What Drives Online Popularity: Author, Content or Sharers? Estimating Spread Dynamics with Bayesian Mixture Hawkes

Pio Calderon$^{(\boxtimes)[0000-0002-8747-8917]}$ and Marian-Andrei Rizoiu$^{[0000-0003-0381-669X]}$

University of Technology Sydney, Australia
piogabrielle.b.calderon@student.uts.edu.au
marian-andrei.rizoiu@uts.edu.au

Abstract. The spread of content on social media is shaped by intertwining factors on three levels: the source, the content itself, and the pathways of content spread. At the lowest level, the popularity of the sharing user determines its eventual reach. However, higher-level factors such as the nature of the online item and the credibility of its source also play crucial roles in determining how widely and rapidly the online item spreads. In this work, we propose the Bayesian Mixture Hawkes (BMH) model to jointly learn the influence of source, content and spread. We formulate the BMH model as a hierarchical mixture model of separable Hawkes processes, accommodating different classes of Hawkes dynamics and the influence of feature sets on these classes. We test the BMH model on two learning tasks, cold-start popularity prediction and temporal profile generalization performance, applying to two real-world retweet cascade datasets referencing articles from controversial and traditional media publishers. The BMH model outperforms the state-of-the-art models and predictive baselines on both datasets and utilizes cascade- and item-level information better than the alternatives. Lastly, we perform a counter-factual analysis where we apply the trained publisher-level BMH models to a set of article headlines and show that effectiveness of headline writing style (neutral, clickbait, inflammatory) varies across publishers. The BMH model unveils differences in style effectiveness between controversial and reputable publishers, where we find clickbait to be notably more effective for reputable publishers as opposed to controversial ones, which links to the latter's overuse of clickbait.

Keywords: Hawkes process · hierarchical model · mixture model

1 Introduction

Social media platforms have played an increasingly important role as distribution hubs for content. In 2023, it was reported that 69% of the U.S. adult population use social media as a news source [7], implying a significant shift in how information is consumed. Understanding how content propagates on these platforms
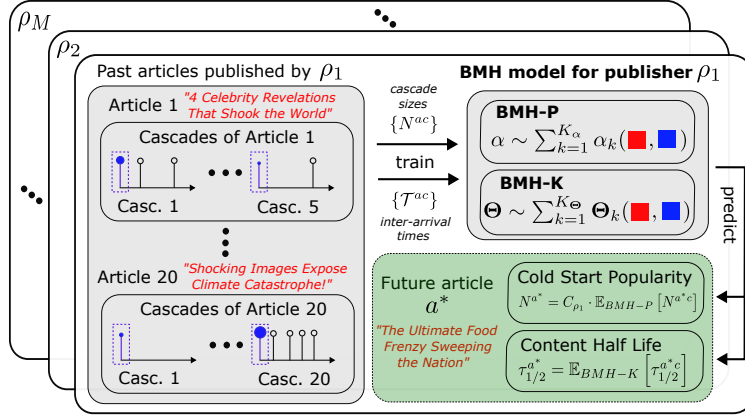
Fig. 1: An *intuitive plate diagram* for the BMH model. *Left:* The BMH model is trained using a historical dataset: a collection of M publishers $\{\rho_1, \ldots, \rho_M\}$, items for each publisher (i.e. articles), and a set of diffusion cascades for each item. Each diffusion cascade consists of a timeline of events, here represented by a set of lollipops. *Upper Right:* The BMH is a publisher-level model that maps cascade features (shown in blue color) and article features (in red color) to a mixture of Hawkes processes. *Lower Right:* The trained BMH model (with the historical follower count distribution) can be used to infer spread dynamics of future articles based on their headlines.

– both the size and speed of dissemination – is vital since the impact is intrinsically tied to the level of online engagement the content receives. To command attention in today's digital age, it is not sufficient to craft high-quality content alone, but rather high-quality content that resonates with social media.

The spread of content online is influenced by factors at varying levels. At the lowest level, the breadth of a *diffusion cascade*, referring to the sequence of content shares triggered by a user, often hinges on the user's popularity as reflected by their follower count [1]. If a highly followed user shares an online item, it reaches a broader audience, increasing the likelihood that it will be shared. However, the cascade's growth is not solely dependent on user popularity. The nature and category of the shared content play crucial roles, as various topics may engage audiences in different ways [27,32]. For news dissemination, the way an article headline is written, particularly the use of clickbait tactics to create an *information gap* to exploit the audience's curiosity [33], significantly impacts the total attention (i.e. *popularity*) the news article receives. Beyond cascade- and item-level factors, the reputation of the online item's source also affects how widely and quickly information spreads [24]. An article from a reputable source like The New York Times may spread more quickly and be taken more seriously than an article from a controversial, lesser-known blog due to the former's established credibility. Accurately modeling diffusion cascades of online content requires an approach that jointly considers these factors at different levels.

In this work, we address two open questions related to jointly modeling the influence of the source, item- and cascade-level factors on online content spread.

The first research question examines how these three levels influence the spread of online content. While prior studies have explored the effects of cascade features [30] and item-level variations [12], a comprehensive framework that jointly considers the three levels has yet to be developed. Our first question is: **Can we build a model for the spread dynamics of online content that accounts for the intertwining influence of its source, the content itself, and cascade-level factors?** To tackle this, we propose the Bayesian Mixture Hawkes (BMH) model, a novel source-level hierarchical mixture model of separable Hawkes processes that models diffusion cascades' size and temporal profile as a function of cascade- and item-level features. The left half of Fig. 1 showcases how the source-level BMH model learns across both the cascade and item levels from a hierarchically structured dataset (i.e., a set of items, cascade groups for each item, and feature sets attached to each). The BMH model is capable of learning different classes of Hawkes process dynamics, taking into account the ability of online content to trigger varied responses, from highly popular to largely unnoticed cascades, as well as those that fade quickly or diminish over time. The BMH learns the influence of feature sets on these classes in two ways: the location of each class in the Hawkes parameter space and the membership probability of each cascade belonging to each class. The trained BMH model can then be used to predict future items' popularity and spread dynamics from the same source (see the right half of Fig. 1). We test the BMH model on two hierarchical retweet cascade datasets that reference articles from controversial and reputable media publishers [12] and on two tasks: cold-start popularity prediction and temporal profile generalization performance. We show that the BMH outperforms the state-of-the-art in item popularity prediction (Dual Mixture Model [12]), Empirical Bayes approach [30] and predictive baselines for both tasks and datasets, and that the BMH model jointly leverages cascade- (i.e., the follower count of the seed user) and article-level (i.e., the article headline embedding vector) information better than the benchmarks. Furthermore, our model ablation highlights the role of the initiating user in shaping the cascade dynamics related to controversial media, a factor less critical for cascades linked to reputable media. This distinction mirrors the diverse pathways of online information dissemination: controversial media often circulate within topical social groups [3,11], with the initial endorser serving to validate the content, while for reputable media the publisher's reputation is the most important factor.

Our second open question relates to learning differences in the spread dynamics across news publishers: **Can we uncover across-publisher differences in how headline writing style (neutral, clickbait, inflammatory) affects published content's popularity and temporal profile?** We run a counterfactual analysis using the trained publisher-level BMH models and a labeled set of article headlines [13] to show the variation of headline style effectiveness across publishers. We find that the BMH model is able to capture nuanced publisher behavior, such as the effectiveness of inflammatory headlines for tabloids. The

BMH model also unveils differences in the success of clickbait between controversial and reputable outlets, linking to existing research on clickbait fatigue and the diminishing relationship between clickbait effectiveness and volume [33,17].

The main contributions of the work are as follows:

1. The Bayesian Mixture Hawkes (BMH) model[1], a novel hierarchical mixture model of the joint influence of cascade- and item-level features on online item spread dynamics. On two news datasets, we show that the BMH outperforms the state-of-the-art and baselines in cold-start popularity prediction and temporal profile generalization performance.
2. A counter-factual analysis showing how headline writing style affects published content's spread dynamics. Using the BMH model we learn the differences in the effectiveness of headlines across publishers and show general trends across controversial and reputable media outlets.

Related Work. In recent years, generative models, and specifically the Hawkes process [10], have been employed to model online information diffusion given their dual *predictive* and *interpretable* capabilities [2,35,9,18]. However, the Hawkes process cannot incorporate feature sets in its base form since it relies only on observed temporal sequences to fit the model parameters. Numerous modifications to incorporate feature sets have been proposed to enhance model fit and predictive capabilities. A hybrid approach introduced in [20] integrates the Hawkes process with a scaling factor trained on cascade-level features to improve retweet cascade size prediction. The Empirical Bayes (EB) method [30] utilizes historical retweet sequences to link cascade features and the prior distribution of Hawkes process parameters, leading to better forecasting. The parametric Hawkes process [14] models the branching factor, i.e. the expected number of offsprings from a parent event, as a linear combination of event-level features. Lastly, the Tweedie-Hawkes process [15] improves on this by combining the Hawkes process with the Tweedie distribution to more realistically model the effect of event-level features on the branching factor. The proposed BMH model is a hierarchical model and can incorporate two levels of feature sets: the cascade- and the item (i.e., cascade-group)-level, which previous work does not cover.

Another relevant area is mixtures of point processes, employed when the data is suspected to be generated from multiple dynamical classes (i.e., parameter sets). In [34], the Hawkes process was combined with the Dirichlet distribution to model clusters of cascades. An online learning framework was introduced in [8] to fit mixtures of multivariate Hawkes processes to learn the interaction network across a set of actors. [29] introduces a generative model for mixtures of more complex point processes by using recurrent neural networks. Closest to our work is the Dual Mixture Model (DMM) [12], a generative model for cascade groups. Each cascade is sampled from a mixture of separable Hawkes processes learned jointly with their mixture probabilities. To the best of our knowledge, including

---

[1]The Stan/CmdStanPy implementation of the BMH model is available at `https://github.com/behavioral-ds/bayesian-mixture-hawkes/`.

feature sets into mixture models of point processes has not been explored: the BMH model solves this by learning the influence of features on the mixture components.

2 Preliminaries

We discuss two point process models that form the foundation of the BMH model. Section 2.1 presents the Hawkes Process (HP) [10], a temporal point process model that displays self-exciting behavior. Section 2.2 introduces the Dual Mixture Model (DMM) [12], an approach to jointly model groups of cascades. An introduction to Bayesian hierarchical modeling, which we employ to model hierarchical data, is included in Section 1.3 of the Online Appendix [5].

2.1 Hawkes Process

The Hawkes process (HP) [10] is a temporal point process widely used to model phenomena that display self-excitation, i.e., the likelihood of an event increases as more events occur. The HP is specified using the conditional intensity function $\lambda(t|\mathcal{H})$, the event rate at any time t conditioned on the history $\mathcal{H} = \{t_j | t_j < t\}$ of past events up to that point, i.e. $\lambda(t|\mathcal{H}) = \mu + \sum_{j=1}^{N} \alpha \cdot g(t - t_j | \boldsymbol{\Theta})$. For brevity, we drop the condition on the event history and write $\lambda(t|\mathcal{H})$ as $\lambda(t)$. Under this parametrization, a Hawkes process $\mathcal{HP}(\mu, \alpha, \boldsymbol{\Theta}|g)$ is identified with the parameters μ, α and $g(\cdot|\boldsymbol{\Theta}) : \mathbb{R}^+ \to \mathbb{R}^+$. The parameter $\mu \geq 0$ is the arrival rate of events triggered by external sources, the branching factor $\alpha \geq 0$ is the expected number of offsprings generated by a single parent event which controls the level of self-excitation from previous events, and the memory kernel $g(\cdot|\boldsymbol{\Theta})$ models the temporal decay of influence of previous events on future events controlled by the parameter set $\boldsymbol{\Theta}$. In this work, we utilize the power law kernel parametrized by $\boldsymbol{\Theta} = \{\theta, d\}$, given by $g(t|\theta, d) = \theta \cdot d^\theta \cdot (t + d)^{-(1+\theta)}$. Other common choices for the memory kernel are the exponential kernel $g(t|\theta) = \theta \cdot e^{-\theta t}$ and the Reyleigh kernel $g(t|\theta) = e^{-\frac{1}{2}\theta \cdot t^2}$. We focus on the power law as it has been shown in [20] to outperform these alternatives in popularity prediction. HP estimation and prediction is discussed in detail in Sec. 1.1 of the Appendix [5].

Given a collection of *complete* cascades $\mathbb{H} = \{\mathcal{H}_i\}$ where each $\mathcal{H}_i$ is completely observed (i.e. terminal time $T_i \to \infty$), and assuming no exogenous events (i.e. $\mu = 0$, which occurs for instance with Twitter retweet cascades, where all retweets are considered to be spawned by the original tweet), the HP log-likelihood $\mathcal{L}(\alpha, \boldsymbol{\Theta}|\mathbb{H})$ splits into two log-likelihoods [12],

$$\mathcal{L}(\alpha, \boldsymbol{\Theta}|\mathbb{H}) = \mathcal{L}(\alpha|\mathbb{H}) + \mathcal{L}(\boldsymbol{\Theta}|\mathbb{H}), \tag{1}$$

$$\mathcal{L}(\alpha|\mathbb{H}) = \sum_{\mathcal{H}_i \in \mathbb{H}} \log\left[\alpha^{N_i - 1} e^{-N_i \alpha}\right], \quad \mathcal{L}(\boldsymbol{\Theta}|\mathbb{H}) = \sum_{\mathcal{H}_i \in \mathbb{H}} \sum_{t_j \in \mathcal{H}_i, j \geq 1} \log \sum_{t_z < t_j} g(t_j - t_z | \boldsymbol{\Theta}),$$

where we set $N_i = |\mathcal{H}_i|$. Under this case, Hawkes process estimation splits into two independent problems, hence the term *separable Hawkes process*. The first

problem (popularity estimation) utilizes the cascade sizes $\{N_i\}$ to estimate the branching factor α by maximizing $\mathcal{L}(\alpha|\mathbb{H})$. It was shown in [12] that maximizing $\mathcal{L}(\alpha|\mathbb{H})$ is equivalent to maximizing $\sum_{\mathcal{H}_i \in \mathbb{H}} \log \mathbb{B}(N_i|\alpha)$, where $\mathbb{B}(\cdot|\alpha)$ is the Borel distribution [4]. The second problem (kernel estimation) uses the interevent-time distribution $\mathcal{T} = \{t_j - t_z\}_{t_z < t_j, t_j \in \mathcal{H}, \mathcal{H} \in \mathbb{H}}$ to estimate $\boldsymbol{\Theta}$ by maximizing $\mathcal{L}(\boldsymbol{\Theta}|\mathbb{H})$.

2.2   Dual Mixture Model

Maximizing Eq. (1) yields the best-fitting Hawkes parameter set $\{\alpha, \boldsymbol{\Theta}\}$. However, this approach assumes that all cascades stem from a singular parameter set, an assumption which may not hold if there are multiple dynamical classes of cascade behavior. The Dual Mixture Model (DMM) [12] was proposed to model a cascade group $\mathbb{H}$ with a mixture of K separable Hawkes processes of different parameter sets to account for different dynamical classes. Under separability, the DMM splits into two submodels: the Borel mixture model (BMM) for popularity estimation and the kernel mixture model (KMM) for kernel estimation. The BMM assumes that there exist K popularity classes accounting for the cascade sizes $\{N_i\}$, where the ith class is represented by the branching factor α_i^* with probability p_i^B, i.e. $M^B = \{(\alpha_i^*, p_i^B)\}_{i=1}^K$. Similarly, the KMM assumes that there are K kernel classes accounting for the interevent-time distribution $\mathcal{T}$, where the jth class is represented by the kernel parameter set $\boldsymbol{\Theta}_j^*$ with probability p_j^g, i.e. $M^g = \{(\boldsymbol{\Theta}_j^*, p_j^g)\}_{i=1}^K$. The DMM is the Cartesian product of M^B and M^g, i.e. $M = \{(\alpha_i^*, \boldsymbol{\Theta}_j^*, p_i^B \cdot p_j^g)|(\alpha_i^*, p_i^B) \in M^B, (\boldsymbol{\Theta}_j^*, p_j^g) \in M^g\}$. DMM estimation and prediction is discussed in detail in Section 1.2 of the Online Appendix [5].

3   Bayesian Mixture Hawkes (BMH) Model

In this section, we develop the Bayesian Mixture Hawkes (BMH) model, a hierarchical mixture model of separable Hawkes processes to learn the effect of cascade-level and item-level features on cascade spread dynamics. We first describe the dataset structure that the BMH model is tailored to handle, then discuss the BMH model's objectives and the approach we adopt to address each. We then present the two components of the BMH: the popularity submodel in Section 3.1 and the kernel submodel in Section 3.2.

Assume that we are given the following dataset. First, we have a collection of items, denoted as $\mathcal{A}$, from a shared source ρ, where each item $a \in \mathcal{A}$ is characterized by the feature vector $\overrightarrow{y}^a \in \mathbb{R}^{N_y}$. If ρ is a news publisher, then $\mathcal{A}$ can represent a collection of news articles and $\overrightarrow{y}^a$ the embedding vector for article a's headline. Second, we have a set of complete cascades $\mathbb{H}^a$ for each item $a \in \mathcal{A}$, where cascade $\mathcal{H}^{ac} \in \mathbb{H}^a$ has size N^{ac}, interevent distribution $\mathcal{T}^{ac}$, and is described by the feature vector $\overrightarrow{x}^{ac} \in \mathbb{R}^{N_x}$. In our news example, $\mathbb{H}^a$ can represent discussions on Twitter related to article a, which we obtain by collecting all retweet cascades initiated with a tweet linking article a's URL. The feature vector $\overrightarrow{x}^{ac}$ can be taken as the follower count of the cascade's initiator.

Table 1: Summary of important quantities and notation.

Parameter	Interpretation	Real-World Mapping
$a/\mathcal{A}$	item/s produced by source ρ	news article/s from publisher ρ
$\mathcal{H}^{ac}/\mathbb{H}^a$	cascade/s related to item a	retweet cascade/s for article a
$\overrightarrow{y}^a$	item-level features of a	headline embedding for article a
N^a	item popularity of a	overall tweet count for article a
$\overrightarrow{x}^{ac}$	cascade-level features of $\mathcal{H}^{ac}$	# followers of $\mathcal{H}^{ac}$ seed user
N^{ac}	cascade size of $\mathcal{H}^{ac}$	
$\mathcal{T}^{ac}$	intereevent-time distribution of $\mathcal{H}^{ac}$	
$(\alpha^{ac}, \boldsymbol{\Theta}^{ac})$	HP parameter set generating $\mathcal{H}^{ac}$	
$\tau_{1/2}^{ac}$	diffusion half-life of $\mathcal{H}^{ac}$	
$K_\alpha/K_{\boldsymbol{\Theta}}$	# of BMH-P/-K classes	
$z_{\alpha,k}^{ac}/z_{\boldsymbol{\Theta},k}^{ac}$	class k membership probability	
$\delta_{\alpha,k}/\delta_{\theta,k}$	baseline logit(α), log(θ) for class k	
$\delta_{z_{\alpha,k}}/\delta_{z_{\boldsymbol{\Theta},k}}$	baseline class k mem. probability	
$\overrightarrow{\gamma}_{\alpha,k}/\overrightarrow{\gamma}_{\theta,k}$	effect of $\overrightarrow{y}^a$ on class k center	
$\overrightarrow{\gamma}_{z_{\alpha,k}}/\overrightarrow{\gamma}_{z_{\boldsymbol{\Theta},k}}$	effect of $\overrightarrow{y}^a$ on class k mem. prob.	
$\overrightarrow{\beta}_{\alpha,k}/\overrightarrow{\beta}_{\theta,k}$	effect of $\overrightarrow{x}^{ac}$ on class k center	
$\overrightarrow{\beta}_{z_{\alpha,k}}/\overrightarrow{\beta}_{z_{\boldsymbol{\Theta},k}}$	effect of $\overrightarrow{x}^{ac}$ on class k mem. prob.	

We model the generative process of $\mathcal{H}^{ac}$ using a separable power-law HP with parameter set $(\alpha^{ac}, \boldsymbol{\Theta}^{ac})$, i.e. $\mathcal{H}^{ac} \sim \mathcal{HP}(\alpha^{ac}, \boldsymbol{\Theta}^{ac}|g)$. We construct the BMH as a model for $(\alpha^{ac}, \boldsymbol{\Theta}^{ac})$ with three goals: (1) jointly learn across the item set $\mathcal{A}$, (2) learn the relationship between $\overrightarrow{y}^a$ and $(\alpha^{ac}, \boldsymbol{\Theta}^{ac})$, and (3) learn the link between $\overrightarrow{x}^{ac}$ and the same parameters. We handle goal (1) by using a two-level Bayesian hierarchical model to jointly fit across each item $a \in \mathcal{A}$ and to tie together cascade- and item-level information. For goals (2) and (3), we consider a mixture of separable HPs with K_α classes for α^{ac} and $K_{\boldsymbol{\Theta}}$ classes for $\boldsymbol{\Theta}^{ac}$. We learn the influence of $\overrightarrow{y}^a$ and $\overrightarrow{x}^{ac}$ on $\{\alpha^{ac}, \boldsymbol{\Theta}^{ac}\}$ through the centers and membership probabilities of the K_α popularity classes and $K_{\boldsymbol{\Theta}}$ kernel classes.

Due to the separability of the underlying HP, the BMH divides into two independent models: (1) BMH-P, the *popularity* submodel for α^{ac}, and (2) BMH-K, the *kernel* submodel for $\boldsymbol{\Theta}^{ac}$. Table 1 lists the notation for important variables in the BMH and the mapping to real-world quantities in the datasets in Section 4.

3.1 BMH-P, the Popularity Submodel

The branching factor α^{ac} is modeled as the mixture random variable

$$\mathrm{logit}(\alpha^{ac}) = \delta_{\alpha,k}^a + \overrightarrow{\gamma}_{\alpha,k} \cdot \overrightarrow{y}^a, \tag{2}$$

with membership probability $z_{\alpha,k}^{ac}$ $(k = 1, \ldots, K_\alpha)$,

$$z_{\alpha,k}^{ac} = \frac{\exp(\delta_{z_{\alpha,k}}^a + \overrightarrow{\beta}_{z_{\alpha,k}}^a \cdot \overrightarrow{x}^{ac} + \overrightarrow{\gamma}_{z_{\alpha,k}} \cdot \overrightarrow{y}^a)}{\sum_{k'=1}^{K_\alpha} \exp(\delta_{z_{\alpha,k'}}^a + \overrightarrow{\beta}_{z_{\alpha,k'}}^a \cdot \overrightarrow{x}^{ac} + \overrightarrow{\gamma}_{z_{\alpha,k'}} \cdot \overrightarrow{y}^a)}. \tag{3}$$
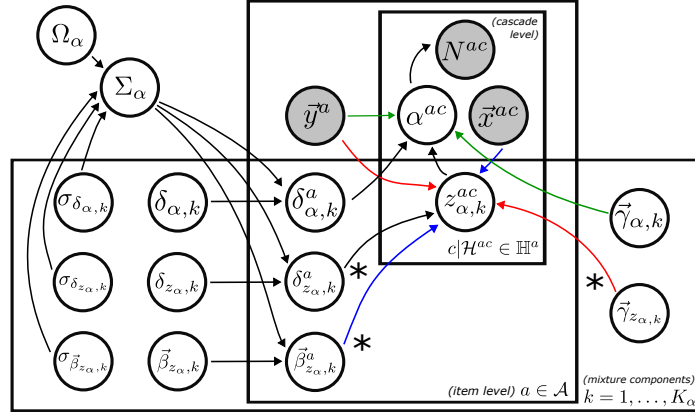
Fig. 2: Plate diagram of the BMH-P model. Shaded nodes are observables while empty nodes are latent variables. Paired colored edges indicate source nodes appearing as a product in the target node. For instance, the green edges indicate that $\overrightarrow{\gamma}_{\alpha,k}$ and $\overrightarrow{y}^a$ appear as $\overrightarrow{\gamma}_{\alpha,k} \cdot \overrightarrow{y}^a$ in the expression for α^{ac} in Eq. (2). The same concept holds for the blue and red edges. Edges marked with * indicate dependence of the target node on the source node indexed with k and the entire set $\{1, \cdots, K_\alpha\}$. For instance, in Eq. (3) $z_{\alpha,k}^{ac}$ depends on $\overrightarrow{\beta}_{z_{\alpha,k}}^{a}$ (see the numerator) and $\overrightarrow{\beta}_{z_{\alpha,k'}}^{a}$ for $k' \in \{1, \cdots, K_\alpha\}$ (see the denominator).

The intercept $\delta_{\alpha,k}^a$ in Eq. (2) sets the centering of $\mathrm{logit}(\alpha^{ac})$ for popularity class k. In Eq. (3), we designate $k = 1$ as the reference class (i.e. $\delta_{z_{\alpha,1}}^a = \overrightarrow{\beta}_{z_{\alpha,1}}^{a} = \overrightarrow{\gamma}_{z_{\alpha,1}} = 0$); parameters for $k > 1$ control deviation from class $k = 1$. The intercept $\delta_{z_{\alpha,k}}^a$ controls the baseline proportion of class k'. The influence of item features on $\mathrm{logit}(\alpha^{ac})$ and class k membership are estimated by $\overrightarrow{\gamma}_{\alpha,k}$ and $\overrightarrow{\gamma}_{z_{\alpha,k}}$, respectively, while the influence of cascade features on class k membership is estimated by $\overrightarrow{\beta}_{z_{\alpha,k}}^{a}$. Note that $\overrightarrow{\gamma}_{\alpha,k}, \overrightarrow{\gamma}_{z_{\alpha,k}}$ are shared across $\mathcal{A}$ while $\overrightarrow{\beta}_{z_{\alpha,k}}^{a}$ is estimated per a.

For brevity, we collect the parameter vector specific to item a as $\overrightarrow{p}_\alpha^a = [\delta_{\alpha,1}^a, \ldots, \delta_{\alpha,K_\alpha}^a, \delta_{z_{\alpha,2}}^a, \ldots, \delta_{z_{\alpha,K_\alpha}}^a, \overrightarrow{\beta}_{z_{\alpha,2}}^{a}, \ldots, \overrightarrow{\beta}_{z_{\alpha,K_\alpha}}^{a}]^\intercal$. We link item a with $\mathcal{A}$ by assuming that $\overrightarrow{p}_\alpha^a$ is drawn from a source-level multivariate normal (MVN) distribution with mean $\overrightarrow{p}_\alpha$ and covariance matrix Σ_α,

$$\overrightarrow{p}_\alpha^a \sim \mathrm{MVN}\left(\overrightarrow{p}_\alpha, \Sigma_\alpha\right), \quad \Sigma_\alpha = \mathrm{D}_\alpha \cdot \Omega_\alpha \cdot \mathrm{D}_\alpha, \quad \mathrm{D}_\alpha = \mathrm{diag}(\sigma_{\overrightarrow{p}_\alpha}), \qquad (4)$$

where Ω_α is a correlation matrix and $\sigma_{\overrightarrow{p}_\alpha}$ is a vector of standard deviations corresponding to $\overrightarrow{p}_\alpha$.

The plate diagram for the BMH-P model is shown in Fig. 2. Variable pairs that appear as a product term are colored green, red and blue in Eqs. (2) and (3), visualized in Fig. 2 as source nodes with green, red and blue edges.

Inference and Prediction. Let $\mathcal{P}_\alpha$ be the parameter set for the BMH-P model. From the set of cascade sizes $\{N_{ac}\}_{\mathcal{H}^{ac} \in \mathbb{H}^a, a \in \mathcal{A}}$, we estimate the posterior distribution $\mathbb{P}(\mathcal{P}_\alpha | \{N_{ac}\}_{ac}) \propto \exp(\mathcal{L}(\mathcal{P}_\alpha | \{N_{ac}\}_{ac})) \cdot \mathbb{P}(\mathcal{P}_\alpha)$, where $\mathbb{P}(\mathcal{P}_\alpha)$ is the prior for $\mathcal{P}_\alpha$ and $\mathcal{L}(\mathcal{P}_\alpha | \{N_{ac}\}_{ac})$ is the log-likelihood of $\mathcal{P}_\alpha$ given the cascade sizes (derived in Section 2.2 of the Online Appendix [5]). Informative priors have to be set on $\{\delta_{\alpha,k}, \delta_{z_{\alpha,k}}\}$ to identify the K_α classes in the α parameter space. $\delta_{\alpha,k}$ and $\delta_{z_{\alpha,k}}$ identify the center and baseline proportion of the kth class, respectively. Weakly informative priors are set for the other parameters in $\mathcal{P}_\alpha$. We implement[1] the BMH-P model in Stan[6], which uses the No-U-Turn Sampler (NUTS), a Hamiltonial Monte Carlo technique, to sample the posterior distribution $\mathbb{P}(\mathcal{P}_\alpha | \{N_{ac}\}_{ac})$. We use CmdStanPy [31] to run Stan code through Python.

Using the average cascade count for items in $\mathcal{A}$, denoted as $\hat{C}_\rho$, and the empirical distribution of the cascade feature vector $\overrightarrow{x}^{ac}$, denoted as $\hat{f}_\rho(x)$, the fitted BMH-P model can be used to estimate the cold-start popularity $\hat{N}^{a^*}$ of an out-of-sample item a^* with feature vector $\overrightarrow{y}^{a^*}$:

$$\hat{N}^{a^*} \approx \hat{C}_\rho \cdot \sum_{x=0}^{\infty} \sum_{k=1}^{K_\alpha} z_{\alpha,k}^{a^*;c} \cdot \left(\delta_{\alpha,k}^{a^*} + \overrightarrow{\gamma}_{\alpha,k} \cdot \overrightarrow{y}^{a^*} \right) \cdot \hat{f}_\rho(x), \tag{5}$$

where we assume that $\overrightarrow{x}^{ac} = x \in \mathbb{N}$ (see Section 2.2 of the Appendix [5]).

3.2   BMH-K, the Kernel Submodel

Under the power-law, the kernel parameter set generating $\mathcal{H}^{ac}$ is $\boldsymbol{\Theta}^{ac} = [\theta^{ac}, d^{ac}]^\mathsf{T}$. We model $\boldsymbol{\Theta}^{ac}$ as a pair of mixture random variables taking the value

$$\log(\theta^{ac}) = \delta_{\theta,k}^a + \overrightarrow{\gamma}_{\theta,k} \cdot \overrightarrow{y}^a, \quad \log(d^{ac}) = \delta_{d,k}^a \tag{6}$$

with probability $z_{\boldsymbol{\Theta},k}^{ac}$ $(k = 1, \ldots, K_{\boldsymbol{\Theta}})$, where

$$z_{\boldsymbol{\Theta},k}^{ac} = \frac{\exp(\delta_{z_{\boldsymbol{\Theta},k}}^a + \overrightarrow{\beta}_{z_{\boldsymbol{\Theta},k}}^a \cdot \overrightarrow{x}^{ac} + \overrightarrow{\gamma}_{z_{\boldsymbol{\Theta},k}} \cdot \overrightarrow{y}^a)}{\sum_{k'=1}^{K_{\boldsymbol{\Theta}}} \exp(\delta_{z_{\boldsymbol{\Theta},k'}}^a + \overrightarrow{\beta}_{z_{\boldsymbol{\Theta},k'}}^a \cdot \overrightarrow{x}^{ac} + \overrightarrow{\gamma}_{z_{\boldsymbol{\Theta},k'}} \cdot \overrightarrow{y}^a)}. \tag{7}$$

In Eq. (7) we designate $k = 1$ as the reference class (i.e. $\delta_{z_{\boldsymbol{\Theta},1}}^a = \overrightarrow{\beta}_{z_{\boldsymbol{\Theta},1}}^a = \overrightarrow{\gamma}_{z_{\boldsymbol{\Theta},1}} = 0$).

Let $\overrightarrow{p}_{\boldsymbol{\Theta},k}^a = [\delta_{\theta,k}^a, \delta_{d,k}^a]^\mathsf{T}$ and $\overrightarrow{p}_{z_{\boldsymbol{\Theta}}}^a = \left[\delta_{z_{\boldsymbol{\Theta},2}}^a, \ldots, \delta_{z_{\boldsymbol{\Theta},K_{\boldsymbol{\Theta}}}}^a, \overrightarrow{\beta}_{z_{\boldsymbol{\Theta},2}}^a, \ldots, \overrightarrow{\beta}_{z_{\boldsymbol{\Theta},K_{\boldsymbol{\Theta}}}}^a \right]^\mathsf{T}$. The complexity of estimating two parameters (i.e. θ^{ac}, d^{ac}) makes it challenging to estimate a joint source-level MVN distribution as we did for BMH-P. To simplify, we assume independence of $(\delta_{\theta,k}^a, \delta_{d,k}^a)$ across classes. For each kernel class k, we assume $\overrightarrow{p}_{\boldsymbol{\Theta},k}^a$ is drawn from a source-level MVN distribution with mean $\overrightarrow{p}_{\boldsymbol{\Theta},k} = [\delta_{\theta,k}, \delta_{d,k}]^\mathsf{T}$ and covariance matrix $\Sigma_{\boldsymbol{\Theta},k}$. Lastly, we assume $\overrightarrow{p}_{z_{\boldsymbol{\Theta}}}^a$ is drawn from an MVN distribution with mean $\overrightarrow{p}_{z_{\boldsymbol{\Theta}}}$ and covariance matrix $\Sigma_{z_{\boldsymbol{\Theta}}}$.

$$\overrightarrow{p}_{\boldsymbol{\Theta},k}^a \sim \mathrm{MVN}\left(\overrightarrow{p}_{\boldsymbol{\Theta},k}, \Sigma_{\boldsymbol{\Theta},k} \right), \quad \Sigma_{\boldsymbol{\Theta},k} = \mathrm{D}_{\boldsymbol{\Theta},k} \cdot \Omega_{\boldsymbol{\Theta},k} \cdot \mathrm{D}_{\boldsymbol{\Theta},k}, \quad \mathrm{D}_{\boldsymbol{\Theta},k} = \mathrm{diag}(\sigma_{\overrightarrow{p}_{\boldsymbol{\Theta},k}})$$

$$\overrightarrow{p}_{z_{\boldsymbol{\Theta}}}^a \sim \mathrm{MVN}\left(\overrightarrow{p}_{z_{\boldsymbol{\Theta}}}, \Sigma_{z_{\boldsymbol{\Theta}}} \right), \quad \Sigma_{z_{\boldsymbol{\Theta}}} = \mathrm{D}_{z_{\boldsymbol{\Theta}}} \cdot \Omega_{z_{\boldsymbol{\Theta}}} \cdot \mathrm{D}_{z_{\boldsymbol{\Theta}}}, \quad \mathrm{D}_{z_{\boldsymbol{\Theta}}} = \mathrm{diag}(\sigma_{z_{\boldsymbol{\Theta}}}),$$
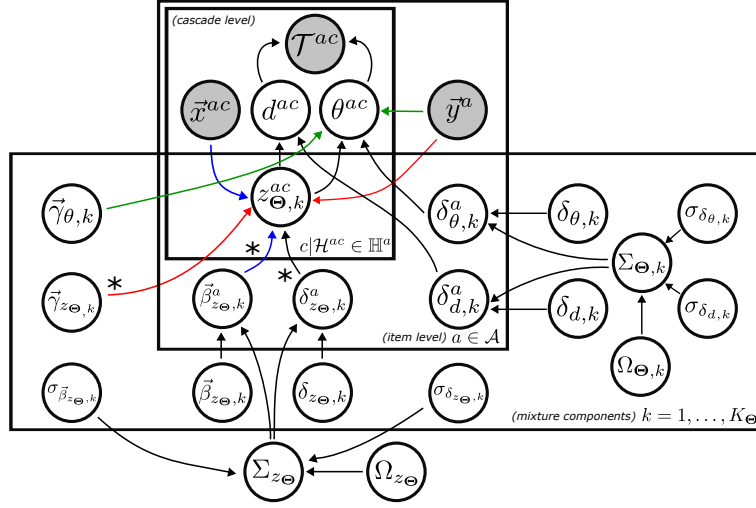
Fig. 3: Plate diagram of the BMH-K model. Shaded nodes are observables while empty nodes are latent variables. Paired colored edges indicate source nodes appearing as a product in the target node. For instance, the green edges indicate that $\overrightarrow{\gamma}_{\theta,k}$ and $\overrightarrow{y}^a$ appear as the product $\overrightarrow{\gamma}_{\theta,k} \cdot \overrightarrow{y}^a$ in the expression for θ^{ac} in Eq. (6). The same concept holds for the blue and red edges. Edges marked with * indicate dependence of the target node on the source node indexed with k and the entire set $\{1, \cdots, K_{\Theta}\}$. For instance, in Eq. (7) $z_{\Theta,k}^{ac}$ depends on $\overrightarrow{\beta}^a_{z_{\Theta,k}}$ (see the numerator) and $\overrightarrow{\beta}^a_{z_{\Theta,k'}}$ for $k' \in \{1, \cdots, K_{\Theta}\}$ (see the denominator).

where $\sigma_{\overrightarrow{p}_{\Theta,k}}, \sigma_{\overrightarrow{p}_{z_{\Theta}}}$ are standard deviation vectors and $\Omega_{\Theta,k}, \Omega_{z_{\Theta}}$ are correlation matrices.

The plate diagram for the BMH-K model is shown in Fig. 3. Variable pairs that appear as a product term are colored green, red and blue in Eqs. (6) and (7), visualized in Fig. 3 as source nodes with green, red and blue edges.

Inference and Prediction. Let $\mathcal{P}_{\Theta}$ be the parameter set for the BMH-K model. From the interevent-time distributions $\{\mathcal{T}^{ac}\}_{ac}$, we estimate the posterior distribution $\mathbb{P}(\mathcal{P}_{\Theta}|\mathcal{T}^{ac}) \propto \exp(\mathcal{L}(\mathcal{P}_{\Theta}|\{\mathcal{T}^{ac}\}_{ac}) \cdot \mathbb{P}(\mathcal{P}_{\Theta})$. The log-likelihood of $\mathcal{P}_{\Theta}$ given $\{\mathcal{T}^{ac}\}_{ac}$ is derived in Section 2.3 of Online Appendix [5]. Informative priors have to be set on $\{\delta_{\theta,k}, \delta_{d,k}, \delta_{z_{\Theta,k}}\}$ to identify the K_{Θ} classes in the (θ, d) parameter space. $(\delta_{\theta,k}, \delta_{d,k})$ and $\delta_{z_{\Theta,k}}$ identify the center and baseline proportion of the kth class, respectively. Weakly informative priors are set for the other parameters in $\mathcal{P}_{\Theta}$. Similar to the BMH-P model, we implementt[1] the BMH-K model in Stan and CmdStanPy to sample from the posterior distribution $\mathbb{P}(\mathcal{P}_{\Theta}|\mathcal{T}^{ac})$.

Table 2: Statistics of the predictive evaluation datasets.

	$CNIX - Fit$	$CNIX - Test$	$RNIX - Fit$	$RNIX - Test$
#publishers	41	41	28	28
#articles	72,009	40,506	2,682	18,116
#cascades	4,620,509	1,874,729	244,596	460,504
#tweets	42,546,067	18,235,185	1,573,909	5,139,967

The BMH-K model predicts the half-life $\hat{\tau}_{1/2}^{a^*}$ of an out-of-sample item a^* as (see Section 2.3 of the Online Appendix [5]),

$$\hat{\tau}_{1/2}^{a^*} \approx \sum_{x=0}^{\infty} \sum_{k=1}^{K_\Theta} z_{\Theta,k}^{a^*,c} \cdot e^{\delta_{d,k}^{a^*}} \cdot \left[2^{\exp\left(\delta_{\theta,k}^{a^*} + \overrightarrow{\gamma}_{\theta,k} \cdot \overrightarrow{y}^{a^*}\right)} - 1 \right] \cdot \hat{f}_\rho(x). \tag{8}$$

4   Predictive Evaluation

In this section, we introduce two evaluation datasets (Section 4.1) and assess the BMH model's performance on two tasks: cold-start popularity prediction (Section 4.2) and temporal profile generalization performance (Section 4.3), i.e. evaluating the likelihood of the interevent distribution of future cascades.

4.1   Datasets

We use two datasets from [12] for predictive evaluation, consisting of collections of Twitter retweet cascades that link articles from online news sources. The Controversial News Index ($CNIX$) dataset consists of retweet cascades mentioning articles from 41 online news publishers known for controversial content, such as `https://www.breitbart.com/`. Conversely, the Reputable News Index ($RNIX$) follows the same structure as the $CNIX$ dataset but gathers cascades linked to articles from 28 reputable publishers, such as `https://www.news.com.au/`. The tweets for both datasets were collected by the QUT Digital Media Research Centre by retrospectively querying the Twitter search endpoint for URL mentions of the articles between June 30, 2017 and Dec 31, 2019. In Table 1 we link quantities in these datasets with variables in the BMH model.

Both *CNIX* and *RNIX* are temporally split into *Fit* (i.e. training) and *Test* (i.e. evaluation) datasets. The first contains tweets published from Jun 30, 2017 to Jan 1, 2019, while the second contains tweets from Feb 1, 2019 to Dec 31, 2019. A one-month gap between *Fit* and *Test* ensures that cascades in the training data are finished before the test period. Table 2 shows summary statistics.

We use the standardized 32-D embedding of a's headline (i.e. PCA-reduced, *all-MiniLM-L6-v2* [26]) as our article feature vector $\overrightarrow{y}^a$, and the standardized log-follower count of the cascade's seed user as the cascade feature vector $\overrightarrow{x}^{ac}$.

4.2   Cold-Start Popularity Prediction

Our first task is evaluating the ability of the BMH-P model to predict cold-start popularity of unpublished content. With publisher ρ's trained BMH-P model, we predict the future popularity N^{a^*} of an out-of-sample article a^* with Eq. (5). To guide the selection of the number of mixture components K_α, we fit the BMM to each publisher in $RNIX$. We observe that the BMM-fitted $\{\alpha_i^a\}$ distribution is bimodal, corresponding to clusters of popular and unpopular cascades. See Section 3.1 of the Appendix [5] for full details. Using this result, we fit a BMH-P model for each publisher in $CNIX$ and $RNIX$ in Stan with the hyperparameter $K_\alpha = 2$. The full set of priors for the BMH-P model is listed in Section 3.2 of the Appendix [5]. Note that we use a Laplace prior on $\overrightarrow{\gamma}_{\alpha,1}, \overrightarrow{\gamma}_{\alpha,2}, \overrightarrow{\gamma}_{z_{\alpha,2}}$ to impose regularization given the high dimensionality of the article feature vector ($|\overrightarrow{y}^a| = 32$) we consider.

To evaluate the predictive power of $\overrightarrow{x}^{ac}$ and $\overrightarrow{y}^a$, apart from the full model as developed in Section 3.1 (which we call $\alpha(\overrightarrow{y}^a) + z(\overrightarrow{x}^{ac}, \overrightarrow{y}^a)$) we fit three simpler variants of BMH-P: (1) $\alpha(\overrightarrow{y}^a) + z(\overrightarrow{y}^a)$, where we set $\overrightarrow{x}^{ac} = 0$ in Eq. (3); (2) $\alpha(\varnothing) + z(\overrightarrow{y}^a)$, where set $\overrightarrow{x}^{ac} = 0$ in Eq. (3) and $\overrightarrow{y}^{ac} = 0$ in Eq. (2); and (3) $\alpha(\varnothing) + z(\varnothing)$, where we set $\overrightarrow{x}^{ac} = 0$ in Eq. (3) and $\overrightarrow{y}^{ac} = 0$ in Eqs. (2) and (3).

We compare the performance of the BMH-P model to three approaches: (1) the DMM [12], (2) the empirical Bayes (EB) approach [30], and (3) feature-based cascade-size (CR) regression models (i.e. a neural network with one hidden layer of 100 nodes) built using scikit-learn [25]. For EB and CR, we fit two variants: one using only article features (i.e. EB(y) and CR(y)) and another using both cascade and article features (i.e. EB(x,y) and CR(x,y)). We report the Average Relative Error (ARE) over the set of articles in the $Test$ datasets. Let N^a and $\hat{N}^a$ be the actual and predicted popularity of article a, then $ARE(a) = \frac{|\hat{N}^a - N^a|}{N^a}$.

Results. In the top half of Table 3, we summarize cold-start popularity prediction performance of the model variants for $CNIX/RNIX$. In both datasets the variants with only article-level features $\overrightarrow{y}^a$ and without the cascade-level features $\overrightarrow{x}^{ac}$ show minimal performance gain ($RNIX$) or even worse performance ($CNIX$) over the no-feature $\alpha(\varnothing) + z(\varnothing)$ model. The full model $\alpha(\overrightarrow{y}^a) + z(\overrightarrow{x}^{ac}, \overrightarrow{y}^a)$ significantly outperforms each simpler variant, highlighting the importance of the seed user's popularity as a predictor of final popularity [1].

We compare the performance of the best-performing BMH-P model with the benchmarks in the top row of Fig. 4(a) and Fig. 4(b). We can see that the BMH-P model outperforms each benchmark based on median performance. We note that in each task, the benchmarks that only have article features ($CR(y)$ and $EB(y)$) outperform the corresponding benchmarks that also include cascade features ($CR(x,y)$ and $EB(x,y)$). However, our ablation results show that the best-performing BMH-P model includes both the cascade and article features. This implies that the added structure of the BMH-P model jointly leverages the article- and cascade-level information better than the benchmarks.

Table 3: Popularity prediction and generalization results. We show the median $(25^{th}, 75^{th}$ quantiles) for BMH with different feature components removed. Lower ARE/NLL mean better performance. The best score across variants is in bold.

Popularity (ARE)	$CNIX$	$RNIX$
$\alpha(\varnothing) + z(\varnothing)$	0.707 (0.334, 1.513)	0.644 (0.335, 0.921)
$\alpha(\varnothing) + z(\overrightarrow{y}^a)$	0.708 (0.336, 1.497)	0.666 (0.339, 1.033)
$\alpha(\overrightarrow{y}^a) + z(\overrightarrow{y}^a)$	0.738 (0.370, 1.316)	0.643 (0.325, 0.953)
$\alpha(\overrightarrow{y}^a) + z(\overrightarrow{x}^{ac}, \overrightarrow{y}^a)$	**0.646** (0.313, 0.935)	**0.635** (0.342, 0.932)

Generalization (NLL)	$CNIX$	$RNIX$
$\theta(\varnothing) + z(\varnothing)$	-3.841 (-5.293, -2.717)	-2.564 (-3.231, -2.031)
$\theta(\varnothing) + z(\overrightarrow{y}^a)$	-3.782 (-4.873, -2.683)	-2.550 (-3.226, -1.988)
$\theta(\overrightarrow{y}^a) + z(\overrightarrow{y}^a)$	-3.649 (-4.816, -2.617)	**-2.689** (-3.492, -2.117)
$\theta(\overrightarrow{y}^a) + z(\overrightarrow{x}^{ac}, \overrightarrow{y}^a)$	**-4.013** (-5.766, -2.714)	-2.645 (-3.450, -2.063)

4.3   Temporal Profile Generalization Performance

Our second task is evaluating the performance of the BMH-K model in capturing the inter-arrival distribution of future cascades of unpublished articles. Given publisher ρ's trained BMH-K model, we calculate the log-likelihood $\mathcal{L}(\mathcal{P}_{\Theta}|\{\mathcal{T}^{a^*c}\})$ of the inter-arrival distribution $\{\mathcal{T}^{a^*c}\}$ of an unpublished article a^*.

To guide the selection of the number of mixture components K_{Θ}, we fit the KMM to each publisher in $RNIX$. We observe that the KMM-fitted $\{\theta_i^a, d_I^a\}$ distribution is trimodal, corresponding to clusters of usual, slow- and fast-diffusing cascades cascades. See Section 3.1 of the Appendix [5] for full details. Using this result, we fit a BMH-K model for each publisher in $CNIX$ and $RNIX$ in Stan with the hyperparameter $K_{\Theta} = 3$. The full set of priors for the BMH-K model is listed in Section 3.3 of the Appendix [5]. Note that we use a Laplace prior on $\overrightarrow{\gamma}_{\Theta,2}, \overrightarrow{\gamma}_{\Theta,3}, \overrightarrow{\gamma}_{z\Theta,2}, \overrightarrow{\gamma}_{z\Theta,3}$ to impose regularization given the high dimensionality of the article feature vector ($|\overrightarrow{y}^a| = 32$) we consider.

In addition to the full BMH-K model developed in Section 3.2 (which we call $\theta(\overrightarrow{y}^a) + z(\overrightarrow{x}^{ac}, \overrightarrow{y}^a)$) we fit three progressively simpler variants analogous to the ablation for the BMH-P model: $\theta(\overrightarrow{y}^a) + z(\overrightarrow{y}^a)$, $\theta(\varnothing) + z(\overrightarrow{y}^a)$, and $\theta(\varnothing) + z(\varnothing)$. To evaluate performance, we calculate the loglikelihood $\mathcal{L}(\mathcal{P}_{\Theta}|\{\mathcal{T}^{ac}\})$ of inter-arrival times $\{\mathcal{T}^{ac}\}_{a \in \mathcal{A}}$ over articles in the $Test$ datasets. Since we are evaluating on likelihood, we use generative models as benchmarks: the DMM, EB(y), EB(x,y), and publisher-level joint HP (see Sec. 1.1 of the Appendix [5]).

Results. In the lower half of Table 3, we see that for $CNIX$ each additional model component improves the log-likelihood, and that the full model $\alpha(\overrightarrow{y}^a) + z(\overrightarrow{x}^{ac}, \overrightarrow{y}^a)$ has the best performance. For $RNIX$ we observe that the variant without the seed user follower count, i.e., $\theta(\overrightarrow{y}^a) + z(\overrightarrow{y}^a)$, has the best performance. This finding suggests that in cascades related to reputable media articles, the seed user is not as influential in determining how long a cascade
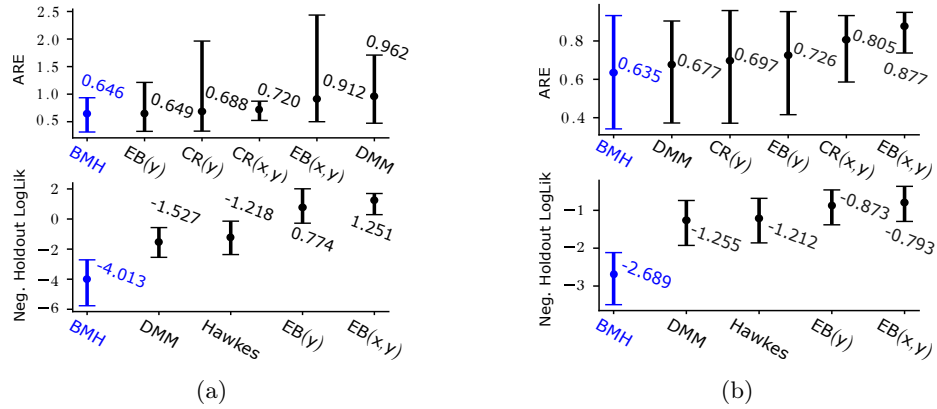
Fig. 4: Predictive performance for (a) CNIX and (b) RNIX. The dots indicate the median and the error bars give the $25^{th}/75^{th}$ quantiles. We compare the BMH with the DMM [12], EB [30], cascade-size (CR) models, and the joint HP.

unfolds. In contrast, for controversial media articles, the seed user plays a significant role. We posit this is because the more fringe messaging in controversial media spreads through topical social groups (like conspiracy theorists, QAnon sympathizers and far-right supporters) [3,11]. As a result, the first endorser is particularly important to legitimize content within the group. This is in contrast with the publicizing of traditional media articles on social media, where the most important factor is the publisher's reputation. In the bottom row of Fig. 4(a) and Fig. 4(b), we see that similar to the popularity prediction task, the BMH-K model outperforms all benchmarks on median performance for both datasets.

5   What-If? Headline Style Profiling

This section performs a counter-factual analysis to show that BMH successfully captures the relationship between headline writing style (i.e. neutral, clickbait or inflammatory) and content popularity and half-life. We run a 'What-If?' experiment, taking headlines of different writing styles and using the trained BMH models to infer how these headlines would perform under different publishers.

We utilize $HEADLINES$, a dataset of $1,227$ article headlines collected using the news aggregation platform The Daily Edit [13]. The headlines come from four topics (Top Stories, Australia, Finance, and Climate Change) and six media sources (Daily Telegraph, Sky News, Sunday Morning Herald, The Guardian, news.com.au). Each headline was examined and sorted into one of three categories based on its informational and emotional content: neutral (N=727), clickbait (N=438) and inflammatory (N=62). Neutral headlines are detailed and appropriate, avoiding unnecessary information or emotive language, e.g. *'Australia's top military officer in the UK speaks ahead of Queen's funeral.'* Clickbait lacks informational and/or emotive quality without being misleading or in-
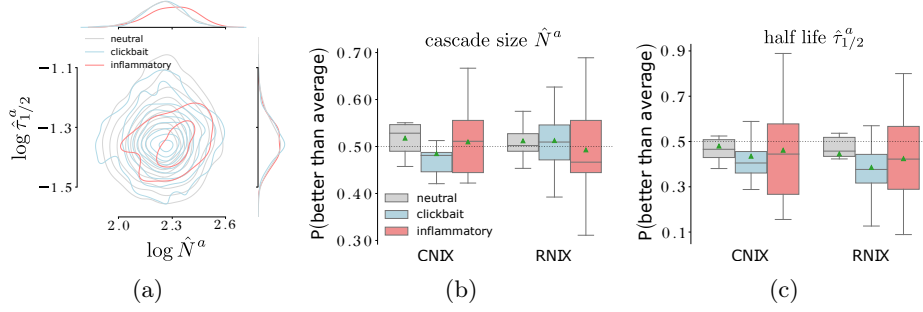
Fig. 5: (a) Distribution of predicted half-life $\log \hat{\tau}^a_{1/2}$ vs. cascade size $\log \hat{N}^a$ for each article in $HEADLINES$ using the `news.com.au` BMH model. (b and c) Probability that an article performs better than the publisher average, for each headline style across $CNIX$ and $RNIX$: (b) cascade size $\hat{N}^a$; (c) half life $\hat{\tau}^a_{1/2}$.

flammatory, often designed to attract attention, e.g. *'Bizarre sight spotted amid Aussie floods.'* Inflammatory headlines contain unnecessary details, often on serious topics, and may include inappropriate emotional language or details that reinforce negative stereotypes, e.g. *'Absolutely disgraceful': AFL fans blasted.'*

We use the trained publisher-level BMH models in Section 4 to predict performance of article headlines for each publisher in $CNIX$ and $RNIX$: expected cascade size (Eq. (5)) (setting $\hat{C}_\rho = 1$) and half-life (Eq. (8)). We use the variants that include only item features (i.e. $\alpha(\overrightarrow{y}^a) + z(\overrightarrow{y}^a)$ for BMH-P and $\theta(\overrightarrow{y}^a) + z(\overrightarrow{y}^a)$ for BMH-K) since cascade features are not available in this counter-factual setting.

Results. We apply the trained the BMH-P/-K models of each publisher ρ in $\{CNIX, RNIX\}$ to each of the $1,227$ article headlines in $HEADLINES$ to infer the article's peformance if it were published under ρ. We summarize the predictions with a publisher-level performance heatmap ($\log \hat{N}^a$ vs. $\log \hat{\tau}^a_{1/2}$), where we differentiate the performance of neutral, clickbait and inflammatory headlines by aggregating the predictions of each headline style as contour plots. Fig. 5a exemplifies the performance heatmap for the $RNIX$ publisher `news.com.au`. For this news source, we see that inflammatory headlines appear to have much higher popularity than neutral or clickbait headlines, while there is not much difference in half-life across headline styles. This is somewhat expected, as this publisher is known for its tabloid tendencies, focusing on "celebrity gossip, travel, lifestyle, sport, business, technology, money, and real estate", according to Media Bias Fact Check (MBFC) [19]. MBFC also rates its factual reporting as "MOSTLY FACTUAL" due to the occasional use of poor sources. We observe differences in the patterns for the headline styles across publishers (see Section 4.1 of the Online Appendix [5]), implying that effective headlines for one publisher might not be effective for another, and that the BMH model learns these differences.

To summarise the differences across the categories $CNIX$ and $RNIX$, we compute the probability that each headline performs better – has a larger predicted cascade size or longer predicted half-life based on the BMH– than the publisher average based on the publisher's historical data. In Figs. 5b and 5c we show the distribution of these probabilities for each category and headline style.

We have three observations for the popularity probabilities in Fig. 5b. First, we see that for $CNIX$, neutral headlines are effective (i.e. median better-than-average probability $> 50\%$). In contrast, clickbait headlines are ineffective (i.e. median better-than-average probability $< 50\%$). We link this result to the known inverse U-shaped relationship between clickbait volume and audience engagement [33], where too little or too much clickbait leads to suboptimal attention, suggesting the existence of a *sweet spot* for clickbait use. The over-prevalence of clickbait in controversial media outlets results in clickbait fatigue among readers [17], leading to diminished effectiveness of clickbait headlines observed in Fig. 5b.

Second and interestingly, we see that for $RNIX$ clickbait tends to perform better than neutral headlines. This is explained by Rony et al [28], who show that traditional news-oriented media consist of only 22% clickbait headlines while unreliable media consists of 39% clickbait based on a large sample of headlines. Since reputable media publishers have lower clickbait volume than controversial outlets, they are closer to the sweet spot for clickbait usage, retaining its effectiveness for drawing audience engagement. We do see a larger variance for clickbait for $RNIX$ compared to $CNIX$, suggesting that clickbait effectiveness is inconsistent and may not resonate universally, linking to the fact that clickbait strategies are only successful with certain audience segments [21].

Third, we observe large variance of performance for inflammatory headlines in both categories, indicative of the polarizing nature of this headline style. Inflammatory headlines tend to perform better in controversial outlets.

For the half-life probabilities (Fig. 5c), we see similar results, except that neutral headlines in both categories have higher half-life than clickbait, demonstrating the ephemerality of clickbait [16] irrespective of where it is published.

6   Conclusion and Future Work

This paper proposes the Bayesian Mixture Hawkes (BMH) model, a hierarchical mixture model of Hawkes processes capable of learning the influence of item- and cascade-level features on spread dynamics. We demonstrate the applicability of the BMH model on two retweet cascade datasets that reference articles from reputable and controversial online news sources and show that the BMH model outperforms benchmark models in cold-start popularity prediction and temporal profile generalization performance. We apply the trained BMH models to a dataset of article headlines written in different headline styles and show differences in performance of headline styles across reputable and controversial outlets.

Limitations and Future Work. We use the Hawkes process as the building block of the BMH model since it does not require the branching structure of diffusion cascades for inference. This choice is driven by data limitations on Twitter, where the branching structure of content shares is not accessible.

We propose two improvements. First, the BMH model assumes that α and Θ depend only on cascade- and content-level features. We can allow α and Θ to vary per event by including event-level features, which can be achieved by using the parametric Hawkes process [14] or Tweedie-Hawkes [15]. Second, the BMH model assumes a fixed number of popularity/kernel classes, obtained empirically by pre-fitting with the DMM. We can learn the manifest number of components directly from the data by assuming an infinite number of components via nonparametric Bayesian methods, such as using a Dirichlet Process prior [22].

We aim to develop the BMH model as a cold-start headline optimization tool by combining it with generative AI (e.g. ChatGPT [23]). The system would work in a 'generate-then-evaluate' loop, where headlines are generated automatically by ChatGPT and then we apply the BMH model to rank the generations.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyone's an influencer: quantifying influence on twitter. In: WSDM 2011
2. Bao, P.: Modeling and predicting popularity dynamics via an influence-based self-excited hawkes process. In: CIKM 2016
3. Booth, E., Lee, J., Rizoiu, M.A., Farid, H.: Conspiracy, misinformation, radicalisation: understanding the online pathway to indoctrination and opportunities for intervention. Journal of Sociology (feb 2024)
4. Borel, E.: Sur l'emploi du theoreme de Bernoulli pour faciliter le calcul d'une infinite de coefficients. CR Acad. Sci. Paris (1942)
5. Calderon, P., Rizoiu, M.A.: Appendix: What Drives Online Popularity: Author, Content or Sharers? `https://arxiv.org/pdf/2406.03390.pdf#page=19` (2024)
6. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. Journal of statistical software **76**(1) (2017)
7. Center, P.R.: Pew research center (2023), `https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/`
8. Ghassemi, M., Dalmasso, N., Lamba, S., Potluru, V., Balch, T., Shah, S., Veloso, M.: Online learning for mixture of multivariate hawkes processes. In: ICAIF 2022
9. Gomez-Rodriguez, M., Balduzzi, D., Schölkopf, B.: Uncovering the temporal dynamics of diffusion networks. In: ICML 2011 (2011)
10. Hawkes, A.G.: Spectra of some self-exciting and mutually exciting point processes. Biometrika **58**(1), 83–90 (1971)

11. Johns, A., Bailo, F., Booth, E., Rizoiu, M.A.: Labelling, shadow bans and community resistance: did meta's strategy to suppress rather than remove covid misinfo and conspiracy theory on facebook slow the spread? Media International Australia
12. Kong, Q., Rizoiu, M.A., Xie, L.: Describing and predicting online items with reshare cascades via dual mixture self-exciting processes. In: CIKM 2020 (2020)
13. Lee, J., Booth, E., Farid, H., Rizoiu, M.A.: Misinformation is not about Bad Facts: An Analysis of the Production and Consumption of Fringe Content (mar 2024), http://arxiv.org/abs/2403.08391
14. Li, L., Zha, H.: Learning parametric models for social infectivity in multi-dimensional hawkes processes. In: AAAI 2014
15. Li, T., Ke, Y.: Tweedie-hawkes processes: Interpreting the phenomena of outbreaks. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)
16. Liao, Y., Wang, S., Han, E., Lee, J., Lee, D.: Characterization and early detection of evergreen news articles. In: ML and Knowledge Discovery in Databases (2020)
17. Lischka, J., Garz, M.: Clickbait news & algorithmic curation: Game theory framework of the relation bet. journalism, users and platforms. New Media & Society
18. Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.F., Cha, M.: Detecting rumors from microblogs with recurrent neural networks. In: IJCAI 2016 (2016)
19. Media Bias Fact Check: News.com.au – Bias and Credibility (2024), https://mediabiasfactcheck.com/news-com-au/
20. Mishra, S., Rizoiu, M.A., Xie, L.: Feature driven and point process approaches for popularity prediction. In: CIKM 2016. pp. 1069–1078
21. Mukherjee, P., Dutta, S., De Bruyn, A.: Did clickbait crack the code on virality? Journal of the Academy of Marketing Science **50**(3), 482–502 (2022)
22. Navarro, D.J., Griffiths, T.L., Steyvers, M., Lee, M.D.: Modeling individual differences using dirichlet processes. Journal of mathematical Psychology **50**(2), 101–122 (2006)
23. OpenAI: Chatgpt (2023), https://openai.com/chatgpt, software tool
24. Parikh, S.B., Patil, V., Makawana, R., Atrey, P.K.: Towards impact scoring of fake news. In: MIPR 2019. IEEE
25. Pedregosa, F., et.al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
26. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: EMNLP 2019
27. Rizoiu, M.A., Xie, L., Sanner, S., Cebrian, M., Yu, H., Van Hentenryck, P.: Expecting to be HIP: Hawkes Intensity Processes for Social Media. In: WWW 2017
28. Rony, M.M.U., Hassan, N., Yousuf, M.: Diving deep into clickbaits: Who use them to what extents in which topics with what effects? (2017)
29. Sharma, A., Ghosh, A., Fiterau, M.: Generative sequential stochastic model for marked point processes. In: ICML Time Series Workshop (2019)
30. Tan, W.H., Chen, F.: Predicting the popularity of tweets using internal and external knowledge: an empirical bayes type approach. AStA **105**(2), 335–352 (2021)
31. Team, S.D.: Cmdstanpy (0.9.76) (2023), https://pypi.org/project/cmdstanpy
32. Tsagkias, M., Weerkamp, W., de Rijke, M.: Predicting the volume of comments on online news stories. In: CIKM 2009 (2009)
33. W, Z., W, D., Y, B., et al.: Seeing is not always believing: an exploratory study of clickbait in wechat. Internet Research **30**(3), 1043–1058 (2020)
34. Xu, H., Zha, H.: A dirichlet mixture model of hawkes processes for event sequence clustering. Advances in neural information processing systems **30** (2017)
35. Zhao, Q., Erdogdu, M.A., He, H.Y., Rajaraman, A., Leskovec, J.: Seismic: A self-exciting point process model for predicting tweet popularity. In: SIGKDD 2015

Supplementary Material for *What Drives Online Popularity: Author, Content or Sharers? Estimating Spread Dynamics with Bayesian Mixture Hawkes*

Pio Calderon$^{(\boxtimes)\,[0000-0002-8747-8917]}$ and Marian-Andrei Rizoiu$^{[0000-0003-0381-669X]}$

University of Technology Sydney, Australia
piogabrielle.b.calderon@student.uts.edu.au
marian-andrei.rizoiu@uts.edu.au

1 Background Material

1.1 Hawkes Process

Inference. Given a cascade $\mathcal{H}$ of length N, i.e. an ordered collection of time stamps $\{t_i\}_{i=1}^N$ observed until some terminal time $T \geq t_N$, we can estimate the parameters of the Hawkes process $(\mu^*, \alpha^*, \boldsymbol{\Theta}^*)$ that generated the data by maximizing the log-likelihood function,

$$\mathcal{L}(\mu, \alpha, \boldsymbol{\Theta}|\mathcal{H} = \{t_j\}_{j=1}^N) = \sum_{j=1}^N \log \lambda(t_j; \mu, \alpha, \boldsymbol{\Theta}) - \int_0^T \lambda(s; \mu, \alpha, \boldsymbol{\Theta}) \mathrm{d}s. \quad (1)$$

This approach can be extended to the case of a collection of cascades $\mathbb{H} = \{\mathcal{H}_i\}$, where the best-fitting Hawkes process is obtained by maximizing the sum of the log-likelihood functions,

$$\mathcal{L}(\mu, \alpha, \boldsymbol{\Theta}|\mathbb{H}) = \sum_{\mathcal{H}_i \in \mathbb{H}} \mathcal{L}(\mu, \alpha, \boldsymbol{\Theta}|\mathcal{H}_i). \quad (2)$$

Prediction. The fitted Hawkes process can be leveraged to predict the cascade size $\hat{N}$ of a new cascade $\mathcal{H} \in \mathbb{H}$:

$$\hat{N} = \mathbb{E}[N|\alpha] = \mathrm{logit}(\alpha). \quad (3)$$

1.2 Dual Mixture Model

Inference. Given the pre-defined number of components K, we obtain the Borel mixture model M^B by maximizing the following log-likelihood function,

$$\mathcal{L}_{BMM} = \sum_{\mathcal{H}_i \in \mathbb{H}} \log \sum_{k=1}^K p_k^B \mathbb{B}(N_i|\alpha_k^*). \quad (4)$$

Note that the DMM is not formulated as a Bayesian model in [2] and the Expectation-Maximization (EM) algorithm [4] is employed to maximize $\mathcal{L}_{BMM}$. Similarly, the kernel mixture model M^g is obtained by applying the EM algorithm to the kernel log-likelihood

$$\mathcal{L}_{KMM} = \sum_{\mathcal{H}_i \in \mathbb{H}} \log \sum_{k=1}^{K} p_k^g f^g(\mathcal{H}_i | \boldsymbol{\Theta}_k^*), \tag{5}$$

where $f(\mathcal{H}_i | \boldsymbol{\Theta}) = \prod_{t_j \in \mathcal{H}_i} \sum_{t_z < t_j} g(t_j - t_z | \boldsymbol{\Theta})$.

Cold Start Popularity Prediction. Assume that we are given a collection of related cascade groups $\{\mathbb{H}_a\}_{a \in \mathcal{A}}$. For instance, suppose $\mathcal{A}$ is a set of news articles from a common online publisher ρ and $\mathbb{H}_a$ is the set of retweet cascades discussing article a. Give a yet-to-be-published article $a^* \notin \mathcal{A}$, we wish to model its popularity $\hat{N}^{a^*}$ by learning from historical data $\{\mathbb{H}_a\}_{a \in \mathcal{A}}$.

To do this, we can construct a publisher-level popularity model M_ρ^B by fitting an independent BMM M_a^B (with K_a classes) to each $\mathbb{H}_a$ and then collecting these as a mixture M_ρ^B over $\mathcal{A}$, i.e.,

$$M_\rho^B = \bigcup_{a \in \mathcal{A}} M_a^B = \bigcup_{a \in \mathcal{A}} \left\{ \left(\alpha_1^a, \frac{p_1^{B,a}}{|\mathcal{A}|} \right), \dots, \left(\alpha_{K_a}^a, \frac{p_{K_a}^{B,a}}{|\mathcal{A}|} \right) \right\}, \tag{6}$$

where $(\alpha_i^a, p_i^{B,a}) \in M_a^B$. We can estimate the cold-start popularity of a new article a^* as

$$\hat{N}^{a^*} = \hat{C}_\rho \cdot \mathbb{E}_{M_\rho^B} \left[\text{logit}(\alpha) \right] = \hat{C}_\rho \cdot \sum_{a \in \mathcal{A}} \sum_{i=1}^{K_a} \text{logit}(\alpha_i^a) \cdot \frac{p_i^{B,a}}{|\mathcal{A}|}, \tag{7}$$

where $\hat{C}_\rho$ is an estimate of the cascade count of article a^*, which we can take as the average cascade count of articles in $\mathcal{A}$.

1.3   Bayesian Hierarchical Modeling

Let θ be a parameter set of a generative process $\mathcal{P}$ and $\mathcal{D}$ be a sample from $\mathcal{P}$. Bayesian inference involves (1) quantifing our prior belief on θ through a prior distribution $\mathbb{P}(\theta)$, which could be uninformative or based on expert opinion, and then (2) updating $\mathbb{P}(\theta)$ using the data $\mathcal{D}$, with the likelihood function $\mathcal{L}(\mathcal{D}|\theta)$ serving as our weight on θ. Our result is the posterior distribution $\mathbb{P}(\theta|\mathcal{D})$, which combines our beliefs on θ based on our prior and the data:

$$\mathbb{P}(\theta|\mathcal{D}) \propto \mathcal{L}(\mathcal{D}|\theta) \cdot \mathbb{P}(\theta). \tag{8}$$

One advantage of Bayesian inference is its ability to accommodate the hierarchical structure of our dataset. For example, suppose that we have N data points $\{x_i\}$ which are sampled from some generative model $\mathcal{P}(\theta)$. Additionally,

we are given information that each data point belong to one of m related groups. We can handle this information in three ways.

First, we ignore it and assume that all groups are drawn from the same generative model, i.e. $x_i \sim \mathcal{P}(\theta)$. This approach ignores variability across groups.

Second, we assume that the groups are independent from one another and fit a separate θ_j for each group, i.e. $x_i \sim \mathcal{P}(\theta_j)$. This approach ignores the fact that the groups are related.

The third approach, Bayesian hierarchical modeling, offers a compromise between these two by allowing variation across groups. Here, we assume that each group j has its own θ_j parameter, and $x_{j[i]} \sim \mathcal{P}(\theta_j)$, where $j[i]$ is read as 'the group data point i belongs to'. We assume that $\{\theta_j\}$ are not independent but are samples from a group-level distribution $\mathcal{Q}$ parametrized by a group-level parameter θ_{group}, i.e. $\theta_j \sim \mathcal{Q}(\theta_{group})$. Under this hierarchical framework, $\mathcal{Q}(\theta_{group})$ acts a prior for each parameter θ_j. Specifying a prior distribution for θ_{group} completes the Bayesian hierarchical model. Our posterior is a joint distribution over each group's parameter θ_j and the group-level parameter θ_{group}.

2   Additional Material for Main Text Section 3

2.1   Complete Table of Notation

In Table 1 we show the full set of notation, BMH model parameters, their interpretation and real-world mapping.

2.2   BMH-P Model

Assumptions. In Main Text Eqs. (2) and (3) we assume that the item-level features $\vec{y}^a$ influence the location (i.e. mean) and membership probability of each popularity class k, while the cascade-level features $\vec{x}^{ac}$ influence only the membership probability. As a concrete example, if we have two popularity classes (popular and unpopular), $\mathcal{A}$ being a set of articles, $\vec{y}^a$ being the headline embedding vector of article a, and $\vec{x}^{ac}$ the follower count of the initiator of cascade c, our assumptions imply how large a cascade will turn out to be (Main Text Eq. (2)) is influenced only by article content $\vec{y}^a$, but whether a cascade will be popular or not is influenced by both article content $\vec{y}^a$ and follower count $\vec{x}^{ac}$.

Likelihood Function. The log-likelihood of $\mathcal{P}_\alpha$ given the set of cascade sizes $\{N_{ac}\}_{ac}$ can be derived as:

$$
\begin{aligned}
\mathcal{L}(\mathcal{P}_\alpha|\{N_{ac}\}_{ac}) &= \log \mathbb{P}(\{N_{ac}\}_{ac}|\mathcal{P}_\alpha) \\
&= \log \prod_{a\in\mathcal{A}} \prod_{\mathcal{H}^{ac}\in\mathbb{H}^a} \mathbb{P}(N_{ac}|\mathcal{P}_\alpha) \\
&= \log \prod_{a\in\mathcal{A}} \prod_{\mathcal{H}^{ac}\in\mathbb{H}^a} \mathbb{P}(N_{ac}|\alpha^{ac}) \\
&\overset{(a)}{=} \log \prod_{a\in\mathcal{A}} \prod_{\mathcal{H}^{ac}\in\mathbb{H}^a} \mathbb{B}(N_{ac}|\alpha^{ac}) \\
&\overset{(b)}{=} \log \prod_{a\in\mathcal{A}} \prod_{\mathcal{H}^{ac}\in\mathbb{H}^a} \sum_{k=1}^{K_\alpha} z_{\alpha,k}^{ac} \cdot \mathbb{B}(N_{ac}|\text{inv-logit}(\delta_{\alpha,k}^a + \vec{\gamma}_{\alpha,k}\cdot\vec{y}^a)) \\
&= \sum_{\mathcal{H}^{ac}\in\mathbb{H}^a, a\in\mathcal{A}} \log \sum_{k=1}^{K_\alpha} z_{\alpha,k}^{ac}\cdot\mathbb{B}(N_{ac}|\text{inv-logit}(\delta_{\alpha,k}^a + \vec{\gamma}_{\alpha,k}\cdot\vec{y}^a)),
\end{aligned}
$$

where in (a) we use the fact that the cascade size of a Hawkes process is Borel-distributed with parameter α^{ac} and in (b) we note that the BMH-P model specifies a^{ac} as a mixture over the K_α classes, weighted by the membership probabilities $\{z_{\alpha,k}^{ac}\}$.

Cold-Start Popularity. First, from our dataset $\mathcal{A}$, compute $\hat{C}_\rho$ as the average cascade count for an article in $\mathcal{A}$ and $\hat{f}_\rho(\vec{x}^{ac}|\vec{y}^a)$ as the empirical probability density of the cascade feature vector $\vec{x}^{ac}$ given item feature vector $\vec{y}^a$.

Second, from Main Text Eq. (4) we draw the parameter set $\vec{p}_\alpha^{u*}$ for the out-of-sample item a^*. Consider an arbitrary cascade c of item a^* with feature vector $\vec{x}^{a^*c}$. The expected cascade size of c is given by the expectation of Main Text Eq. (2) over the K_α popularity classes $\mathbb{E}_{z_{\alpha,k}^{a^*c}}[\mathbb{E}[N^{a^*c}]|\vec{x}^{a^*c}, \vec{y}^{a^*}]$. Since c is arbitrary, we need to average $\vec{x}^{a^*c}$ out. Hence, our expected cascade size is $\mathbb{E}_{\vec{x}^{a^*c}}\mathbb{E}_{z_{\alpha,k}^{a^*c}}\left[\mathbb{E}[N^{a^*c}]|\vec{x}^{a^*c}, \vec{y}^{a^*}\right]$. Our estimate $\hat{N}^{a^*}$ of item a^*'s popularity is then given by

$$
\begin{aligned}
\hat{N}^{a^*} &= \hat{C}_\rho \cdot \mathbb{E}_{\vec{x}^{a^*c}}\mathbb{E}_{z_{\alpha,k}^{a^*c}}\left[\mathbb{E}[N^{a^*c}]|\vec{x}^{a^*c}, \vec{y}^{a^*}\right] \\
&= \hat{C}_\rho \cdot \mathbb{E}_{\vec{x}^{a^*c}}\mathbb{E}_{z_{\alpha,k}^{a^*c}}\left[\text{logit}(\alpha^{a^*c})|\vec{x}^{a^*c}, \vec{y}^{a^*}\right] \\
&\overset{(a)}{=} \hat{C}_\rho \cdot \mathbb{E}_{\vec{x}^{a^*c}} \sum_{k=1}^{K_\alpha} z_{\alpha,k}^{a^*,c} \cdot \left(\delta_{\alpha,k}^{a^*} + \vec{\gamma}_{\alpha,k}\cdot\vec{y}^{a^*}\right) \\
&\overset{(b)}{=} \hat{C}_\rho \cdot \int \sum_{k=1}^{K_\alpha} z_{\alpha,k}^{a^*,c} \cdot \left(\delta_{\alpha,k}^{a^*} + \vec{\gamma}_{\alpha,k}\cdot\vec{y}^{a^*}\right) \cdot f_\rho(\vec{x}^{a^*c}|\vec{y}^{a^*}) \cdot d\vec{x}^{a^*c} \\
&\overset{(c)}{\approx} \hat{C}_\rho \cdot \int \sum_{k=1}^{K_\alpha} z_{\alpha,k}^{a^*,c} \cdot \left(\delta_{\alpha,k}^{a^*} + \vec{\gamma}_{\alpha,k}\cdot\vec{y}^{a^*}\right) \cdot \hat{f}_\rho(\vec{x}^{a^*c}) \cdot d\vec{x}^{a^*c},
\end{aligned}
$$

where in (a) we use the fact that the BMH-P model specifies a^{ac} as a mixture over the K_α classes, weighted by the membership probabilities $\{z_{\alpha,k}^{ac}\}$, in (b) we marginalize over the unobserved cascade-level features $\vec{x}^{a^*c}$ in a cold-start setup, and in (c) we use the simplification $f_\rho(\vec{x}^{ac}|\vec{y}^a) \approx \hat{f}_\rho(x)$ as detailed in the main text.

To simplify this expression, we impose two additional assumptions on the feature vectors. First, assume our cascade feature vector is one-dimensional, discrete and nonnegative (for instance, this may be the follower count of the seed user). This simplifies our probability density $\hat{f}_\rho(\vec{x}^{ac}|\vec{y}^a)$ into a probability mass function over $x \in \mathbb{N} \cup \{0\}$, converting the integral over $\vec{x}^{ac}$ into a sum. Second, in practice we usually will not have enough variance across $\vec{y}^a$ to build $\hat{f}_\rho(\vec{x}^{ac}|\vec{y}^a)$ reliably, and so we assume that $\vec{x}^{ac}$ is independent of $\vec{y}^a$. These two assumptions allow us to write $\hat{f}_\rho(\vec{x}^{ac}|\vec{y}^a) \approx \hat{f}_\rho(x)$. Our expression simplifies to

$$\hat{N}^{a^*} \approx \hat{C}_\rho \cdot \sum_{x=0}^{\infty} \sum_{k=1}^{K_\alpha} z_{\alpha,k}^{a^*,c} \cdot \left(\delta_{\alpha,k}^{a^*} + \vec{\gamma}_{\alpha,k} \cdot \vec{y}^{a^*} \right) \cdot \hat{f}_\rho(x), \tag{9}$$

2.3 BMH-K Model

Assumptions. In Main Text Eqs. (6) and (7), we assume that the item-level features $\vec{y}^a$ influence the location of θ^{ac} and not d^{ac}. We found that including influence of item-level features $\vec{y}^a$ on both parameters leads to identifiability issues in the BMH-K model. We assume $\vec{y}^a$ influence the location (i.e. mean) and membership probability of each popularity class k, while the cascade-level features $\vec{x}^{ac}$ influence only the membership probability. As a concrete example, if we have two kernel classes (slow and fast), $\mathcal{A}$ being a set of articles, $\vec{y}^a$ being the headline embedding vector of article a, and $\vec{x}^{ac}$ the follower count of the initiator of cascade c, our assumptions imply the speed at which a cascade will diffuse (Main Text Eq. (6)) is influenced only by article content $\vec{y}^a$, but whether a cascade will be slow or fast is influenced by both article content $\vec{y}^a$ and follower count $\vec{x}^{ac}$.

Likelihood Function. The log-likelihood of $\mathcal{P}_{\boldsymbol{\Theta}}$ given the set of interevent-time distributions $\{\mathcal{T}^{ac}\}_{ac}$ is

$$
\mathcal{L}(\mathcal{P}_{\boldsymbol{\Theta}}|\{\mathcal{T}^{ac}\}_{ac}) = \log \mathbb{P}(\{\mathcal{T}^{ac}\}_{ac}|\mathcal{P}_{\boldsymbol{\Theta}})
$$

$$
\overset{(a)}{=} \log \prod_{\mathcal{H}^{ac} \in \mathbb{H}^a, a \in \mathcal{A}} \left[\prod_{t_j \in \mathcal{H}^{ac}, j \geq 1} \sum_{t_z < t_j} g(t_j - t_z|\theta^{ac}, d^{ac}) \right]
$$

$$
\overset{(b)}{=} \log \prod_{\mathcal{H}^{ac} \in \mathbb{H}^a, a \in \mathcal{A}} f(\mathcal{H}^{ac}|\theta^{ac}, d^{ac})
$$

$$
\overset{(c)}{=} \log \prod_{\mathcal{H}^{ac} \in \mathbb{H}^a, a \in \mathcal{A}} \left[\sum_{k=1}^{K_{\boldsymbol{\Theta}}} z_{\boldsymbol{\Theta},k}^{ac} \cdot f(\mathcal{H}^{ac}|e^{\delta_{\theta,k}^a + \vec{\gamma}_{\theta,k} \cdot \vec{y}^a}, e^{\delta_{d,k}^a}) \right]
$$

$$
= \sum_{\mathcal{H}^{ac} \in \mathbb{H}^a, a \in \mathcal{A}} \log \sum_{k=1}^{K_{\boldsymbol{\Theta}}} z_{\boldsymbol{\Theta},k}^{ac} \cdot f(\mathcal{H}^{ac}|e^{\delta_{\theta,k}^a + \vec{\gamma}_{\theta,k} \cdot \vec{y}^a}, e^{\delta_{d,k}^a}),
$$

where in (a) we make use of the likelihood for the interevent-time distribution for separable Hawkes processes as derived in [2], in (b) we set $f(\mathcal{H}|\theta, d) = \prod_{t_j \in \mathcal{H}} \sum_{t_z < t_j} g(t_j - t_z|\theta, d)$, and in (c) we use the fact that the BMH-K model specifies $\boldsymbol{\Theta}^{ac}$ as a mixture over the $K_{\boldsymbol{\Theta}}$ classes, weighted by the membership probabilities $\{z_{\boldsymbol{\Theta},k}^{ac}\}$.

Half-Life Prediction. Under the BMH-K model, the half-life of an out-of-sample item a^* can be expressed as

$$
\hat{\tau}_{1/2}^{a^*} = \mathbb{E}_{\vec{x}^{a^*c}} \mathbb{E}_{z_{\boldsymbol{\Theta},k}^{a^*c}} \left[\tau_{1/2}^{a^*}|\vec{x}^{a^*c}, \vec{y}^{a^*} \right]
$$

$$
\overset{(a)}{=} \mathbb{E}_{\vec{x}^{a^*c}} \mathbb{E}_{z_{\boldsymbol{\Theta},k}^{a^*c}} \left[d^{a^*c} \cdot (2^{\theta^{a^*c}} - 1)|\vec{x}^{a^*c}, \vec{y}^{a^*} \right]
$$

$$
\overset{(b)}{=} \mathbb{E}_{\vec{x}^{a^*c}} \sum_{k=1}^{K_{\boldsymbol{\Theta}}} z_{\boldsymbol{\Theta},k}^{a^*,c} \cdot e^{\delta_{d,k}^{a^*}} \cdot \left[2^{\exp(\delta_{\theta,k}^{a^*} + \vec{\gamma}_{\theta,k} \cdot \vec{y}^{a^*})} - 1 \right]
$$

$$
\overset{(c)}{\approx} \sum_{x=0}^{\infty} \sum_{k=1}^{K_{\boldsymbol{\Theta}}} z_{\boldsymbol{\Theta},k}^{a^*,c} \cdot e^{\delta_{d,k}^{a^*}} \cdot \left[2^{\exp(\delta_{\theta,k}^{a^*} + \vec{\gamma}_{\theta,k} \cdot \vec{y}^{a^*})} - 1 \right] \cdot \hat{f}_\rho(x),
$$

where in (a) we use the expression for the half-life of a Hawkes process under the power law $g(t) = \theta \cdot d^\theta \cdot (t+d)^{-(1+\theta)}$ (i.e. by solving $\tau_{1/2}$ such that $\int_0^{\tau_{1/2}} g(t)dt = \frac{1}{2}$), in (b) we use the fact that the BMH-K model specifies $\boldsymbol{\Theta}^{ac}$ as a mixture over the $K_{\boldsymbol{\Theta}}$ classes, weighted by the membership probabilities $\{z_{\boldsymbol{\Theta},k}^{ac}\}$, and in (c) we marginalize over the unobserved cascade-level features $\vec{x}^{a^*c}$ and use the simplification $f_\rho(\vec{x}^{ac}|\vec{y}^a) \approx \hat{f}_\rho(x)$ as detailed in the main text.
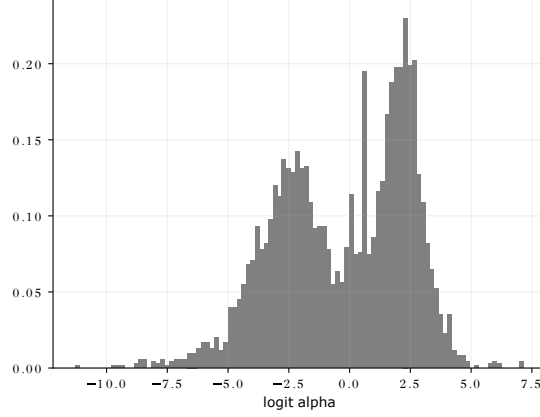
Fig. 1: Distribution of DMM-estimated logit(α) across $RNIX$ publishers. We note the bimodality of the distribution, with the modes corresponding to low and high cascade sizes. Based on this observation we set $K_\alpha = 2$ for the BMH-P model.

3   Additional Material for Main Text Section 4

3.1   Selection of K_α and K_Θ

To guide the selection of the number of mixture components for the BMH-P (i.e. K_α) and BMH-K (i.e. K_Θ) models in Main Text Section 4, we fit the DMM [2] to each publisher in $RNIX$. Given that the EM algorithm is very sensitive to initial conditions, we use 10 random EM initializations and select the output that yields the highest log-likelihood.

We collect the distribution of parameter estimates for logit(α) across publishers in Fig. 1. We see two modes for α, corresponding to cascade groups with low and high sizes, prompting us to set $K_\alpha = 2$ in Main Text Section 4.

We collect the distribution of parameter estimates for $(\log(c), \log(\theta))$ across publishers in the upper plot of Fig. 2, where we see three modes for the kernel parameters. From the lower plot of Fig. 2, we can interpret these modes as belonging to usual, fast and slow cascade groups, prompting us to set $K_\Theta = 3$ in Main Text Section 4.

3.2   Prior Specification for the BMH-P Model

The full set of priors for the BMH-P model implementation is given below. Informative priors are set for $\delta_{\alpha,1}, \delta_{\alpha,2}, \delta_{z_{\alpha,2}}$ based on the observations in Section 3.1. Weakly informative priors are set for the other parameters. We use a Laplace prior on $\vec{\gamma}_{\alpha,1}, \vec{\gamma}_{\alpha,2}, \vec{\gamma}_{z_{\alpha,2}}$ to impose regularization given the high dimensionality of the article feature vector ($|\vec{y}^a| = 32$) we consider.
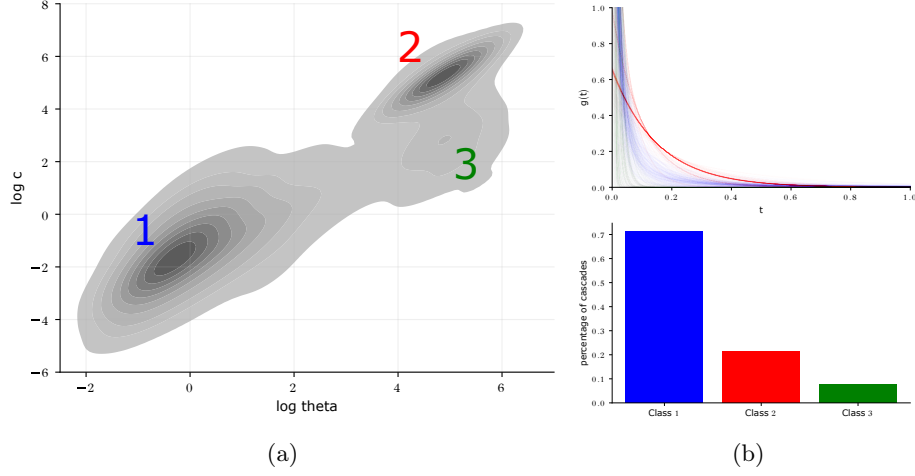
Fig. 2: (a) Distribution of DMM-estimated $(\log(c), \log(\theta))$ across $RNIX$ publishers. We observe the trimodality of the distribution, with the modes corresponding to usual (labeled 1), slow (labeled 2) and fast (labeled 3) cascades. Based on this observation we set $K_\Theta = 3$ for the BMH-K model. (b) In the top plot, we show samples of the power law kernel g for the three classes. In the bottom plot, we show the distribution of cascades for each class.

$$\delta_{\alpha,1} \sim \mathcal{N}(-2, 0.5)$$
$$\delta_{\alpha,2} \sim \mathcal{N}(2, 0.5)$$
$$\delta_{z_{\alpha,2}} \sim \mathcal{N}(-1.39, 0.5)$$
$$\vec{\beta}_{z_{\alpha,2}} \sim \mathcal{N}(0, 0.1)$$
$$\vec{\gamma}_{\alpha,1}, \vec{\gamma}_{\alpha,2}, \vec{\gamma}_{z_{\alpha,2}} \sim \text{Laplace}(0, 0.01)$$
$$\Omega_\alpha \sim \text{LKJCorr}(2)$$
$$\sigma_{\delta_{\alpha,1}}, \sigma_{\delta_{\alpha,2}}, \sigma_{\delta_{z_{\alpha,2}}} \sim \mathcal{N}(0, 1)$$
$$\sigma_{\vec{\beta}_{z_{\alpha,2}}} \sim \mathcal{N}(0, 0.1)$$

3.3   Prior Specification for the BMH-K Model

The full set of priors for the BMH-K model implementation is given below. Informative priors are set for $\delta_{\theta,1}, \delta_{d,1}, \delta_{\theta,2}, \delta_{d,2}, \delta_{\theta,3}, \delta_{d,3}, \delta_{z_{\Theta,2}}, \delta_{z_{\Theta,3}}$ based on the observations in Section 3.1. Weakly informative priors are set for the other parameters. We use a Laplace prior on $\vec{\gamma}_{\Theta,2}, \vec{\gamma}_{\Theta,3}, \vec{\gamma}_{z_{\Theta,2}}, \vec{\gamma}_{z_{\Theta,3}}$ to impose regularization given the high dimensionality of the article feature vector ($|\vec{y}^a| = 32$) we

consider. For $\Omega_{\boldsymbol{\Theta},1}, \Omega_{\boldsymbol{\Theta},2}, \Omega_{\boldsymbol{\Theta},3}$, we set a LKJCorr(0.5) prior (i.e. higher weights on the tails of [0,1]) as (θ, d) for any given Hawkes fit are correlated.

$$\delta_{\theta,1} \sim \mathcal{N}(-0.41, 0.5)$$
$$\delta_{d,1} \sim \mathcal{N}(-1.37, 1)$$
$$\delta_{\theta,2} \sim \mathcal{N}(4, 0.5)$$
$$\delta_{d,2} \sim \mathcal{N}(4.805, 0.5)$$
$$\delta_{\theta,3} \sim \mathcal{N}(4, 0.5)$$
$$\delta_{d,3} \sim \mathcal{N}(1, 0.5)$$
$$\delta_{z_{\boldsymbol{\Theta}},2}, \delta_{z_{\boldsymbol{\Theta}},3} \sim \mathcal{N}(-2, 1)$$
$$\vec{\beta}_{z_{\boldsymbol{\Theta}},2}, \vec{\beta}_{z_{\boldsymbol{\Theta}},3} \sim \mathcal{N}(0, 0.1)$$
$$\vec{\gamma}_{\boldsymbol{\Theta},2}, \vec{\gamma}_{\boldsymbol{\Theta},3}, \vec{\gamma}_{z_{\boldsymbol{\Theta}},2}, \vec{\gamma}_{z_{\boldsymbol{\Theta}},3} \sim \text{Laplace}(0, 0.01)$$
$$\Omega_{\boldsymbol{\Theta},1}, \Omega_{\boldsymbol{\Theta},2}, \Omega_{\boldsymbol{\Theta},3} \sim \text{LKJCorr}(0.5)$$
$$\Omega_{z_{\boldsymbol{\Theta}}} \sim \text{LKJCorr}(2)$$
$$\sigma_{\delta_{\theta,1}}, \sigma_{\delta_{\theta,2}}, \sigma_{\delta_{\theta,3}}, \sigma_{\delta_{d,1}}, \sigma_{\delta_{d,2}}, \sigma_{\delta_{d,3}}, \sigma_{\delta_{z_{\boldsymbol{\Theta}},2}}, \sigma_{\delta_{z_{\boldsymbol{\Theta}},3}} \sim \mathcal{N}(0, 1)$$
$$\sigma_{\vec{\beta}_{z_{\boldsymbol{\Theta}},2}}, \sigma_{\vec{\beta}_{z_{\boldsymbol{\Theta}},3}} \sim \mathcal{N}(0, 0.1)$$

3.4 Implementation Details.

We use the Python implementation of Stan [1] to run both the BMH-P and BMH-K models. We run for 4 chains, adapt delta set to 0.9, 500 warmup iterations and 500 post-warmup iterations. To speed up convergence we implement non-centered parametrization [3] for each of the normally distributed priors.

4 Additional Material for Main Text Section 5

4.1 Performance Heatmaps for CNIX and RNIX

We show performance heatmaps for a selection of CNIX publishers in Fig. 3 and RNIX publishers in Fig. 4. Note the different patterns of which headline style works for each publisher, implying that the BMH model picks up subtle differences of what is effective across publishers.

References

1. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. Journal of statistical software **76**(1) (2017)

2. Kong, Q., Rizoiu, M.A., Xie, L.: Describing and predicting online items with reshare cascades via dual mixture self-exciting processes. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. pp. 645–654 (2020)
3. Papaspiliopoulos, O., Roberts, G.O., Sköld, M.: A general framework for the parametrization of hierarchical models. Statistical Science pp. 59–73 (2007)
4. Tomasi, C.: Estimating gaussian mixture densities with em–a tutorial. Duke University pp. 1–8 (2004)

Table 1: Table of Notation.

Parameter	Interpretation	Real-World Mapping
Source-Level		
ρ	source of items	news publisher
$\mathcal{A}$	set of items produced by ρ	news articles from publisher ρ
$f_\rho(\cdot)$	follower count distribution	
$\hat{C}_\rho$	cascade count estimate	
Item-Level		
$a \in \mathcal{A}$	item produced by ρ	news article
$\mathbb{H}^a$	set of cascades related to item a	retweet cascades for article a
$\vec{y}^a$	item-level features	headline embedding for article a
N^a	item popularity	overall tweet count for article a
$\tau^a_{1/2}$	content half-life	
Cascade-Level		
$\mathcal{H}^{ac} \in \mathbb{H}^a$	cascade related to item a	retweet cascade for article a
$\vec{x}^{ac}$	cascade-level features	follower count of seed user
N^{ac}	cascade size	
$\tau^{ac}_{1/2}$	cascade half-life	
$\mathcal{T}^{ac}$	intereevent-time distribution	
α^{ac}	Hawkes branching factor	
$\boldsymbol{\Theta}^{ac}$	Hawkes kernel parameters	
BMH-P		
K_α	# of BMH-P mixture classes	
$\vec{\gamma}_{\alpha,k}$	effect of $\vec{y}^a$ on center of class k	
$\vec{\gamma}_{z_{\alpha,k}}$	effect of $\vec{y}^a$ on membership probability of class k	
$\delta^a_{\alpha,k}/\delta_{\alpha,k}$	item-/ publisher-level baseline value of $\mathrm{logit}(\alpha)$ for class k	
$\vec{\beta}^a_{\alpha,k}/\vec{\beta}_{\alpha,k}$	effect of $\vec{x}^{ac}$ on center of class k	
$z^{ac}_{\alpha,k}$	mem. probability for class k	
$\delta^a_{z_{\alpha,k}}/\delta_{z_{\alpha,k}}$	item-/ publisher-level mem. prob. softmax baseline for class k	
$\vec{\beta}^a_{z_{\alpha,k}}/\vec{\beta}_{z_{\alpha,k}}$	effect of $\vec{x}^{ac}$ on membership probability for class k	
$\vec{p}^a_\alpha/\vec{p}_\alpha$	item-/ pub.-level parameter vector	
$\Sigma_\alpha/\Omega_\alpha$	cov./ corr. matrix for $\vec{p}_\alpha$	
BMH-K		
$K_{\boldsymbol{\Theta}}$	# of BMH-K mixture classes	
$\vec{\gamma}_{\theta,k}$	effect of $\vec{y}^a$ on center of class k	
$\vec{\gamma}_{z_{\boldsymbol{\Theta},k}}$	effect of $\vec{y}^a$ on membership probability of class k	
$\delta^a_{\theta,k}/\delta_{\theta,k}$	item-/ publisher-level baseline value of $\log(\theta)$ for class k	
$\vec{\beta}^a_{\theta,k}/\vec{\beta}_{\theta,k}$	effect of $\vec{x}^{ac}$ on center of class k	
$z^{ac}_{\boldsymbol{\Theta},k}$	mem. probability for class k	
$\delta^a_{z_{\boldsymbol{\Theta},k}}/\delta_{z_{\boldsymbol{\Theta},k}}$	item-/ publisher-level mem. prob. softmax baseline for class k	
$\vec{\beta}^a_{z_{\boldsymbol{\Theta},k}}/\vec{\beta}_{z_{\boldsymbol{\Theta},k}}$	effect of $\vec{x}^{ac}$ on membership probability for class k	
$\vec{p}^a_{\boldsymbol{\Theta},k}/\vec{p}_{\boldsymbol{\Theta},k}$	item-/ pub.-level kernel parameter baseline values for class k	
$\vec{p}^a_{z_{\boldsymbol{\Theta}}}/\vec{p}_{z_{\boldsymbol{\Theta}}}$	item-/ pub.-level membership probability parameters for class k	
$\Sigma_{\boldsymbol{\Theta},k}/\Omega_{\boldsymbol{\Theta},k}$	cov./ corr. matrix for $\vec{p}_{\boldsymbol{\Theta},k}$	
$\Sigma_{z_{\boldsymbol{\Theta}}}/\Omega_{z_{\boldsymbol{\Theta}}}$	cov./ corr. matrix for $\vec{p}_{z_{\boldsymbol{\Theta}}}$	

Fig. 3: Peformance heatmaps for a selection of CNIX publishers.

theage.com.au

crikey.com.au

(a)   (b)

thewest.com.au

afr.com

(c)   (d)

theaustralian.com.au

skynews.com.au
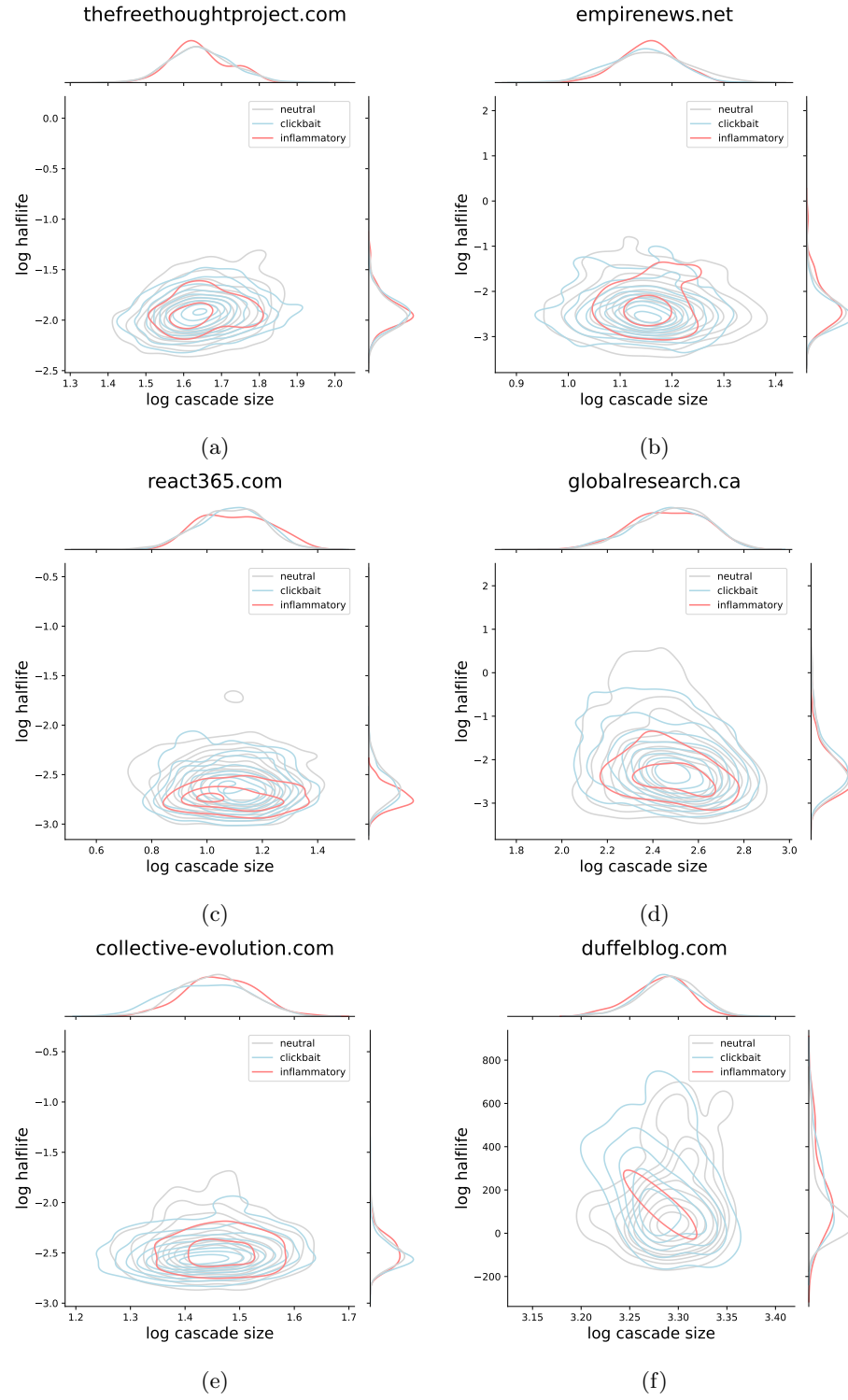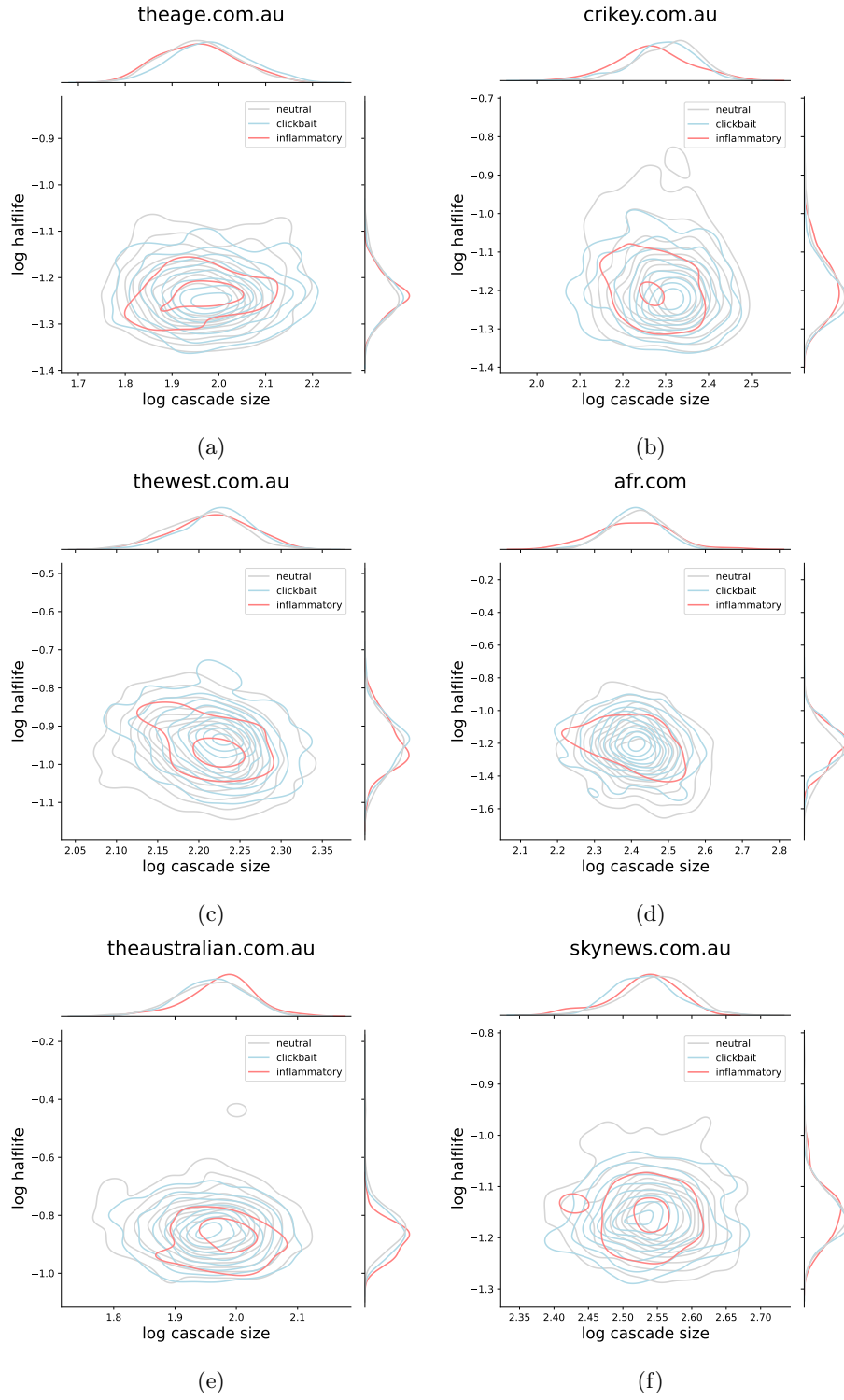
(e)   (f)

Fig. 4: Peformance heatmaps for a selection of RNIX publishers.