Incomplete Gamma Integrals for Deep Cascade Prediction using Content, Network, and Exogenous Signals

Subhabrata Dutta, Shravika Mittal, Dipankar Das, Soumen Chakrabarti, and Tanmoy Chakraborty

Abstract—The behavior of information cascades (such as retweets) has been modeled extensively. While point process-based generative models have long been in use for estimating cascade growths, deep learning has greatly enhanced diverse feature integration. We observe two significant temporal signals in cascade data that have not been emphasized or reported to our knowledge. First, the popularity of the cascade root is known to influence cascade size strongly; but the effect falls off rapidly with time. Second, there is a measurable positive correlation between the novelty of the root content (with respect to a streaming external corpus) and the relative size of the resulting cascade. Responding to these observations, we propose GammaCas, a new cascade growth model as a parametric function of time, which combines deep influence signals from content (e.g., tweet text), network features (e.g., followers of the root user), and exogenous event sources (e.g., online news). Specifically, our model processes these signals through a customized recurrent network, whose states then provide the parameters of the cascade rate function, which is integrated over time to predict the cascade size. The network parameters are trained end-to-end using observed cascades. GammaCas outperforms seven recent and diverse baselines significantly on a large-scale dataset of retweet cascades coupled with time-aligned online news — it beats the best baseline with 18.98% increase in terms of Kendall’s τ correlation and 35.63% reduction in Mean Absolute Percentage Error. Extensive ablation and case studies unearth interesting insights regarding retweet cascade dynamics.

Index Terms—Cascade prediction, social network, exogenous signals, Twitter.

1 INTRODUCTION

(Re)sharing is a common way in which content spreads in social networks. A root user posts some content (such as a photo or an article) and then friends or followers of that user share it with their friends, and so on, resulting in a cascade. In such a cascade tree, information flows from the root to the leaves. In case of Twitter, resharing is called retweeting. The size, duration, and intensity of a reshare cascade are important indicators of user engagement at various levels: within the topic, the community, or the social media platform at large. Modeling user engagement is useful in political discourse mining, market trend analysis, and user-persona detection.

Predicting the progression of a cascade, given early observations at its onset, is known to be a challenging problem [1,4]. Early approaches [5,6] relied on three types of features (network structure, root content, and initial observations along time) for modeling the growth of reply trees. Self-exciting point processes [7,8] were also employed as generative models. Recently, exogenous influence has been incorporated [9,10]. Neural methods, particularly graph embedding-based techniques, are quickly becoming popular [11,12].

Different existing approaches suffer from specific limitations. Extensive feature engineering provides remarkable performance over specific platforms. But they fail to generalize as importance and interdependence of different features vary sharply over different platforms. Pure point-process based models, however simple and explainable, do not take important signals of cascade growth (e.g., content-based features). They rely completely on the numerical growth of the cascade over the observed time to predict future behavior. Previous studies [13] as well as our experiments suggest that the predictions of such models are often adversely affected by noise in the observed cascade. Prior neural models often heavily depend on the graph structure of the cascade growth. In most platforms, however, only the cascade participants are observable and not the exact cascade formation path (i.e, if a retweeter is a common follower of two previous retweeters, it is ambiguous to decide which one of them is the predecessor in the cascade graph). Moreover, most of these approaches do not model cascade growth as an explicit function of the prediction horizon. They need to be trained separately for predicting on different prediction horizons.

Our point of departure is the recognition of certain delicate temporal dynamics that existing cascade prediction methods seem unable to exploit, despite their rapidly increasing sophistication. As an example, Fig. 1(a) shows that, although the root user’s popularity (follower count) is initially strongly predictive of cascade growth rate, the effect is not stationary, but rapidly fades with time. As another example, Fig. 1(b) shows a scatter of cascade sizes (logarithmic) achieved in 15 minutes against the content
similarity between the root tweet and a body of news articles published shortly before and after the root tweet. It hints at a certain “novelty premium” — text that is not mere repetition of current news enjoys greater cascade rates.

Guided by observations like the ones narrated above, we present GammaCas, a novel deep model for cascade prediction. We directly model the gradient of cascade growth as a trainable neural function of content, network, and exogenous features. Specifically, we monitor network (popularity) features evolving through time, and feed (continuous forms of) these features into a novel LSTM variant, whose hidden states are then mapped to parameters that dictate the gradient of cascade growth. Textual and exogenous features modulate how LSTM states influence the temporal process parameters.

The gradient of cascade growth is then integrated over the past to predict the size of the cascade at a given time beyond the observation horizon. Inspired by many natural growth processes [15, 16], we model cascade trajectory as an incomplete gamma function by integrating its temporal derivative numerically. This allows us, during training, to back-propagate prediction errors and train all model weights end to end.

We report on extensive experiments using 342,111 resharing cascades from Twitter, temporally aligned with 206,180 news articles published online on 5,138 news sources. We compare GammaCas against several recent competitive approaches: a basic Hawkes process, SEISMIC [17], TiDeH, a time-dependent Hawkes Process [7], NeuralPointProcess [18], CasPred [2], DeepHawkes [12], DeepCas [11] and ChatterNet [10]. GammaCas achieves lower mean absolute percentage prediction error compared to these prior systems. It is more stable and robust to variations in prediction horizons, compared to some prior systems. Another benefit of GammaCas’s transparent network design is that, by correlating observable features against the parameters involved in the time integration, we get additional insights into the factors that govern cascade dynamics.

Fig. 1. (a) Correlation between root user’s follower count and cascade size at different time after the arrival of the tweet. All three correlation coefficients indicate a decreasing influence of the root follower as the cascade grows further in time. (b) The “novelty premium”: tweets that are not mere repetition of current news enjoy greater cascade rates. Average unigram and bigram similarity between a tweet and the news articles published within 12 hrs. before its arrival is plotted against the log of cascade growth (starting 15 mins.). The later value signifies the virality of the tweet among its first responders. We observe a weakly negative correlation (−0.09 Spearman’s ρ) but with p-value < 10−5.

2 Related Work

Prior works in the field of information cascade modeling can be broadly distinguished into two categories: Macro cascade modeling focused on the overall growth and structural properties of a cascade (e.g., retweet count prediction) and Micro cascade modeling which investigates the behavior of individual agents participating in the cascade (e.g., retweeter prediction) [19, 20]. Our work specifically aligns with the macro category.

Feature-driven cascade modeling. Among the earliest of works, Cheng et al. [2] studied the structural and temporal properties of resharing cascades and came up with a feature-driven strategy to devise a classification problem: after observing a cascade reaching a size k, what is its probability of reaching size nk? Bakshy et al. [21] attempted to identify potential influencers in a feature-driven approach to predict information cascades. To explore richer feature set of cascade dynamics, Krishnan et al. [22] conceptualized cascades as information flow along forests as opposed to the usual tree structure. Most of the feature-driven approaches have revolved around temporal features [2, 23], structural and network features [21, 24], user features [21, 25] and exogenous signals over observable cascade progress to learn parametric representation of cascade growth at a future time. GammaCas achieves a Kendall’s τ correlation of 0.63 (25.06 Mean Absolute Percentage Error) between predicted and actual size of the cascade at 24 hours after only 6 hours of early observation.

• We collect and contribute a large-scale dataset of recent retweet cascades with a temporally aligned stream of online news articles.
• We compare GammaCas with several recent baselines for cascade size prediction developed upon generative, feature-driven, and neural network-based approaches. While GammaCas outperforms each of these baselines by a significant margin, we also investigate the behaviors of these baseline models on our dataset.
• We perform in-depth ablation and case study using GammaCas to investigate into the different signals influencing its parameter estimation. We present insights from these experiments which may be of independent interest.

Organization: The rest of the paper is organized as follows:
• We review the relevant literature on cascade and popularity prediction in Section 2 focusing on point-process and neural methods that incorporate different influence signals.
• GammaCas is presented in Section 3 with detailed descriptions of its various functional components.
• In Section 4 we describe the dataset preparation, training protocols of GammaCas, baseline methods and ablation variants of GammaCas.
• We present experimental results in Section 5.
• We conclude with important observations and possible future direction in Section 6.

Reproducibility: To encourage reproducible research, we present detailed hyper-parameter configurations in Section 4.2. Moreover, we supplement our submission with dataset and source code of GammaCas, available at: https://github.com/LCS2-IIITD/GammaCas

We conclude our major contributions as follows:
• We propose GammaCas, a novel framework for reshare cascade prediction which incorporates content, network and exogenous signals over observable cascade progress to learn parametric representation of cascade growth at a future time. GammaCas achieves a Kendall’s τ correlation of 0.63 (25.06 Mean Absolute Percentage Error) between predicted and actual size of the cascade at 24 hours after only 6 hours of early observation.

• We collect and contribute a large-scale dataset of recent retweet cascades with a temporally aligned stream of online news articles.

• We compare GammaCas with several recent baselines for cascade size prediction developed upon generative, feature-driven, and neural network-based approaches. While GammaCas outperforms each of these baselines by a significant margin, we also investigate the behaviors of these baseline models on our dataset.

• We perform in-depth ablation and case study using GammaCas to investigate into the different signals influencing its parameter estimation. We present insights from these experiments which may be of independent interest.

Organization: The rest of the paper is organized as follows:
• We review the relevant literature on cascade and popularity prediction in Section 2 focusing on point-process and neural methods that incorporate different influence signals.
• GammaCas is presented in Section 3 with detailed descriptions of its various functional components.
• In Section 4 we describe the dataset preparation, training protocols of GammaCas, baseline methods and ablation variants of GammaCas.
• We present experimental results in Section 5.
• We conclude with important observations and possible future direction in Section 6.

Reproducibility: To encourage reproducible research, we present detailed hyper-parameter configurations in Section 4.2. Moreover, we supplement our submission with dataset and source code of GammaCas, available at: https://github.com/LCS2-IIITD/GammaCas

2 Related Work

Prior works in the field of information cascade modeling can be broadly distinguished into two categories: Macro cascade modeling focused on the overall growth and structural properties of a cascade (e.g., retweet count prediction) [2, 7] and Micro cascade modeling which investigates the behavior of individual agents participating in the cascade (e.g., retweeter prediction) [19, 20]. Our work specifically aligns with the macro category.

Feature-driven cascade modeling. Among the earliest of works, Cheng et al. [2] studied the structural and temporal properties of resharing cascades and came up with a feature-driven strategy to devise a classification problem: after observing a cascade reaching a size k, what is its probability of reaching size nk? Bakshy et al. [21] attempted to identify potential influencers in a feature-driven approach to predict information cascades. To explore richer feature set of cascade dynamics, Krishnan et al. [22] conceptualized cascades as information flow along forests as opposed to the usual tree structure. Most of the feature-driven approaches have revolved around temporal features [2, 23], structural and network features [21, 24], user features [21, 25] and exogenous signals over observable cascade progress to learn parametric representation of cascade growth at a future time. GammaCas achieves a Kendall’s τ correlation of 0.63 (25.06 Mean Absolute Percentage Error) between predicted and actual size of the cascade at 24 hours after only 6 hours of early observation.

• We collect and contribute a large-scale dataset of recent retweet cascades with a temporally aligned stream of online news articles.

• We compare GammaCas with several recent baselines for cascade size prediction developed upon generative, feature-driven, and neural network-based approaches. While GammaCas outperforms each of these baselines by a significant margin, we also investigate the behaviors of these baseline models on our dataset.

• We perform in-depth ablation and case study using GammaCas to investigate into the different signals influencing its parameter estimation. We present insights from these experiments which may be of independent interest.

Organization: The rest of the paper is organized as follows:
• We review the relevant literature on cascade and popularity prediction in Section 2 focusing on point-process and neural methods that incorporate different influence signals.
• GammaCas is presented in Section 3 with detailed descriptions of its various functional components.
• In Section 4 we describe the dataset preparation, training protocols of GammaCas, baseline methods and ablation variants of GammaCas.
• We present experimental results in Section 5.
• We conclude with important observations and possible future direction in Section 6.

Reproducibility: To encourage reproducible research, we present detailed hyper-parameter configurations in Section 4.2. Moreover, we supplement our submission with dataset and source code of GammaCas, available at: https://github.com/LCS2-IIITD/GammaCas

2 Related Work

Prior works in the field of information cascade modeling can be broadly distinguished into two categories: Macro cascade modeling focused on the overall growth and structural properties of a cascade (e.g., retweet count prediction) [2, 7] and Micro cascade modeling which investigates the behavior of individual agents participating in the cascade (e.g., retweeter prediction) [19, 20]. Our work specifically aligns with the macro category.

Feature-driven cascade modeling. Among the earliest of works, Cheng et al. [2] studied the structural and temporal properties of resharing cascades and came up with a feature-driven strategy to devise a classification problem: after observing a cascade reaching a size k, what is its probability of reaching size nk? Bakshy et al. [21] attempted to identify potential influencers in a feature-driven approach to predict information cascades. To explore richer feature set of cascade dynamics, Krishnan et al. [22] conceptualized cascades as information flow along forests as opposed to the usual tree structure. Most of the feature-driven approaches have revolved around temporal features [2, 23], structural and network features [21, 24], user features [21, 25] and exogenous signals over observable cascade progress to learn parametric representation of cascade growth at a future time. GammaCas achieves a Kendall’s τ correlation of 0.63 (25.06 Mean Absolute Percentage Error) between predicted and actual size of the cascade at 24 hours after only 6 hours of early observation.

• We collect and contribute a large-scale dataset of recent retweet cascades with a temporally aligned stream of online news articles.

• We compare GammaCas with several recent baselines for cascade size prediction developed upon generative, feature-driven, and neural network-based approaches. While GammaCas outperforms each of these baselines by a significant margin, we also investigate the behaviors of these baseline models on our dataset.

• We perform in-depth ablation and case study using GammaCas to investigate into the different signals influencing its parameter estimation. We present insights from these experiments which may be of independent interest.
content features [26]. While feature-based approaches have produced seminal insights regarding the dynamics of cascade growth, they require heavily curated manual feature engineering that are exclusively platform-dependent.

**Generative models for cascade prediction.** An alternative emerging approach that has seen significant success involve generative models that perceive cascades as temporal event arrival sequences, generates random arrival sequences conditioned on certain parameters, and finally maximizes a chosen likelihood function between the observed and generated sequences [21, 17]. Shen et al. [27] sought to model item popularity over complex networks using a Reinforced Poisson Process model. Crane and Sornette [28] described the view dynamics of YouTube as an epidemic modeled by a self-exciting Hawkes Process. Multiple studies reported using Hawkes Process or its modified variations to predict retweet cascade size [4, 5, 17, 29]. In a cross-platform setting, Rizoiu et al. [8] used a Hawkes process to model popularity growth of content in one platform controlled by endorsement provided in other platforms. Mishra et al. [13] combined a feature-driven approach with Hawkes process for popularity prediction. Rizoiu et al. [30] proposed a hybrid of epidemic and self-excitations models to analyse diffusion cascades. Although not often applied to cascade modeling, recent advances have been used to model more complex dynamics of temporal point processes using neural networks [18, 31]. Other than point-process models, a few others explored epidemic models [26, 32]. Bass model [33, 34]. Survival Analysis [35, 36], Jump Processes [9], etc. Despite their explainable behavior and zero need for heavy feature engineering, generative models are susceptible to adverse influences from outliers [13] and found less powerful at making precise predictions [12].

**Neural network based methods.** Recently, neural models have facilitated more powerful representations of two major components of cascade predictions: recurrent neural architectures can learn the complex temporal dynamics of early observation without constrained approximations [37] and graph learning methods render the integration of complex structural properties to be seamless [11]. In their proposed model DeepCas, Li et al. [11] sought to learn the structural properties of observed retweet cascade using random walk embeddings of the cascade graph and aggregated the dynamics using gated recurrent units with attention. DeepHawkes was proposed by Mishra et al. [12] to translate the explainable behavior of Hawkes Process into the representational superiority of neural networks to predict retweet and citation cascades. In the absence of explicit knowledge about a social or information network like Reddit, Dutta et al. [10] proposed ChatterNet to model the growth of reply cascades; their model integrates exogenous and endogenous influence to learn textual representations of content using time-evolving convolution kernels and aggregates the observed cascade growth using LSTMs. One implementation challenge regarding most of these models is their lack of flexibility to migrate to different observation/prediction horizons without retraining. Moreover, in most of the cases, the superior representation power of neural network-based models is shadowed by the lack of explainability and the inability to produce actionable insights from the learned representations.

**Exogenous influence over cascade growth.** While the mentioned works mostly focus on driving factors of cascade growth implicit to the cascade and the platform, signals exogenous to the platform determine the virality and popularity of content heavily [10, 35]. Prior works seeking to identify the influence of exogenous event arrivals have explored point process with self and external excitation to model observed event sequences [39]. De et al. [40] attempted to demarcate opinion diffusion in Twitter under the influence of exogenous influence from endogenous ones. Broxton et al. [41] investigated the influence of external information sources on virality of online video content. Cascade predictions based on cross domain influences are specialized scenarios of modeling and exploiting signals external to a platform, i.e, predicting YouTube view cascades from Twitter cascades [42]. Dutta et al. [10] employed a similar strategy to incorporate exogenous signals; with the target domain being Reddit, their source domain of external influence was free-flowing new-streams on online news portals.

Given this vast prior development in modeling cascade dynamics, our proposed **GammaCas** model seeks to deliver a generalizable, flexible model for cascade growth prediction, similar to the generative family while incorporating the powerful representation capability of neural methods in an end-to-end fashion to capture the temporal, network-based, content-based and exogenous influences on the cascade growth.

**Differences between ChatterNet and GammaCas.** Among the discussed models for cascade and popularity predictions, ChatterNet seeks to use a set of influence signals similar to ours. It predicts the future chatter intensity under a submission on Reddit, defined as the number of comments posted under that submission. However, there are some key differences as follows: (i) Owing to the closed definition of Reddit’s communities (i.e., subreddits), the original design of ChatterNet is able to characterize endogenous influences in terms of contemporary submissions posted within that subreddit. This is not at all possible for a Twitter-like open platform. Instead, **GammaCas** uses the social network information (i.e, follower count of users) to model the endogenous influence. ChatterNet is not developed to handle such information because Reddit does not provide any. (ii) Being a purely deep learning based model like DeepCas [11], ChatterNet does not learn the prediction function as explicitly dependent on the prediction horizon. Therefore, a new training setup is needed for each different prediction horizon. **GammaCas** overcomes this lack of flexibility by learning a parametric estimation of retweet arrival intensity and then performing numerical integration of the said intensity function over the prediction horizon. This novel hybrid of deep feature learning with numerical function approximation empowers **GammaCas** with the flexibility that, once trained, it may predict for arbitrary prediction horizons.

## 3 Proposed Model

In this section, we describe GammaCas in detail. It has many modules which may appear complex, but we will justify their utility through ablation in Section 4.
3.1 Preliminaries and problem definition

Let $G = \{U, E\}$ be a directed graph representing the social network of Twitter, where $U$ is the set of vertices representing the users and $e_{ij} \in E$ if $u_j$ follows $u_i$ for any $u_i, u_j \in U$. Therefore, the follower count of any given user $u_i$ translates to the out-degree of the corresponding node in $G$.

Given a tweet $\tau$ posted by a user $u$ at time $t_0$, its retweet cascade at time $t > t_0$ can be defined as an ordered sequence of retweet arrival timestamps along with the corresponding retweeter, $R^*_t = \{(t_i, u_i) | t_i > t_j \text{ for } i > j, t_i \leq t\}$. The exogenous event signals within any time frame $[t, t + \Delta t]$ are represented as the sequence of news articles $N(t, t + \Delta t) := \{(n_j, t_j) | t_j \leq t < t + \Delta t\}$, where $n_j$ is an article published at $t_j$.

For any given cascade $R^*_t$, we define the early observation window to be $(t_0, t_0 + \Delta_{obs})$. A model would estimate the future growth of the cascade upon observing the dynamics within this observation period. We also define a prediction horizon $[t_0, t_0 + \Delta_p]$, $\Delta_p > \Delta_{obs}$, so that our problem translates to predicting $|R^*_t|_{t_0 + \Delta_p}$ upon observing $R^*_{\Delta_{obs}}, \tau$, and $N(t_0 - \Delta_{obs}, t_0 + \Delta_{obs})$. Henceforth, for the sake of brevity, we will consider $t_0 = 0$ in general.

Notation. Table 1 summarizes important notations and denotations. While describing GammaCas, we use bold lowercased symbols to denote vector inputs and outputs, and bold uppercased symbols to denote sequences of vectors as well as the trainable parameters of GammaCas.

3.2 Parametric estimation of cascade growth

As Zhao et al. [17] suggested, such a cascade can be either in a supercritical stage (rate of cascade growth is increasing) or in a subcritical stage (rate of cascade growth is decreasing) at different points of time, depending on multiple factors like the relevance of the content expressed by the piece of tweet, out-degree of the nodes participated in the cascade by that time, inter-arrival time of retweets, etc. Extending discrete-valued $R^*_t$ to a continuous, real-valued map of time, we can redefine these two stages as $d|R^*_t| \geq 0$ (supercritical) or $d|R^*_t| < 0$ (subcritical). Such a rate of growth can be modeled as a simple product of two functions of time,

$$\frac{d|R^*_t|}{dt} = \Psi_1(t) \Psi_2(t) \quad (1)$$

constrained with the following conditions: i) $\Psi_1(t), \Psi_2(t) > 0$, ii) $\frac{d\Psi_1}{dt} > 0, \frac{d\Psi_2}{dt} < 0$ and iii) $\lim_{t \to +\infty} \Psi_1(t) \Psi_2(t) = 0$. The first condition ensures a monotonic growth of the cascade, while the second and third conditions ensure a possible initial supercritical growth followed by a mandatory subcritical growth.

Simple choices for such functions would be a polynomial $\Psi_1$ and an exponentially decaying $\Psi_2$. Concretely, we can approximate Eq. 1 in a parametric form as follows:

$$\frac{d|R^*_t|}{dt} = A \tau^\gamma e^{-\lambda \tau} \quad (2)$$

where $A, \gamma,$ and $\lambda$ are arbitrary constants.

The choice of such a function restrains $\frac{d|R^*_t|}{dt}$ to a single “hill”-shaped curve corresponding to a single supercritical
and single subcritical phase, whereas real cascades may have multiple consecutive super- and subcritical phases. The growth rate of such cascades can be easily approximated as:

$$\frac{d|\mathcal{R}_t^r|}{dt} = \sum_k A_k(t - \phi_k) \gamma_k e^{-\lambda_k(t - \phi_k)}$$

(3)

where \(A_k, \lambda_k, \gamma_k\) correspond to the growth parameters of the \(k\)-th “hill” in the cascade growth and \(\phi_k\) correspond to its starting time.

However, we are interested in the size of the cascade after a finite amount of time \(T\), which is given by

$$|\mathcal{R}_T^r| = \int_0^T \sum_k A_k(t - \phi_k) \gamma_k e^{-\lambda_k(t - \phi_k)} dt$$

$$= \sum_k A_k \int_0^T (t - \phi_k) \gamma_k e^{-\lambda_k(t - \phi_k)} dt$$

$$= \sum_k \frac{A_k}{\lambda_k^{1+\gamma_k}} \left[ -\Gamma(\gamma_k + 1, \lambda(t - \phi_k)) \right]_{t=0}^{t=T}$$

$$= \sum_k \frac{A_k}{\lambda_k^{1+\gamma_k}} \left[ \Gamma(\gamma_k + 1, -\lambda_k \phi_k) - \Gamma(\gamma_k + 1, \lambda_k(T - \phi_k)) \right]$$

(4)

where \(\Gamma(s, z)\) are incomplete Gamma functions. Since any generalized incomplete Gamma function can be represented as a finite sum of modified Bessel functions of the first kind [23], the above form is equivalent to a single Gamma function with suitably chosen values of the arbitrary constants. Therefore, we choose to model cascade growth as a parametric function

$$|\mathcal{R}_{\Delta_p}^r| = \int_0^{\Delta_p} A^r t e^{-\lambda t}$$

(5)

where the parameters, in turn, are estimated as (neural) functions of \(\mathcal{R}_{\Delta_{obs}}^r, \tau, \) and \(N(t_0, t_0 + \Delta_{obs})\).

### 3.3 Capturing temporal dynamics of retweet arrival

At any time \(t\), the rate of cascade growth \(\frac{d|\mathcal{R}_t^r|}{dt}\) directly depends on the retweets arriving within \((t, t + \Delta)\) interval. The exact number of retweets (we denote it as \(C_{\Delta t+}^r\)) arrived within this interval directly adds to the size of the cascade. Moreover, each of the new retweets expands the number of potential future retweeters (i.e., susceptible nodes) by the out-degree of the current retweeter.

To capture this temporal dynamics within the early observation phase, we quantize the observation window into \(M\) consecutive, equal-sized bins of size \(\Delta_{obs}\) (i.e., \(\Delta_{obs} = M\Delta_o\)), where \(M\) is an application-driven hyperparameter. We denote the total number of retweets arrived within the \(m\)-th bin as \(C_m^r\), where \(m \in [M]\). We aggregate the additional amount of susceptible nodes created within the \(m\)-th bin as \(C_m^f = \sum_{j \in outdegree(u_i)} \forall (v_j, u_i) \in \mathcal{R}_{\Delta_{obs}}^r/\mathcal{R}_{\Delta_{obs}}^{r(m-1)}\) .

Furthermore, as shown in Figure 2 (5), we apply trainable normalization on the integer elements of \(C_m^r\) and \(C_m^f\) to avoid gradient saturation in the subsequent layers of our framework. The resulting values are denoted as \(r_m\) and \(f_m\), respectively.

The sequences \(\{r_m\}_{m=1}^M \) and \(\{f_m\}_{m=1}^M\) represent the temporal dynamics of cascade growth within the observation window, and a simple choice of architecture to model it would be from the Recurrent Neural Network (RNN) family. While LSTMs have been successfully applied to model temporal dependencies over long sequences, we modify information flow along the LSTM gates according to the intuitive knowledge of the retweet arrival dynamics. As shown in Figure 2 (1), the modified LSTM cell in our architecture instantiates the following six operations:

$$x_g = \sigma(W_g[r_m : h_{m-1}] + B_g)$$

(6)

$$x_{in} = \sigma(W_{in}[r_m : h_{m-1}] + B_{in})$$

(7)

$$x_c = \tanh(W_c[r_m : h_{m-1}] + B_c)$$

(8)

$$x_f = \sigma(W_{f} f_m + B_f)$$

(9)

$$c_m = c_{m-1} \odot x_g + x_{in} \odot x_c \odot x_f$$

(10)

$$h_m = h_{m-1} \odot \tanh(W_h c_m + B_h)$$

(11)

where \([\cdot]\) denotes concatenation; \(\odot\) denotes the Hadamard product; \(\sigma\) signifies the sigmoid non-linearity; \(c_m\) and \(h_m\) correspond to the cell and hidden state of the LSTM after the \(m\)-th timestep (observation bin) respectively; \(W_g, W_{in}, W_c, W_{f}, W_{h}\) are the learnable weight matrices, and \(B_g, B_{in}, B_c, B_{f}, B_h\) are the learnable bias matrices.

Equations 6 [7] and 11 correspond to the operations performed by the original LSTM cell. However, Equation 9 generates a modulation signal \(x_f\) from the out-degree of the participating user nodes at that step to control the contribution of their retweets. Equation 10 takes this modulation into account to update the cell state for the current step. Moreover, this modification decreases the size of the parameter space compared to the original LSTM. Assuming the state size to be \(s\), the four weight matrices of the original LSTM cell would incur a total of \(12 \times s\) number of weight and bias parameters, while the modified one uses \(10 \times s\) parameters due to split inputs.

### 3.4 Processing textual content

We take every piece of text (tweet or news) as a sequence of words and compute a single vector representation of the text relevant to the downstream task, as shown in Figure 2 (2).

We use a trainable embedding layer to map each word \(u_i\) to a \(d\)-dimensional vector \(v_i \in \mathbb{R}^d\), converting a piece of text into a sequence of vectors \(V\). Typical content-sharing platforms like Twitter incur heavy traffic, with millions of textual pieces arriving each second. To speed up the processing, we intend to maintain parallel operations on \(V\). Consequently, we do not use any sequential architecture involving variants of RNN to encode the representation.
Instead, we compute positional encoding vector $p_i \in \mathbb{R}^d$ as
\[
p^{(j)}_i = \begin{cases} 
\sin(\omega_j) & \text{if } j \text{ is even} \\
\cos(\omega_j) & \text{otherwise}
\end{cases}
\]
where $i, j \in \mathbb{N}, \omega_k = L^{-\frac{2k}{d}}, L$ is the maximum length of the input text sequence in the corpus, and $p^{(j)}_i$ denotes the $j$-th element of the vector $p_i$. The embedded sequence of words, $V$ is then transformed to a position encoded sequence $V' = \{v'_i = vi + p_i\}$. 

Next, for every token position, we compute an attention weight $\alpha_i$ using a feed-forward layer followed by a softmax activation:
\[
\alpha_i = \frac{e^{s_i}}{\sum_i e^{s'_i}}, \quad (12)
\]
where $s_i = W_s v_i + B_s$, $W_s$ and $B_s$ are learnable weight and bias matrices, respectively. We compute the final representation of the text as weighted aggregation of $V'$ as $\sum_i \alpha_i v'_i$. Intuitively, Equation (12) generates a word-wise attention weight sequence, which modulates the contribution of each word in the final representation of the text.

We also experimented with more complex text encoding methods like Transformer encoder, Bi-LSTM encoder, and BERT. These models incurred higher training/inference cost in terms of memory and time with no significant improvement over our proposed method. As Dutta et al. [10] suggested, popularity of a content in social media is majorly governed by simpler textual features like topic, polarity, etc. which can be easily captured by simpler models, and sophisticated NLP methods tend to be overkill. Furthermore, the per-word weights, $\alpha_i$, computed by this proposed approach further serve to explain the effects of the textual content of the tweet on the growth of the resulting retweet cascade.

3.5 News-tweet attention as exogenous influence

For a given tweet $\tau$, and a sequence of news $N$, the text processing module outputs a single vector $x_\tau$ and a sequence of vectors $\{n_j\}$, respectively. As exogenous influence on cascade growth varies for tweets expressing different topics, we amalgamate the two signals to compute the final influence, as shown in Figure 2(3). We compute an attention weight between the tweet representation $x_\tau$ and a news representation $n_j$ as
\[
\beta_{\tau,j} = \text{softmax}_j \left( x_\tau^T n_j \right), \quad (13)
\]
The scaling component $d^{-0.5}$ reduces the chance of $\text{softmax}(\cdot)$ reaching saturation. Similar to the text processing module, the final representation of the exogenous influenced tweet text is computed as $x_{\tau,N} = \sum_j \beta_{\tau,j} n_j$.

3.6 Computing cascade growth parameters

The cascade growth parameters $A, \gamma$, and $\lambda$ (see Equation 5) are computed from the textual representation $x_{\tau,N}$ and the observed cascade dynamics encoded by the modified LSTM, $h_m$ (see Equation 11). We hypothesize that while the growth and decay parameters, $\gamma$ and $\lambda$, can be estimated from observing the retweet arrivals exclusively, the scaling parameter $A$ is dependent on the tweet text and the exogenous influence.

We map $h_i$ to three separate non-negative scalars, $A'_m, \gamma_m$ and $\lambda_m$, using three parallel feed-forward layers as follows:
\[
A'_m = \text{relu}(W_A h_m + B_A), \quad (14)
\]
\[
\gamma_m = \text{relu}(W_s h_m + B_s), \quad (15)
\]
\[
\lambda_m = \text{softplus}(W_A h_m + B_A), \quad (16)
\]
We choose these activations experimentally. While $\text{relu}(\cdot)$ is the most straightforward activation function to ensure non-negative output, GammaCas suffers from the zero-gradient problem of ReLU while computing $\lambda_i$.

Next, we compute a modulation parameter emerging from the tweet and the exogenous signals as another non-negative scalar value and scale $A'_i$ as follows:
\[
A_m = A'_m \text{relu}(W_{x_{\tau,N}} + B_u), \quad (17)
\]
where $W_{x_{\tau,N}}$ and $B_u$ are learnable parameters of a feed-forward layer.

3.7 Final prediction

From Equations 14, 15 and 16 we estimate the cascade growth parameters for each observation bin. We apply average-pooling from these three sequences to get the cascade size parameters $A, \gamma$, and $\lambda$. For a given prediction horizon $\Delta_p$, the predicted size of the cascade can then be found by solving the integration in Equation 5 We use 4-th order Runge-Kutta method with fixed number of steps to solve this integration numerically and predict the cascade size at $\Delta_p$ as $\bar{Y}_{\Delta_p}$.

Learning to estimate the aggregate parameters of cascade growth at some prediction horizon is the primary task which GammaCas is designed for. However, within the observation window, a fine-grained prediction modeling of retweet arrival is supposed to help the model learn more robustly. We use a joint learning strategy in an autoregressive setting. At the $m$-th observation bin, we have already estimated the parameters $A_m, \gamma_m$, and $\lambda_m$. From these, we predict the aggregate retweet arrival at the $(m+1)$-th bin as $\bar{C}_{\tau,m+1} = A_m \gamma_m \lambda_m \Delta_p$. The gradient from the loss can be back-propagated through the quadrature [45] as mentioned below.

Loss/cost function. We use two different loss functions to train the model in the joint learning setting. As future cascade size varies largely, we use the Mean Absolute Percentage Error between the predicted and actual cascade size at a prediction horizon $\Delta_{pr}$ as suggested by Dutta et al. [10]. For the autoregressive task of predicting retweet arrival in the next observation window, we use Mean Squared Error loss. The final loss function therefore becomes:
\[
J = \frac{||R_{\Delta_p} - \bar{Y}_{\Delta_p}||}{|R_{\Delta_p}|} + \zeta \sum_{m=1}^{M} (C'_{m+1} - \bar{C}'_{m+1})^2 / M, \quad (18)
\]
where $\zeta < 1$ is a hyperparameter to set the relative importance of the autoregressive gradient.

4 Experimental Setup

In this section, we present the dataset used in the experiments, the baselines and ablation variants of GammaCas considered for the comparison.
After cleaning and tokenization, we set a maximum length for tweet and news headline texts to be 30 and 36, respectively. We set the word embedding dimension $d$ to be 256. To initialize the embedding layer (see Section 3.4), we use pre-trained word vectors which we train on the tweets and news corpus jointly using Word2Vec [47]. The state-size of the modified LSTM layer (Section 3.3) is set as 16.

We set the parameter $\zeta$ in Equation 18 experimentally on the development set; varying it from 0.1 to 0.6 with a step size of 0.05, we found the best configuration to be $\zeta = 0.25$.

For mini-batch training of GammaCas, we use a batch size of 256 (after varying it from 64 to 512 with a step size of 64). GammaCas is optimized using the Adam optimizer with a learning rate 0.0005 (optimal value found between the search range 0.0005−0.005 with step 0.0005). We trained GammaCas for a total of 40 epochs. All of the mentioned hyperparameter search was done using 10% of the training data as a development set. We use MAPE loss for predicting cascade size at 24 hrs. prediction horizon as the search criteria.

GammaCas is implemented using Tensorflow v2.0.0-beta0 on a Intel Xeon Silver 4116 2.10GHz CPU with 48 cores and 64 GB RAM.

4.3 Baseline methods

To compare the performance of GammaCas, we implement a diverse set of baselines from generative, feature-based, and neural network-based families of frameworks.

4.3.1 Generative baselines

We implement the following three self-excitation process-based models:

- **Hawkes**. We implement a univariate Hawkes Process-based model with the exponential kernel, optimized using maximum log-likelihood estimation to provide a basic generative baseline for future cascade size prediction on our data.

- **SEISMIC**, proposed by Zhao et al. [17], uses a self-exciting point process for retweet cascade prediction combined with the exposure provided by a user’s follower base.

- **TiDeH**, a time-dependent Hawkes Process [2], looks at how a cascade evolves with time considering the network structure and aging of information.

4.3.2 Feature-driven baseline

Following the work of Cheng et al. [2], we implement CasPred to predict whether a given cascade will reach a particular size, exploiting rich, hand-crafted temporal and textual features of the cascade. We implement two versions of the model as our baseline – CasPred (org) which uses a subset of the original features used, applicable to our setting, and CasPred (add) which uses additional features proposed by Dutta et al. [10].

4.3.3 Neural network baselines

We consider the following three recent neural architectures as baselines:

- **NNPP** or Neural Network Point Process [18] is an RNN-based method for generalized modeling of temporal point processes.
TABLE 2
Comparison with the baselines and the variants of GammaCas. (↓: lower
class value is better). CasPred versions do not predict the actual size of
future cascades; hence metrics other than step-τ are unapplicable for
these two baselines. SEISMIC and TiDeH emerge as the best
generative baselines in terms of correlation and MAPE, respectively.
GammaCas outperforms the rest of the neural network baselines in both
metrics. t/s signifies average inference time per sample.

| Model            | τ | μ   | MAPE (%) | Step-τ | t/s (ms) |
|------------------|---|-----|----------|--------|----------|
| Hawkes           | 0.202 | 0.277 | 110.25 | 0.231 | 196.72  |
| SEISMIC          | 0.532 | 0.572 | 188.86 | 0.522 | 67.80   |
| TiDeH            | 0.306 | 0.403 | 77.90   | 0.379 | 14.50   |
| NNPP             | 0.344 | 0.427 | 79.12   | 0.379 | 6.23    |
| DeepHawkes       | 0.315 | 0.411 | 71.57   | 0.326 | 11.23   |
| DeepCas          | 0.350 | 0.476 | 60.69   | 0.419 | 9.14    |
| ChatterNet       | 0.342 | 0.455 | 63.69   | 0.404 | 8.77    |
| CasPred (org)    | -   | -    | -       | 0.231 | 0.01    |
| CasPred (add)    | -   | -    | -       | 0.300 | 0.02    |
| GammaCas-LSTM    | 0.597 | 0.769 | 35.78   | 0.688 | 5.54    |
| GammaCas-CO      | 0.625 | 0.784 | 24.16   | 0.741 | 1.08    |
| GammaCas-text    | 0.627 | 0.789 | 24.01   | 0.742 | 2.19    |
| **GammaCas**     | **0.633** | **0.793** | **25.06** | **0.744** | **5.40** |

DeepHawkes [12] is an end-to-end deep learning framework that combines the predictive power of models based on
neural network architectures and interpretability of cascades provided by the Hawkes Process.

DeepCas [11] is a neural network model for predicting cascade growth. It learns a representation of cascade networks by sampling node sequences through random walks processes, thereby leveraging the structural information of the network.

ChatterNet [10] is a neural network model to predict social chatter intensity leveraging on exogenous and endogenous influence combination. To apply it in our setting, we remove the endogenous influence module, resulting in a single LSTM layer integrating exogenous signals from news. Moreover, we incorporate aggregated follower count at each observation bin (similar to GammaCas) in addition to retweet arrival.

4.4 Ablation variants
We seek to investigate the contributions of different components of GammaCas in the overall performance by ablation. We explore the following three ablation variants:

**GammaCas-text**. We take away the contribution of exogenous influence in this variation by removing the scaled dot-product attention between news and tweet. In this variation, the modulation parameter μ in Section 3.6 is computed by applying the feed-forward layer transformation on the tweet text representation Xtext only.

**GammaCas-CO**. In this variation, contributions from the tweet content as well as the exogenous influence are ablated; retweet growth parameters are estimated from the cascade growth dynamics in the observation window alone, using the modified LSTM layer.

**GammaCas-LSTM**. To investigate the gain in modeling capacity enforced by the modifications we applied on LSTM gates in Section 3.3, we replace it with the original LSTM layer with rest of the components unchanged.

5 RESULTS AND DISCUSSION
The growth of a retweet cascade is a stochastic process that is hard to predict, as random events may shift the growth

dynamics of a cascade even after a sufficient observation window. It is important for a model to decide which tweets possess the potential to generate a larger cascade compared to another even when the predicted sizes may not be in range with the actual cascade sizes in a future time. For this reason, we compare GammaCas, its variants, and all the baselines with three evaluation metrics — **Mean Absolute Percentage Error (MAPE)** to estimate the difference in predicted and actual sizes; **Kendall’s τ** and **Spearman’s ρ correlation** between the predicted and actual set of cascade sizes to estimate the models’ ability to rank tweets according to their potential to generate cascades. As CasPred predicts whether a cascade will reach a certain size range instead of predicting the actual size, we compute **step-wise Kendall’s τ correlation** [10] between the predicted range and the actual range.

5.1 Overall performance
In Table 2 we present the performance of GammaCas, its ablation variants, and baselines to predict cascade size at 24 hrs. prediction horizon upon 6 hrs. observation window.

5.1.1 Comparison among baselines
All the purely generative models (SEISMIC, TiDeH, and Hawkes) yield high MAPE (i.e., poor performance) across all prediction horizons. After investigating the actual predictions made by these three models, we find that these models often overestimate the future cascade size by a large margin (often to an order of $10^5$–$10^7$). Though excluding such cases results in a performance comparable to GammaCas, the fraction of such overestimating instances is high enough (> 20%) to cause performance instability.

Among the generative baselines, in terms of correlation coefficients, SEISMIC emerges as the best performing generative baseline, while TiDeH stands as best in terms of MAPE.

All the three neural network-based baselines perform closely with respect to all the evaluation metrics, with DeepCas emerging as the best performing one. ChatterNet suffers
from the tailoring we had to introduce for the sake of making it applicable to retweet cascade prediction in a different problem setting altogether. Neural network-based model of temporal point processes is able to model cascade growth better compared to simple generative models. However, NNPP does not take any other features except the retweet-arrival statistics. This explains its limitation compared to DeepCas. In terms of consistent performance on variable-sized cascades and MAPE, we consider DeepCas to be the best performing baseline altogether.

5.1.2 Comparing GammaCas with baselines

From the lowermost block of Table 2 it is evident that GammaCas and all its ablation variants perform better than all the baselines by a substantial margin in terms of correlation and absolute error (18.98\% increase in Kendall’s \( \tau \) from SEISMIC and 35.63 absolute reduction in MAPE compared to DeepCas). In Figure 4 we plot how the performance of four highly-ranked competing models, namely SEISMIC, TiDeH, DeepCas, and GammaCas, are influenced by the actual size of the cascade at 24 hours prediction horizon. The overshooting problem of SEISMIC and TiDeH is evident from these plots as well.

All the ablation variants perform closely to GammaCas; the common signal present in all these models is the temporal dynamics of retweet arrival within the observation window. One may trivially decide this to be the most important signal for modeling cascade growth dynamics. However, we can observe significant improvement of correlation measures once we introduce the exogenous influence-modulated signals. Interestingly, the overall MAPE error decreases slightly with some ablated variants. We investigate the influence of tweet content and exogenous signals later in Section 5.4 while diagnosing GammaCas predictions.

The design choice we made to introduce extra gating mechanism to LSTM cell to model retweet arrival dynamics evidently brings performance gain. As seen in Table 2 GammaCas-LSTM (with all signals included) is outperforming GammaCas as well as rest of the ablation variants.

We also investigate the latency of prediction for all the models in Table 2. Generative models usually take longer to predict per sample as they use the observation window to estimate the parameters using a likelihood measure. As CasPred solely depends on a manually engineered feature set and does not need any temporal processing (thereby reducing the number of operations), it emerges as the fastest inferring model. Among the rest, GammaCas is an order of magnitude faster than the models which show comparable accuracy. As expected, ablated variants with no news-tweet attention or textual features are faster than full GammaCas.

5.2 Variation with observation window

As past studies suggested [2], a sufficient amount of early observation is necessary to estimate the future size of a cascade. GammaCas offers the flexibility of using different observation windows due to its temporal processing of the input along with an incremental estimation of the growth parameters. In Figure 5 we show the variation of performance of GammaCas for multiple observation windows. Evidently, a larger observation window helps predict the future cascade size with better accuracy. However, even with a shorter observation window (4 hours), GammaCas outperforms all the baseline models in terms of correlation and absolute percentage error.

Splitting the cascade dynamics within the observation window into successive bins of retweet arrival and aggregate follower counts serves as a uniform discretization of the irregular arrival processes. Intuitively, a smaller temporal bin width would result in a more accurate approximation of time, leading to superior performance. This is also evident in Figure 6 where we plot \( \tau \) (left) and MAPE (right) of GammaCas for predicting cascade sizes at different prediction horizons when using different bin widths (5, 15, 30, 45 and 60 mins.). While with narrower bins, the performance drop from near to distant prediction horizons is steep, it effectively flattens with the higher error rate in longer bins. However, narrow bins result in a longer sequence of input, resulting in longer recurrence relations to be captured and higher training/testing cost.

5.3 Variation with prediction horizons

The quality of fit for the estimated parameters of a monotone function of time is judged by how they fit at different future horizons. We vary the prediction horizon and observe the evaluations for GammaCas, its ablation variants, and the best-performing baseline, DeepCas. As shown in Figure 7 GammaCas and its ablation variants produce a more stable performance over different horizons, compared to DeepCas. While in terms of correlation, GammaCas shows...
to send a surprise to many of you

Tweet-I: I will be doing a major giveaway soon. Keep your notifications ON. I am going to send a surprise to many of you.
Tweet-II: Why I homeschool my kids. Who vets these lunatics allowed to teach? No one.

In Figure 7, we present the variation of performance of GammaCas, its ablation variants, and the best performing baseline DeepCas, on different prediction horizons.

Fig. 8. Predicted and actual cascade sizes for two tweets by GammaCas and GammaCas-CO. Underlined words in the tweets are those attaining higher attention weights. In both the cases, GammaCas-CO meets a very low value of $\lambda$ entangled with a low value of $A$, which leads to overshooting the cascade size.

an initial performance drop as the prediction horizon increases, we can see an almost consistent MAPE over all the horizons. Moreover, models like DeepCas need to be trained and tested for each prediction horizon separately, while GammaCas offers a flexible prediction setting much similar to its generative counterparts, adding significance to the judgment of parameter utility.

5.4 Diagnostic experiments on GammaCas

In Section 5, we provided intuitive justifications of our design decisions. To look for the potential presence of more profound connections between different influencing signals and the cascade growth parameters that GammaCas attempts to model, we look into individual predictions as well as the overall distribution of parameters.

In Figure 8, we present two example tweets, actual sizes of the cascades they generate, and the predicted sizes by GammaCas and GammaCas-CO over different prediction horizons. While Tweet-I was from a popular social media influencer addressing their fan-base (no exogenous influence), Tweet-II was regarding a teacher passing abusive remarks to students in the context of COVID-19 (triggered by exogenous event). In both cases, GammaCas-CO, in the absence of content-based signals, underestimates $A$ and to fit the observed retweet arrivals, underestimates the decay parameter $\lambda$ as well. This leads to overshooting the actual cascade size by a large margin. The low value of $\lambda$ also sets a longer supercritical phase of the cascades. On the other hand, GammaCas estimates a much higher value of $\lambda$ with larger $\lambda$ decay, providing a better approximation of the future cascade size.

We extract the attention values $\alpha_i$ (see Equation 12) for each token (other than stopwords) of the tweets. In Figure 8, we mark the words receiving significant attention. It is evident that certain topic-signaling and positive/negative sentiment words put a higher contribution constituting the signals deciding cascade growth.

To investigate the effects of follower count of the root users and exogenous influence on the cascade growth parameters estimated by GammaCas, we plot one-to-one mappings between them in Figure 9. We compute the correlations between each pair of variables in terms of Spearman’s $\rho$ and the corresponding $p$-value.

In Figure 9, we present two example tweets, actual sizes of the cascades they generate, and the predicted sizes by GammaCas and GammaCas-CO over different prediction horizons. While Tweet-I was from a popular social media influencer addressing their fan-base (no exogenous influence), Tweet-II was regarding a teacher passing abusive remarks to students in the context of COVID-19 (triggered by exogenous event). In both cases, GammaCas-CO, in the absence of content-based signals, underestimates $A$ and to fit the observed retweet arrivals, underestimates the decay parameter $\lambda$ as well. This leads to overshooting the actual cascade size by a large margin. The low value of $\lambda$ also sets a longer supercritical phase of the cascades. On the other hand, GammaCas estimates a much higher value of $\lambda$ with larger $\lambda$ decay, providing a better approximation of the future cascade size.

We extract the attention values $\alpha_i$ (see Equation 12) for each token (other than stopwords) of the tweets. In Figure 8, we mark the words receiving significant attention. It is evident that certain topic-signaling and positive/negative sentiment words put a higher contribution constituting the signals deciding cascade growth.

To investigate the effects of follower count of the root users and exogenous influence on the cascade growth parameters estimated by GammaCas, we plot one-to-one mappings between them in Figure 9. We compute the correlations between each pair of variables in terms of Spearman’s $\rho$ and the corresponding $p$-value.

In Figure 9, we present two example tweets, actual sizes of the cascades they generate, and the predicted sizes by GammaCas and GammaCas-CO over different prediction horizons. While Tweet-I was from a popular social media influencer addressing their fan-base (no exogenous influence), Tweet-II was regarding a teacher passing abusive remarks to students in the context of COVID-19 (triggered by exogenous event). In both cases, GammaCas-CO, in the absence of content-based signals, underestimates $A$ and to fit the observed retweet arrivals, underestimates the decay parameter $\lambda$ as well. This leads to overshooting the actual cascade size by a large margin. The low value of $\lambda$ also sets a longer supercritical phase of the cascades. On the other hand, GammaCas estimates a much higher value of $\lambda$ with larger $\lambda$ decay, providing a better approximation of the future cascade size.

We extract the attention values $\alpha_i$ (see Equation 12) for each token (other than stopwords) of the tweets. In Figure 8, we mark the words receiving significant attention. It is evident that certain topic-signaling and positive/negative sentiment words put a higher contribution constituting the signals deciding cascade growth.

To investigate the effects of follower count of the root users and exogenous influence on the cascade growth parameters estimated by GammaCas, we plot one-to-one mappings between them in Figure 9. We compute the correlations between each pair of variables in terms of Spearman’s $\rho$ and the corresponding $p$-value.
of diffusion of contents posted by them are usually high, pointing towards a possible positive reinforcement of $A$ in the future cascades they cause.

As opposed to the follower count, similarity of a tweet with news articles published in the past 6 hrs. shows a weakly negative (yet statistically significant) correlation with all three of the parameters. In this case, the effect is strongest in the case of both $\gamma$ and $\lambda$, pointing towards a slow growth as well as decay when the similarity is high, and vice versa. This weakly negative correlation is consistent with our findings shown in Figure 1(b) in Section 1, where we observed a similar weakly negative impact of similarity between a tweet and past news on the cascade size. Again, a plausible intuition behind this might be that the potential of a tweet be the genesis of a large cascade is facilitated if it brings new, hitherto unknown information.

6 Conclusion and Future Work

We presented GammaCas, a new deep cascade prediction architecture that combines content, network, and exogenous signals into a transparent, parameterized time integral. Prediction loss can be back-propagated to the feature-processing networks. We prepared a large-scale dataset of retweet cascades and time-aligned news texts, and provided insightful findings on the dynamics of cascade growth. GammaCas provides a better and more robust cascade size prediction compared to recent competitive baselines on different prediction horizons with varying early observation window. Investigations on parametric functions and feature representations learned by GammaCas provide a meaningful interpretation of relations between cascade dynamics and various input features related to exogenous influences obtained from online news articles, the textual content of a tweet, degree distribution of cascade participant nodes, retweet arrival, etc.

As a future extension, one may intend to introduce multimodal signals introduced by richer metadata of the tweet (images, memes, videos, etc.). Information cascades formed from a tweet are not limited to simple retweet trees as well. For example, link to an existing tweet may be posted as standalone tweets. When such a tweet gets retweeted, this practically forms an extended information cascade of the original tweet. These complex dynamics makes the cascade modeling problem intrinsically challenging. Moreover, information cascades in general goes beyond the resharing (via retweet, quote, or links) mechanism. A certain news (and specially in the present day scenario, a fake one) may form cascades of diffusion via independent tweets. Modeling such dynamics using the various signals we used is likely to provide further insights.

References

[1] M. J. Salganik, P. S. Dodds, and D. J. Watts, “Experimental study of inequality and unpredictability in an artificial cultural market,” Science, vol. 311, no. 5762, pp. 854–856, 2006.

[2] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, “Can cascades be predicted?” in WWW ’14. New York, NY, USA: Association for Computing Machinery, 2014, p. 925–936.

[3] D. J. Watts, Everything is Obvious: How Common Sense Fails Us. Random House LLC, 2012.

[4] J. M. Hofman, A. Sharma, and D. J. Watts, “Prediction and explanation in social systems,” Science, vol. 355, no. 6324, pp. 486–488, 2017.

[5] K. Wang, M. Bansal, and J.-M. Frahm, “Retweet wars: Tweet popularity prediction via dynamic multimodal regression,” in WACV, 2018, pp. 1842–1851.

[6] Z. Zhao, L. Meng, J. Xiao, M. Yang, F. Wu, D. Cai, X. He, and Y. Zhuang, “Attentional image retweet modeling via multi-faceted ranking network learning,” in IJCAI, 2018, pp. 3184–3190.

[7] R. Kobayashi and R. Lambiotte, “Tideh: Time-dependent hawkes process for predicting retweet dynamics,” in ICWSM, 2016, pp. 191–200.

[8] M.-A. Rizoiu, L. Xie, S. Sanner, M. Cebrian, H. Yu, and P. Van Hentenryck, “Expecting to be hip: Hawkes intensity processes for social media popularity,” in WWW, 2017, p. 735–744.

[9] J. Jia and A. R. Benson, “Neural jump stochastic differential equations,” in NeurIPS, 2019, pp. 9847–9858.

[10] S. Dutta, S. Masud, S. Chakrabarti, and T. Chakraborty, “Deep exogenous and endogenous influence combination for social chatter intensity prediction,” in SIGKDD, 2020, p. 1999–2008.

[11] C. Li, J. Ma, X. Guo, and Q. Mei, “Deepcas: An end-to-end predictor of information cascades,” in WWW, 2017, pp. 577–586.

[12] Q. Cao, H. Shen, K. Cen, W. Ouyang, and X. Cheng, “Deep Hawkes: Bridging the gap between prediction and understanding of information cascades,” in CIKM, ser. CIKM ’17, 2017, p. 1149–1158.

[13] S. Mishra, M.-A. Rizoiu, and L. Xie, “Feature driven and point process approaches for popularity prediction,” in CIKM. New York, NY, USA: Association for Computing Machinery, 2016, p. 1069–1078.

[14] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] J. Milbrandt and M. Yau, “A multimoment bulk microphysics parameterization. part i: Analysis of the role of the spectral shape parameter,” Journal of the atmospheric sciences, vol. 62, no. 9, pp. 3051–3064, 2005.

[16] A. Seifert and K. D. Beheng, “A two-moment cloud microphysics parameterization for mixed-phase clouds. part 1: Model description,” Meteorology and atmospheric physics, vol. 92, no. 1-2, pp. 45–66, 2006.

[17] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, “Seismic: A self-excitation point process model for predicting tweet popularity,” in SIGKDD, 2015, p. 1513–1522.

[18] T. Omi, n. ueda, and K. Aihara, “Fully neural network based model for general temporal point processes,” in NeurIPS, vol. 32, 2019.

[19] J. Wang, V. W. Zheng, Z. Liu, and K. C. Chang, “Topological recurrent neural network for diffusion prediction,” in ICDM. IEEE Computer Society, 2017, pp. 475–484.

[20] M. R. Islam, S. Muthiah, B. Adhikari, B. A. Prakash, and N. Ramakrishnan, “Deepdiffuse: Predicting the ‘who’ and ‘when’ in cascades,” in ICDM, 2018, pp. 1055–1060.

[21] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, “Everyone’s an influencer: Quantifying influence on twitter,” in ICWSM, 2011, p. 65–74.

[22] S. Krishnan, P. Butler, R. Tandon, J. Leskovec, and N. Ramakrishnan, “Seeing the forest for the trees: New approaches to forecasting cascades,” in WebSci, 2016, p. 249–258.

[23] S. Petrovic, M. Osborne, and V. Lavrenko, “RT to win! predicting message propagation in twitter,” in CIWSM, L. A. Adamic, R. Baeza-Yates, and S. Counts, Eds., 2011.

[24] L. Weng, F. Menczer, and Y. Ahn, “Predicting success-
ful memes using network and community structure,” in ICWSM, 2014.

[25] T. Zaman, E. B. Fox, E. T. Bradlow et al., “A bayesian approach for predicting the popularity of tweets,” The Annals of Applied Statistics, vol. 8, no. 3, pp. 1583–1611, 2014.

[26] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusiev, and A. Kustarev, “Prediction of retweet cascade size over time,” in CIKM, 2012, p. 2325–2338.

[27] H. Shen, D. Wang, C. Song, and A. Barabási, “Modeling and predicting popularity dynamics via reinforced poisson processes,” in AAAI, 2014, pp. 291–297.

[28] R. Crane and D. Sornette, “Robust dynamic classes revealed by measuring the response function of a social system,” PNAS, vol. 105, no. 41, pp. 15649–15653, 2008.

[29] P. Bao, H.-W. Shen, X. Jin, and X.-Q. Cheng, “Modeling and predicting popularity dynamics of microblogs using self-excited hawkes processes,” in WWW, 2015, p. 9–10.

[30] M.-A. Rizoiu, S. Mishra, Q. Kong, M. Carman, and L. Xie, “Sir-hawkes: Linking epidemic models and hawkes processes to model diffusions in finite populations,” in WWW, 2018, p. 419–428.

[31] H. Mei and J. Eisner, “The neural hawkes process: A neurally self-modulating multivariate point process,” in NeurIPS, 2017, pp. 6754–6764.

[32] Q. Wang, Z. Lin, Y. Jin, S. Cheng, and T. Yang, “Esis: Emotion-based spreader-ignorant-stifter model for information diffusion,” Knowledge-Based Systems, vol. 81, pp. 46–55, 2015.

[33] Y. Yan, Z. Tan, X. Gao, S. Tang, and G. Chen, “Sth-bass: A spatial-temporal heterogeneous bass model to predict single-tweet popularity,” in DASFAA, 2016, pp. 18–32.

[34] X. Gao, Z. Zheng, Q. Chu, S. Tang, G. Chen, and Q. Deng, “Popularity prediction for single tweet based on heterogeneous bass model,” IEEE TKDE, pp. 1–1, 2019.

[35] J. G. Lee, S. Moon, and K. Salamatian, “An approach to model and predict the popularity of online contents with explanatory factors,” in WI-IAT, 2010, p. 623–630.

[36] X. Gao, X. Jia, C. Yang, and G. Chen, “Using survival theory in early pattern detection for viral cascades,” IEEE TKDE, pp. 1–1, 2020.

[37] Y. Wang, H. Shen, S. Liu, J. Gao, and X. Cheng, “Cascade dynamics modeling with attention-based recurrent neural network,” in IJCAI, 2017, pp. 2985–2991.

[38] S. A. Myers, C. Zhu, and J. Leskovec, “Information diffusion and external influence in networks,” in SIGKDD, 2012, p. 33–41.

[39] S. Koyama and S. Shinomoto, “The statistical physics of discovering exogenous and endogenous factors in a chain of events,” arXiv preprint arXiv:2003.00659, 2020.

[40] A. De, S. Bhattacharya, and N. Ganguly, “Demarcating endogenous and exogenous opinion diffusion process on social networks,” in WWW, ser. WWW '18, 2018, p. 549–558.

[41] T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer, “Catching a viral video,” J. Intell. Inf. Syst., vol. 40, no. 2, pp. 241–259, 2013.

[42] S. D. Roy, T. Mei, W. Zeng, and S. Li, “Towards cross-domain learning for social video popularity prediction,” IEEE Trans. Multimed., vol. 15, no. 6, pp. 1255–1267, 2013.

[43] E. Veling, “The generalized incomplete gamma function as sum over modified bessel functions of the first kind,” Journal of computational and applied mathematics, vol. 235, no. 14, pp. 4107–4116, 2011.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in NIPS, 2017, pp. 5998–6008.

[45] A. Wenerkel and G. Louppe, “Unconstrained monotonic neural networks,” in NeurIPS, 2019, pp. 1545–1555.

[46] F. Hamborg, N. Meuschke, C. Breitinger, and B. Gipp, “news-please: A generic news crawler and extractor,” in ICIS, March 2017, pp. 218–223.

[47] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in NIPS, 2013, pp. 3111–3119.

Subhabrata Dutta is a PhD student in the Department of Computer Science and Engineering, Jadavpur University. His research interests include Social Computing, Natural Language Processing, and Machine Learning.

Dipankar Das is an Assistant Professor in the Department of Computer Science and Engineering, Jadavpur University and Visveswaraya Young Faculty, Ministry of Electronics and Information Technology (MeitY), Government of India. He is presently leading four research projects of DRDO, SERB, DST, UGC, Govt. of India. His research interests are in the area of Natural Language Processing / Computational Linguistics, Emotion and Sentiment Analysis, Search Engine and Information Extraction, Machine Learning, Deep Learning, Social Network Analysis, Data Science (Big Data) and so on.

Soumen Chakrabarti is a Professor of Computer Science at IIT Bombay. He works on knowledge graphs, question answering, and social networks. He has published extensively in WWW, SIGKDD, ACL, EMNLP, IJCAI, AAAI, VLDB, SIGIR, ICDE and other conferences. He won the best paper award at WWW 1999. He was coauthor on the best student paper at ECML 2008. His work on keyword search in databases got the 10-year influential paper award at ICDE 2012. He got his PhD from University of California, Berkeley and worked at IBM Almaden, CMU and Google in the past. He received the Bhatnagar Prize in 2014 and the Jagadis Bose Fellowship in 2019.

Tanmoy Chakraborty is an Assistant Professor and a Ramanujan Fellow at the Dept. of CSE, IIT-Delhi, India, where he leads a research group, Laboratory for Computational Social Systems (LCS2). His primary research interests include Social Computing and Natural Language Processing. He has received several awards including Faculty Awards from Google, IBM and Accenture; Early Career Research Award, DAAD Faculty Fellowship. He is a member of ACM and IEEE. More details at http://faculty.iitd.ac.in/~tanmoy/