Information diffusion analysis in online social networks based on deep representation learning

Chen, X.

Chapter 1

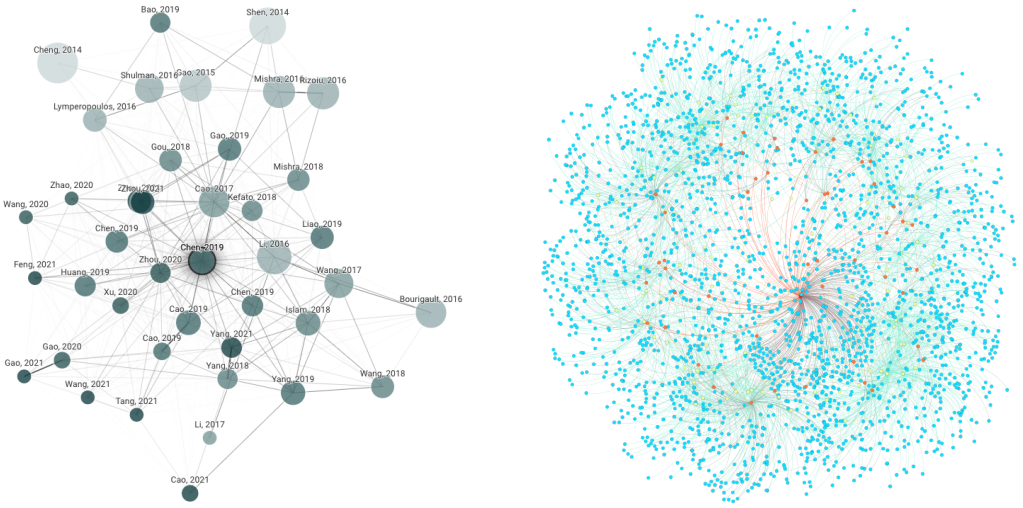
Introduction

1.1 Background

According to Pew Research[1], as of 2017 approximately 88% of American adults have either free or paid Internet access at home, and about 81% obtain news from online platforms (e.g., news websites/apps, social media, or both). We can see that with the rapid development of Internet technology, the way people share information has gradually democratized. Especially, the rise of large online social network (OSN) platforms, such as Twitter[2], Sina Weibo[3], Facebook[4], etc., have provided an environment for free information creation and distribution while substantially changing how people acquire and share information. In a nutshell, everyone can generate and share various contents and communicate on topics of mutual interest on these platforms without hindrance. Such activities facilitate the fast diffusion of information (both true and false) in various contexts, for example, the spread of rumors in News, the propagation of marketing campaigns, the diffusion of innovative technological achievements, and so on, which spur the phenomenon of information cascades. Information cascade is formed as information or innovative ideas propagated among users [1], which is ubiquitous and has been identified in various settings: e.g., paper citations [2], blogging space [3, 4], email forwarding [5, 6], as well as in social medias (e.g., Twitter [7, 8] and Sina Weibo[9]). Figure 1.1 shows the example for citation cascade and retweet cascade, respectively.

Understanding how information spreads through OSNs, as well as what elements drive the success of information diffusion, and making forecasts about the population size that information can affect, has significant implications for a wide range of real-world applications [11], such as marketing viral discrimination [12], user behavior prediction [13], media advertising [14], social recommendation [15], and fake

[1]https://www.journalism.org/2016/07/07/pathways-to-news/
[2]https://twitter.com/
[3]https://weibo.com/
[4]https://www.facebook.com/



(a) Citation cascade **(b)** Retweet cascade

Figure 1.1: Examples of cascade for paper citation and tweet retweet. (a) A citation cascade for CasCN [10], a screenshot from the website – Connected Paper (https://www.connectedpapers.com/). (b) A retweet casdcade from Weibo.

news detection [16], etc. Therefore, analyzing the diffusion of a given piece of information is urgently expected by academia and industry. Among various information diffusion analysis tasks, modeling and predicting the information cascade turns out to be of utmost importance since it enables controlling (or accelerating) information spreading in various scenarios, which is the first research problem we try to tackle in this thesis.

Furthermore, OSNs are a double-edged sword. On the one hand, OSNs brought enormous convenience to people's daily life. On the other, the proliferation of fake news, rumors and false information has had strong and negative societal and economic consequences. The explosive spread of false information can pose a threat to the credibility of legitimate online platforms and resources and has a serious negative impact on both individuals and society [17], with the potential consequences to destabilize nations, affect the fairness of competition [18], and shock the stock market [19]. Take the more recent event as an example. In the global effort to contain the COVID-19 pandemic, misinformation abounds and flourishes on the Internet [20, 21], and people have been led to believe that COVID-19 can be cured by ingesting fish tank cleaning products or that 5G networks generate radiation that triggers the virus. Such misinformation not only causes panic among citizens but could potentially undercut collective efforts to control the pandemic. Thus, detecting rumors on OSNs as early as possible is a necessary, urgent and socially beneficial task, which is the second research problem studied in this thesis.

In the past decade, researchers made enormous efforts and remarkable progress in trying to find effective solutions to the above two problems, i.e., information cas-

cade modeling and rumor detection. Specifically, prior works tried to solve these two tasks by extracting various hand-crafted features (e.g., content features [22, 23, 24], user features [25, 26], and social context features [27, 28, 29], etc.) and incorporating the power of machine learning models (e.g., random forest, naive Bayes, and support vector machines). However, these well-designed features required extensive domain knowledge and thus are hard to be generalized to new domains. Besides that, researchers also proposed point process-based approaches [2, 30] to do information cascade modeling, focusing on modeling the intensity function of the arrival process for each message independently. These methods demonstrate an enhanced interpretability but are still unable to fully leverage the information encoded in the cascade for a satisfactory prediction. And as for rumor detection, some researchers leveraged the inter-entity relations and constructed the so-called credibility network [31, 32] to find the truth against conflicting information, but the performance of these models heavily relies on the quality of the credibility network used.

Inspired by the recent success of deep representation learning in computer vision and natural language processing, recently, researchers began to employ deep learning techniques to develop models for information cascade modeling [8, 9] and rumor detection [33, 34]. These models aim to learn more powerful high-level feature representations directly from raw data via various deep learning techniques, which improve model performance and alleviate the heavy manual effort in conventional methods at the same time. However, the existing deep learning-based approaches in information cascade modeling and rumor detection still face some limitations, such as incomplete feature extraction, inefficient feature fusion, absence of fine-grained feature learning, and so on. The core work of this thesis is to eliminate these problems, focusing on the development of **deep learning-based models** to solve the problem of **information cascade modeling** and **rumor detection**.

1.2 Research Questions and Contributions

This thesis is structured in two parts: Part I will introduce deep learning-based models for information cascade modeling task. And Part II will focus on the topic of rumor detection.

Information cascades modeling is accomplished via specific prediction tasks, which are categorized into two levels: Micro-level and Macro-level. (1) At micro-level, local patterns of social influence are studied – e.g., inferring the action status of a user [35, 36]. The methods predict the likelihood of a user propagating a particular piece of information, or forecast when the next propagation might occur given a certain information cascade [36]. (2) At macro-level, typical studies include cascade size prediction [8, 9, 35, 37, 38] and outbreak prediction (above a certain threshold) [1, 7, 37, 39], both cascade size (popularity) prediction and outbreak prediction are aiming

to estimate the future size (popularity) of the diffusion cascade. Because micro-level tasks requires the complete diffusion network/social network, we need to know the historical interactions (retweet/social relationship (follower/following)) among all users, which is extremely large and hard to acquire. Therefore, in this thesis, we focus on **macro-level information cascade prediction**, and the research questions and corresponding contributions of Part I are listed as follows:

Research Question 1 (RQ 1) *Can we develop an effective deep learning-based model to capture structural and temporal features from the observed cascade graph for macro-level information cascade prediction?*

Earlier deep learning-based approaches either only extracted temporal features from the diffusion paths [9] but ignored the structural features or learned structural features from the global graph [8]. The key challenge of **RQ 1** is how to learn structural and temporal features at the same time when only given the observed cascade graph. Our key contribution to address this research question is the design of a novel graph-based neural network called **Recurrent Cascades Convolutional Network (CasCN)**. CasCN is introduced in Chapter 4 and based on the following publication[10], that to the best of our knowledge, is the first work to study this problem :

- **Chen, X.**, Zhou, F., Zhang, K., Trajcevski, G., Zhong, T., Zhang, F.: Information diffusion prediction via recurrent cascades convolution. In: 2019 IEEE 35th International Conference on Data Engineering. ICDE '19 (2019) 770–781

CasCN is designed based on the long short-term memory network [40] and the graph convolutional network [41], which captures structural-temporal features from a sequence of subgraphs – sampled from the observed cascade graph. CasCN also introduces a new way to calculate the Laplacian, allowing it to handle directed graphs. We demonstrate significant advantages of using CasCN compared with earlier state-of-the-art methods both in terms of feature learning and predictive accuracy.

Research Question 2 (RQ 2) *How can we improve upon earlier deep learning-based models that learn the latent representation for the observed cascade graph from a multi-scale perspective to predict the future size of this cascade?*

Existing works demonstrate the effectiveness to extract structural-temporal features for macro-level information cascade prediction. However, they still fail to take the higher-order and position information for each node into consideration in modeling the structure of information diffusion, which carries out our second research question **RQ 2**. To address RQ 2, we introduce a novel graph-based model – **Multi-scale Cascades model (MUCas)** in Chapter 5 based on the following publication [42]:

- **Chen, X.**, Zhang, F., Zhou, F., Bonsangue, M.: Multi-scale graph capsule with influence attention for information cascades prediction. International Journal of Intelligent Systems 37 (2022) 2584–2611

MUCas makes full use of the direction-scale, high-order-scale, position-scale, and dynamic-scale of cascades via a newly designed multi-scale graph capsule network (MUG-Caps) and the influence-attention mechanism. And the experiments on two real-worlds datasets demonstrate the superiority of the proposed model on macro-level information cascades prediction.

Having introduced deep learning-based models for macro-level information cascade prediction task in Part I of this thesis (Chapter 4 and 5). In Part II, we will explore the possibility to detect rumors by extracting various diffusion patterns. Besides that, we also investigate the importance of users in the diffusion of rumors by developing participant-level rumor detection models.

Research Question 3 (RQ 3) *Can we detect rumors at an early stage by learning various diffusion patterns from the information diffusion?*

Rumors are created by mimicking the real news, which aims to mislead the public, making it difficult to be detected by using textual and visual features. A recent empirical study [43] demonstrated that rumors and non-rumors show different diffusion patterns. The key challenge of **RQ 3** is how to develop an effective deep learning-based model to explore the full-scale diffusion patterns of rumors, i.e., from both macroscopic and microscopic perspectives. In Chapter 6, we design a novel diffusion-based rumor detection model to solve **RQ 3**, called **Macroscopic and Microscopic-aware Rumor Detection model (MMRD)**, this chapter is based on the following publication[44]:

- **Chen, X.**, Zhou, F., Zhang, F., Bonsangue, M.: Modeling microscopic and macroscopic information diffusion for rumor detection. International Journal of Intelligent Systems 36 (2021) 5449–5471

MMRD leverages graph neural networks to learn the macroscopic diffusion of rumor propagation and capture microscopic diffusion patterns using bidirectional recurrent neural networks while taking into account the user-time series. Moreover, it leverages knowledge distillation technique to create a more informative student model and further improve the model performance. Experiments conducted on two real-world data sets demonstrate that our method achieves significant accuracy improvements over the state-of-the-art baseline models on rumor detection.

Research Question 4 (RQ 4) *Can we improve the model performance by developing effective rumor detection models at the participant level?*

Users are the main contributor to rumor spreading in online social networks, modeling the rumor spreading at a more fine-grained participant-level rather than event-level. This hypothesis improve detection accuracy and is at the core of research

question **RQ 4**. To answer RQ 4, we proposed two participant-level models in Chapter 7, i.e., **Participant-level Rumor Detection model (PLRD)** and **User-aspect Multi-view Learning with Attention for Rumor Detection model (UMLARD)**. This chapter is based on the following publications [45, 46]:

- **Chen, X.**, Zhou, F., Zhang, F., Bonsangue, M.: Catch me if you can: A participant-level rumor detection framework via fine-grained user representation learning. Information Processing & Management 58 (2021) 102678
- **Chen, X.**, Zhou, F., Trajcevski, G., Bonsangue, M.: Multi-view Learning with Distinguishable Feature Fusion for Rumor Detection. Knowledge-Based Systems 240 (2022) 108085

PLRD aims to learn the users' social homophily, social influence, susceptibility, and temporal features for rumor detection, while UMLARD exploits different embedding methods to learn the view-specific high-level representations of a given post from the hierarchical diffusion process and user profiles. Both models are designed at participant-level, and experimental results show their superiority over state-of-the-art methods.

1.3 Thesis Outline

The overall organization of this thesis is as follows. In Chapter 1, we give a brief introduction to the thesis' background and its motivation, the main research questions and contributions, and the overview of this thesis. Chapter 2 provides a comprehensive literature review for existing methods in information cascade modeling and rumor detection, and focuses more on deep learning-based methods. In Chapter 3 presents some general definitions, which are throughout the whole dissertation and problem definitions of macro-level information cascades prediction and rumor detection. Besides that, Chapter 3 also briefly introduces some related technical supports.

The main content of this thesis related to the research questions is divided into two parts. Part I (Chapter 4 and Chapter 5) introduces two deep learning-based models for macro-level information cascades prediction. And Part II (Chapter 6 and Chapter 7) focuses on the task of rumor detection.

Specifically, in Chapter 4, we target **RQ 1** and propose the first graph-based neural network, which extracts structural-temporal features from a sequence of subgraphs and makes a prediction of the incremental size of the cascade. Chapter 5 address **RQ 2** and introduces a multi-scale graph capsule network to fully explore the direction-scale, high-order-scale, position-scale, and dynamic-scale information from the observed cascade graph.

Chapter 6 is related to **RQ 3** and introduces a novel deep learning model for rumor detection by exploring the microscopic and macroscopic diffusion patterns. In

Chapter 7, we target **RQ 4** and propose two participant-level model for rumor detection.

At last, Chapter 8 concludes the contributions of the thesis and discusses possible future work.