambda_i(t,u)$. Based on this assumption, the likelihood of the function of $\mathcal{D}$ is described as follows:

$$P(\mathcal{D}|\mathcal{O},\mathcal{C},\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\sigma},\boldsymbol{\Theta},\boldsymbol{\gamma}) = \prod_{i\in\mathcal{I}} \exp\left(-\int_0^T \sum_{u\in\mathcal{U}} \lambda_i(t,u)dt\right) \prod_{j=1}^{J_i} \lambda_i(t_{ij}^{\mathrm{p}}, u_{ij}^{\mathrm{p}}), \tag{6}$$

where $\boldsymbol{\alpha} = \{\alpha_l\}_{l\in\mathcal{I}\cup\mathcal{U}}$, $\boldsymbol{\omega} = \{\omega_i\}_{i\in\mathcal{I}}$, $\boldsymbol{\sigma} = \{\sigma_i\}_{i\in\mathcal{I}}$, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_u\}_{u\in\mathcal{U}_+}$, $\boldsymbol{\theta}_u = \{\theta_{uu'}\}_{u'\in\mathcal{U}\backslash u}$, and $\boldsymbol{\gamma} = \{\gamma_{\mathrm{p}}, \gamma_{\mathrm{o}}, \gamma_{\mathrm{d}}\}$. Here, $\mathcal{U}_+$ denotes $\mathcal{U}\cup\{0,-1,-2\}$, where 0, −1, and −2 represent virtual creators who are used for original work attractiveness, original work popularity, and derivative work popularity, respectively. The term $\exp\left(-\int_0^T \sum_{u\in\mathcal{U}} \lambda_i(t,u)dt\right)$ represents the probability that no creator posts i's derivative work between time 0 and time T. The integral part can be analytically calculated as follows:

$$\begin{aligned}\int_0^T \sum_{u\in\mathcal{U}} \lambda_i(t,u)dt = \alpha_i T &+ \frac{1}{\gamma_{\mathrm{p}}} \sum_{j=1}^{J_i} \alpha_{u_{ij}} \left(1 - \mathrm{e}^{-\gamma_{\mathrm{p}}(T - t_{ij}^{\mathrm{p}})}\right) \\ &+ \frac{\omega_i}{\gamma_{\mathrm{o}}} \sum_{k=1}^{K_i^o} rb(r_{ik}^o)\left(1 - \mathrm{e}^{-\gamma_{\mathrm{o}}(T - t_{ik}^{\mathrm{o}})}\right) \\ &+ \frac{\sigma_i}{\gamma_{\mathrm{d}}} \sum_{k=1}^{K_i^c} rb(r_{ik}^c)\left(1 - \mathrm{e}^{-\gamma_{\mathrm{d}}(T - t_{ik}^{\mathrm{c}})}\right).\end{aligned} \tag{7}$$

Following Iwata *et al.* [8], we introduce latent variables $z_{ij} \in \{0, 1, \cdots, \left|\mathcal{D}_{it\backslash u}\right| + |\mathcal{O}_{it}| + |\mathcal{C}_{it}|\}$ to indicate the index of the latent trigger of the jth derivative work posting event of original work i. The terms $z_{ij} = 0$, $\left|\mathcal{D}_{it\backslash u}\right| + 1 \leq z_{ij} \leq \left|\mathcal{D}_{it\backslash u}\right| + |\mathcal{O}_{it}|$, $\left|\mathcal{D}_{it\backslash u}\right| + |\mathcal{O}_{it}| + 1 \leq z_{ij} \leq \left|\mathcal{D}_{it\backslash u}\right| + |\mathcal{O}_{it}| + |\mathcal{C}_{it}|$ indicate that the event was triggered due to the influence of original work attractiveness, original work popularity, and derivative work popularity, respectively, and $z_{ij} = j'$ $(1 \leq j' \leq \left|\mathcal{D}_{it\backslash u}\right|)$ indicates that the event was triggered due to the influence of the creator who posted the j'th derivative work of i. By using the latent variables, the derivative work post rate in Eq. (5) can be written as $\lambda_i(t,u) = \sum_z \lambda_i(t,u,z)$, where

$$\lambda_i(t,u,z) = \begin{cases} f_i(u) & \text{if } \; z = 0 \\ g_{(i,t_{iz}^{\mathrm{p}},u_{iz}^{\mathrm{p}})}(t,u) & \text{if } \; 1 \leq z \leq \left|\mathcal{D}_{it\backslash u}\right| \\ h_{o(i,t_{iz'}^{o},r_{iz'}^{o})}(t,u) & \text{if } \; \left|\mathcal{D}_{it\backslash u}\right| + 1 \leq z \leq \left|\mathcal{D}_{it\backslash u}\right| + |\mathcal{O}_{it}| \\ h_{d(i,t_{iz''}^{c},r_{iz''}^{c})}(t,u) & \text{if } \; \left|\mathcal{D}_{it\backslash u}\right| + |\mathcal{O}_{it}| + 1 \leq z. \end{cases} \tag{8}$$

Here, $z' = z - \left|\mathcal{D}_{it\backslash u}\right|$ and $z'' = z - \left|\mathcal{D}_{it\backslash u}\right| - |\mathcal{O}_{it}|$.

By combining Eqs. (6), (7), and (8), the joint distribution of $\mathcal{D}$ and latent variables $\mathcal{Z} = \{\{z_{ij}\}_{j=1}^{J_i}\}_{i\in\mathcal{I}}$ is given by:

$$\begin{aligned}&P(\mathcal{D},\mathcal{Z}|\mathcal{O},\mathcal{C},\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\sigma},\boldsymbol{\Theta},\boldsymbol{\gamma}) \\ &= \prod_{i\in\mathcal{I}} \exp\left[\alpha_i T + \frac{1}{\gamma_{\mathrm{p}}} \sum_{j=1}^{J_i} \alpha_{u_{ij}} \left(1 - \mathrm{e}^{-\gamma_{\mathrm{p}}(T - t_{ij}^{\mathrm{p}})}\right)\right. \\ &+ \frac{\omega_i}{\gamma_{\mathrm{o}}} \sum_{k=1}^{K_i^o} rb(r_{ik}^o)\left(1 - \mathrm{e}^{-\gamma_{\mathrm{o}}(T - t_{ik}^{\mathrm{o}})}\right) \\ &\left.+ \frac{\sigma_i}{\gamma_{\mathrm{d}}} \sum_{k=1}^{K_i^c} rb(r_{ik}^c)\left(1 - \mathrm{e}^{-\gamma_{\mathrm{d}}(T - t_{ik}^{\mathrm{c}})}\right)\right] \prod_{j=1}^{J_i} \lambda_i(t_{ij}^{\mathrm{p}}, u_{ij}^{\mathrm{p}}, z_{ij}).\end{aligned} \tag{9}$$

We assume a Gamma prior for each of the original work attractiveness scores α_i as follows:

$$P(\alpha_i|a,b) = \frac{1}{\Gamma(a)} b^a \alpha_i^{a-1} \exp(-b\alpha_i), \tag{10}$$

where a and b are hyperparameters. In this study, following Iwata *et al.* [8], we set $a = b = 1$. We also assume a Gamma prior for each creator influence α_u, original work popularity ω_i, and derivative work popularity σ_i. In addition, we assume a Dirichlet prior over $\boldsymbol{\theta}_u$, $u \in \mathcal{U}_+$ as follows:

$$P(\boldsymbol{\theta}_u|\beta) = \frac{\Gamma(\beta\,|\mathcal{U}|)}{\Gamma(\beta)^{|\mathcal{U}|}} \prod_{u'\in\mathcal{U}\backslash u} \theta_{uu'}^{\beta-1} \tag{11}$$

We use a Gamma prior for $\boldsymbol{\alpha}$, $\boldsymbol{\omega}$, and $\boldsymbol{\sigma}$, and a Dirichlet prior for $\boldsymbol{\Theta}$ to analytically calculate the marginalization over the parameters. The marginalized joint distribution is computed by integrating out those parameters as follows:

$$\begin{aligned}&P(\mathcal{D},\mathcal{Z}|\mathcal{O},\mathcal{C},\boldsymbol{\gamma},\beta,a,b) \\ &= \iiiint P(\mathcal{D},\mathcal{Z}|\mathcal{O},\mathcal{C},\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\sigma},\boldsymbol{\Theta},\boldsymbol{\gamma})P(\boldsymbol{\alpha}|a,b) \\ &\quad \times P(\boldsymbol{\omega}|a,b)P(\boldsymbol{\sigma}|a,b)P(\boldsymbol{\Theta}|\beta)d\boldsymbol{\alpha}d\boldsymbol{\omega}d\boldsymbol{\sigma}d\boldsymbol{\Theta} \\ &\propto \exp\left(-\sum_{i\in\mathcal{I}} \sum_{j:z_{ij}\neq 0} \eta(z_{ij})(t_{ij}^{\mathrm{p}} - t_{iz_{ij}})\right) \\ &\times \prod_{i\in\mathcal{I}} \prod_{k=1}^{K_i^o} rb(r_{ik}^o)^{M_{ik}} \prod_{i\in\mathcal{I}} \prod_{k=1}^{K_i^c} rb(r_{ik}^c)^{N_{ik}} \\ &\times \prod_{i\in\mathcal{I}} \frac{\Gamma(L_i + a)}{(T + b)^{L_i + a}} \prod_{u\in\mathcal{U}} \frac{\Gamma(L_u + a)}{(R_u + b)^{L_u + a}} \\ &\times \prod_{i\in\mathcal{I}} \frac{\Gamma(M_i + a)}{(R_i^o + b)^{M_i + a}} \prod_{i\in\mathcal{I}} \frac{\Gamma(N_i + a)}{(R_i^c + b)^{N_i + a}}\end{aligned}$$

Table 2 Parameters of proposed model

Symbol	Description
α_i	original work attractiveness of i, $\alpha_i \geq 0$
α_u	influence of u, $\alpha_u \geq 0$
ω_i	popularity of i, $\omega_i \geq 0$
σ_i	popularity of i's derivative work, $\sigma_i \geq 0$
$\theta_{u'u}$	transition probability from u' to u, $\theta_{u'u} \geq 0$, $\sum_{u\in\mathcal{U}\backslash u'} \theta_{u'u} = 1$
$\gamma_{\mathrm{p}}, \gamma_{\mathrm{o}}, \gamma_{\mathrm{d}}$	decay parameter, $\gamma_{\mathrm{p}} \geq 0$, $\gamma_{\mathrm{o}} \geq 0$, $\gamma_{\mathrm{d}} \geq 0$

$$\times \left(\frac{\Gamma(\beta\,|\mathcal{U}|)}{\Gamma(\beta)^{|\mathcal{U}|}}\right)^{|\mathcal{U}_+|} \prod_{u\in\mathcal{U}_+} \frac{\prod_{u'\in\mathcal{U}\backslash u}\Gamma(L_{uu'}+\beta)}{\Gamma(L_u+\beta\,|\mathcal{U}|)}, \tag{12}$$

where

$$\eta(z_{ij}) = \begin{cases} \gamma_{\mathrm{p}} & \text{if } \ 1 \le z_{ij} \le \left|\mathcal{D}_{it\backslash u}\right| \\ \gamma_{\mathrm{o}} & \text{if } \ \left|\mathcal{D}_{it\backslash u}\right| + 1 \le z_{ij} \le \left|\mathcal{D}_{it\backslash u}\right| + |O_{it}| \\ \gamma_{c} & \text{if } \ \left|\mathcal{D}_{it\backslash u}\right| + |O_{it}| + 1 \le z_{ij}. \end{cases} \tag{13}$$

Here, M_{ik} and N_{ik} are the number of posting events triggered by the kth ranked event of i and kth ranked event of i's derivative work, respectively. The terms $M_i = \sum_{k=1}^{K_i^o} M_{ik}$ and $N_i = \sum_{k=1}^{K_i^c} N_{ik}$ represent the total number of posting events triggered by ranked events of i and ranked events of i's derivative work, respectively. Furthermore, $L_i = \sum_{j=1}^{J_i} \delta(z_{ij}, 0)$ is the number of posting events triggered by i's attractiveness, where $\delta(x, y) = 1$ if $x = y$, and $\delta(x, y) = 0$ otherwise. The term $L_{uu'} = \sum_{i\in\mathcal{I}} \sum_{j=1}^{J_i} \delta(u_{iz_{ij}}, u)\delta(u_{ij}, u')$ is the number of posting events where u triggered u''s post, and $L_u = \sum_{u'\in\mathcal{U}\backslash u} L_{uu'}$ is the total number of posting events triggered by u. Here, $u_{iz_{ij}} = 0$ if $z_{ij} = 0$, $u_{iz_{ij}} = -1$ if $\left|\mathcal{D}_{it\backslash u}\right| + 1 \le z_{ij} \le \left|\mathcal{D}_{it\backslash u}\right| + |O_{it}|$, and $u_{iz_{ij}} = -2$ if $\left|\mathcal{D}_{it\backslash u}\right| + |O_{it}| + 1 \le z_{ij}$, which represent virtual creators. In addition, R_u for each creator $u \in \mathcal{U}$ is given by:

$$R_u = \frac{1}{\gamma_{\mathrm{p}}} \sum_{t\in\mathcal{D}_u} \left(1 - \mathrm{e}^{-\gamma_{\mathrm{p}}(T-t)}\right), \tag{14}$$

where $\mathcal{D}_u$ is a set of timestamps of derivative work posting events of u, and R_i^o and R_i^c for i are computed as follows:

$$R_i^o = \frac{1}{\gamma_{\mathrm{o}}} \sum_{k=1}^{K_i^o} rb\left(r_{ik}^o\right)\left(1 - \mathrm{e}^{-\gamma_{\mathrm{o}}\left(T-t_{ik}^o\right)}\right), \tag{15}$$

$$R_i^c = \frac{1}{\gamma_c} \sum_{k=1}^{K_i^c} rb\left(r_{ik}^c\right)\left(1 - \mathrm{e}^{-\gamma_c\left(T-t_{ik}^c\right)}\right). \tag{16}$$

Based on the marginalized joint distribution in Eq. (12), we developed a stochastic EM procedure for the iteration. In the E-step, given the current state of all but one variable z_{ij}, the new latent assignment of z_{ij} is sampled from the following probability:

$$\begin{aligned} &P(z_{ij} = y|\mathcal{D}, \mathcal{Z}_{\backslash ij}, O, C, \boldsymbol{\gamma}, \beta, a, b) \\ &\quad\propto \frac{P\left(\mathcal{D}, \mathcal{Z}_{\backslash ij}, z_{ij} = y|O, C, \boldsymbol{\gamma}, \beta, a, b\right)}{P\left(\mathcal{D}_{\backslash ij}, \mathcal{Z}_{\backslash ij}|O, C, \boldsymbol{\gamma}, \beta, a, b\right)}, \end{aligned} \tag{17}$$

where $y \in \{0, 1, \cdots, \left|\mathcal{D}_{it\backslash u}\right| + |O_{it}| + |C_{it}|\}$, and $\backslash ij$ represent the procedure excluding the jth derivative work posting event of i.

In the M-step, we estimate the decay parameters $\boldsymbol{\gamma}$ and Dirichlet parameter β by maximizing the logarithm of the joint likelihood in Eq. (12). We estimate $\boldsymbol{\gamma}$ using Newton's method. For example, the update rule of γ_{p} is given by:

$$\gamma_{\mathrm{p}} \leftarrow \gamma_{\mathrm{p}} - \frac{\partial S(\gamma_{\mathrm{p}})/\partial\gamma_{\mathrm{p}}}{\partial^2 S(\gamma_{\mathrm{p}})/\partial^2\gamma_{\mathrm{p}}}, \tag{18}$$

where $S(\gamma_{\mathrm{p}})$ is given by:

$$\begin{aligned} S(\gamma_{\mathrm{p}}) = &-\gamma_{\mathrm{p}} \sum_{i\in\mathcal{I}} \sum_{j:1\le z_{ij}\le\left|\mathcal{D}_{it\backslash u}\right|} (t_{ij}^{\mathrm{p}} - t_{iz_{ij}}) \\ &- \sum_{u\in\mathcal{U}} (L_u + a)\log(R_u + b). \end{aligned} \tag{19}$$

The β is estimated using the fixed point iteration method [23]. The update rule is given by:

$$\beta \leftarrow \beta \frac{\sum_{u\in\mathcal{U}_+} \sum_{u'\in\mathcal{U}\backslash u} (\Psi(L_{uu'} + \beta) - \Psi(\beta))}{|\mathcal{U}| \sum_{u\in\mathcal{U}_+} (\Psi(L_u + \beta\,|\mathcal{U}|) - \Psi(\beta\,|\mathcal{U}|))}, \tag{20}$$

where Ψ is the digamma function.

Finally, we can make the point estimates of the integrated out parameters as follows:

$$\hat{\alpha}_i = \frac{L_i + a}{T + b}, \quad \hat{\alpha}_u = \frac{L_u + a}{R_u + b}, \tag{21}$$

where $\hat{\alpha}_i$ and $\hat{\alpha}_u$ can be used to find attractive original works and influential creators, respectively,

$$\hat{\omega}_i = \frac{M_i + a}{R_i^o + b}, \quad \hat{\sigma}_i = \frac{N_i + a}{R_i^c + b}, \tag{22}$$

where $\hat{\omega}_i$ and $\hat{\sigma}_i$ can be used to find popular original works and popular derivative works, respectively, and

$$\hat{\theta}_{uu'} = \begin{cases} \frac{L_{uu'}+\beta}{L_u+\beta|\mathcal{U}|} & \text{if } \ u = 0 \lor u = -1 \lor u = -2 \\ \frac{L_{uu'}+\beta}{L_u+\beta(|\mathcal{U}|-1)} & \text{otherwise,} \end{cases} \tag{23}$$

which can be used to analyze influences between creators including virtual creators.

5. Quantitative Experiments

In this section, we answer the following research questions based on our quantitative experimental results:

RQ1 Is adopting three factors, which are original work attractiveness, original work popularity, and derivative work popularity, effective to model derivative creation activity? (Sect. 5.2)

RQ2 What kinds of ranking bias methods are effective to model derivative creation activity? (Sect. 5.3)

5.1 Dataset

In our experiments, we used derivative creation activity data of music content on Niconico, which is one of the most popular video sharing Web services in Japan. On Niconico, any user can upload and view videos, and derivative creation activity of music content occurs frequently: according to Songrium, as of the end of May 2019, more than 300,000 original song videos and more than 680,000 derivative videos had been uploaded to Niconico. Most original songs are created using singing synthesizer software called VOCALOID [25]; we restricted ourselves to

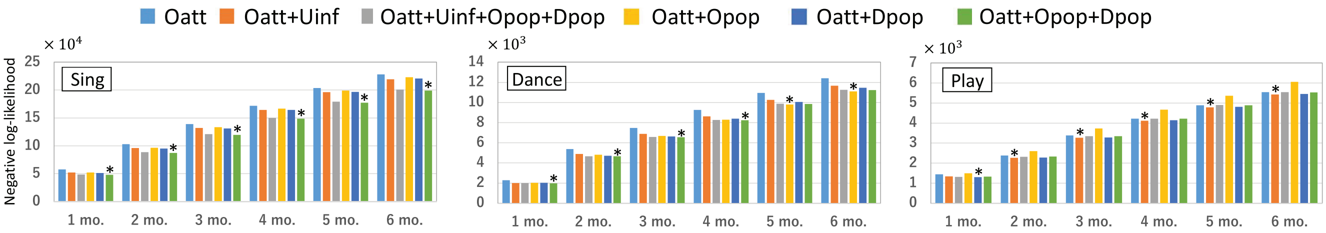

Fig. 2 Negative logarithm of likelihood of each model. Vertical axis and horizontal axis represent negative log-likelihood and test periods (*e.g.*, "1 mo." is first set of test data), respectively.

Table 3 Statistics of our dataset

Category	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{O}\rvert$	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{C}\rvert$	$\lvert\mathcal{U}\rvert$
Sing	4,035	64,973	199,320	67,627	18,715
Dance	396	30,925	9,420	22,954	1,153
Play	583	38,726	5,526	20,492	692

original song videos of this type. With respect to derivative works, Niconico maintains three categories of derivative works: (1) sing: covering an original song, (2) dance: dancing to an original song, and (3) play: playing an original song on a musical instrument such as a guitar or piano. We crawled original songs (*i.e.*, original works) posted between 1/1/2010 and 3/31/2013 and their derivative works posted between 1/1/2010 and 9/30/2013. Data between 1/1/2010 and 3/31/2013 were used as training data and data between 4/1/2013 and 9/30/2013 were used as test data. In each category, we eliminated original works that had fewer than two derivative works and creators who posted fewer than three derivative works during the training period.

We also collected ranking data. On Niconico, users can see the top 100 daily ranking of original songs and the top 100 daily rankings for derivatives in each of the sing, dance, and play categories. Ranking data on one day is created based on several statistics of the previous day (*e.g.*, view count and comment count) so that the ranking data represent the work's aggregated popularity. We crawled the top 100 ranking data in each of the original song and three derivative content categories between 1/1/2010 and 9/30/2013. Since only daily ranking data is available on Niconico, the timestamp in all our experiments is measured in days.

Table 3 lists the statistics of the dataset used in the experiments.

5.2 Combination of Factors

5.2.1 Settings

[Comparison Models] Recall that we introduced four factors that can be used to model derivative creation activity. Hereafter, let Oatt, Uinf, Opop, and Dpop denote original work attractiveness, creator influence, original work popularity, and derivative work popularity, respectively. As mentioned in Sect. 3, we hypothesize that a model adopting Oatt, Opop, and Dpop is the most effective. To evaluate this hypothesis, the following six models were compared: (1) Oatt, (2) Oatt+Uinf, (3) Oatt+Uinf+Opop+Dpop, (4) Oatt+Opop, (5) Oatt+Dpop, and (6) Oatt+Opop+Dpop, where Oatt+Uinf, for example, represents the model that combines the factors of Oatt and Uinf. Among the six models, (2) corresponds to SCPP [8] and (6) is our proposed model.

[Evaluation Metric] Predictive performance is one of the most commonly used metrics to evaluate the appropriateness of a learned model [8], [26]. Predictive performance is computed using the negative logarithm of the likelihood for posting events (t, u) during the test period from T to T'. The logarithm of the likelihood is given by:

$$L = \sum_{i\in\mathcal{I}}\left(-\int_{T}^{T'}\sum_{u\in\mathcal{U}}\lambda_i(t,u)dt\right)\sum_{(t,u)\in\mathcal{D}_i^{\text{test}}}\log\lambda_i(t,u), \quad (24)$$

where $\mathcal{D}_i^{\text{test}}$ is the test data for i. When the value of $-L$ is small, the predictive performance is high. To examine the influence of the length of the test period on our results, we examined spans of test data from one month (4/1/2013 to 4/30/2013) up to six months (4/1/2013 to 9/30/2013). In every case, the model was trained using data between 1/1/2010 and 3/31/2013, and the test period began on 4/1/2013.

5.2.2 Experimental Results

Figure 2 shows the negative log-likelihood during each test period in each category. During each test period, the model that achieved the best performance is marked with "*". In the "sing" category, Oatt+Opop+Dpop exhibited the best result for all test periods. In the "dance" and "play" categories, it did not perform the best for all test periods, but it stably exhibited high performance in all categories during all test periods. To evaluate the stability of the models, we computed the average rank of each model over six test periods in each category. We also computed the average rank of each model over 3 categories × 6 test periods = 18 test periods. Table 4 lists the results. We can see that Oatt+Opop+Dpop achieved the highest average rank over 18 test periods with relatively small standard deviation. Although Oatt+Uinf exhibited the best results for five test periods in the "play" category, the results in Table 4 indicate the instability of the model because its average ranks in both "sing" and "dance" categories were low. By comparing Oatt+Uinf+Opop+Dpop with Oatt and Oatt+Uinf, we can show the usefulness of adopting all four factors rather than adoption only Oatt or Oatt+Uinf. Similarly,

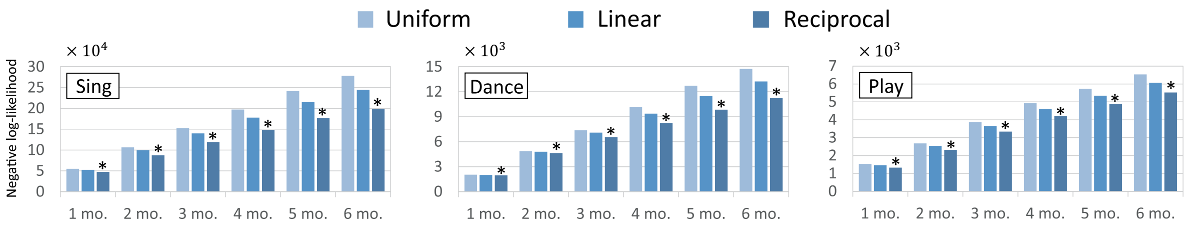

Fig. 3 Negative logarithm of likelihood of each ranking bias method.

Table 4 Average rank of each model over six test periods in each category. "All" means average rank over 18 test periods. Numbers in parentheses indicate standard deviations.

Model	Sing	Dance	Play	All
Oatt	6.0 (0.0)	6.0 (0.0)	4.5 (0.76)	5.5 (0.83)
Oatt+Uinf	3.7 (0.47)	4.7 (0.75)	1.5 (1.12)	3.28 (1.56)
Oatt+Uinf+Opop+Dpop	2.0 (0.0)	2.3 (0.47)	3.8 (1.07)	2.72 (1.04)
Oatt+Opop	5.0 (0.0)	3.0 (1.53)	6.0 (0.0)	4.67 (1.53)
Oatt+Dpop	3.3 (0.47)	3.7 (0.47)	1.8 (0.37)	2.94 (0.91)
Oatt+Opop+Dpop	1.0 (0.0)	1.3 (0.47)	3.3 (0.47)	1.89 (1.1)

comparison results between Oatt+Opop, Oatt+Dpop, and Oatt+Opop+Dpop indicate the usefulness of adopting both Opop and Dpop. Finally, by comparing Oatt+Opop+Dpop and Oatt+Uinf+Opop+Dpop, we can conclude that our proposed model adopting Oatt, Opop, and Dpop is the most effective to model derivative creation activity.

5.3 Ranking Bias Method Comparison

5.3.1 Settings

As mentioned in Sect. 3.2.3, our model uses the reciprocal rank method to bias the content ranking based on the creators' browsing behaviors of a ranked list (hereafter, "Reciprocal"). To evaluate its effectiveness, we compared it with the following two methods. The first method, "Linear," linearly decreases the ranking bias:

$$rb(r^o_{ik}) = \frac{101 - r^o_{ik}}{100}. \tag{25}$$

Here, $rb(r^c_{ik})$ is also computed in the same manner. With this method, it is assumed that content influence does not dramatically decrease when the content position in the ranking decreases compared to Reciprocal. The second method, "Uniform," does not take into account the ranking bias: $rb(r^o_{ik}) = rb(r^c_{ik}) = 1$ regardless of the content rank. With this method, it is assumed that all ranked content has equal influence.

We used the negative logarithm of the likelihood as an evaluation metric, as in Sect. 5.2.1.

5.3.2 Experimental Results

Figure 3 shows the comparison results of the three ranking bias methods. Reciprocal outperformed the other two methods for all test periods in all categories. In addition, Linear always outperformed Uniform: this result indicates the usefulness in considering the rank position of content. Based on these results, we conclude that Reciprocal, which reflects the creators' browsing behaviors, is the most effective for modeling derivative creation activity.

Algorithm 1 Calculate degrees of three factors for *j*th derivative work posting event of *i*

```
Require: $P(z_{ij}|\mathcal{D}, \mathcal{Z}_{\setminus ij}, O, C, \boldsymbol{\gamma}, \boldsymbol{\beta}, a, b)$
$E_f \Leftarrow P(z_{ij} = 0|\mathcal{D}, \mathcal{Z}_{\setminus ij}, O, C, \boldsymbol{\gamma}, \boldsymbol{\beta}, a, b)$, $E_{h_o} \Leftarrow 0$, $E_{h_d} \Leftarrow 0$, $y \Leftarrow 1$
while $y \le |O_{it}| + |C_{it}|$ do
    if $y \le |O_{it}|$ then
        $E_{h_o} \Leftarrow E_{h_o} + P(z_{ij} = y|\mathcal{D}, \mathcal{Z}_{\setminus ij}, O, C, \boldsymbol{\gamma}, \boldsymbol{\beta}, a, b)$
    else
        $E_{h_d} \Leftarrow E_{h_d} + P(z_{ij} = y|\mathcal{D}, \mathcal{Z}_{\setminus ij}, O, C, \boldsymbol{\gamma}, \boldsymbol{\beta}, a, b)$
    end if
    $y \Leftarrow y + 1$
end while
return $E_f, E_{h_o}, E_{h_d}$
```

6. Qualtitative Experiments

In this section, we report on the qualitative analysis results in terms of (1) category characteristics, (2) temporal development of factors that trigger derivative work posting events, (3) creator characteristics, (4) N-th order derivative creation process, and (5) original work ranking.

6.1 Category Characteristics

By using the posterior distribution of latent variables in Eq. (17), we can analyze the impact of each of the three factors (Oatt, Opop, and Dpop) that trigger derivative work posting events in a category. Algorithm 1 shows the pseudocode for computing the strengths of the three factors for the *j*th derivative work posting event of *i*. In the pseudo-code, E_f, E_{h_o}, and E_{h_d} correspond to the strength of original work attractiveness, original work popularity, and derivative work popularity, respectively, where $E_f + E_{h_o} + E_{h_d} = 1$. By summing E_f of all derivative works of all original works in a category, we can obtain the strength of Oatt in the category. In the same manner, we can obtain the strength of E_{h_o} and E_{h_d} in a category.

Table 5 lists the ratios of the three factors during the training period. The ratios of the three factors vastly differed from one category to another. In the "sing" category, the ratios of Opop and Dpop were both high, while that of Oatt was low. These results indicate that the creators in this category are susceptible to fads and put a high priority on content popularity. In the "dance" category, the ratio of Dpop

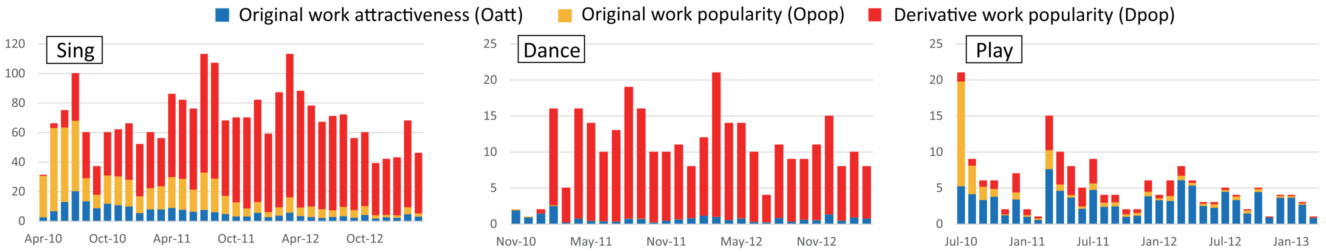

Fig. 4 Estimated number of derivative work posting events triggered by each of three factors per month. Vertical axis represents the number of derivative works posted in a month.

Table 5 Ratios of estimated factors (%)

Factor	Sing	Dance	Play
Original work attractiveness (Oatt)	14.6	17.3	42.5
Original work popularity (Opop)	40.0	21.7	40.0
Derivative work popularity (Dpop)	45.4	61.0	17.5

was higher than those of the other two factors. In this category, not all creators can compose their own choreography. Hence, a creator often posts a derivative work in which the creator dances to the original song with the original choreography. After that, other creators also post derivative works in which they imitate the choreography. The results in Table 5 show that our model described this category's characteristics well. In the "play" category, the ratio of Oatt was high compared to those of the other two categories. This indicates that creators in this category often play their favorite original songs without being affected by fads.

6.2 Temporal Development of Factors

By using Algorithm 1, we can also analyze the temporal development of factors that trigger the derivative work posting events of each original work. In this section, we report on the temporal development of factors per month. Given original work i, we computed the sum of E_f for each i's derivative work posting event every month. Similarly, we computed the sum of E_{h_o} and E_{h_d} every month.

Figure 4 shows example results for three original songs that we selected for this evaluation. In each category, we show the temporal development of factors for one original work. The horizontal axis represents months in the training period and the vertical axis represents the number of derivative works posted in a month. The first month in the horizontal axis is the month when the original work's first derivative work was posted. The blue, orange, and red bars indicate the number of posting events caused by the factor of Oatt, Opop, and Dpop, respectively. Again, we can observe the characteristics of each category. In the "sing" category, in the early period of derivative creation activity, Oatt and Opop had large influences. We could estimate that some derivative works created in the period became popular, after that, other creators who put a high priority on Dpop also posted the original work's derivative work: this is why the influence of Dpop increased as time proceeded. We also observed that Oatt had some influence even at the end period of the graph. This indicates the possibility that some creators happened to find the original song (*e.g.*, by keyword search) and decided to cover the song because they were attracted to the song. In the "dance" category, a limited number of creators who can compose original choreography posted derivative works in the early period (see blue bars in the first two months). After that, many other creators were influenced by such derivative works and posted new derivative works. The influence of Opop was low throughout the period. In the first half of the period in the "play" category, many creators who put a high priority on Opop or Dpop posted this original work's derivative works; while in the last half, creators who put a high priority on Oatt kept posting derivative works.

6.3 Creator Characteristics

Given a creator, we aggregated E_f, E_{h_o}, and E_{h_d} for each of his/her derivative work posting events into the three factors and normalized their sum to 1. This enables us to analyze the ratio at which a creator is influenced by each of three factors. Figure 5 shows the results where each dot represents a creator. In each category, we plot the top 250 creators in terms of the number of derivative works for visibility. In the "sing" category, although most creators put a high priority on Opop and/or Dpop, creators in circle A put a high priority on Oatt. For a creator in circle A, it would be useful to recommend original works similar to the original works he/she used in the past to encourage more derivative work creation. In the "dance" category, creators in circle B were likely to be those who can compose original choreography. We can also use this finding for recommendation. In the "play" category, creators in circle C put a high priority on Dpop contrary to other creators. For these creators, it would be helpful to recommend popular content in the derivative work ranking.

6.4 N-th Order Derivative Creation Process

By using the posterior distribution of latent variables, we can visualize the derivative creation process of an original work. To visualize the process, for each derivative work, we detected y, which is the maximum value in Eq. (17). When y was equal to 0 or indicated the index of the original work's ranked event, derivative work creation was triggered by the original work; when y indicated the index of the ranked

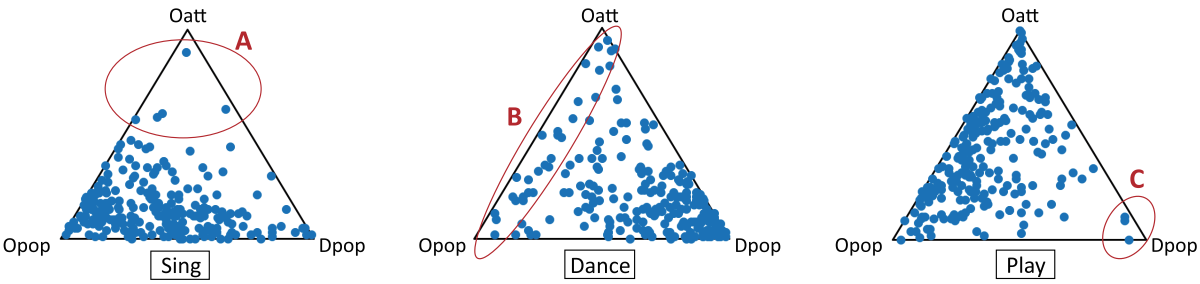

Fig. 5 Creator distributions based on ratio at which a creator is influenced by each of three factors. Each dot represents one creator.

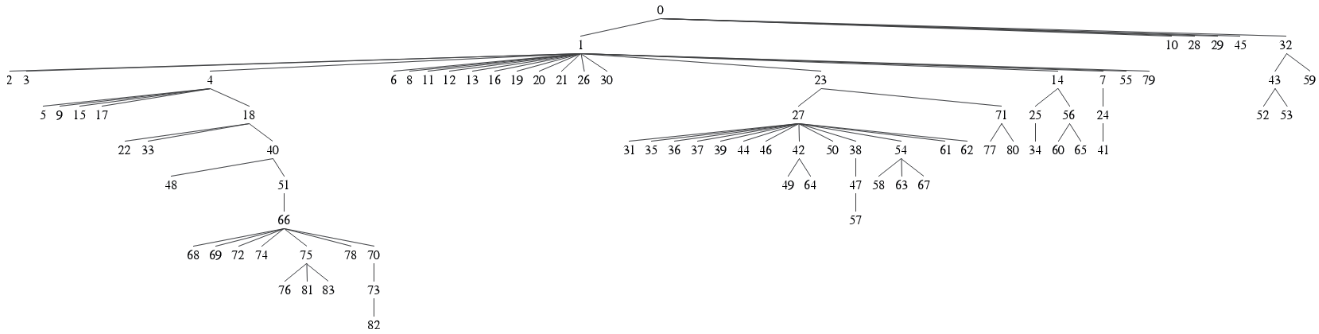

Fig. 6 Derivative works creation process of an original work in "dance" category. 0 represents the original work, and the number $j \geq 1$ give the indices of the derivative works.

event of the j'th derivative work, derivative work creation was triggered by the j'th derivative work.

Figure 6 shows the derivative creation process of an original work in the "dance" category. In the figure, 0 represents an original work, and $j \geq 1$ represents the jth derivative work. An edge between numbers indicates that the lower content creation was triggered by the upper one. In this derivative creation process, the 1st derivative work played an important role because it triggered many derivative creations. We also observed that 9th order derivative creation (82nd derivative work) occurred in this process.

6.5 Original Work Ranking

Finally, we show that we can generate original work ranking in a new light based on the estimated parameters. Popularity is an important metric when an original work ranking is generated. For example, it is common to rank videos based on their view count. The ranking based on view count reflects the popularity among consumers or viewers. It is also possible to consider the popularity among creators, where one of the simplest ways is to rank original works according to the total number of derivative works. However, this popularity includes various kinds of factors: in our case, Oatt, Opop, and Dpop. On the other hand, by using our model, we can generate original work rankings based on each of the three factors. In this section, due to space limitations, we discuss the ranking based on Oatt.

Given a category and rank threshold l, we generate the top l original work ranking, denoted as R_1, where original works are ranked in descending order of the number of derivative works during the test period. We also generate the original work ranking, denoted as R_2, where l original works in R_1 are re-ranked in descending order of α_i. Recall that α_i in Eq. (21) represents i's attractiveness at the end of the test period. Then Spearman's rank correlation between R_1 and R_2 is computed.

Table 6 Spearman's rank correlations between ranking based on number of derivative works and that based on original work attractiveness. "All" means ranking that takes into account derivative works in all three categories.

l	Sing	Dance	Play	All
10	0.170	-0.0909	-0.316	-0.333
30	-0.0372	0.238	0.202	0.272
50	-0.199	0.286	0.0832	0.00101
100	-0.0439	0.105	0.357	0.0883

In addition, we generate two original work rankings considering all three categories at once: (1) the ranking where original works are ranked based on the sum of the number of derivative works in three categories, and (2) the ranking where original works are ranked based on the sum of α_i in three categories.

Table 6 lists the results. Overall, the rank correlations were low. These results indicate that original work ranking based on original work attractiveness largely differs from that based on the number of derivative works, and our proposed model enables consumers to see the ranked list of original works in a new light.

7. Conclusion

We proposed a probabilistic model for inferring latent factors and their influences in derivative creation activity. The model incorporates three factors: (1) original work attractiveness, (2) original work popularity, and (3) derivative work popularity. Our model takes into account content popularity obtained from content ranking data. Our experimental results using real-world derivative creation data showed that our model adopting all three factors achieved the best result in terms of the log likelihood. With respect to content popularity, we showed that using reciprocal rank in content ranking that reflects creators' browsing behaviors achieved the best result. In our qualitative experiments, we showed the usefulness of our model in various aspects including category characteristic analysis (*e.g.*, original work attractiveness has a large influence in the "play" category), visualizing the spread of derivative works creation activity, etc. The limitation of our model is that it presupposes the existence of the ranking information. Such information can be easily obtained in other domains such as 3D models and recipes since the popularity-based rankings of original and derivative works are usually generated from some statistics such as the view count and the favorite count of each work. We therefore believe that our model is general and independent of content types.

For future work, we are interested in extending our model by considering additional factors. For example, some original works may be often used to create derivative works during Christmas. Considering such seasonality [27] is one possible direction to extend our model.

Acknowledgments

This work was supported in part by JST ACCEL Grant Number JPMJAC1602, Japan.

References

[1] M. Hamasaki, H. Takeda, and T. Nishimura, "Network analysis of massively collaborative creation of multimedia contents: Case study of hatsune miku videos on nico nico douga," UXTV, pp.165–168, Oct. 2008.

[2] C. Cayari, "The YouTube effect: How YouTube has provided new ways to consume, create, and share music," International Journal of Education & the Arts, vol.12, no.6, pp.1–28, 2011.

[3] L.A. Liikkanen and A. Salovaara, "Music on YouTube: user engagement with traditional, user-appropriated and derivative videos," Computers in Human Behavior, vol.50, pp.108–124, Sept. 2015.

[4] S. Papadimitriou and E.E. Papalexakis, "Towards laws of the 3d-printable design web," WebSci, pp.255–256, June 2014.

[5] M. Goto, "Grand challenges in music information research," Dagstuhl Follow-Ups: Multimodal Music Processing, vol.3, pp.217–225, 2012.

[6] L.A. Granka, T. Joachims, and G. Gay, "Eye-tracking analysis of user behavior in www search," SIGIR, pp.478–479, July 2004.

[7] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," SIGIR, pp.154–161, Aug. 2005.

[8] T. Iwata, A. Shah, and Z. Ghahramani, "Discovering latent influence in online social activities via shared cascade poisson processes," KDD, pp.266–274, Aug. 2013.

[9] K. Tsukuda, M. Hamasaki, and M. Goto, "Why did you cover that song?: Modeling n-th order derivative creation with content popularity," CIKM, pp.2239–2244, Oct. 2016.

[10] K. Eto, M. Hamasaki, K. Watanabe, Y. Kawasaki, and T. Nishimura, "Modulobe: A creation and sharing platform for articulated models with complex motion," ACE, pp.305–308, Dec. 2008.

[11] G. Cheliotis and J. Yew, "An analysis of the social structure of remix culture," C&T, pp.165–174, June 2009.

[12] M. Hamasaki and M. Goto, "Songrium: A music browsing assistance service based on visualization of massive open collaboration within music content creation community," WikiSym, pp.4:1–4:10, Aug. 2013.

[13] X. Song, Y. Chi, K. Hino, and B.L. Tseng, "Information flow modeling based on diffusion rate for prediction and ranking," WWW, pp.191–200, May 2007.

[14] X. Song, B.L. Tseng, C.Y. Lin, and M.T. Sun, "Personalized recommendation driven by information flow," SIGIR, pp.509–516, Aug. 2006.

[15] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," ICDM, pp.599–608, 2010.

[16] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Learning continuous-time information diffusion model for social behavioral data analysis," ACML, pp.322–337, 2009.

[17] A. Simma and M.I. Jordan, "Modeling events with cascades of poisson processes," UAI, pp.546–555, July 2010.

[18] Y. Tanaka, T. Kurashima, Y. Fujiwara, T. Iwata, and H. Sawada, "Inferring latent triggers of purchases with consideration of social effects and media advertisements," WSDM, pp.543–552, Feb. 2016.

[19] K. Tsukuda, K. Ishida, M. Hamasaki, and M. Goto, "Songrium Derivation Factor Analysis: A web service for browsing derivation factors by modeling n-th order derivative creation," IEICE Trans. Inf. & Syst., vol.E101-D, no.4, pp.1096–1106, April 2018.

[20] A.L. Barabási and R. Albert, "Emergence of scaling in random networks," Science, vol.286, no.5439, pp.509–512, Oct. 1999.

[21] D.L. Snyder and M.I I. Miller, Random Point Processes in Time and Space, Springer, 1991.

[22] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," ICML, pp.561–568, June 2011.

[23] S. Myers and J. Leskovec, "On the convexity of latent social network inference," in Advances in Neural Information Processing Systems 23, pp.1741–1749, Dec. 2010.

[24] G. Last and A. Brandt, Marked point processes on the real line : the dynamic approach, Probability and its applications, Springer-Verlag, 1995.

[25] H. Kenmochi and H. Ohshita, "Vocaloid - commercial singing synthesizer based on sample concatenation," INTERSPEECH, pp.4009–4010, 2007.

[26] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver, "Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering," RecSys, pp.79–86, Sept. 2010.

[27] H. Kim, N. Takaya, and H. Sawada, "Tracking temporal dynamics of purchase decisions via hierarchical time-rescaling model," CIKM, pp.1389–1398, Nov. 2014.

Kosetsu Tsukuda received the Ph.D. degrees in Informatics from Kyoto University, Japan in 2014. He is currently a Researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. His research interests lie in the areas of data mining, information recommendation, and information retrieval regarding user-generated content and music content. He has received nine awards including IPSJ Computer Science Research Award for Young Scientists and IPSJ Yamashita SIG Research Award.

Masahiro Hamasaki received the Ph.D. degree in informatics from the Graduate University for Advanced Studies (SOKENDAI), Japan in 2005. He is currently the Leader of the Media Interaction Group at the National Institute of Advance Industrial Science and Technology (AIST), Japan. His research interests include Web mining, semantic Web, and social media analysis. He is a member of the JSAI, the IPSJ, and ACM.

Masataka Goto received the Doctor of Engineering degree from Waseda University in 1998. He is currently a Prime Senior Researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. Over the past 27 years he has published more than 270 papers in refereed journals and international conferences and has received 51 awards, including several best paper awards, best presentation awards, the Tenth Japan Academy Medal, and the Tenth JSPS PRIZE. In 2016, as the Research Director he began OngaACCEL Project, a 5-year JST-funded research project (ACCEL) on music technologies.