Why Did You Cover That Song?: Modeling N-th Order Derivative Creation with Content Popularity

Kosetsu Tsukuda, Masahiro Hamasaki, and Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST), Japan
{k.tsukuda, masahiro.hamasaki, m.goto}@aist.go.jp

ABSTRACT

Many amateur creators now create derivative works and put them on the web. 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 behavior. Our main contributions are three-fold: (1) to the best of our knowledge, this is the first study modeling derivative creation activity, (2) by using a real-world dataset of music-related derivative work creation to evaluate our model, we showed the effectiveness of adopting all three factors to model derivative creation activity and considering creators' browsing behavior, and (3) we carried out qualitative experiments and showed that our model is useful in analyzing derivative creation activity in terms of category characteristics, temporal development of factors that trigger derivative work posting events, etc.

Keywords

Derivative creation; latent variable model; user generated content

1. INTRODUCTION

The era when only professional creators were able to create and provide content on the web has passed; now amateur creators who used to be just consumers can also easily create and provide content. Such content is known as user generated content (UGC). 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* [5] or 2nd generation work. For example, on YouTube[1], there are many videos in which amateur creators dance to an existing song or perform a cover of it. On Thingiverse[2], which is a web service where amateur creators can share 3D model data intended for a 3D printer, it is popular for creators to download original 3D model data created by others, modify it, and upload their new version. 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 [4]."

We know that derivative creation is popular, but *why* are individual derivative works created? There are various factors that inspire the creation of derivative works. 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, there are three triggering factors: (1) original work's attractiveness, (2) original work's popularity, and (3) derivative work's popularity. Based on this assumption, we propose a model to estimate the factors that triggered derivative work creation. Since the relative influence of the three factors varies among creators, 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 [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) algorithm [6], we can obtain the latent triggers for derivative work posts.

Our main contributions in this paper are as follows. First, 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. Second, 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 behavior is the most effective to model derivative creation activity. Third, we carried out qualitative experiments in terms of (1) category characteristics, (2) temporal development of factors that trigger the derivative work posting events, and (3) N-th order derivative creation process and showed that our model can be used to analyze derivative work creation activity.

2. RELATED WORK

[Analysis of Derivative Creation Activity] A limited number of studies have investigated derivative creation activity. Eto *et al.* [3] 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 [1] examined

[1] http://www.youtube.com

[2] http://www.thingiverse.com

CIKM'16 , October 24 - 28, 2016, Indianapolis, IN, USA

ISBN 978-1-4503-4073-1/16/10…$15.00

DOI: http://dx.doi.org/10.1145/2983323.2983674

the remixing activity in the ccMixter online music community[3]. They reported that derivative creation greatly boosted the output of a community as well as increased the diversity of the output. Hamasaki *et al.* [5] analyzed derivative creation activity on a video sharing web service called Niconico[4]. They used explicit citation information between an original work and its derivative works and discussed certain statistics (*e.g.*, the number of works derived from an original work). 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.

[Modeling Influences in Social Communities] Since estimating influences among users in social activities is useful for various applications, such as influential user detection [14] and personalized recommendation [13], 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 [16]. Although discrete time is assumed with this model, Saito *et al.* [11] 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.* [6] proposed a model that discovers latent influences between users without a network. Although the cascade Poisson process [12] models a sequence of cascading events, the model proposed by Iwata *et al.*, which is called the shared cascade Poisson process (SCPP), can handle multiple sequences of adoption events for multiple items by sharing parameters. Tanaka *et al.* [15] 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.* [15], 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 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.

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 [6, 15]. 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, 3, 5]. Instead, it seemed that the *rich-get-richer* phenomenon exists in the activity [5]. Hence, we assume that the popularity of original and derivative works 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

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^p_{ij}, u^p_{ij}) denote the jth derivative work posting event of original work i. More specifically, creator $u^p_{ij} \in \mathcal{U}$ 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: $t^p_{ij} \leq t^p_{ij'}$ 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 $\mathcal{D}_i = \{(t^p_{ij}, u^p_{ij})\}_{j=1}^{J_i}$. 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 $(t^{\text{o}}_{ik}, r^{\text{o}}_{ik})$ 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: $t^{\text{o}}_{ik} \leq t^{\text{o}}_{ik'}$ 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 $\mathcal{O}_i = \{(t^{\text{o}}_{ik}, r^{\text{o}}_{ik})\}_{k=1}^{K^o_i}$. Therefore, a set of ranked events of all original works is given by $\mathcal{O} = \{\mathcal{O}_i\}_{i \in \mathcal{I}}$.

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 $(t^{\text{c}}_{ik}, r^{\text{c}}_{ik})$ 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 $\mathcal{C}_i = \{(t^{\text{c}}_{ik}, r^{\text{c}}_{ik})\}_{k=1}^{K^c_i}$. Note that $\mathcal{C}_i$ includes ranked events of all i's derivative works. Finally, a set of ranked events of all derivative works of all original works is given by $\mathcal{C} = \{\mathcal{C}_i\}_{i \in \mathcal{I}}$.

3.2 Factors

3.2.1 Original Work Attractiveness (Oatt)

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. 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 Figure 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 [2]. Based on these assumptions, we model the rate at which creator u posts i's derivative work triggered by i's attractiveness as follows: $f_i(u) = \alpha_i \theta_{0u}$, where $\alpha_i \geq 0$ is the original work attractiveness. The $\theta_{0u} \geq 0$ 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 Figure 1(a), the height of the blue line corresponds to $\alpha_i \theta_{0u}$.

3.2.2 Creator Influence (Cinf)

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

[3] http://ccmixter.org

[4] http://www.nicovideo.jp

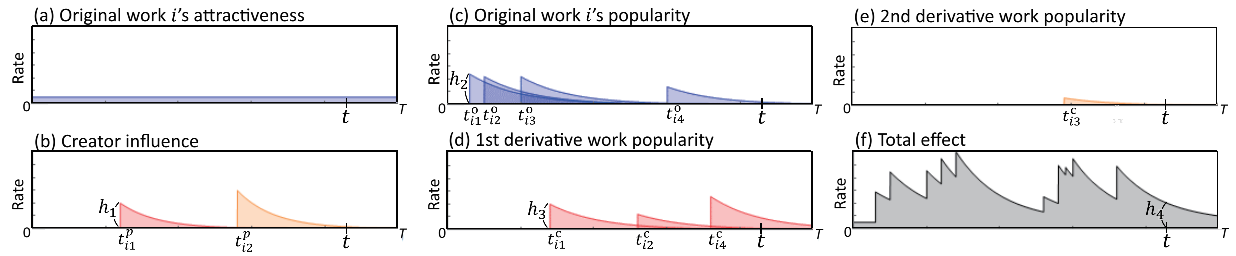

Figure 1: Rate at which creator u posts original work i's derivative work at time t.

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 [10]. 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. $g_{(i,t',u')}(t,u) = \alpha_{u'}\theta_{u'u}\mathrm{e}^{-\gamma_p(t-t')}$ if $t' < t$, and 0 otherwise, 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 $\sum_{u \in \mathcal{U} \setminus u'} \theta_{u'u} = 1$, where $\mathcal{U} \setminus u'$ is the set of creators excluding u'. Hence, $\alpha_{u'}\theta_{u'u}$ means the influence of u' on u. Finally, $\mathrm{e}^{-\gamma_p(t-t')}$ models the decay of influence over time with decay parameter $\gamma_p \geq 0$. 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 Figure 1(b), two 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.

3.2.3 Original Work Popularity (Opop)

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 Section 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 [7]. In addition, we assume that each creator has a different preference for original work popularity. 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: $h_{o(i,t',r')}(t,u) = rb(r')\omega_i\theta_{-1u}\mathrm{e}^{-\gamma_o(t-t')}$ if $t' < t$, and 0 otherwise, 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 [7]. Based on the examination behavior, we compute the rank bias as $rb(r') = \frac{1}{r'}$. In Section 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, $\mathrm{e}^{-\gamma_o(t-t')}$ models the decay of influence over time with decay parameter $\gamma_o \geq 0$.

In Figure 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 Figure 1(c) at t^o_{i1}. Then, the influence decreases as time proceeds.

3.2.4 Derivative Work Popularity (Dpop)

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'. Based on similar assumptions and the hypothesis described in Section 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. $h_{d(i,t',r')}(t,u) = rb(r')\sigma_i\theta_{-2u}\mathrm{e}^{-\gamma_d(t-t')}$ if $t' < t$, and 0 otherwise, 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, $\mathrm{e}^{-\gamma_d(t-t')}$ models the decay of influence over time with decay parameter $\gamma_d \geq 0$.

Figure 1(d) and (e) show the influences of i's first and second derivative work popularity, respectively. Let r' be the rank of the first ranked event in Figure 1(d). The influence of the first ranked event is $rb(r')\sigma_i\theta_{-2u}$, which corresponds to h_3 in Figure 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 Sections 3.2.1 to 3.2.4, the rate at which u posts i's derivative work at t is given by:

$$\lambda_i(t,u) = f_i(u) + \sum_{(t',u') \in \mathcal{D}_{it\setminus u}} g_{(i,t',u')}(t,u) + \sum_{(t',r') \in \mathcal{O}_{it}} h_{o(i,t',r')}(t,u) + \sum_{(t',r') \in \mathcal{C}_{it}} h_{d(i,t',r')}(t,u), \quad (1)$$

where $\mathcal{D}_{it\setminus u} = \{(t',u')|(t',u') \in \mathcal{D}_i \text{ and } t' < t \wedge u' \neq u\}$, $\mathcal{O}_{it} = \{(t',r')|(t',r') \in \mathcal{O}_i \text{ and } t' < t\}$, and $\mathcal{C}_{it} = \{(t',r')|(t',r') \in \mathcal{C}_i \text{ and } t' < t\}$. Here, $\lambda_i(t,u)$ corresponds to h_4 in Figure 1(f).

4. INFERENCE

Given $\mathcal{D}$, $\mathcal{O}$, and $\mathcal{C}$, we infer the model parameters by using the stochastic EM algorithm. Following Iwata *et al.* [6], we assume that a set of i's derivative work posting events $\mathcal{D}_i$ is generated from a marked point process [9] at a rate of $\lambda_i(t,u)$. Based on this assumption, the likelihood of the function of $\mathcal{D}$ is given by:

$$P(\mathcal{D}|\mathcal{O},\mathcal{C},\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\sigma},\boldsymbol{\Theta},\boldsymbol{\gamma}) = \prod_{i \in \mathcal{I}} \exp\left(-\int_0^T \sum_{u \in \mathcal{U}} \lambda_i(t,u)dt\right) \prod_{j=1}^{J_i} \lambda_i(t^p_{ij}, t^p_{ij}), \quad (2)$$

where $\boldsymbol{\alpha} = \{\alpha_l\}_{l \in \mathcal{I} \cup \mathcal{U}}$, $\boldsymbol{\omega} = \{\omega_i\}_{i \in \mathcal{I}}$, $\boldsymbol{\sigma} = \{\sigma_i\}_{i \in \mathcal{I}}$, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_u\}_{u \in \mathcal{U}_+}$, $\boldsymbol{\theta}_u = \{\theta_{uu'}\}_{u' \in \mathcal{U}\setminus u}$, and $\boldsymbol{\gamma} = \{\gamma_p, \gamma_o, \gamma_d\}$. Here, $\mathcal{U}_+$ denotes $\mathcal{U} \cup \{0, -1, -2\}$, where 0, -1, and -2 represent virtual creators for Oatt, Opop, and Dpop, 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.

Following Iwata *et al.* [6], 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 Oatt, Opop, and Dpop, respectively. $z_{ij} = j'$ $(1 \leq j' \leq \left|\mathcal{D}_{it\backslash u}\right|)$ 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 Equation (1) can be written as $\lambda_i(t, u) = \sum_z \lambda_i(t, u, z)$, where $\lambda_i(t, u, z) = f_i(u)$ if $z = 0$, $g_{(i, t^p_{iz}, u^p_{iz})}(t, u)$ if $1 \leq z \leq \left|\mathcal{D}_{it\backslash u}\right|$, $h_{o(i, t^o_{iz'}, r^o_{iz'})}(t, u)$ if $\left|\mathcal{D}_{it\backslash u}\right| + 1 \leq z \leq \left|\mathcal{D}_{it\backslash u}\right| + |\mathcal{O}_{it}|$, and $h_{d(i, t^c_{iz''}, r^c_{iz''})}(t, u)$ if $\left|\mathcal{D}_{it\backslash u}\right| + |\mathcal{O}_{it}| + 1 \leq z$. Here, $z' = z - \left|\mathcal{D}_{it\backslash u}\right|$ and $z'' = z - \left|\mathcal{D}_{it\backslash u}\right| - |\mathcal{O}_{it}|$.

Since the integral part in Equation (2) can be analytically calculated, by combining the above equations, the joint distribution of $\mathcal{D}$ and latent variables $\mathcal{Z} = \{\{z_{ij}\}_{j=1}^{J_i}\}_{i \in \mathcal{I}}$ is given by:

$$\begin{aligned}
&P(\mathcal{D}, \mathcal{Z} | \mathcal{O}, \mathcal{C}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\sigma}, \boldsymbol{\Theta}, \boldsymbol{\gamma}) \\
&= \prod_{i \in \mathcal{I}} \exp \Bigg[\alpha_i T + \frac{1}{\gamma_p} \sum_{j=1}^{J_i} \alpha_{u_{ij}} \left(1 - \mathrm{e}^{-\gamma_p (T - t^p_{ij})}\right) \\
&+ \frac{\omega_i}{\gamma_o} \sum_{k=1}^{K^o_i} rb(r^o_{ik}) \left(1 - \mathrm{e}^{-\gamma_o (T - t^o_{ik})}\right) \\
&+ \frac{\sigma_i}{\gamma_d} \sum_{k=1}^{K^c_i} rb(r^c_{ik}) \left(1 - \mathrm{e}^{-\gamma_d (T - t^c_{ik})}\right) \Bigg] \prod_{j=1}^{J_i} \lambda_i(t^p_{ij}, t^p_{ij}, z_{ij}). \quad (3)
\end{aligned}$$

We assume a Gamma prior for each of the original work attractiveness scores α_i as $P(\alpha_i | a, b) \propto \alpha_i^{a-1} \exp(-b\alpha_i)$, where a and b are hyperparameters. In this study, following Iwata *et al.* [6], we set $a = b = 1$. We also assume a Gamma prior for each creator influence α_u, influence of original work popularity ω_i, and influence of derivative work popularity σ_i. In addition, we assume a Dirichlet prior over $\boldsymbol{\theta}_u$, $u \in \mathcal{U}_+$ as $P(\boldsymbol{\theta}_u | \beta) \propto \prod_{u' \in \mathcal{U} \backslash u} \theta_{uu'}^{\beta - 1}$. 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:

$$\begin{aligned}
&P(\mathcal{D}, \mathcal{Z} | \mathcal{O}, \mathcal{C}, \boldsymbol{\gamma}, \beta, a, b) \\
&= \iiiint P(\mathcal{D}, \mathcal{Z} | \mathcal{O}, \mathcal{C}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \boldsymbol{\sigma}, \boldsymbol{\Theta}, \boldsymbol{\gamma}) P(\boldsymbol{\alpha} | a, b) \\
&\quad \times P(\boldsymbol{\omega} | a, b) P(\boldsymbol{\sigma} | a, b) P(\boldsymbol{\Theta} | \beta) d\boldsymbol{\alpha} d\boldsymbol{\omega} d\boldsymbol{\sigma} d\boldsymbol{\Theta}. \quad (4)
\end{aligned}$$

Based on the marginalized joint distribution, 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:

$$\begin{aligned}
&P(z_{ij} = y | \mathcal{D}, \mathcal{Z}_{\backslash ij}, \mathcal{O}, \mathcal{C}, \boldsymbol{\gamma}, \beta, a, b) \\
&\propto \frac{P\left(\mathcal{D}, \mathcal{Z}_{\backslash ij}, z_{ij} = y | \mathcal{O}, \mathcal{C}, \boldsymbol{\gamma}, \beta, a, b\right)}{P\left(\mathcal{D}_{\backslash ij}, \mathcal{Z}_{\backslash ij} | \mathcal{O}, \mathcal{C}, \boldsymbol{\gamma}, \beta, a, b\right)}, \quad (5)
\end{aligned}$$

where $y \in \{0, 1, \cdots, \left|\mathcal{D}_{it\backslash u}\right| + |\mathcal{O}_{it}| + |\mathcal{C}_{it}|\}$, and $\backslash ij$ represents 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 Equation (4). Following Iwata *et al.* [6], $\boldsymbol{\gamma}$ is estimated using Newton's method and β is estimated using the fixed point iteration method [10]. Finally, we can make the point estimates of the integrated out parameters α_i, α_u, ω_i, σ_i, and $\theta_{uu'}$.

Table 1: 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

5. QUANTITATIVE EXPERIMENTS

In this section, we answer the following two research questions: (1) is adopting three factors (Oatt, Opop, and Dpop) effective to model derivative creation activity? (Section 5.2) and (2) what kinds of ranking bias methods are effective to model derivative creation activity? (Section 5.3)

5.1 Dataset

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: as of the end of April 2016, more than 140,000 original song videos and more than 590,000 derivative videos had been uploaded to Niconico. Most original songs are created using singing synthesizer software called VOCALOID; we restricted ourselves to original song videos of this type. 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 6/30/2013. Data between 1/1/2010 and 3/31/2013 were used as training data and data between 4/1/2013 and 6/30/2013 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 6/30/2013. Since only daily ranking data is available on Niconico, the timestamp in all our experiments is measured in days.

Table 1 lists the statistics of the dataset used in the experiments.

5.2 Combination of Factors

[Comparison Models] As mentioned in Section 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+Cinf, (3) Oatt+Cinf+Opop+Dpop, (4) Oatt+Opop, (5) Oatt+Dpop, and (6) Oatt+Opop+Dpop, where Oatt+Cinf, for example, represents the model that combines the factors of Oatt and Cinf. Among the six models, (2) corresponds to SCPP [6] 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 [6, 8]. 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 $L = \sum_{i \in \mathcal{I}} \left(-\int_T^{T'} \sum_{u \in \mathcal{U}} \lambda_i(t, u) dt\right) \sum_{(t,u) \in \mathcal{D}_i^{\text{test}}} \log \lambda_i(t, u)$, 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

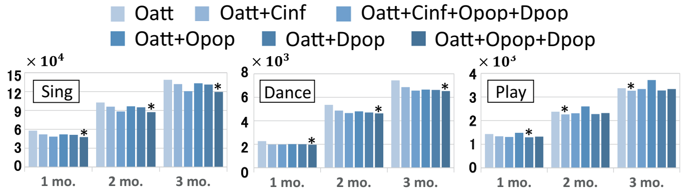

Figure 2: Negative logarithm of likelihood of each model. Vertical axis and horizontal axis represent negative log-likelihood and test periods (*e.g.*, "1 mo." is first set of test data), respectively.

data from one month (4/1/2013 to 4/30/2013) up to three months (4/1/2013 to 6/30/2013).

[Results] Figure 2 shows the negative log-likelihood during each test period in each category. The model that achieved the best performance is marked with "*". In the "sing" and "dance" categories, Oatt+Opop+Dpop exhibited the best result for all test periods. In the "play" category, it did not perform the best, but it stably exhibited high performance in all categories during all test periods. Although other models exhibited the best results in the "play" category, they were unstable because their average ranks in both the "sing" and "dance" categories were low. From these results, we 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

[Settings] As mentioned in Section 3.2.3, our model uses the reciprocal rank method to bias the content ranking based on the creators' browsing behavior 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, $rb(r^o_{ik}) = \frac{101-r^o_{ik}}{100}$. 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 Section 5.2.

[Results] Figure 3 shows the 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 behavior, is the most effective for modeling derivative creation activity.

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, and (3) N-th order derivative creation process.

6.1 Category Characteristics

By using the posterior distribution of latent variables in Equation (5), 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 pseudo-code for computing the strengths of the three factors for the jth derivative work posting event of i. In the pseudo-code, E_f, E_{h_o}, and E_{h_d} correspond to the strength of Oatt, Opop, and Dpop, 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.

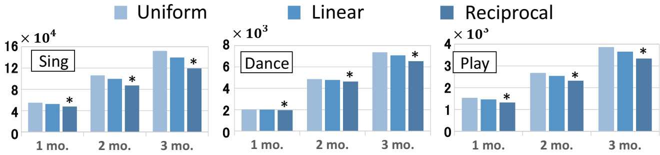

Figure 3: Negative logarithm of likelihood of each ranking bias method.

Algorithm 1 Calculate degree of three factors for jth derivative work posting event of i

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

Table 2: 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

Table 2 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 was high. 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 show that our model described this category's characteristics well. In the "play" category, the ratio of Oatt was high. 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. This section reports 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 the 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 on 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 Oatt, Opop, and Dpop, respectively. Again, we can observe the characteristics

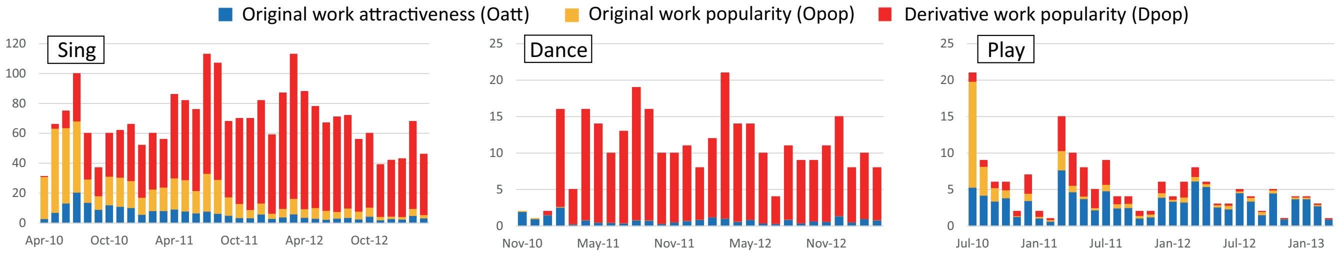

Figure 4: Estimated number of derivative work posting events triggered by each of three factors per month. Vertical axis represents the number of derivative works posted in a month.

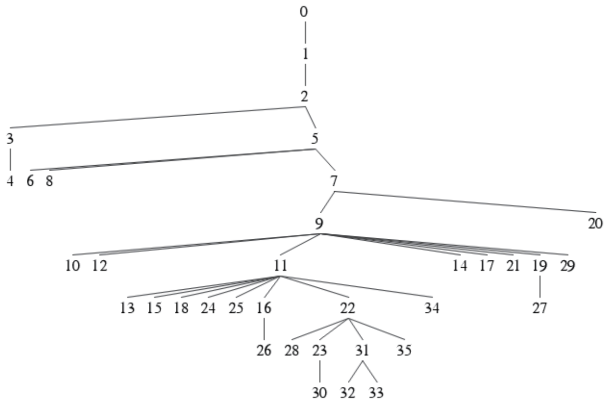

Figure 5: Derivative work creation process of an original work in "dance" category. 0 represents the original work, and $j \geq 1$ represents the jth derivative work.

of each category. In the "sing" category, in the early period of the derivative creation activity, Oatt and Opop had a large influence. We 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 covered the song. In the "dance" category, a limited number of creators who can compose original choreography posted derivative works in the early period (see the blue bars in the first two months). After that, many other creators were influenced by such derivative works and posted new derivative works. In the first half of the test 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 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 Equation (5). 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 event of the j'th derivative work, derivative work creation was triggered by the j'th derivative work.

Figure 5 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 9th and 11th derivative works played an important role because they triggered many derivative creations. We also observed that 10th order derivative creation (30th, 32nd, and 33rd derivative works) occurred in this process.

7. CONCLUSION

We proposed a model for inferring latent factors and their influences in derivative creation activity. For future work, we are interested in applying our model to other derivative creation data such as data on Thingiverse. We are also 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 is one possible way to extend our model.

Acknowledgements

This work was partly supported by MEXT KAKENHI Grant Number (15H06887) and CREST, JST.

8. REFERENCES

[1] G. Cheliotis and J. Yew. An analysis of the social structure of remix culture. In *C&T*, pages 165–174, 2009.

[2] L. S. Donald and I. M. Michael. *Random Point Processes in Time and Space*. Springer, 1991.

[3] K. Eto *et al.* Modulobe: A creation and sharing platform for articulated models with complex motion. In *ACE*, pages 305–308, 2008.

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

[5] M. Hamasaki *et al.* Network analysis of massively collaborative creation of multimedia contents: Case study of hatsune miku videos on nico nico douga. In *UXTV*, pages 165–168, 2008.

[6] T. Iwata *et al.* Discovering latent influence in online social activities via shared cascade poisson processes. In *KDD*, pages 266–274, 2013.

[7] T. Joachims *et al.* Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*, pages 154–161, 2005.

[8] A. Karatzoglou *et al.* Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering.

[9] G. Last and A. Brandt. *Marked point processes on the real line : the dynamic approach*. Springer, 1995.

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

[11] K. Saito *et al.* Learning continuous-time information diffusion model for social behavioral data analysis. In *ACML*, pages 322–337, 2009.

[12] A. Simma and M. I. Jordan. Modeling events with cascades of poisson processes. In *UAI*, pages 546–555, 2010.

[13] X. Song *et al.* Personalized recommendation driven by information flow. In *SIGIR*, pages 509–516, 2006.

[14] X. Song *et al.* Information flow modeling based on diffusion rate for prediction and ranking. In *WWW*, pages 191–200, 2007.

[15] Y. Tanaka *et al.* Inferring latent triggers of purchases with consideration of social effects and media advertisements. In *WSDM*, pages 543–552, 2016.

[16] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *ICDM*, pages 599–608, 2010.