Multi-scale Information Diffusion Prediction with Reinforced Recurrent Networks

Cheng Yang[1,2,3*], **Jian Tang**[4,5,6], **Maosong Sun**[1,2,3], **Ganqu Cui**[1,2,3] and **Zhiyuan Liu**[1,2,3]
[1] Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2] Institute for Artificial Intelligence, Tsinghua University, Beijing, China
[3]State Key Lab on Intelligent Technology and Systems, Tsinghua University, Beijing, China
[4]Mila-Quebec Institute for Learning Algorithms, Canada
[5]HEC Montréal, Canada
[6]Canadian Institute for Advanced Research (CIFAR)

Abstract

Information diffusion prediction is an important task which studies how information items spread among users. With the success of deep learning techniques, recurrent neural networks (RNNs) have shown their powerful capability in modeling information diffusion as sequential data. However, previous works focused on either microscopic diffusion prediction which aims at guessing the next influenced user or macroscopic diffusion prediction which estimates the total numbers of influenced users during the diffusion process. To the best of our knowledge, no previous works have suggested a unified model for both microscopic and macroscopic scales. In this paper, we propose a novel multi-scale diffusion prediction model based on reinforcement learning (RL). RL incorporates the macroscopic diffusion size information into the RNN-based microscopic diffusion model by addressing the non-differentiable problem. We also employ an effective structural context extraction strategy to utilize the underlying social graph information. Experimental results show that our proposed model outperforms state-of-the-art baseline models on both microscopic and macroscopic diffusion predictions on three real-world datasets.

1 Introduction

The prediction of information diffusion, also known as *cascade* prediction, has been studied over a wide range of applications, such as product adoption [Leskovec *et al.*, 2007; Watts and Dodds, 2007; Aral and Walker, 2012], epidemiology [Wallinga and Teunis, 2004], social networks [Lappas *et al.*, 2010; Dow *et al.*, 2013] and the spread of news and opinions [Liben-Nowell and Kleinberg, 2008; Leskovec *et al.*, 2009]. Recent works [Li *et al.*, 2017; Cao *et al.*, 2017; Islam *et al.*, 2018; Wang *et al.*, 2018b] on diffusion prediction took advantage of the success of deep learning techniques by modeling information diffusion as sequential data based on recurrent neural networks (RNNs) and achieved promising performances. Existing works [Wang *et al.*, 2017a; Wang *et al.*, 2018b] also explored the social graph information which is available when information diffusion spreads through a social network service for diffusion prediction.

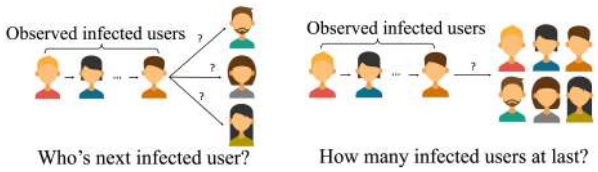

Figure 1: Illustrative examples for microscopic next infected user prediction (left) and macroscopic cascade size prediction (right). Conventionally, we usually use "infected" to indicate that a user is "influenced" by an information item.

However, as shown in Fig. 1, previous works focused on either microscopic diffusion prediction which aims at guessing the next infected user or macroscopic diffusion prediction which estimates the total numbers of infected users during the diffusion process. To the best of our knowledge, no previous works have suggested a unified model for both microscopic and macroscopic scales. A unified model can utilize more information in the training data especially for macroscopic diffusion prediction, *e.g.* previous works [Li *et al.*, 2017; Cao *et al.*, 2017] considered cascade size prediction as a regression problem and discarded the information of detailed infected users and the ordering of their infections.

Also, the handling of social graph information in existing works [Wang *et al.*, 2017a; Wang *et al.*, 2018b] could be suboptimal. [Wang *et al.*, 2017a] only explored user pairs connected by direct social links and [Wang *et al.*, 2018b] computed the similarities of all users and suffered from high time and space complexity which is square to the number of users.

In this paper, we propose a novel multi-scale diffusion prediction model by endowing a microscopic cascade model the ability to predict macroscopic cascade properties, *i.e.* the final size of a cascade, by a reinforcement learning (RL) framework. We further employ an efficient and effective structure context extraction method which was initially proposed for semi-supervised graph classification [Hamilton *et al.*, 2017] to utilize the social graph information. Experimental results

*This work was partially done during the first author's visit to MILA. Correspondence to: Cheng Yang <albertyang33@gmail.com>, Jian Tang <jian.tang@hec.ca>.

show that our proposed model achieves 10% and 12% relatively improvement against state-of-the-art baseline methods on microscopic and macroscopic diffusion predictions, respectively.

To sum up, the contributions of this work are 3-fold:

(1) We innovatively propose the problem of multi-scale diffusion prediction by jointly considering the microscopic and macroscopic diffusion prediction tasks.

(2) We propose a novel reinforcement learning framework to enable a microscopic cascade model for macroscopic diffusion predictions. We also employ a novel structural context extraction algorithm to further take advantage of underlying social graph information.

(3) Experimental results show that our proposed model outperforms state-of-the-art baseline methods on both microscopic and macroscopic diffusion predictions on three real-world datasets.

2 Related Works

Our work is related to microscopic and macroscopic diffusion prediction algorithms based on deep learning techniques. We further group them into embedding-based and RNN-based methods according to their models. We summarize related work in Table **??**. "History" and "Graph" indicate whether a method uses the ordering or social graph of infected users. "Micro" and "Macro" are short for microscopic and macroscopic tasks.

Method	History	Graph	Micro	Macro
Embedded IC			✓	
Inf2vec		✓	✓	
DeepCas	✓	✓		✓
DeepHawkes	✓			✓
CYAN-RNN	✓		✓	
TopoLSTM	✓	✓	✓	
DeepDiffuse	✓		✓	
NDM	✓		✓	
SNIDSA	✓	✓	✓	
this work	✓	✓	✓	✓

Table 1: Summary of related works.

2.1 Embedding-based Methods

Embedding-based methods target on microscopic level predictions by extending IC-model [Kempe *et al.*, 2003] which assumed an independent diffusion probability for every user pair. [Bourigault *et al.*, 2014; Gao *et al.*, 2017] projected users into real-valued user embeddings and simplified the IC model by assuming that infected users are determined only by the source user. Embedded IC [Bourigault *et al.*, 2016] modeled the diffusion probability between two users by a function of their user embeddings instead of assigning a real number to each user pair. Inf2vec [Feng *et al.*, 2018] further considered global user similarity context as an extension. However, embedding-based methods failed to model the infection history, *e.g.* the ordering of infected users, for next infected user prediction and have been shown to be suboptimal choices in the experiments of recent RNN-based approaches [Wang *et al.*, 2017a; Wang *et al.*, 2018b; Yang *et al.*, 2018].

2.2 RNN-based Methods

For macroscopic cascade models, DeepCas [Li *et al.*, 2017] sampled sequences from social graph and observed cascades, and then employed RNN to encode the sequences and predict the eventual size of a cascade. DeepHawkes [Cao *et al.*, 2017] explored Hawkes self-exciting point process based on RNN architecture to utilize the infection timestamp information instead of social graph.

For microscopic cascade models, TopoLSTM [Wang *et al.*, 2017a] extended the standard LSTM model by structuring the hidden states as a directed acyclic graph extracted from the social graph. CYAN-RNN [Wang *et al.*, 2017b], DAN [Wang *et al.*, 2018a] and DeepDiffuse [Islam *et al.*, 2018] all employed RNN and attention model to utilize the infection timestamp information. NDM [Yang *et al.*, 2018] built a microscopic cascade model based on self-attention and convolution neural networks motivated by two heuristic assumptions. SNIDSA [Wang *et al.*, 2018b] computed pairwise similarities of all user pairs and incorporated the structural information into RNN by a gating mechanism. However, to the best of our knowledge, no previous works have suggested a unified model for both microscopic and macroscopic scales.

3 Method

In this section, we will first formalize the diffusion prediction problem on both microscopic and macroscopic scales. Then we will propose a novel structural context extraction algorithm which was originally introduced for semi-supervised graph classification [Hamilton *et al.*, 2017] to build an RNN-based microscopic cascade model. We further incorporate the ability of macroscopic prediction, *i.e.* estimating the eventual size of a cascade, into the model by reinforcement learning. Finally, we will present the overall algorithm and implementation details.

3.1 Problem Formalization

Given user set V, cascade set C, each cascade $c_i \in C$ is a sequence of users $\{v_1^i, v_2^i \ldots, v_{|c_i|}^i\}$ ranked by their infection timestamps where $|c_i|$ is the size of cascade c_i, *i.e.* the number of users infected by the corresponding item. In this paper, we only keep the ordering of users and ignore the exact timestamps as previous works did [Wang *et al.*, 2017a; Wang *et al.*, 2018b; Yang *et al.*, 2018] and leave the modeling of timestamps as future works. Also, an underlying social graph $G = (V, E)$ among users will be available when information diffusion occurs on a social network service. The social graph G will be considered as additional structural inputs for diffusion prediction.

In this work, we target on both microscopic and macroscopic diffusion predictions which focus on fine-grained short-term modeling and coarse-grained long-term estimation, respectively. We formalize the problems as below.

Microscopic Diffusion Prediction aims at predicting the next infected user v_{k+1}^i given previously infected users

$\{v_1^i, v_2^i \ldots, v_k^i\}$ in cascade c_i for $k = 1, 2 \ldots, |c_i| - 1$.

Macroscopic Diffusion Prediction aims at predicting the eventual size $|c_i|$ of cascade c_i, *i.e.* the total number of infected users, given the first K infected users $\{v_1^i, v_2^i \ldots, v_K^i\}$.

3.2 Microscopic Cascade Modeling

In this subsection, we will first present the Gated Recurrent Unit (GRU), a variant of RNNs, as the basis of microscopic cascade modeling. Then we introduce the structural context extraction algorithm to utilize the social graph information. Afterward, we show how to combine these features for microscopic diffusion prediction.

Recurrent Networks

RNNs have shown their effectiveness in sequential data modeling in many areas, such as natural language processing [Mikolov *et al.*, 2010]. Previous works [Wang *et al.*, 2017a; Wang *et al.*, 2018b; Islam *et al.*, 2018] also explored RNNs for cascade modeling and achieved promising results. Specifically, we employ Gated Recurrent Unit (GRU) as the basis of our model. Given the cascade sequence $\{v_1, v_2 \ldots, v_k\}$, GRU takes user v_t as input and computes a hidden state h_t at each step $t = 1, 2 \ldots, k$. We use Eq. 1 to denote the computation process of t-th step hidden state h_t.

$$h_t = GRU(h_{t-1}, x_{v_t}), \tag{1}$$

where $x_{v_t} \in \mathbb{R}^d$ is the d-dimensional embedding of user v_t.

The hidden state $h_k \in \mathbb{R}^d$ encodes the history information of all previously infected users $\{v_1, v_2 \ldots, v_k\}$ in the cascade. Now we go on to encode the structural information to take advantage of the underlying social graph.

Structural Context Extraction

For each user v, we assume $f_v^{(0)}$ as its user features which can be obtained from user profiles or pretrained node embeddings. Our goal is to incorporate the structural information into the feature vector $f_v^{(0)}$. Inspired by the recent success in semi-supervised graph learning [Kipf and Welling, 2017; Hamilton *et al.*, 2017], we employ an efficient structural context extraction algorithm based on neighborhood sampling for the purpose.

Formally, we first sample Z users $\{u_1, u_2 \ldots, u_Z\}$ from v and its neighbors $N(v)$. Then we update the feature vector $f_v^{(0)}$ by aggregating the neighborhood features by Eq. 2.

$$f_v^{(1)} = relu(W \cdot \frac{1}{Z} \sum_{k=1}^{Z} f_{u_k}^{(0)} + b) \tag{2}$$

where u_k is uniformly sampled from user set $\{v\} \cup N(v)$ for $k = 1, 2 \ldots Z$, W, b are weight matrix and bias vector, and activation function $relu(\cdot) = max(\cdot, 0)$.

The updated user feature vector $f_v^{(1)}$ encodes structural information by aggregating features from v's first-order neighbors. As shown in Fig. 2, Eq. 2 can also be processed recursively to explore a larger neighborhood of user v. Empirically, a two-step neighborhood exploration is time-efficient and enough to give promising results. We will use f_v to denote the final user features for simplification.

Though initially proposed for semi-supervised graph classification, we find the algorithm suitable for our problem.

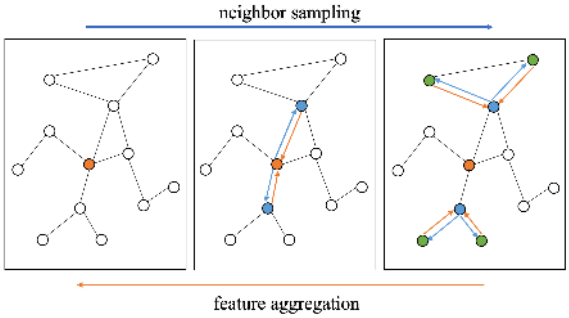

Figure 2: An illustrative example of structural context extraction of the orange node by neighbor sampling and feature aggregation. (Best viewed in color)

Compared with previous works [Wang *et al.*, 2018b] which takes $O(|V|^2)$ space and time complexity for structural context extraction, the neighborhood sampling and aggregation strategy only take $O(|V|)$ time and constant space complexity. Also, the algorithm can explore the two-step neighborhood besides directly connected neighbors.

Microscopic Diffusion Prediction

Now we go back to the next infected user prediction problem by combining the GRU and structural context. Given previously infected users $\{v_1^i, v_2^i \ldots, v_k^i\}$ in cascade c_i, the k-step hidden state h_k^i of GRU encodes the sequential history and user feature vectors $f_{v_1^i}, f_{v_2^i}, \ldots f_{v_k^i}$ encode the underlying social graph information.

Intuitively, the structural context of recent infected users should contribute to the next infected user prediction because the diffusion may spread through the social links. Since we ignore the exact timestamps, we define "recent infected users" as the last m users $\{v_{k-m+1}^i, v_{k-m+2}^i \ldots, v_k^i\}$ where m is a hyper-parameter controlling this window size. We further employ mean pooling[1] to aggregate the feature vectors as $s_k^i = mean(f_{v_{k-m+1}^i}, f_{v_{k-m+2}^i}, \ldots f_{v_k^i})$.

Finally, the probability of the next infected user is computed as

$$p_k^i = softmax(W_p \cdot concat(h_k^i, s_k^i) + b_p), \tag{3}$$

where $p_k^i \in \mathbb{R}^{|V|}$ is the multinomial probability distribution over all users, $concat(\cdot, \cdot)$ is the concatenation operation and W_p, b_p are weight matrix and bias vector, respectively.

The training objective of microscopic diffusion prediction is to maximize the log-likelihood of all cascades

$$J_{micro}(\Theta) = \sum_{i=1}^{|C|} \sum_{k=1}^{|c_i|-1} \log p_k^i[v_{k+1}^i], \tag{4}$$

where $p[j]$ indicates the j-th dimension of vector p and Θ denotes all parameters in the microscopic cascade model.

[1] We also investigate other aggregation strategies such as attention mechanism or concatenation. We find that other options will lead to model overfitting and suboptimal performances.

3.3 Macroscopic Cascade Modeling

The key of this work lies in how to endow a microscopic cascade model the ability to predict macroscopic cascade properties, *i.e.* the size of a cascade. We divide our method into four steps: (a) encode observed K users by a microscopic cascade model; (b) enable the microscopic cascade model to predict the size of a cascade by cascade simulations; (c) use Mean-Square Log-Transformed Error (MSLE) as the supervision signal for macroscopic predictions; and (d) employ a reinforcement learning framework to update parameters through policy gradient algorithm. The overall workflow is illustrated in Fig. 3.

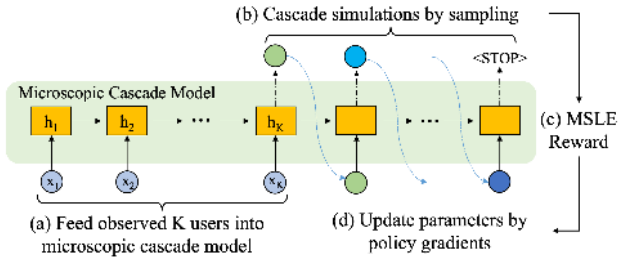

Figure 3: The workflow of adopting microscopic cascade model for macroscopic size prediction by reinforcement learning.

Encoding Observed Users

We feed observed K users of c_i into the microscopic cascade model and get the last hidden state h_K^i as shown in step (a) of Fig. 3. Also, we explicitly encode the positional information which makes the model aware of how many users have been inputted into GRU at each step. In specific, we assign a positional embedding $POS_t \in \mathbb{R}^{d_{pos}}$ for each step $t = 1, 2 \dots maxlen$ where $maxlen$ is the maximum length of cascades. At the t-th step of GRU in Eq. 1, we concatenate the user embedding x_{v_t} and positional embedding POS_t as the input vector instead.

Cascade Simulation for Macroscopic Prediction

To adopt the microscopic cascade model for cascade size prediction in step (b), we firstly append a virtual user <STOP> to the end of every cascade and ask the model to predict it as well. To estimate the size of a cascade given the first K infected users, we recursively sample a user according to the predicted probability distribution in Eq. 3, take it as the next input and make further predictions. Once the <STOP> signal is predicted, we can count the users have been predicted already as the final size of the cascade. Such cascade simulation will be processed for multiple times to reduce the variance of estimations.

Supervision Signals for Macroscopic Prediction

Though the modified microscopic cascade model can predict the size of cascades by simulations, the model still has no supervision signals to guide the way towards better performances. In this paper, we employ Mean Square Log-transformed Error (MSLE), which was used in previous works [Li *et al.*, 2017; Cao *et al.*, 2017] as the evaluation metric of cascade size prediction, for the supervision signal in step (c) of Fig. 3: MSLE$= \frac{1}{|C|}\sum_{i=1}^{|C|}(\log(|c_i|)-\log(pred_i))^2$ where $pred_i$ is the predicted size of cascade c_i.

However, the sampling operation used in cascade size estimation is non-differentiable and makes it impossible to update the parameters by backward propagation. To overcome this problem, we put the simulation process into the reinforcement learning (RL) framework and then employ policy gradient to update the parameters as shown in step (d) of Fig. 3.

Policy Gradient for Parameter Updating

We map the GRU and its hidden state (including structural context) to the *agent* and *state* concepts in RL. The *action* at each step is to choose the next infected user and the *policy* which decides the probability of actions given the current state is defined by Eq. 3. When the <STOP> action is chosen, a *reward* will be given as the opposite number of MSLE[2].

Formally, for the size prediction of cascade c_i, the first K users of c_i are fed into the microscopic cascade model and the last hidden state h_K^i is used for the initial state of RL. For every action sequence $seq = \{a_1, a_2 \dots a_{maxlen}\}$ where a_j is the selected user in the j-th action, we can compute the reward $reward(seq, c_i)$ by the opposite number of MSLE. Then we aim to maximize the expectation of reward for cascade c_i as

$$J_{RL}^i(\Theta) = \sum_{seq} \Pr(seq; \Theta, h_K^i) reward(seq, c_i), \tag{5}$$

where $\Pr(seq; \Theta, h_K^i)$ is the probability of choosing action sequence seq and can be decomposed into the product of the probability of each action a_j. Note that seq is in the space of $|V|^{maxlen}$ and can not be enumerated to compute J_{RL}^i. Instead, the gradients of J_{RL}^i can be computed by REINFORCE algorithm [Williams, 1992]:

$$\begin{aligned}
\nabla_\Theta J_{RL}^i &= \sum_{seq} \nabla_\Theta \Pr(seq; \Theta, h_K^i) \cdot reward(seq, c_i) \\
&= \sum_{seq} \Pr(seq; \Theta, h_K^i) \nabla_\Theta \log \Pr(seq; \Theta, h_K^i) \cdot reward(seq, c_i) \\
&= \mathbb{E}_{seq}[\nabla_\Theta \log \Pr(seq; \Theta, h_K^i) \cdot reward(seq, c_i)] \\
&\simeq \frac{1}{M} \sum_{m=1}^{M} \nabla_\Theta \log \Pr(seq_m; \Theta, h_K^i) \cdot reward(seq_m, c_i),
\end{aligned} \tag{6}$$

where seq_m for $m = 1, 2 \dots M$ are M samples from $\Pr(seq; \Theta, h_K^i)$ and the expectation over the whole action sequence is approximated by Monte Carlo simulations at the last step. Finally, parameters Θ will be updated by gradient ascent to maximize the expectation of reward, *i.e.* the supervision signal for macroscopic prediction.

3.4 Implementation Details

For the joint training of both microscopic and macroscopic objectives, we first train the microscopic cascade model for 10 epochs as warming up. The warming up phase will ensure the next infected user prediction p_k^i generates high-quality

[2]If the <STOP> action is not chosen when the maximum length $maxlen$ is reached, we assume this action sequence predicts the size as $2 \times maxlen$.

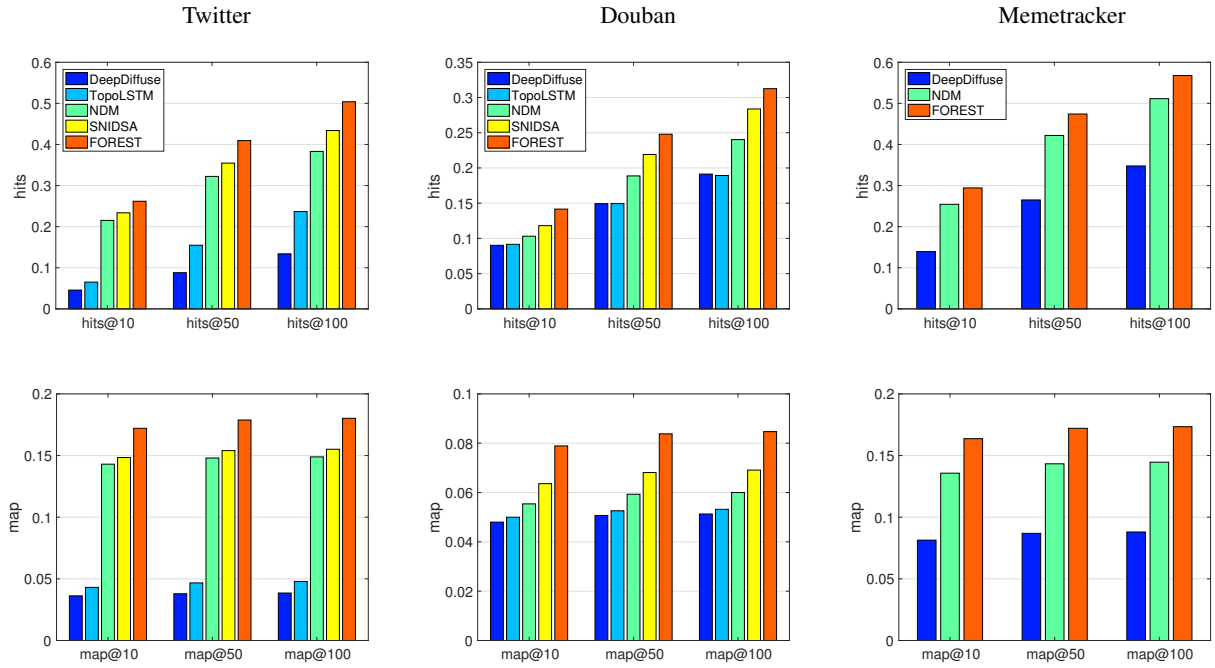

Figure 4: Experimental results on microscopic diffusion prediction. For both MAP and HITS metrics, scores are the higher the better.

cascade simulations and help the RL training converge more quickly. Then we iteratively update the parameters to maximize the microscopic and macroscopic objectives until we reach the best performance on validation data. We employ Adam [Kingma and Ba, 2015] optimizer for gradient ascent.

For hyper-parameter settings, the dimension of hidden state and user feature vector $d = 64$, controlling window size $m = 3$, neighbors sampled in structural context extraction $Z_1 = 25$, $Z_2 = 10$ for first-order and second-order aggregation, the dimension of positional embedding $d_{pos} = 8$ and training data are grouped into mini-batches with batch size 16.

We name our method as reinFOrced REcurrent networks with STructural context (FOREST). The source code of this paper can be found at https://github.com/albertyang33/FOREST.

4 Experiments

We conduct experiments on both microscopic and macroscopic cascade predictions to demonstrate the effectiveness of our proposed model.

4.1 Datasets

Twitter [Hodas and Lerman, 2014] dataset records the tweets containing URLs during October 2010. Each URL is interpreted as an information item spreading among users.

Douban [Zhong *et al.*, 2012] is a Chinese social website where users can update their book reading statuses and follow the statuses of other users. Each book is considered as an information item and a user is infected if she reads the book.

Memetracker [Leskovec *et al.*, 2009] collects a million of news stories and blog posts from online websites and track the most frequent quotes and phrases, *i.e.* memes, to analyze the migration of memes among people. Each meme is regarded as an information item and each URL of websites is treated as a user. Note that this dataset has no underlying social graph.

We randomly sample 80% of cascades for training, 10% for validation and the rest 10% for test. The statistics of datasets are listed in Table 2.

Dataset	# Users	# Links	# Cascades	Avg. Length
Twitter	12,627	309,631	3,442	32.60
Douban	23,123	348,280	10,602	27.14
Memetracker	4,709	-	12,661	16.24

Table 2: Statistics of datasets.

4.2 Baselines

For a thorough comparison, we consider five very recent baseline methods on both microscopic and macroscopic cascade predictions. Microscopic cascade prediction models:

TopoLSTM [Wang *et al.*, 2017a] extends the standard LSTM model by structuring the hidden states as a directed acyclic graph which is extracted from the social graph.

DeepDiffuse [Islam *et al.*, 2018] employs embedding technique and attention model to utilize the infection timestamp information. We replace the timestamps by infection steps as we ignore the exact timestamps in the datasets.

NDM [Yang *et al.*, 2018] builds a microscopic cascade model based on self-attention and convolution neural networks to alleviate the long-term dependency problem.

SNIDSA [Wang *et al.*, 2018b] computes pairwise similarities of all user pairs and incorporates the structural information into RNN by a gating mechanism.

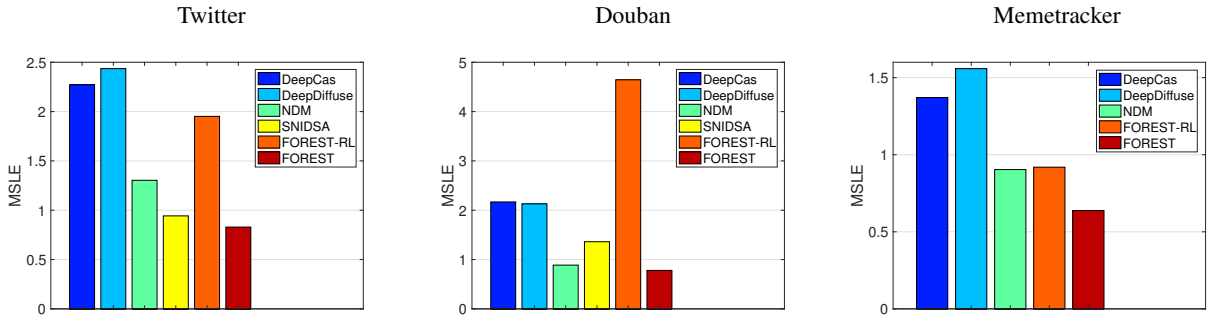

Figure 5: Experimental results on macroscopic diffusion prediction. The scores are the lower the better. FOREST-RL is the variant of FOREST by removing the RL part. We omit the results of TopoLSTM because its scores are larger than 10 and cannot be fit into the figure.

Macroscopic cascade prediction models:

DeepCas [Li *et al.*, 2017] is the state-of-the-art cascade size prediction algorithm which considers both cascade information and underlying social graph.

4.3 Experimental Settings

For microscopic prediction, we consider the next infected user prediction as a retrieval task by ranking the uninfected users by their infection probabilities in Eq. 3. We report the Mean Average Precision (MAP) and HITS scores. The same settings are used in [Wang *et al.*, 2017a; Islam *et al.*, 2018].

For macroscopic prediction, the first $K = 5$ users in a cascade are given to predict the size of a cascade. We use MSLE which is detailed in section 3.3 for evaluation. The same evaluation metric is used in [Li *et al.*, 2017]. Also, all microscopic baselines are adopted for macroscopic cascade size prediction by appending an additional <STOP> signal to the training cascades.

For Twitter and Douban datasets, we use pretrained DeepWalk [Perozzi *et al.*, 2014] embedding with dimension $d = 64$ as initial user feature vectors $f_v^{(0)}$. We excludes TopoLSTM and SNIDSA for Memetracker because of the absence of underlying social graph.

4.4 Experimental Results

Figure 4 and 5 present the experimental results on microscopic and macroscopic diffusion predictions, respectively. We have the following observations:

(1) FOREST consistently outperforms all state-of-the-art baseline methods on microscopic diffusion prediction by a relative improvement of more than 10% in terms of HITS and MAP scores. Compared with TopoLSTM and SNIDSA, the improvements mostly come from the encoding of structural context. The structural context extraction algorithm of FOREST considers second-order neighborhoods while previous works only considered the first-order neighbors.

(2) FOREST consistently outperforms other baselines including DeepCas, the state-of-the-art macroscopic diffusion prediction method, on cascade size prediction by a relative improvement of more than 12% in terms of MSLE. Compared with FOREST-RL where the reinforcement training of macroscopic size prediction is removed, FOREST achieves more promising and robust performances by incorporating macroscopic supervision signals for parameter training.

(3) Compared with DeepCas, microscopic cascade models are able to take advantage of more information in the training data, *i.e.* the detailed infected users and the ordering of their infections. Thus microscopic cascade models are able to give comparable and even better results on macroscopic diffusion prediction task. This finding would encourage microscopic cascade models to replace macroscopic ones in future works.

(4) We omit the ablation study of FOREST-RL and FOREST in microscopic prediction because the benefit from RL for microscopic prediction is not as significant as that for macroscopic prediction (30%-80% relative improvement in Fig. 5). This is because reinforcement learning focuses on long-term modeling and microscopic prediction (next infected user prediction) is short-term. We will consider a long-term microscopic prediction setting for future work.

5 Conclusion

In this paper, we propose a novel multi-scale diffusion prediction model, FOREST, for both microscopic and macroscopic predictions. In specific, we adopt microscopic cascade models for macroscopic cascade size prediction by a reinforcement learning framework which incorporates macroscopic supervision and achieves promising performances. Also, we employ an effective structural context extraction method to utilize the underlying social graph information. Experimental results on next infected user and cascade size predictions demonstrate the effectiveness of our method.

For future works, an intriguing direction is to consider the exact timestamp information which is ignored in this work for modeling. We will also investigate the possibility of applying the RL module on other microscopic cascade models.

Acknowledgments

This research is jointly supported by the NSFC project under the grant no. 61661146007 and the NExT++ project, the National Research Foundation, Prime Minister's Office, Singapore under its IRC@Singapore Funding Initiative. Jian Tang is supported by the Natural Sciences and Engineering Research Council of Canada, as well as the Canada CIFAR AI Chair Program.

References

[Aral and Walker, 2012] Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 2012.

[Bourigault *et al.*, 2014] Simon Bourigault, Cedric Lagnier, Sylvain Lamprier, Ludovic Denoyer, and Patrick Gallinari. Learning social network embeddings for predicting information diffusion. In *Proceedings of WSDM*. ACM, 2014.

[Bourigault *et al.*, 2016] Simon Bourigault, Sylvain Lamprier, and Patrick Gallinari. Representation learning for information diffusion through social networks: an embedded cascade model. In *Proceedings of WSDM*. ACM, 2016.

[Cao *et al.*, 2017] Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. Deephawkes: Bridging the gap between prediction and understanding of information cascades. In *Proceedings of CIKM*. ACM, 2017.

[Dow *et al.*, 2013] P Alex Dow, Lada A Adamic, and Adrien Friggeri. The anatomy of large facebook cascades. *ICWSM*, 2013.

[Feng *et al.*, 2018] Shanshan Feng, Gao Cong, Arijit Khan, Xiucheng Li, Yong Liu, and Yeow Meng Chee. Inf2vec: Latent representation model for social influence embedding. In *Proceedings of ICDE*, 2018.

[Gao *et al.*, 2017] Sheng Gao, Huacan Pang, Patrick Gallinari, Jun Guo, and Nei Kato. A novel embedding method for information diffusion prediction in social network big data. *IEEE Transactions on Industrial Informatics*, 2017.

[Hamilton *et al.*, 2017] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of NeurIPS*, 2017.

[Hodas and Lerman, 2014] Nathan O Hodas and Kristina Lerman. The simple rules of social contagion. *Scientific reports*, 4:4343, 2014.

[Islam *et al.*, 2018] Mohammad Raihanul Islam, Sathappan Muthiah, Bijaya Adhikari, B Aditya Prakash, and Naren Ramakrishnan. Deepdiffuse: Predicting the'who'and'when'in cascades. In *Proceedings of ICDM*, 2018.

[Kempe *et al.*, 2003] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of SIGKDD*, 2003.

[Kingma and Ba, 2015] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, 2015.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of ICLR*, 2017.

[Lappas *et al.*, 2010] Theodoros Lappas, Evimaria Terzi, Dimitrios Gunopulos, and Heikki Mannila. Finding effectors in social networks. In *Proceedings of SIGKDD*, pages 1059–1068. ACM, 2010.

[Leskovec *et al.*, 2007] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

[Leskovec *et al.*, 2009] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of SIGKDD*. ACM, 2009.

[Li *et al.*, 2017] Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei. Deepcas: An end-to-end predictor of information cascades. In *Proceedings of WWW*, pages 577–586. International World Wide Web Conferences Steering Committee, 2017.

[Liben-Nowell and Kleinberg, 2008] David Liben-Nowell and Jon Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the national academy of sciences*, 105(12):4633–4638, 2008.

[Mikolov *et al.*, 2010] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

[Perozzi *et al.*, 2014] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of SIGKDD*, pages 701–710. ACM, 2014.

[Wallinga and Teunis, 2004] Jacco Wallinga and Peter Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of epidemiology*, 2004.

[Wang *et al.*, 2017a] Jia Wang, Vincent W Zheng, Zemin Liu, and Kevin Chen-Chuan Chang. Topological recurrent neural network for diffusion prediction. In *ICDM*, pages 475–484. IEEE, 2017.

[Wang *et al.*, 2017b] Yongqing Wang, Huawei Shen, Shenghua Liu, Jinhua Gao, and Xueqi Cheng. Cascade dynamics modeling with attention-based recurrent neural network. In *Proceedings of IJCAI*, 2017.

[Wang *et al.*, 2018a] Zhitao Wang, Chengyao Chen, and Wenjie Li. Attention network for information diffusion prediction. In *Proceedings of WWW*, 2018.

[Wang *et al.*, 2018b] Zhitao Wang, Chengyao Chen, and Wenjie Li. A sequential neural information diffusion model with structure attention. In *Proceedings of CIKM*, 2018.

[Watts and Dodds, 2007] Duncan J Watts and Peter Sheridan Dodds. Influentials, networks, and public opinion formation. *Journal of consumer research*, 34(4):441–458, 2007.

[Williams, 1992] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[Yang *et al.*, 2018] Cheng Yang, Maosong Sun, Haoran Liu, Shiyi Han, Zhiyuan Liu, and Huanbo Luan. Neural diffusion model for microscopic cascade prediction. *arXiv preprint arXiv:1812.08933*, 2018.

[Zhong *et al.*, 2012] Erheng Zhong, Wei Fan, Junwei Wang, Lei Xiao, and Yong Li. Comsoc: Adaptive transfer of user behaviors over composite social network. In *Proceedings of SIGKDD*, 2012.