Decoupling Representation and Regressor for Long-Tailed Information Cascade Prediction

Fan Zhou
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
fan.zhou@uestc.edu.cn

Liu Yu
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
liu.yu@std.uestc.edu.cn

Xovee Xu*
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
xovee@ieee.org

Goce Trajcevski
Iowa State University
Ames, Iowa, USA
gocet25@iastate.edu

ABSTRACT

Effectively predicting the size of information cascades is crucial for understanding the evolution of many social applications, such as influence maximization and fake news detection. Conventional methods face the challenge of data imbalance which, in turn, yields unsatisfactory prediction performance. To prevent the loss functions or metrics from being affected by extreme values and assure numerical stability, previous works reformulate the problem definitions or adopt other types of evaluation metrics. However, solving the regression prediction of information cascades from a long-tailed distribution perspective is under explored. In this paper, we propose a general decoupling prediction solution – first extracting the representation, then fine-tuning the regressor, which combines the original prediction value and weighted bias generated by a sub-network (SUB) that we designed. Our experiments conducted on long-tailed benchmarks demonstrate that our method significantly improves the prediction accuracy over state-of-the-art methods and mitigates the long-tailed cascade prediction problem.

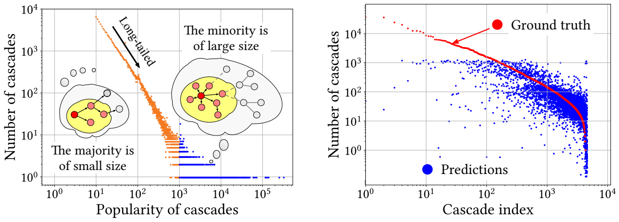
CCS CONCEPTS

• **Information systems** → **Social networks**.

KEYWORDS

Information diffusion; information cascade; long-tailed distribution

ACM Reference Format:
Fan Zhou, Liu Yu, Xovee Xu, and Goce Trajcevski. 2021. Decoupling Representation and Regressor for Long-Tailed Information Cascade Prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3463104



(a) Cascade size distribution (b) Predictions vs. ground truth

Figure 1: Motivation: Long-tailed impact

1 INTRODUCTION

Sharing content through social networks such as Twitter, Facebook, and Weibo has become an important way for people to discover and consume information. The users' behavior contributes to the rapid dissemination of information by forming an information cascade [23]. Effective prediction of information cascades size after a certain time-period has become a critical task for understanding influence maximization, fake news detection, and hate speech diffusion [12].

Related work. In recent years, information cascade modeling and prediction has spurred significant research interest. One group of existing works focused on hand-crafted feature engineering [13, 16], which requires extensive human domain knowledge. Moreover, some features only apply to certain platform or particular type of information being diffused, thus they are hard to be generalized or adapted to new domains. Another group of works explore the diffusion mechanisms of information [9, 11, 15, 20], utilizing various pattern-based models, e.g., Poisson and Hawkes point processes, to fit the intensity functions of the arrival process for incoming events. Although mathematically reliable and with demonstrated enhanced interpretability, these methods rely on long-term observation dependencies and are still incapable to fully leverage the characteristics encoded in the cascades for a satisfactory prediction. Recent deep learning-based models [2, 3, 5, 10, 17, 22, 24] achieved significant improvements by leveraging recurrent neural networks (RNN) and graph neural networks (GNN) to model and learn cascading and topological patterns in information spreading.

Challenge. While the existing methods achieve promising results in information cascade prediction, they face several non-trivial drawbacks. What motivates this work is the observation that there is a lack of methodology to improve predictive performance in the

*Corresponding author

SIGIR '21, July 11–15, 2021, Virtual Event, Canada
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463104

settings of long-tailed distribution of the datasets. Conventional methods for cascade prediction face the challenge of data imbalance impact. Consider Weibo dataset as an example: Figure 1(a) shows the cascade long-tailed size distribution of Weibo dataset; and Figure 1(b) compares the ground truth and the predictions made by DeepCas [10]. We can observe that feeding the heavy-tailed data directly into prediction model makes the instance-rich (or head) data dominate the training procedure, influencing model's predictions to be conservative and distributed in the relatively middle range and lowering the prediction performance. To prevent the loss functions or metrics from being affected by extreme values and assure numerical stability, previous works reformulate the problem definitions by either treating it as a class balanced binary classification, by predicting whether a cascade will exceed the median size of all cascades [6], or adopting other evaluation metrics to train the model, e.g., r^2 and k-top coverage [14, 20].

Present work. Inspired by the long-tailed distribution of visual recognition [8], we propose a general decoupling solution for *regression task*: we first train the backbone to extract information cascade representations, and then fine-tune the regressor which combines original prediction values and weighted biases generated by a specifically designed sub-network. Our main contributions are as threefold:

- We introduce novel consideration of the impact of long-tailed distribution and separate the whole training into two stages: representation extractor and regressor.
- We design a novel probabilistic sub-network to adaptively adjust the weighted bias for different popularity classes, facilitating the regressor to rectify the model prediction.
- Our method can be easily incorporated into existing models. Extensive experiments are conducted on public datasets and baselines, and the results show that our method (both the decoupling scheme and sub-network) significantly improves the prediction performance and explains the prediction well.

2 PRELIMINARIES AND SYSTEM OVERVIEW

Problem Statement. Information cascade popularity prediction aims at predicting the future size of cascade by observing its early stage evolution. Let C denote an event of interest which, starting at a time-instant t_0, is propagated through a network. For N observed cascades $C_i(t_o) (1 \leq i \leq N)$, the popularity prediction can be formalized as a regression problem solved by minimizing the following loss function:

$$\mathcal{L}_{\text{loss}}(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \left(\log_e \hat{P}_i(t_p) - \log_e P_i(t_p) \right)^2, \quad (1)$$

$$\hat{P}_i(t_p) = \text{Model}_{\Theta}\left(C_i(t_o)\right), \quad (2)$$

where t_o is the observation time and t_p is the prediction time, $P_i(t_p) = |C_i(t_p)|$ is the ground truth popularity of cascade $C_i(t_p)$, Θ are model trainable parameters. Traditional baselines use unmodified original data (with long-tailed distribution) for training.

Long-tailed prediction. Since infrequent cascades (e.g., outbreak tweets) are few when training, models trained with unbalanced data are prone to under-fit the uncommon cascades. However, in practice, we expect predictions not to be affected by extreme values/outliers and the model generalizes well to all sizes of cascades.

Prediction process. Given $X = \{x_i = C_i(t_o), y_i = P_i(t_p)\}, i \in \{1, 2, \ldots, n\}$ be the training set, where x_i is the observed information cascade, y_i is the popularity (label) of x_i. Let n_j denote the number of training samples for class j, and R denote the number of classes, then the total number of training samples $n = \sum_{j=1}^{R} n_j$. For a specific baseline, e.g., DeepHawkes [2] or VaCas [24], we denote $z_i = f(x_i; \theta)$ as the representation of cascade x_i, where $f(x_i; \theta)$ is implemented by baseline with parameter θ except for the final dense layers (called backbone). The final popularity $\hat{y}_i = \hat{P}_i(t_p)$ is predicted by a regression function $g(z_i) = \mathbf{W}^{\top} z_i + b$, where $\mathbf{W}$ is the weight matrix and b is bias.

Sampling strategies. To address the representation learning problem of imbalanced data, various strategies are designed to re-balance the data distribution [1, 4, 18]. Let p_j denote the probability of sampling a data point from class j. Considering $p_j = n_j^q / (\sum_{r=1}^{R} n_r^q), q \in [0, 1]$, we have the following four basic sampling strategies:

• *Instance-balanced sampling.* One of the most common strategies is to allow samples in the training set to have equal select probability, i.e., letting $q = 1$, the probability becomes $p_j^{\text{IB}} = n_j / (\sum_{r=1}^{R} n_r)$.

• *Class-balanced sampling.* Samples in different classes have equal select probability, corresponding to $q = 0$ which yields $p_j^{\text{CB}} = 1/R$. In this case, predictions are not skewed to instance-rich classes.

• *Squared-root sampling.* As a trade-off strategy between instance- and class-balanced samplings, the q is set to $1/2$, i.e., probability $p_j^{\text{SR}} = \sqrt{n_j} / (\sum_{r=1}^{R} \sqrt{n_r})$.

• *Progressively-balanced sampling.* This strategy combines the characteristics of previously presented strategies and is utilized by recent models [1, 7, 8]. The sampling probability of class j is defined as $p_j^{\text{PB}}(e) = \left(1 - \frac{e}{E}\right) p_j^{\text{IB}} + \frac{e}{E} p_j^{\text{CB}}$, where e is the current epoch and E is a hyper-parameter which controls the total number of epochs.

Overview. We now present the details of our framework, which consists of three main components (cf. Figure 2).

(i) The input data are divided into R classes in according to the cascade popularity. Several sampling strategies in Section 2 are used to re-sample the cascade datasets.

(ii) We train the backbone network (which learns the structural or temporal characteristics of information cascades) until convergence, fix its parameters and fine-tune the regressor (decoupling it from cascade representation, cf. Section 3.1).

(iii) We design a novel probabilistic sub-network (SUB) facilitating the regressor for final prediction. SUB rectifies the weighted bias of the predicted popularity in different classes for obtaining more accurate predictions (cf. Section 3.2).

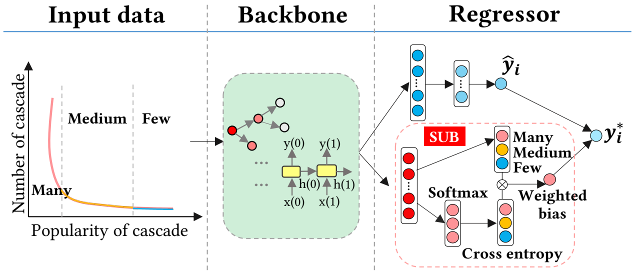


Figure 2: The overall architecture of our decoupling scheme

3 LONG-TAILED INFORMATION CASCADE POPULARITY PREDICTION

Previous *re-sampling* strategies are effective for joint training in classification long-tailed tasks, however, sometimes unexpectedly damage the representation learning to some extent [21], making them unfeasible for regression tasks. We propose a decoupling of the learned representation and regressor for long-tailed information cascade popularity prediction.

3.1 Decoupling Representation and Regressor

We first consider decoupling the learned representation from the regression perspective in long-tailed cascade prediction. Three approaches are utilized to re-train the regressor aiming to rectify decision boundaries through fine-tuning, enabling the regressor to distinguish different cascade classes and predict more accurately.

Joint Training (Joint). As for information cascade popularity prediction, the regressor weights $\mathbf{W}$ and b are usually trained jointly with the backbone parameters θ for extracting the representation $f(x_i;\theta)$ by minimizing the loss function between the ground truth y_i and prediction $\mathbf{W}^\top f(x_i;\theta)+b$. This is also a typical baseline for long-tailed information cascade prediction.

Regressor Re-Training (rRT). We first keep the cascade representations fixed, then randomly re-initialize and optimize the regressor weights $\mathbf{W}$ and b for a small number of epochs using class-balanced sampling strategy.

η-Normalized Regressor (η-norm). Norms of regressor weights tend to be similar in **rRT** after fine-tuning the regressor. In order to let the norms be more distinguishable, we rectify the imbalance of decision boundaries by adjusting the regressor weight norms through a re-scaling procedure: $\tilde{w_i} = w_i/\|w_i\|^\eta$. In this case, we keep both the cascade representations and regressor weights fixed and only learn the scaling factors η on the training set using class-balanced sampling.

3.2 Sub-Network for Two-Stage Prediction

Based on the decoupling scheme introduced above, our method can be readily used for popularity prediction. However, although the performance improvements are non-trivial and the trained model largely mitigates the long-tailed problem (cf. Table 2), we find that the predictions are still prone to be influenced by instance-rich data, which can cause performance degradation.

Hence, we resort to reformulating the prediction process in two stages: firstly, predict the popularity weighted bias by a specifically designed sub-network (SUB); next, rectify the final popularity according to the previously predicted bias. As shown in the right of Figure 2, in addition to the original regressor, we use two branches of MLPs after the backbone network to adaptively adjust the weighted bias for different classes. One of the new branches represents the bias of cascade in different classes b_r, and the other one through the softmax represents the probability of the bias p_r. Then the weighted bias $b_{\mathrm{SUB},i} = \sum_{r=1}^{R} b_r p_r$ is obtained and added to the original predicted value to get the final popularity, i.e., $\tilde{y}_i = \hat{y}_i + b_{\mathrm{SUB},i}$. The new loss for training the whole network is:

$$\mathcal{L}(\Theta) = \mathcal{L}_{\mathrm{loss}}(\Theta) + \mathrm{CE}_{\mathrm{loss}}, \tag{3}$$

$$\mathrm{CE}_{\mathrm{loss}} = -\textstyle\sum_{r\in|R|} \log p_r, \tag{4}$$

where $\mathrm{CE}_{\mathrm{loss}}$ represents the cross entropy between true and predicted popularity classes.

4 EXPERIMENTS

We now describe in detail the experimental evaluations.

4.1 Dataset and Experimental Settings

Dataset. We select two information cascade datasets – *Weibo* [2] and *Twitter* [19] – both follow the long-tailed distribution and suffer from imbalanced label distribution (cf. Figure 1(a)). For each dataset, we split it into training (70%), validation (15%), and test (15%) sets. For our decoupling schemes, we divide training data into three classes of popularity in decreasing order: many-shot (20%), medium-shot (60%), and few-shot (20%). The descriptive statistics of datasets are shown in Table 1.

Table 1: Descriptive statistics of two long-tailed datasets

Dataset	Weibo			Twitter		
	few	*medium*	*many*	*few*	*medium*	*many*
# Cascade	4,259	12,776	4,259	1,820	5,459	1,820
Range	10-32	33-225	>225	11-18	19-332	>332
Avg. popularity	22	88	923	14	101	2,075

Evaluation Protocol. Following [20], we employ our framework on baselines with two commonly used metrics, i.e., mean square logarithmic error (MSLE) and mean absolute percentage error (MAPE). We note the base of logarithm is 2.

Parameter Setting. For fair comparison, we follow the default parameter settings of baselines. The observation (prediction) times of Weibo and Twitter are 0.5 (24) hours and 1 (32) days, respectively. The hyper-parameter E is set to 60.

Baselines. We compare our proposed framework with the following state-of-the-art information cascade prediction models:

- *DeepCas* [10] is the first end-to-end deep learning model for cascade prediction by using multiple random walk processes.
- *DeepHawkes* [2] combines both deep learning and Hawkes self-exciting point process, bridging the gap between prediction performance and interpretability.
- *VaCas* [24] integrates the hierarchical diffusion modeling and temporal-structural characteristics of information cascades, while also capturing the diffusion uncertainty.

4.2 Performance Comparison

Table 2 represents the information cascade popularity prediction results between three baselines and our decoupling schemes. It can be easily observed that our proposed schemes outperform all the baselines across both datasets. Specifically, the best performing scheme (η-norm+SUB) yields 9.7%, 11.8%, and 9.1% improvements over DeepCas, DeepHawkes, and VaCas, respectively, in terms of MSLE. The performance improvements of our proposed schemes demonstrate that decoupling the representation learning and regressor is a promising direction towards addressing the long-tailed regression for information cascade prediction.

Analysis of Joint training. Figure 3 shows the MSLE comparison between *Joint* and *decoupling* schemes on two datasets(the left is deepCas on Weibo, the right is VaCas on Twitter). We can see that when using class-balanced sampling strategy for *Joint* training,

Table 2: Overall prediction performance comparison

Dataset	Weibo			Twitter		
Method	*DeepCas*	*DeepHaw.*	*VaCas*	*DeepCas*	*DeepHaw.*	*VaCas*
Plain	3.097	2.556	2.032	7.702	7.216	6.483
Joint	3.054	2.514	1.986	7.655	7.166	6.449
rRT	2.897	2.352	1.954	7.524	7.017	6.289
η-norm	2.806	2.268	1.861	7.417	6.927	6.194
Joint+SUB	2.941	2.361	1.892	7.506	7.015	6.323
rRT+SUB	2.823	2.279	1.873	7.417	6.933	6.224
η-norm+SUB	**2.798**	**2.254**	**1.847**	**7.399**	**6.921**	**6.187**

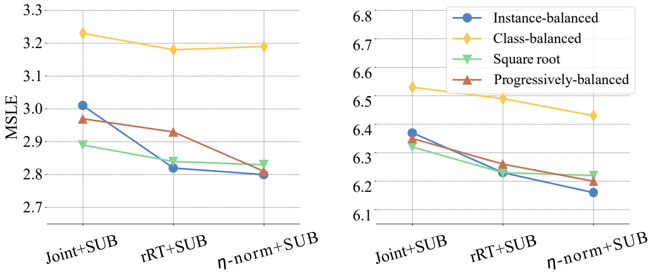


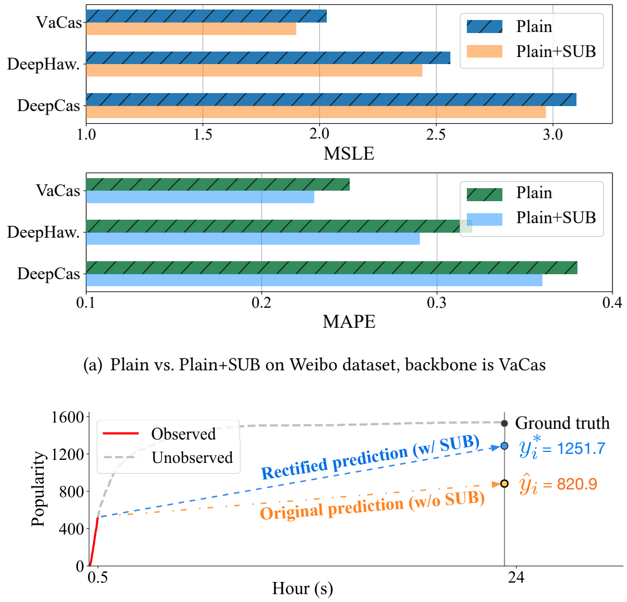
Figure 3: Comparisons between different regressors

there is no performance gains. This is contrary to the conclusion of [8]. We speculate that, different from visual recognition classification tasks, a more balanced sampling strategy might damage the universal representation learning of information cascades by distorting the original data distributions.

Analysis of decoupling schemes. As shown in Figure 3, schemes rRT+SUB and η-norm+SUB outperform the *jointly* trained baselines by large margins. For example, when using the DeepCas with η-norm+SUB, it achieves 4.8% performance improvement compared to Joint+SUB. In addition, class-balanced sampling strategy performs badly on all decoupling schemes. This might be because excessively increasing the tail data harms the model fine-tuning stage. The best performed model is achieved by combining the η-norm+SUB decoupling scheme and instance-balanced sampling strategy. Furthermore, we have following key observations:

- *Regressor*: when we apply the same sampling strategy for representation learning, decoupling schemes rRT/η-norm+SUB always achieve lower prediction error than *Joint*, which is due to their effective re-balancing operations by adjusting the updating process of regressor's weights, which is used to match the long-tailed distribution and the weighted biases generated by sub-network.
- *Information cascade representations*: when we apply the same decoupling scheme, it is rather surprising that the prediction errors of instance-balanced strategy are consistently lower than other sampling strategies. This finding indicates that training with instance-balanced sampling strategy is better for information cascade *representation learning*. Our decoupling scheme eliminates the risk of instance-rich data dominate the regressor fine-tuning stage.

Analysis of sub-network. To investigate the performance of the sub-network (SUB), we apply the sub-network on plain model, i.e., employing the instance-balanced sampling strategy on three baselines for joint training. Figure 4(a) shows the performance comparison on Weibo dataset. We can observe that even without decoupling training, sub-network effectively reduces the prediction errors. The performance boosting may stem from the sub-network's ability to adaptively adjust the weighted bias for many-, medium-, and few-shot classes, allowing the regressor to predict final popularity after determining the weighted bias. Consider Figure 4(b), $\hat{y}_i$ is the prediction of plain model, let y_i^* be the prediction of plain+SUB, after introducing the sub-network. Our framework first specifies the weighted bias based on learned representations (which is a simpler task compared to predict exact popularity), and then the regressor predicts the final popularity by adding the weighted bias. In this way, the model is more robust to extreme values/outliers, avoiding the predictions skewed by instance-rich data when jointly training the whole network.



(a) Plain vs. Plain+SUB on Weibo dataset, backbone is VaCas

(b) Case study: An example information cascade (ID is 72978) from Weibo dataset, observation time is 0.5 hour, prediction time is 24 hours, backbone is VaCas.

Figure 4: Effects of sub-network and intuitive explanations.

5 CONCLUDING REMARKS

We proposed to decouple the representation and regressor for improving information cascade prediction by decreasing the impact of extreme values and outliers. Our framework can be easily implemented on top of existing works towards that purpose. Furthermore, we designed a novel probabilistic sub-network incorporating the regressor, which appears to be more suitable for regression prediction. Experiments conducted on two long-tailed datasets (Weibo and Twitter) demonstrated that our proposed framework effectively improves the prediction performance over the baselines. Our future work will focus on: (i) extending our solution to other long-tailed social applications & data scenarios; and (ii) investigating the method for solving long-tailed problem from learned representations.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (Grant No.62072077), NSF Grant CNS 1646107 and SWIFT 2030249, and National Key R&D Program of China (2019YFB1406202).

REFERENCES

[1] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *NeurIPS*. 18 pages.

[2] Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. 2017. DeepHawkes: Bridging the gap between prediction and understanding of information cascades. In *CIKM*. 1149–1158.

[3] Qi Cao, Huawei Shen, Jinhua Gao, Bingzheng Wei, and Xueqi Cheng. 2020. Popularity prediction on social platforms with coupled graph neural networks. In *WSDM*. 70–78.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.

[5] Xueqin Chen, Fan Zhou, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Fengli Zhang. 2019. Information diffusion prediction via recurrent cascades convolution. In *ICDE*. 770–781.

[6] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted?. In *WWW*. 925–936.

[7] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. 2018. Large scale fine-grained categorization and domain-specific transfer learning. In *CVPR*. 4109–4118.

[8] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2020. Decoupling representation and classifier for long-tailed recognition. In *ICLR*. 16 pages.

[9] Quyu Kong, Marian-Andrei Rizoiu, and Lexing Xie. 2020. Describing and Predicting Online Items with Reshare Cascades via Dual Mixture Self-exciting Processes. In *CIKM*. 645–654.

[10] Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei. 2017. DeepCas: An end-to-end predictor of information cascades. In *WWW*. 577–586.

[11] Ilias N Lymperopoulos. 2021. RC-Tweet: Modeling and predicting the popularity of tweets through the dynamics of a capacitor. *Expert Systems with Applications* 163 (2021), 113785.

[12] Sarah Masud, Subhabrata Dutta, Sakshi Makkar, Chhavi Jain, Vikram Goyal, Amitava Das, and Tanmoy Chakraborty. 2020. Hate is the New Infodemic: A Topic-aware Modeling of Hate Speech Diffusion on Twitter. *arXiv:2010.04377* (2020), 12 pages.

[13] Swapnil Mishra, Marian-Andrei Rizoiu, and Lexing Xie. 2016. Feature driven and point process approaches for popularity prediction. In *CIKM*. 1069–1078.

[14] Amandianeze O Nwana, Salman Avestimehr, and Tsuhan Chen. 2013. A latent social approach to youtube popularity prediction. In *GLOBECOM*. 3138–3144.

[15] Marian-Andrei Rizoiu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. 2018. SIR-Hawkes: Linking epidemic models and Hawkes processes to model diffusions in finite populations. In *WWW*. 419–428.

[16] Oren Tsur and Ari Rappoport. 2012. What's in a hashtag? Content based prediction of the spread of ideas in microblogging communities. In *WSDM*. 643–652.

[17] Yongqing Wang, Huawei Shen, Shenghua Liu, Jinhua Gao, and Xueqi Cheng. 2017. Cascade Dynamics Modeling with Attention-based Recurrent Neural Network.. In *IJCAI*. 2985–2991.

[18] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. 2017. Learning to model the tail. In *NIPS*. 7032–7042.

[19] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2013. Virality prediction and community structure in social networks. *Scientific Reports* 3 (2013).

[20] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *KDD*. 1513–1522.

[21] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. 2020. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*. 9719–9728.

[22] Fan Zhou, Xovee Xu, Ce Li, Goce Trajcevski, Ting Zhong, and Kunpeng Zhang. 2020. A Heterogeneous Dynamical Graph Neural Networks Approach to Quantify Scientific Impact. *arXiv:2003.12042* (2020), 8 pages.

[23] Fan Zhou, Xovee Xu, Goce Trajcevski, and Kunpeng Zhang. 2021. A survey of information cascade analysis: Models, predictions, and recent advances. *Comput. Surveys* 54, 2, Article 27 (2021), 36 pages. https://doi.org/10.1145/3433000

[24] Fan Zhou, Xovee Xu, Kunpeng Zhang, Goce Trajcevski, and Ting Zhong. 2020. Variational information diffusion for probabilistic cascades prediction. In *INFOCOM*. 1618–1627.