Diachronically Like-Minded User Community Detection

by

Hossein Fani

Master of Computer Science, 2009
Amirkabir University of Technology (Polytechic of Tehran)
Bachelor of Computer Science, 2005
Amirkabir University of Technology (Polytechic of Tehran)

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

In the Graduate Academic Unit of Faculty of Computer Science

Supervisor(s):	Weichang Du, Ph.D., Computer Science
	Ebrahim Bagheri, Ph.D., Computer Science
Examining Board:	Michael Fleming, Ph.D., Computer Science
	Arash Habibi Lashkari, Ph.D., Computer Science
	Donglei Du, Ph.D., Business Administration
External Examiner:	James Caverlee, Ph.D., Computer Science and Engineering
	Texas A&M University, United States

This dissertation is accepted

Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

October, 2019

Abstract

Study of users' behaviour, interests, and influence is of interest within the realm of online social networks due to its wide range of applications, such as personalized recommendations and marketing campaigns. However, the proposed approaches are not always scalable to a large number of users and a huge amount of user-generated content. Community-level studies are introduced to facilitate scalability, among other characteristics, highlighting the main properties of the network at a higher collective level. Prior work is mainly focused on the identification of online communities that are formed based on shared links and/or similar content. However, there is little literature on detecting communities that simultaneously share topical and *temporal* similarities. To extract diachronically like-minded user communities who have similar temporal dispositions according to their topics of interest from social content, we put forward two approaches: *i*) multivariate time series analysis, and *ii*) neural embeddings. In the former approach, we model users' temporal topics of interest through multivariate time series, and inter-user affinities are calculated based on pairwise cross-correlation. While simple and

effective, this approach suffers from sparsity in multivariate time series. In the latter method, however, each user is mapped to a dense embedding space and inter-user affinities are calculated based on pairwise cosine similarity.

While the objective of these two proposed approaches is to identify user communities up until the present; in the last step of this thesis, we propose two approaches to identify future communities, i.e., community prediction: *i*) Granger regression, and *ii*) temporal latent space modeling. In Granger regression, we propose to consider both the temporal evolution of users' interests as well as inter-user influence through the notion of *causal dependency*. In the latter method, however, we assume that each user lies in an unobserved latent space, and similar users in the latent space are more likely to be members of the same user community. The model allows each user to adjust her location in the latent space as her topics of interest evolve over time.

Empirically, we demonstrate that our proposed approaches, when evaluated on a Twitter dataset, outperform existing methods under two application scenarios, namely news recommendation and user prediction.

to my ma!

& Reza Rezaei
& Arash, Haniye, Deyar.

تقدیم به مادرم!

و ر- ابهام
و آرش، هانیه، دیار.

Acknowledgements

Foremost, I'd like to thank my advisor, Ebrahim Bagheri, for his wisdom, guidance, support, encouragement, patience, friendship, love, and his *magic wand!* Ebrahim, I am truly honored and fortunate to have you as *my prof.*

I would like to thank my supervisor, Weichang Du. His input, comments, and support have been absolutely invaluable.

I would like to thank my thesis committee members, Michael W. Fleming, Dima Alhadidi, Dr. Arash Habibi Lashkari, and Dr. Donglei Du for their valuable comments and support.

I am thankful to my collaborators: F. Zarrinkalam, J. (Dani) Ghatta, A. Pourmasoomi, Y. (Luna) Feng, X. Zhao, Bashari brothers, Mahdi and Masoud, M. Noorian, A. Mirlohi, H. Hosseini, N. Arabzadeh, K. Poots, M. Kargar, E. Jiang, T. Nguyen, A. Navivala, and M. Khodabakhsh; special thanks to Fattane, *the Angel*, for long-lasting collaborations, and career advice.

I am very grateful to Zhiting Hu for providing me with the implementation of his proposed approach [51].

My deepest gratitude goes to my caring and loving friend, Marina Ameri. Your supports adjourned this journey most gracefully and are duly noted.

Thanks, also, to others who offered to help.

Table of Contents

List of Tables

List of Figures

List of Symbols, Nomenclature or Abbreviations

Chapter 1

Introduction

Online social networks have been an effective medium for communication and social interaction. Predicting users' behaviour, interests, and influence are of interest within this realm due to the wide range of applications such as personalized recommendations and marketing campaigns. However, the proposed approaches are not always scalable to large numbers of users and huge amounts of user-generated content. Community-level studies are introduced to help facilitate scalability, among other characteristics, highlighting the main properties of the network at a higher collective macro level. Therein, information sharing and communication patterns of users in social network platforms lead to the formation of communities that consist of like-minded or similarly behaving users.

In order to support community-level models, various community detection methods have been proposed in the literature. Topology-based commu-

nity detection methods focus on explicit links, e.g., followership on Twitter[1], to detect like-minded users [111, 38]. However, they fall short to identify accurate like-minded user communities due to two main reasons. First, many of the explicit social connections are grounded in other factors, e.g., kinship, that do not necessarily point to inter-user interest similarity [27]. Second, like-minded users are not necessarily explicitly connected to each other. As a result, content (topic)-based approaches are introduced [1, 115]. There are also hybrid approaches that incorporate both topology and content to identify a reliable account of like-minded communities [3, 93]. However, the approaches proposed in [3, 1, 115, 93] do not take into account the fact that like-minded users need to exhibit similar *temporal* behaviour towards similar topics as well. Indeed, users' interests have a dynamic nature and drift over time in online social networks; a user may become interested in a new topic, lose interest in a topic, or change her degree of interest towards a topic [85]. For example, Figure 1.1 shows how the degree of interest of three Twitter users `@joe`[2], `@john` and `@mary`, changes over the *'War in Afghanistan'* topic from mid November to the end of December 2010. The first two users seem to share a similar behavioural pattern towards this topic. However, another user `@mary` does not start posting about the same topic until much later in late December of the same year. While the three users share a similar interest, they do not exhibit this interest in similar time intervals. Contem-

[1] `twitter.com`

[2] These names are being used in this thesis as substitutes for the users' real Twitter screen names.

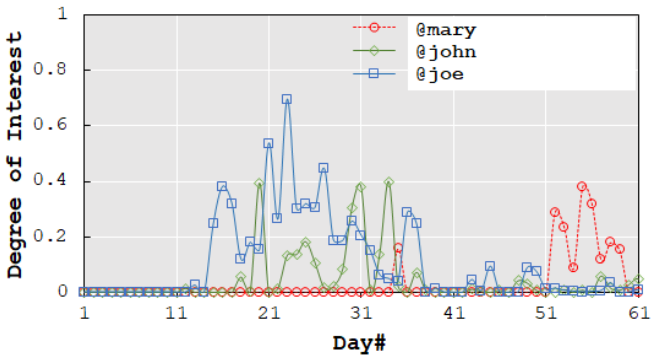

Figure 1.1: Different temporal inclination of three Twitter users with respect to the *'War in Afghanistan'* topic from November till end of December 2010.

porary community detection methods, such as the aforementioned methods, would cluster all these three users in the same community because they do not incorporate the temporal nature of users' topics of interest. This renders it difficult for applications such as news recommender systems to generate recommendations that are temporally sensitive. If the three mentioned users were identified as members of the same community, they would be recommended the same news articles at the end of December on the given topic. However, `@joe` and `@john` have already covered this topic in November and have now moved on and as a result are not interested in it any longer but `@mary` has just become interested in the topic.

The first objective of our work is to identify diachronically like-minded user communities that have similar temporal dispositions with regards to the topics of interest. A diachronic approach (from Greek from $\delta\iota\alpha$-'through' and $\chi\varrho o\nu o\varsigma$ 'time') considers the way in which users' topics of interest has

developed and evolved through time. Specifically, we want to identify the following type of communities: those communities that distinguish between the users who are interested in a similar set of topics this month, e.g. {`@joe`, `@john`}, from those who have the same behavioural pattern towards the same set of topics but in a different time period, e.g., `@mary`; hence, supporting temporality in topic-based community detection methods. We propose two topical and temporal community detection methods, *i*) multivariate time series analysis, and *ii*) neural embeddings of users, that measure inter-user similarity based on temporal topical interests. In the former approach, we model users' temporal topics of interest through multivariate time series and inter-user affinities are calculated based on pairwise cross-correlation between users' multivariate time series. While simple and effective, this approach suffers from sparsity in multivariate time series. In the latter method, however, each user is mapped to a dense vector space (embedding space) and inter-user affinities are calculated based on pairwise cosine similarity between users' vectors in embedding space. Both approaches are completely independent of the underlying topic detection method and are applicable in any textual content sharing networks which include timestamps for the shared content, e.g. tweets, blog posts, news articles and citation networks, just to name a few.

The second objective of our work is to predict latent like-minded user communities in a future yet to be observed time interval. While the objective of the first two proposed methods is to identify user communities up until the

present (now), in the last step of this research we put forward two approaches, namely Granger regression and temporal latent space modeling, to identify future communities, i.e., community prediction. More specifically, given a sequence of users' contributions towards a set of topics of interest from time interval 1 to T, the objective is to predict user communities in a future time interval T + 1. In the Granger regression method, G-regression for short, we propose to consider both the temporal evolution of users' interests as well as inter-user causal influence. We employ the Granger concept of causality to determine the degree of inter-user influence that can be used to identify which users play influential roles in the behavioural evolution of one or more other users. Based on Granger causality, we identify a causing user c to influence the affected user e if and when the past observations of c lead to a more accurate prediction of the behaviour of e above and beyond the information contained in past observation of e alone. Although the proposed G-regression method has shown promising results, it requires building predictive models on a per user basis and, hence, is computationally expensive. Alternatively, we propose a temporal latent space model for user community prediction in online social networks based on a history of users' past and present topics of interest. The model assumes that each user lies in an unobserved latent space, and similar users in the latent space representation are more likely to be members of the same user community. The model allows each user to adjust her location in the latent space as her set of topics of interest evolve over time.

1.1 Research Objectives

Communities provide summarization of network structure and content, highlighting the main properties of the network at a macro level; hence, they give insights into the dynamics and the overall status of the network. Community detection finds application in areas as diverse as sociology, biology, marketing and computer science. In sociology, it helps with understanding the formation of action groups in the real world such as clubs and committees. In biology, community detection methods are used to organize computationally probed protein structure spaces [55]. In marketing, companies can use communities to design targeted marketing, as the Edelman Trust Barometer Report found, 44% of users react to online advertisements if other users in their peer group have already done so. In computer science, how information is disseminated in the network through communities has been studied primarily because community drives like-minded people to connect and encourages them to share more content. Also, communities are employed to discover previously unknown interests of users, also known as alias interest detection, which can be potentially useful in recommender systems to set up efficient recommendations [93]. In a very recent concrete application, Customer Relationship Management (CRM) systems are empowered to tap into the power of social intelligence by looking at the collective behaviour of users within communities in order to enhance client satisfaction and experience. As an example, customers often post their opinions, suggestions,

criticisms or support requests through online social networks such as Twitter or Facebook[3]. Customer service representatives would quickly identify the mindset of the customer that has called into the call center by a series of short questions. For such cases, appropriate techniques are required that would look at publicly available social and local customer data to understand their background so as to efficiently address their needs and work towards their satisfaction. Important data such as the list of influential users within the community, the position of a given user in relation to influential users, the impact of users' opinions on the community, customers' social behavioural patterns, and emergence of social movement patterns are of interest in order to adapt the customer care experience for individual customers [87].

The first part of this research is motivated by the very fact that mindsets are temporal, i.e., people's perceptions of things, events, and world are constantly changing. For example, as shown in Figure 1.1, while the three Twitter users seem to share a similar interest in the *'War in Afghanistan'* topic, they are not aligned in time. `@mary` is not interested in the topic, as she does not start posting about it, until much later in late December of the same year. We do not consider all these three users like-minded as opposed to the state-of-the-art community detection methods because we add a temporal dimension. Our research helps applications such as news recommender systems with recommending the right items at the right point in time. As such, `@joe` and `@john` as a member of the same community would be recom-

[3] `www.facebook.com`

mended relevant news articles in November while `@mary` receives the same articles but later in December as members of a different community.

The second part of this research, i.e., community prediction, allows us to perform forward planning by predicting users' topics of interest in the future and is of high theoretical and practical significance. For instance, while our research in the first part helps applications such as news recommender systems with recommending the right items in the present time, community prediction allows us to know future probable community affiliations for users such that news recommender systems are able to recommend the right items in the future yet to be observed time intervals. For instance, we are interested in determining whether `@mary` continues to be interested in the topic *'War in Afghanistan'* for the next month, i.e., January, or what the probabilities are for `@joe` and `@john` to return to this topic yet again in the future.

This thesis aims to develop methods to form communities of like-minded users based on their temporal and topical interests in the past, present and future. While a large body of literature has been devoted to community detection, many challenges remain to be addressed. In the following, we review some of the challenges and present detailed objectives of this research in terms of research questions.

1.1.1 Research Question 1 (RQ1)

Does the consideration of temporal evolution of users' topics of interest lead to higher quality communities compared to when time

is overlooked?

The overarching goal of this research is to address this research question. We propose two alternative temporal content-based (topical) user community detection methods which aim to latent communities whose members share higher similarity with respect to topics of interest over time. As such, those users who share not only similar topical interests but also share similar temporal behaviour are considered to be like-minded and hence members of the same community. In contrast, those users who are simply dissimilar in topics of interest or share similar topical interests but in different time intervals are not considered like-minded and need to end up in different communities.

Existing methods have often overlooked the temporal nature of users' topics of interest. As a result, they can fall short in temporally sensitive applications such as news recommender systems where it is imperative to recommend relevant news articles at the right time. This needs to be addressed by our proposed methods.

1.1.2 Research Question 2 (RQ2)

Do temporal content-based methods lead to higher quality communities compared to link-based methods?

Instead of inter-user connections (links) in the social network structure, we propose to take users' content similarity into account in our methods for two main reasons: *i*) there are many users on a social network that have similar interests but are not explicitly connected to each other; and, *ii*) an

explicit social connection does not necessarily indicate user interest similarity but could be owing to sociological processes such as conformity, aspiration, and sociability or other factors such as friendship and kinship that do not necessarily point to inter-user interest similarity. There are also some special cases where link-based methods are not applicable like when the network is not available or misleading, e.g., when links are fraudulent because of link-farmers (social capitalists).

The noisy and sparse link information misleads the process of community detection for existing methods which rely only on links and result in a poor set of user communities.

1.1.3 Research Question 3 (RQ3)

Do temporal content-based and link-based methods have synergistic impact on each other and reinforce the quality of the identified communities when applied in tandem?

Earlier non-temporal user community detection methods have already shown improvement when incorporating social network structure (links) with topics of interest (content) compared to those in which links and content are used separately [115, 93]. However, to the best of our knowledge, all existing temporal user community detection methods are only content-based and little has been studied on the effect of social network structure and temporal evolution of user content simultaneously.

In order to address this research question, we simultaneously consider

users' temporal content and their social network structure when identifying user communities, and we embed both users' temporal interests and their social network structure into a dense vector representation using neural embedding mechanisms. The user embeddings derived from two different information sources (modalities), i.e., *i*) temporal content-based embeddings based on users' topics of interest over time, and *ii*) network embeddings based on social network neighborhoods, are then linearly interpolated to build a single final *multimodal* user embedding. The linear interpolation of two user embeddings at the embeddings level allows us to investigate how and to what extent users' dynamic topics of interest and/or users' social network structure contribute to the quality of the inferred user communities.

1.1.4 Research Question 4 (RQ4)

Is it possible to predict future-yet-unobserved content-based communities on social networks?

Our research is among the first to explore the idea of predicting future-yet-unobserved content-based (topical) communities on social networks. We propose a method that learns to represent users within a latent space that preserves users' similarities over time for the task of community prediction. Contrary to the proposed methods to address research questions **RQ1** to **RQ3**, which use users' temporal and topical interests for pairwise similarity to identify user communities up until the present (now), the proposed temporal latent space modeling uses such information for predicting user

communities in the future.

There are temporal link prediction methods which assume that the social network structure is dynamic and changing with time. While suitable for identifying future links between users in a social network structure, such methods inherently fall short when the communities need to take users' content similarity into account, i.e., identifying content-based user communities in the future due to the same reason as in **RQ2**, i.e., noisy and sparse link information.

1.2 Contributions

The concrete contributions of this research are:

1. We have modeled the contribution of each user towards topics using multivariate time series and exploit two-dimensional cross-correlation on such time series on a pairwise basis to find similar users in topics of interest and temporal behaviour. We employed Louvain clustering [15], a graph-based heuristic modularity-based partitioning algorithm, to create final user communities. To find topics from the text stream of the social network, we used state-of-the-art topic detection methods in order to show that our approach and its contributions are independent of topic detection algorithms and perform well on a variety of topic detection methods. We used one graph-based [104] and two probabilistic LDA [14] and ToT [101] methods. According to the

results obtained from a Twitter dataset covering a two-month period, our temporal topic-based community detection method is able to effectively identify user communities that are formed around temporally similar behaviour towards shared topics when compared to the non-temporal approaches. This contribution is to address **RQ1** and **RQ2**. In Chapter 3, Multivariate User Time Series, we provide the details.

2. We have proposed a neural embedding approach to identify temporally like-minded user communities. We modeled the users' temporal contribution towards topics of interest by introducing the notion of regions of like-mindedness between users. These regions cover users who share not only similar topical interests but also similar temporal behaviour. By considering the identified set of regions of like-mindedness as a context, we train a neural network such that the probability of a user in a region is maximized given other users in the same region. The final weights of the neural networks form the low-dimensional vector representation of each user that incorporates both topics of interest and their temporal nature. Finally, we applied the Louvain [15] clustering technique to identify like-minded user communities on a weighted user graph in which the similarity of two users is based on the cosine similarity of their respective vectors. We demonstrated the effectiveness of the user embedding approach on a Twitter dataset in the context of news recommendation and user prediction applications compared to our previous approach and the state of the art. This contribution is to

address **RQ1** and **RQ2**. In Chapter 4, Neural User Embeddings, we explain the details.

3. We have systematically interpolated temporal content-based embeddings and social link-based embeddings to capture both social network connections and temporal content evolution for representing users. We employ neural graph embedding techniques to embed information from users' social network structure into user representations. We build a single set of multimodal embeddings from embeddings of temporal social content and social network structure through their linear interpolation in order to elucidate the contribution of users' temporal content on the one hand, and social network structure, on the other hand, for finding user communities. Having learnt two different user vector representations of users from the temporal social content and social network structure, denoted by $\mathbf{W}_{\mathcal{D}}$ and $\mathbf{W}_{\mathcal{G}}$, respectively, we adopt a linear weighting mechanism to interpolate the embeddings into a single vector representation W, i.e., $\mathbf{W} = \alpha\mathbf{W}_{\mathcal{D}} + (1-\alpha)\mathbf{W}_{\mathcal{G}}$ where α denotes a weighting coefficient to interpolate between temporal content and social network structure in the final user vector representation. For instance, if $\alpha = 0$, the interpolated embeddings lead to the conventional link-based user community detection on the one extreme. On the other extreme, it will solely rely on temporal content if $\alpha = 1$ and becomes a pure temporal content-based method. The effect of embedding interpolation to the overall performance of user community detection is

evaluated by choosing $\alpha \in [0, 1]$. Although simple, linear weighting is uninformed, easy to implement, interpretable, and could achieve competitive performance across a wide span of different data types and domains. We demonstrated the synergistic impact of content-based and link-based user embeddings on a Twitter dataset in the context of news recommendation and user prediction applications. This contribution is to address **RQ3**. In Chapter 4, Neural User Embeddings, we explain the details.

4. We have proposed to incorporate the temporal evolution of users' contents as well as a stricter form of inter-user influence through *causal dependency* for user community prediction in online social networks. We use the Granger concept of causality [42] to determine the degree of inter-user influence that can be used to identify which users play influential roles in the behavioural evolution of one or more other users. Based on Granger causality, we identify a causing user c to influence the affected user e if and when the past observations of c lead to a more accurate prediction of the behaviour of e as opposed to when only the information contained in past observations of e is only used. This leads to an *influence network* in which the edges depict the influence amplitude and direction of its adjacent users. We use the influence network to perform interest prediction. Specifically, given a topic of interest z and a user e, we find e's influential neighbor(s) from the influence network such as c and build a vector autoregression model (VAR) based

on e and c's user-topic contribution time series to predict e's degree of interest toward topic z in the future. Last, a weighted *undirected* graph is formed over the users and their pairwise similarity on predicted degrees of interest in the future on which the Louvain method is applied to find user communities in the future. This contribution is to address **RQ4**. In Chapter 5, User Community Prediction, we explain the details.

5. We have proposed a temporal latent space model for user community prediction in online social networks. Given a set of topics $\mathbb{Z}$ extracted by a topic detection method (e.g., LDA) within T time intervals and a set of users $\mathbb{U}$ we built temporal graphs $G_t = (\mathbb{U}, \mathbb{E}_t, s)$ for each time interval $1 \leq t \leq \mathrm{T}$ whose nodes are users in $\mathbb{U}$ and $\mathbb{E}_t$ is the set of weighted undirected edges whose weights are based on the similarity function s which is defined as the cosine similarity of topic preference vector for the users at time interval t. A topic preference vector for user $u \in \mathbb{U}$ towards topic set $\mathbb{Z}$ at time interval $t : 1 \leq t \leq \mathrm{T}$ is a vector that indicates the preference by user u for each topic $z \in \mathbb{Z}$ at time interval t. The stream of graphs $[G_1..G_t..G_{\mathrm{T}}]$ within time period T could be considered as a dynamic graph $\mathcal{G}$ which is evolving over time. We map each node (user) u at G_t to a low-rank d-dimensional latent space, while imposing the following assumptions: *i*) users change their latent representations over time, *ii*) two users that are close to each other in $\mathcal{G}$ remain close in the latent space, *iii*) two users who are

close in latent space share similar topics of interest with one another more than two distant users. We use the local block coordinate gradient descent (bc-gd) algorithm, proposed by Zhu et al. [120], to predict $G_{\mathrm{T}+1}$ whose set of induced subgraphs form content-based user communities at time interval T+1. The proposed method can also be generalized to make predictions for any time period after T. This contribution is to address **RQ4**. In Chapter 5, User Community Prediction, we explain the details.

6. In the absence of gold standard user communities, we have proposed two application scenarios: news recommendation and user prediction, to quantitatively examine our proposed approaches. In these evaluation strategies, a temporal like-minded user community detection method is considered better *iff* its output communities improve an underlying application. To this end, a Twitter dataset, including ˜3M tweets posted during Nov. and Dec. 2010 have been collected. We build a gold standard dataset for the said applications by collecting news articles to which a user has explicitly linked in her tweets (or retweets). We postulate that users post news articles' URLs since they are interested in the topics of the news article. We build the gold standard from a set of news articles whose URLs have been posted by user u at time t. We see each entry as a triple (u, a, t) consisting of the news article a, user u, and the time t. As a result, $\mathbb{G} = \{(u, a, t) : u \in \mathbb{U}, a \in \mathbb{A}, 1 \leq t \leq \mathrm{T}\}$ forms our gold standard where $\mathbb{U}$ and $\mathbb{A}$ are sets of users and news

articles. Through output user communities, we predict the right news article(s) for the user u at time t, i.e., $(u, ?, t)$, in the news recommendation task. In the user prediction task, however, we predict the poster(s) of a news article a at time t, i.e., $(?, a, t)$. In Chapter 6, Evaluation, we provide further details.

1.3 Thesis Organization

The rest of this thesis is organized as follows:

Chapter 2 - Background and Related Work. This chapter covers preliminary concepts and definitions in the domain of user community detection, neural embeddings, and community prediction. Furthermore, a review of the most related research works which are related to the contributions of this thesis is provided. Specifically, the existing work in this area is surveyed based on a taxonomy of different information sources (modalities), i.e., links, content, and time, for the task of user community detection and prediction.

Chapter 3 - Multivariate User Time Series. The focus of this chapter is to develop a multivariate time series representation of users to model the contributions of each user towards the identified topics over time, which allows us to detect temporal content-based user communities.

Chapter 4 - Neural User Embeddings. In this chapter, we propose our temporal content-based neural embedding method, which learns low-dimensional user representations such that users who exhibit similar temporal

behaviour toward similar topics of interest are closer to each other in the embedding space.

Chapter 5 - User Community Prediction. This chapter goes through the task of community prediction in the future and describes our proposed methods, namely Granger regression and a temporal latent space model, to address content-based (topical) community prediction.

Chapter 6 - Evaluation. This chapter reports on our testbed in terms of the evaluation methodology, datasets, gold standard, and experiments.

Chapter 7 - Conclusion and Future Work. This chapter concludes this thesis with potential directions of improvement based on the limitations of this research.

1.4 Related Publications

- Hossein Fani. "Temporal Formation and Evolution of Online Communities." *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016.*

Chapter 2

Background and Related Work

This chapter provides a concise overview of the definitions, the underlying concepts, history and state of the art in user community detection and prediction.

2.1 Online Social Network

A social network is a net structure made up of social actors, mainly human individuals, and ties between them. Online social networks (OSN) are online platforms that allow social actors, i.e., users, in spatially dispersed locations to build social relations. Online social networks facilitate different modes of communication and present diverse types of social interactions. They not only allow individual users to be connected and share content, but also provide the means for active engagement, which enables users to play

social roles that they regularly undertake in real social settings. Such features have made OSNs a fundamental part of the global online experience, having pulled ahead of email [11]. Given individuals mimic their real world ties and acquaintances in their online social preferences [67], the tremendous amount of information offered by OSNs can be mined through social network analysis (SNA) to help sociometrists, sociologists, and decision makers from many application areas with the identification of actionable insight [22, 23]. For instance, despite the heterogeneity of user bases, and the variety of interactions, most of these networks exhibit common properties, including the small-world and scale-free properties [102, 9]. In addition, some users in the networks are better connected to each other than to the rest. In other words, individuals tend to associate with others who share similar interests in order to communicate news, opinions or other information of interest, as opposed to establishing sporadic connections, a tendency termed *homophily* as a result of which communities emerge on social networks [68].

2.2 User Community

The word community refers to a social context. People naturally tend to form groups, within their work environment, family, or friends. A community is a group of users who share similar interests, consume similar content or interact with each other more than other users in the network. Communities are either explicit or latent. Explicit communities are known in advance

and users deliberately participate in managing explicit communities, i.e., users create, destroy, subscribe to, and unsubscribe from them. For instance, Google's social network platform, Google+[1], had Circles that allowed users to put different people in specific groups. In contrast, in this work, communities are meant to be latent. Members of *latent* communities do not tend to show explicit membership and their similarity of interest lies within their social interactions.

No universally accepted quantitative definition of the community has been formulated yet in the literature. The notion of similarity based on which users are grouped into communities has been addressed differently in social network analysis. In fact, similarity often depends on the specific system at hand or application one has in mind, no matter whether they are explicit connections. The similarity between pairs of users may be with respect to some reference property, based on part of the social network or the whole. Nonetheless, a required property of a community is *cohesiveness*. The more users gather into groups such that they are intra-group close (internal cohesion) and inter-group loose (external incoherence), the more the group would be considered as a community. Moreover, in *partitioned* communities, each user is a member of one and only one community. However, in real networks users may belong to more than one community. In this case, one speaks of *overlapping* communities where each user, being associated with a mixture, contributes partially to several or all communities in the network.

[1]Google+ is no longer available since April 2019.

2.2.1 History

Probably the earliest account of research on community detection dates back to 1927. At the time, Stuart Rice studied the voting themes of people in small legislative bodies (less than 30 individuals). He looked for *blocs* based on the degree of agreement in casting votes within members of a group, called Index of Cohesion, and between any two distinct groups, named Index of Likeness [86]. Later, in 1941, Davis et al. [25] performed a social anthropological study on the social activities of a small city and surrounding county of Mississippi over 18 months. They introduced the concept of *caste* to the earlier studies of community stratification by social class. They showed that there is a system of colored caste which parsed a community through rigid social ranks. The general approach was to partition the nodes of a network into discrete subgroup positions (communities) according to some equivalence definition. Meantime, George Homans showed that social groups could be detected by reordering the rows and the columns of the matrix describing social ties until they form a block-diagonal shape [48]. This procedure is now standard and mainly addressed as blockmodel analysis in social network analysis. Further analysis of community structure was carried out by Weiss and Jacobson in 1955 [103] who searched for work groups within bureaucratic organizations based on attitude and patterns of interactions. The authors collected the matrix of working relationships between members of an agency by means of private interviews. Each worker had been asked to list her workers along with frequency, reason, subject, and the importance of her

contacts with them. In addition to reordering a matrix's rows and columns, work groups were separated by removing the persons working with people of different groups, i.e., liaison person. The concept of *liaison* has been referred to as *betweenness* and is at the root of several modern algorithms for community detection.

2.3 Prior Work

In this thesis, we will focus on identifying and modeling the latent like-minded user communities detected in a given time period on online social networks. As a result, the main research area that is closely related to our proposal is community detection which we review in this section.

Existing community detection approaches can be broadly classified into two categories [38]; *link*-based and *content*-based approaches. Link-based approaches, also known as topology-based, see a social network as a paradigmatic example of a graph, whose nodes are users and edges indicate explicit user relationships. On the other hand, content-based approaches, also known as topic-based, mainly focus on the information content of the users in the social network to detect communities. The goal of content-based approaches is to detect communities formed around the topics extracted from users' information content. *Hybrid* approaches incorporate both topological and topical information to find more meaningful communities with higher quality. Recently, researchers have performed longitudinal studies on the task of

community detection in which the social network is monitored in time intervals over a period of time [50, 51, 64]. The Time dimension results in a new *temporal* form of community detection which is the main motivation of this thesis. The following section includes the details of seminal works in each category.

Herein, we use the terms 'graph' and 'network' interchangeably as well as the terms 'vertex', 'node', and 'user'.

2.3.1 Link Analysis

Link-based user community detection methods are primarily based on the homophily principle [68] where links between users are considered important clues for interest similarity and, as a result, densely connected groups of users imply a user community. In this line of work, the social network is modeled as a graph with nodes representing users and edges representing relationships or interactions. The primary principle considered in this line of work is *connectedness*, which means that connections within each community are dense and connections among different communities are relatively sparse. To this end, primitive graph structures such as components, cliques, k-plexes or other pseudo-clique structures are considered to represent user communities [24, 38, 60]. There are also graph partitioning (clustering) approaches which try to minimize the number of links between user communities so that the users inside one community have more intra-connections than interconnections with other communities. Such approaches are based on iterative

bisection: continuously dividing one group into two groups, while the number of communities which should be in a network is unknown. The GirvanNewman approach [41] is one of the most commonly used methods in link-based user community detection. It partitions the graph by gradually removing links with high betweenness centrality in a descending order. Betweenness centrality for a link is defined as the number of the shortest paths between any pairs of nodes that go through the link in a graph or network. A link with a high betweenness centrality score represents a bridge-like connector between two parts of a network such that the communication between many pairs of nodes through the shortest paths between them is affected by its removal.

Other graph partitioning approaches include *modularity* maximization [15, 77], stochastic models [83, 47, 54], spectral methods [78], max-flow min-cut theory [61], and conductance cut minimization [28], among which the first two categories are widely adopted and thus worth more investigation.

Modularity maximization was first introduced by Newman [77]. The modularity function is defined as the difference between the number of links within user communities and the expected number of such links over all pairs of users. Modularity of a user community is a scalar value between -1 and 1. It is positive if the number of edges within the community exceeds the expected number on a random basis. Intuitively, modularity reflects the concentration of edges within modules (communities) compared with random distribution of links between all nodes regardless of modules. The simplest

formulation for a social network structure with two user communities would be the following:

$$Q = \frac{1}{4m} \sum_{uv} (a_{uv} - \frac{k_u k_v}{2m})(h_i h_j) \tag{2.1}$$

where a_{uv} is the number of edges between users u and v, which will normally be 0 or 1, h_i equals to 1 (or -1) if user u belongs to the first (or second) community, $\frac{k_u k_v}{2m}$ is the expected number of links between users u and v if edges are placed randomly, k_u is the degree of user u and $m = \frac{1}{2} \sum_u k_u$ is the total number of links in the network. Maximizing modularity for all communities has been proven to be NP-hard [16] for which optimization algorithms using greedy algorithms, spectral methods, simulated annealing, sampling techniques, and mathematical programming have been proposed [20]. For a review on different formulations of modularity, in-depth analysis of its corresponding maximization methods, and its problem, namely the resolution limit problem, see [20].

In stochastic models [83, 47, 54], user communities are considered as latent variables and the links between users are derived by a generative process where the probability of a connection (link) for a pair of users u and v is based on the probabilities that u and v belong to the same communities. Formally, $P_{uv} = \sum_{k=1}^{\mathrm{K}} P_{uk} P_{vk}$ where K is the number of communities, P_{uv} is the probability that there is a link between users u and v, and P_{uk} and P_{vk} are probabilities that u and v belong to community k. The user community detection problem can be reformulated as non-negative matrix

factorization, i.e., $\mathbf{A} \sim \mathbf{H}\mathbf{H}^\intercal$ where $\mathbf{H}$ represents the latent variables which show users' membership in communities and $\mathbf{A}$ is the adjacency matrix for the social network structure which is reconstructed (generated) by $\mathbf{H}$. Square of the Frobenius norm and Kullback-Leibleer divergence (KLdivergence) are two alternatives to calculate the reconstruction error at each optimization iteration in order to estimate the user communities, i.e., $\mathbf{H}$.

Neural networks, particularly those with deep structures, have been successfully applied to link-based community detection. Yang et al. [111] have proposed to employ auto-encoders to approximate modularity. Specifically, by defining the modularity matrix $\mathbf{B} = [b_{uv}]$ whose elements are $(a_{uv} - \frac{k_u k_v}{2m})$, Eq. (2.1) is re-written as $Q = \frac{1}{4m}\mathbf{h}^\intercal\mathbf{B}\mathbf{h}$ where $\mathbf{h}$ is a community membership indicator matrix. The modularity matrix is then input to an auto-encoder to learn a new representation that can best approximate the original data. Neural networks bring a nonlinearity advantage compared to linear approximation methods like non-negative matrix factorization (NMF) and eigenvalue decomposition (EVD) into modularity maximization, especially because real-world networks have nonlinear properties, e.g., links between users. In this line of work, graph representation learning methods have also been proposed where a data-driven approach is employed to automatically encode graph (social network structure) elements, i.e., nodes (users), edges (links), or even the entire graph, into a dense low-dimensional vector space followed by a clustering method. Node2vec [43] and DeepWalk [80] employ a second order random walk to sample network neighborhoods in a graph and output

vector representations (embeddings) that maximize the likelihood of preserving topological structure of each node neighborhood in the graph. This is in contrast to previous work which uses hand-engineered statistics like node degrees to extract a network's structural information. This not only saves time and effort in the feature engineering process, but also is agnostic to the downstream task. The embeddings can be easily fed into tasks such as user classification and link prediction. In user community detection, graph representation learning offers an unsupervised way to encode homophily into a vector of real values. More sophisticated methods based on deep autoencoders such as deep neural graph representations (DNGR) [39, 18] and structural deep network embeddings (SDNE) [33, 99] have been also proposed to generate graph embeddings.

Not all link-based methods perform well on large real-world networks that have many complex structural features such as sparsity, heavy tailed degree distributions and small diameters, among others. For empirical comparison of these algorithms in practice, see [100, 62].

Nonetheless, link-based methods inherently fall short when the communities of interest need to take users' content similarity into account. This is mainly due to two reasons: *i*) there are many users on a social network that have similar interests but are not explicitly connected to each other; and, *ii*) explicit social connections do not necessarily indicate user interest similarity but could be owing to sociological processes such as conformity, aspiration, and sociability or other factors such as friendship and kinship

that do not necessarily point to inter-user interest similarity [96, 30]. There are also some special cases where link-based methods are not applicable like when the network is not available [10] or misleading, e.g., when links are fraudulent because of link-farmers (social capitalists) [59].

2.3.2 Content Analysis

With the development of social media, a significant amount of user-generated content, known as social content, is available within user networks. Users communicate and interact with each other in social network websites. Besides the links between users, huge amounts of textual content are generated as well. Along with rich information in social network structure, user graphs can be extended with textual information on nodes. In social networking sites, users maintain profile pages, write comments and share articles. In photo and video sharing sites, users use short texts to tag photos and videos. In microblogging websites, users post their status updates. Therefore, researchers have explored the possibility of utilizing the topical similarity of social content. They have been proposing topic-based community detection methods, irrespective of the social network structure, to build like-minded communities of users [10, 63, 76, 1, 115, 93].

Most of these content-based methods have been inspired by latent Dirichlet allocation (LDA) [14] in one way or another and focused on probabilistic generative models based on textual content [119, 93]. For example, Zhou et al. [119] have modeled communities based on topics of interest through

a community-user-topic generative process to identify user communities. In their work, communities follow multinomial distribution over topics with Dirichlet priors where each user is posting about her topics of interest based on the conditional probability of a topic given each community. Abdelbary et al. [1] have identified users' topics of interest and extracted latent communities based on the topics utilizing Gaussian Restricted Boltzmann Machines. Yin et al. [115] have integrated community detection with topic modeling in a unified generative model to detect communities of users who are coherent in both structural relationships and latent topics. In their framework, a community can be formed around multiple topics and a topic can be shared between multiple communities. Sachan et al. [93] have proposed probabilistic schemes that incorporate users' posts, social connections and interaction types to discover latent user communities in social networks. They have considered three types of interactions: a conventional tweet, a reply tweet and a re-tweet. Author-Topic-Community model [63], Author-Topic model [88] and Community-User-Topic model [119] are other variations of latent Dirichlet allocation (LDA), which are also proposed to identify user communities.

Another class of work attempts to transform the content-based community detection problem into a graph clustering problem [79, 52, 65] where a user distance matrix is computed according to the similarity of their topical interests. The distance matrix is then used to identify clusters of users. The work by Peng et al. [79] is an instance of such techniques that focuses on identifying user communities on SINA Weibo by hierarchically clustering

users based on their relation to the predefined categories available on this social networking platform. Liu et al. [65] have proposed a clustering algorithm based on topic-distance between users to detect topic-based communities in a social tagging network. In this work, LDA is used to extract hidden topics in tags. Huang et al. [52] have built a pairwise similarity matrix for users based on the shortest path on the users' retweet graph. A spectral clustering algorithm has been used to find user communities in order to identify influential users and topical changes in the face of natural disasters. Barbieri et al. [10], however, proposed a network-oblivious probabilistic framework based on stochastic diffusion processes to identify like-minded users. They argue that users adopt topics of interest based on underlying diffusion processes over an unobserved social graph where the diffusion process itself is based on community-level influence.

2.3.3 Using Link and Content Information Jointly

Neither link nor content information alone is satisfactory for accurately determining communities; link information is usually sparse and noisy and often results in a poor partition of networks, and the irrelevant content could significantly mislead the process of community detection [3, 113]. It is therefore important to combine link analysis and content analysis for community detection in social networks. Several lines of work have been proposed to combine link and content information for community detection.

In one line of work, the approaches adopt a generative framework where a

generative link model and a generative content model are combined through a set of shared hidden variables of community memberships [115, 93, 75, 31, 44]. For instance, the work by Yin et al. [115] introduces a generative model which recursively defines a community through an integrated relationship between users' social relation and social topics. Simply put, in this model, communities are formed around multiple correlated topics where each topic can be reused in several communities. Similarly, Sachan et al. [93] also propose generative models for community detection but differently from the work by Yin et al., they consider three types of information, namely topics, social connections and interaction types such as retweeting and replying. In both approaches, a user can be a member of different communities but with varying degrees of membership. Erosheva et al. [35] combine LDA with LDA-Link for network analysis, referred to as the LDA-Link-Word model in this paper. Nallapti et al. [75] combine the mixed membership stochastic block model with LDA, and extend the LDA-Link-Word model by separating the citing documents and cited documents with the LDA-Link-Word model on the citing documents and PLSA model on the cited documents. Other approaches that exploit LDA for combining link and content analysis include [31, 44]. One major problem with these approaches is that they apply a generative model for content analysis, which makes them vulnerable to irrelevant keywords. Differently from generative models, Yang et al. [113] have proposed a non-generative probabilistic model to find user communities in citation networks. They estimate the conditional probability that a user is cited given her

popularity and her membership to a community according to her weighted content vector (topics of interest) so that modeling the absent links, as in generative models, is avoided.

In addition to probabilistic models, some other approaches that have been proposed to combine link and content information include matrix factorization [17, 121] and kernel fusion [71] and graph union [92] for spectral clustering.

Graph representation learning methods, which have been recognized as effective in link-based user community detection, have also been extended to incorporate the rich information associated with nodes (users) in addition to links. For instance, in order to take content into account, Yang et al. have extended DeepWalk (TADW) [109] under the framework of matrix factorization. Liu et al. [66] have employed a bi-directional long short-term memory (LSTM) encoder to fuse the content and links by random walks on the social network. Use of LSTM is motivated since it is able to capture long-range structural information by memories of LSTM. This is in contrast to DeepWalk or node2vec's where limited neighborhood scope has been considered by first and/or the second order random walks.

2.3.4 Temporal Analysis

All the above methods do not incorporate *i*) temporal aspects of users' topics of interests, and/or *ii*) dynamics of social links and undermine the fact that users of communities would ideally show similar contribution or interest

patterns for similar topics and/or similar network neighborhood evolution throughout time. As a matter of fact, many content based and link based methods assume that the structure of the network and the topics discussed by the users remain stable over time, which can be a limiting assumption in practice. While a myriad of work has addressed the dynamics of social links in user community detection, i.e., dynamic community detection [89, 107], there have been a few works that have considered temporal aspects of users' topics of interests as an explicit dimension when identifying user communities in social networks [50, 51, 64].

The work by Hu et al. [50] is among the pioneers to consider temporality in user-generated contents through a generative process, which models how users and topics are related to each other and co-evolve over time. Their model, namely Group Specific Topic over Time (GrosToT), learns a specific time-aware probability distribution known as the community-topic-time distribution addressing how communities and topics are associated with each other over time. Assuming the number of topics K and the number of communities C are known in advance, the model associates Dirichlet distributions for topics over words, communities over users, and topics over communities with different parameters, respectively. Also, a Dirichlet distribution for time is assigned given topic-community pairs. A user is a member of a community according to the assigned community-user distribution. Her tweet is generated based on multinomial distribution over, first, topic-community distribution to select topics and then topic-word distribution to select words.

The timestamp of tweet is obtained by the multinomial distribution of time-topic-community distribution. As seen, the model is based on the idea that there is a tight interrelation between communities and topics. This prevents integration of other topic detection methods for the task of community detection.

Similarly, Liang et al. [64] have proposed a content-based dynamic user community detection based on dynamic multinomial Dirichlet distribution. The model tracks changes of each user's time-varying topic distributions based both on the short texts the user posts during a given time period and on previously estimated distributions. The model can be used in two modes: *i*) as a short-term dependency model that infers a user's current topic distribution based on the user's topic distributions during the previous time period only, and *ii*) as a long-term dependency model that infers a user's current topic distribution based on the user's topic distribution during multiple time periods in the past.

In this thesis, we follow the same underlying hypothesis related to topics and temporality as by Hu et al. and Liang et al., i.e., the evolution of user-generated content is dynamic over time (temporal) and users' interests can evolve over different time intervals. In contrast to dynamic community detection methods, however, we assume that the social network structure is static and remains stable over time. The main reason for this assumption is that the social network structure has a significantly lower pace of change compared to how fast content is generated over time and distributed across

the online social network [74]. In Chapter 3, we model users based on a multivariate time series representation where each of the time series depict to what extent the user has contributed to social topics in consecutive time intervals. This time series representation allows one to compute a user similarity matrix for the users based on the cross-correlation similarity of users' time series, which can then be effectively used to extract clusters of users.

None of the proposed neural embeddings take the time dimension into consideration for the task of user community detection either. Although Benton et al. [12] offer the opportunity to integrate different information types, it is not clear how to integrate temporality, which can be considered to be an aspect, rather than new information type. In Chapter 4, we propose a neural embedding approach to model the users' temporal contribution towards topics of interest by introducing the notion of similarity regions between users. These regions cover users who share not only similar topical interests but also similar temporal behaviour. By considering the identified set of regions as a context, we train a neural network such that the probability of a user in a region is to be maximized given other users in the same region.

Another line of work has employed temporal aspects of users' topics of interests and/or dynamics of social links in order to determine how the user communities of a social network will look like in a future yet-to-be-observed time interval, i.e., *community prediction*. Although considerable research has been devoted to link-based user community prediction [120, 34], rather less attention has been paid on content-based or topical future community

prediction. Appel et al. [5] have employed shared matrix factorization in order to factor links, content, and temporal dimension simultaneously. This approach embeds links and content into a shared latent space at each time interval while taking the temporal continuity into account by using the embedding as a surrogate. The temporal sequence of embeddings is then can be utilized by autoregressive models in order to predict future user communities. Regression techniques such as autoregressive integrated moving average (arima) and support vector regression (svr) that leverage temporal information to predict users' future interests have shown promising results and can be employed to identify user communities in the future based on pairwise content similarity among users [6]. However, they require building predictive models per user and, hence, are computationally expensive.

In Chapter 5, we propose temporal latent space modeling for content-based user networks in order to predict user communities in the future. First, in contrast to link-based community prediction methods like Zhu et al.'s method [120] and the likes that focus on social network structure, our approach employs social content. Second, although we use temporal information to predict future users' topical interests similar to regression methods, we train only one model for all users and, thus, significantly reduce computational cost. Third, contrary to Hu et al. [50] who use users' temporal and topical interests for pairwise similarity to identify user communities up until this month, our task uses such information for predicting user communities in the future which is a step forward compared to the state of the art.

2.4 Summary

In this chapter, we have reviewed concepts in the area of user community detection and prediction. We introduced the definition of user community as well as a short history of user community detection. Further, we reviewed the different community detection methods in social networks. The literature on user community detection is broad and an exhaustive survey of community detection algorithms is beyond the scope of this thesis. However, we provided a systematic view of the closely related methods to this thesis and highlighted the distinguishing aspects of this thesis compared to the state of the art.

2.5 Related Publications

- Hossein Fani, and Ebrahim Bagheri. "Community detection in social networks." *Encyclopedia with Semantic Computing and Robotic Intelligence* 1.01 (2017): 1630001.

Chapter 3

Multivariate User Time Series

User communities in social networks are often identified by considering explicit social connections between users. While such communities can reveal important information about their members such as family or friendship ties and geographical proximity, just to name a few, they do not necessarily succeed at pulling like-minded users that share the same interests together. Therefore, researchers have explored the topical similarity of social content to build like-minded communities of users. In this chapter, following topic-based approaches, we are interested in identifying communities of users that share similar topical interests with similar *temporal* behaviour. More specifically, we tackle the problem of identifying temporal (diachronic) topic-based communities, i.e., communities of users who have a similar temporal inclination towards emerging topics. To do so, we utilize multivariate time series analysis to model the temporal inclination of each user towards emerging

topics.

We capture the users' overall mindset based on the textual content they engage with in online social networks, e.g. tweets on Twitter, using topic modeling approaches. This is motivated by topic modeling approaches which are unsupervised methods for identifying topics from text corpora. Further, our approach is agnostic to the topic detection method. We extract topics of interest by employing seminal topic detection methods, one graph-based and two LDA-based methods.

Each user's topics of interest at each time interval, e.g., day, are based on the degree of her contribution toward each identified topic. A multivariate time series for each individual represents her topics of interest in different time intervals. The inter-user similarity in this proposed representation is then measured by cross-correlation, which has been already explored within the signal processing community for measuring the similarity of a pair of time series. Last, based on the pairwise user similarities, a weighted undirected graph is built, on which the Louvain method, a heuristic graph partitioning algorithm based on modularity optimization, yields our final user communities.

The concrete contributions of our work in this chapter are as follows:

1. We formally represent a user within a temporal-topic space through the use of multivariate time series. The proposed user representation effectively incorporates users' contributions towards the topics over time and is able to seamlessly integrate any topic detection methods and is,

therefore, agnostic to the underlying topic detection method.

2. We show how time series analysis techniques can be used to measure the similarity of pairs of users. This notion of similarity is further used to build a graph of user relations, not based on the users' social interactions, but rather based on their disposition towards similar topics in similar time intervals.

3. We propose a graph representation of user interactions composed from their temporal and topical similarity and demonstrate how graph clustering can be used to identify user communities that consider both temporality and topical similarity when grouping users.

The proposed method in this chapter addresses research questions **RQ1**, i.e., whether the consideration of time plays a role in the quality of the identified communities and **RQ2**, i.e., whether temporal content-based user community detection methods show better performance compared to link-based methods.

3.1 Problem Statement

In our work, we aim at identifying latent temporal communities of users within a specific time period T, based on the temporal inclination of the users towards topics. We incorporate temporal aspects of users' interests and consider the fact that users of like-minded communities would ideally

show similar contribution or interest patterns for similar topics throughout time.

Problem Definition. Given a set of users $\mathbb{U}$, we aim to partition $\mathbb{U}$ into non-overlapping subsets in which each $u \in \mathbb{U}$ is only a member of one subset. More formally, $\mathbb{P} = \{\mathbb{C} : \mathbb{C} \subseteq \mathbb{U}, |\mathbb{C}| > 1\}$ such that $\forall \mathbb{C}_i, \mathbb{C}_{j \neq i} \in \mathbb{P} : \mathbb{C}_i \cap \mathbb{C}_j = \varnothing$. The objective of our work is to identify a configuration for $\mathbb{P}$ such that members of each $\mathbb{C}_i$ in $\mathbb{P}$ show highly similar temporal disposition with regards to active topics on the social network and high dissimilarity with members of any other $\mathbb{C}_{j \neq i} \in \mathbb{P}$.

We divide this problem into two subproblems: *topic detection* and *community detection* in which the output of the first subproblem becomes the input of the second one. We concretely formulate these subproblems and propose our approach in the following.

3.2 Topic Detection

Our proposed community detection method is able to seamlessly integrate any topic detection methods and is, therefore, agnostic to the underlying topic detection method. Hence, the focus of our work in this subproblem is not to propose a new topic detection method but rather to provide a common interface to the existing topic detection techniques for the purpose of temporal topic-based community detection. We highlight this by customizing one graph-based and two probabilistic LDA-based approaches in our work,

as alternatives, to extract topics from documents. Foremost, we introduce the required preliminary definitions.

We view all textual content of a user u generated at time interval t, denoted by $m_{u,t} \subseteq \mathbb{M}$, as a single document. A document m is a vector of N nonnegative integers, where the i-th number shows the occurrence frequency of the i-th term. N is the number of the unique terms in $\mathbb{M}$. A topic z is a vector of N real numbers in [0,1], summing to 1, whose i-th number shows the participation score of the i-th term in forming that topic. Collectively, $\mathbb{Z} = \{z \in [0,1]^{\mathrm{N}} : ||z||_1 = 1\}$ is the set of all topics. Topic distribution of a document is a function $\tau : \mathbb{M} \rightarrow [0,1]^{|\mathbb{Z}|}; \forall m \subseteq \mathbb{M}, ||\tau(m)||_1 = 1$. Intuitively, τ maps a document to a set of topics where $\tau(m)_z$ is the score of topic z for document m.

In the topic detection subproblem, given $\mathbb{M}$ as input, we aim at identifying $\mathbb{Z}$, i.e. the topics formed in the documents posted in time period T which is possible using various existing methods in the literature including topic detection methods introduced in [32, 14, 116, 104].

3.2.1 Graph-based approach (GbT)

According to [104], one can utilize signal processing techniques to detect emerging topics. The fundamental hypothesis behind this topic detection method is that those terms that have correlated frequency within time could be considered to be conceptually related and can, therefore, collectively form a topic. To apply this approach, for any term $w \in \mathbb{W}$ a *term signal* is

constructed. Simply, the term signal shows the number of times the term has been mentioned across all documents in different time intervals of time period T. More specifically, a term signal for term w is a temporally ordered set of integer values, expressed as $\mathbf{X}_w = (x_{w,1}, x_{w,2}, ..., x_{w,\mathrm{T}})$, from discrete observations of term frequencies at T consecutive time intervals, such that $x_{w,t}$ represents the occurrence number of the term w in all documents posted at time interval t.

We can calculate the similarity of two terms v and w, denoted by $d_{\mathbb{W}}(v, w)$, based on the cross-correlation of their term signals as follows:

$$d_{\mathbb{W}}(v, w) = \mathbf{X}_v \star \mathbf{X}_w = \sum_{t=1}^{\mathrm{T}} (\mathbf{X}_v)^*[t]\ \mathbf{X}_w[t] \tag{3.1}$$

where $\mathbf{X}$ represents term signal, $\star$ is the measure of cross-correlation between two term signals, and $(\mathbf{X})^*$ is the complex conjugate of $\mathbf{X}$. Based on this, an undirected weighted term graph $G_{\mathbb{W}} = (\mathbb{V}, \mathbb{E}, g)$ can be formulated such that $\mathbb{V} = \mathbb{W}$, $\mathbb{E} = \{e_{v,w} : \forall v, w \in \mathbb{W}\}$ and the weight function $g : \mathbb{E} \to \mathfrak{R}$ is defined as $g(e_{v,w}) = d_{\mathbb{W}}(v, w)$.

When the graph is constructed, graph partitioning algorithms such as the Louvain Method (LM) [15] can be used to identify highly cohesive subgraphs [104]. Each subgraph represents an emerging topic on the text corpus at a given time period T. Here, each topic z is an induced subgraph G_z of $G_{\mathbb{W}}$ such that $\mathbb{V}_z \subseteq \mathbb{W}$, G_z consists of all the edges of $G_{\mathbb{W}}$ with incident vertices in $\mathbb{V}_z$, and $|\mathbb{V}_z| > 1$.

In accordance with our definition of topic $z \in \mathbb{Z}$, we vectorize G_z to N real numbers, summing to 1. To do so, for $1 \leq i \leq \mathrm{N}$, we define the i^{th} number as the degree centrality of the term w if $w \in \mathbb{V}_z$ and 0 otherwise. Also, we normalize the result by its L^1-norm. Finally, we define topic distribution function $\tau(m)[z] = m \cdot z$ where m is a document, $\cdot$ is the vector dot product, and $z \in \mathbb{Z}$.

3.2.2 LDA-based approaches

LDA assumes that a document is a mixture of topics and implicitly exploits co-occurrence patterns of terms to extract sets of correlated terms as topics of a text corpus [14]. Similar to [49, 105], we see all terms extracted from documents of a user u for each time interval t, i.e., $m_{u,t}$, as a single document $m \in \mathbb{M}$. As another LDA-based approach, we use the Topics over Time (ToT) model [101] which simultaneously captures term co-occurrences and locality of those patterns over time and is hence able to discover more event-specific topics. In both LDA and ToT, $z \in \mathbb{Z}$ is the multinomial distribution of terms specific to topic z and the topic distribution function τ is defined as a Dirichlet distribution with parameter $\boldsymbol{\alpha}$; notationally, $\tau(m) \sim Dir(\boldsymbol{\alpha})$.

After detecting topics $\mathbb{Z}$ from a given document collection $\mathbb{M}$ within a specific time period T and defining topic distribution function τ using one the above topic detection methods, our next goal is to identify communities of users formed on the basis of their temporal relation to the identified topics.

3.3 User Community Detection

We represent the degree of contribution of a user to each topic $z \in \mathbb{Z}$ over multiple time intervals as a vector. Collectively, this forms a multivariate time series for each user u towards all topics in $\mathbb{Z}$, which we refer to as the *user-topic contribution time series*. We calculate the pairwise similarity between two users by computing the similarity between their corresponding user-topic contribution time series. Based on these calculated similarities, we aim at calculating $\mathbb{P}$. However, this would be considered to be a graph partitioning problem which is NP-hard. Thus, we build a weighted graph of users and apply Louvain's heuristic in graph partitioning to detect user communities. Our approach for identifying temporal topic-based communities includes three steps: user representation, user similarity calculation, and user community identification, which are described in details as follows.

3.3.1 User Representation

We model each user's topics of interest and temporal inclination towards the topics through *user-topic contribution time series*. Formally, the user-topic contribution time series of user u for topic set $\mathbb{Z}$ is a temporally ordered vectors of real values in T consecutive time intervals, expressed as $\mathbf{X}_u = (\mathbf{x}_{u,1}, \mathbf{x}_{u,2}, ..., \mathbf{x}_{u,\mathrm{T}})$. At each time interval t, $\mathbf{x}_{u,t}$ is a vector whose elements $x_{uz,t} \in \mathfrak{R}^{[0,1]}$ show the degree of interest for the user u towards the topic z. Assuming there are K topics detected, $\mathbf{x}_{u,t}$ becomes a K-tuple vector and the

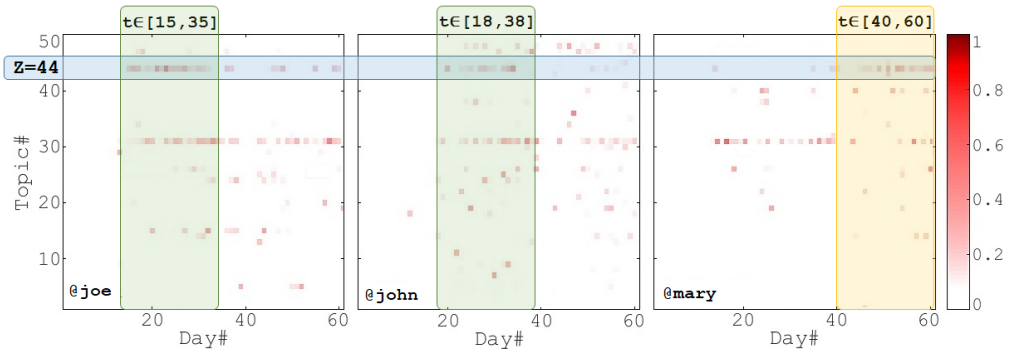

Figure 3.1: The heatmap for the user-topic contribution time series for the three sample Twitter users from November till end of December 2010.

user-topic contribution time series will be a K-variate time series.

It is possible to visualize the topic preference time series of each user by projecting it onto a heatmap, which has been done in Figure 3.1 for the three sample users introduced earlier in Figure 1.1. In this figure, the y-axis represents the topic indices, the x-axis denotes the time intervals, and the cell values show the degree of contribution of the user to that topic. As shown, topic 44 (highlighted horizontally with blue) represents the *'War in Afghanistan'* topic while topic 30 refers to the *'New Year'* topic. As seen in the projection, all three users have shown consistent interest in topic 30 and have started talking about the *'New Year'* topic starting from late November. However, their temporal interest pattern with regards to topic 44 is not as consistent and while `@joe` and `@john` are heavily engaged with this topic in November (as highlighted with the vertical green column), `@mary` only becomes involved with the topic in late December (specified with an orange column on the right most figure of Figure 3.1).

The user-topic contribution time series can be considered to be a good measure for finding the similarity between two users according to our definition of the latent user community. It allows finding like-minded users based on their *temporally*-correlated contributions on similar topics. Based on Figure 3.1, *non*-temporal topic-based approaches group the three users, namely `@joe`, `@john`, and `@mary`, in the same community and consider them like-minded, because they are interested in the same topic, i.e., z_{44}. However, the user `@mary` can be considered to be dissimilar from the other two because the period of time during which she reacts to z_{44} is not the same.

3.3.2 User Similarity

In order to find the similarity of a pair of users, we compute the similarity of their corresponding user-topic contribution time series. For this purpose, we employ the 2-dimensional variation of the cross-correlation measure. The 2-dimensional cross-correlation measure of two matrices $\mathbf{A}_{[\mathrm{C}\times\mathrm{D}]}$ and $\mathbf{B}_{[\mathrm{C}\times\mathrm{D}]}$, denoted by $\mathrm{XC}_{[(2\mathrm{C}-1)\times(2\mathrm{D}-1)]}$, is calculated as follows:

$$\mathrm{XC}[i,j](\mathbf{A},\mathbf{B}) = \sum_{c=0}^{\mathrm{C}-1}\sum_{d=0}^{\mathrm{D}-1} \mathbf{A}[c,d]\ \mathbf{B}^*[c-i,d-j] \tag{3.2}$$

where $\mathbf{B}^*$ denotes the complex conjugate of $\mathbf{B}$. Intuitively, the 2-dimensional cross-correlation slides one matrix over the other and sums up the multiplications of the overlapping elements. A positive row index i corresponds to a downward shift of the rows of $\mathbf{A}$ over $\mathbf{B}$ and a negative column index

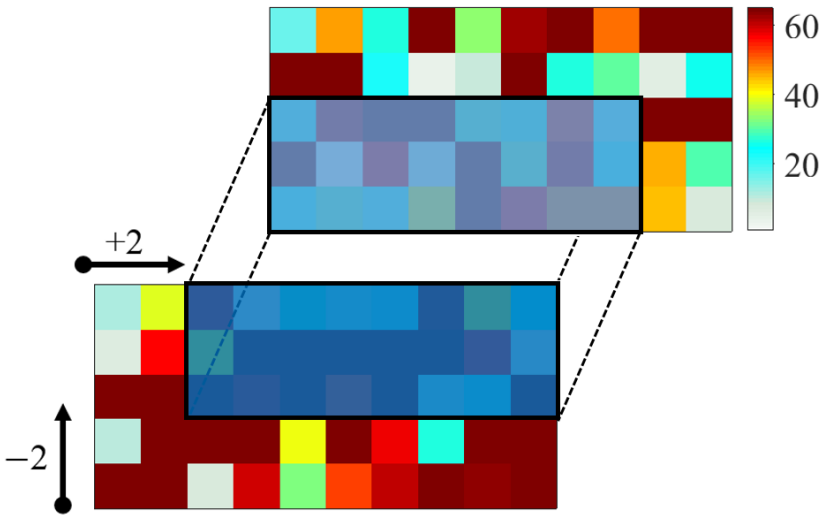

Figure 3.2: 2D cross-correlation in XC$[-2, 2]$.

j indicates a leftward shift of the columns. To make it clearer, Figure 3.2 illustrates how XC$[-2, 2]$ is calculated in two 5×10 sample matrices. A maximum correlation occurs at XC$[0, 0]$ if the two signals are similar without any time shift. We use the normalized value of XC$[0, 0]$ in $\Re^{[0,1]}$ when calculating user similarity distances.

We can represent user-topic contribution time series with respect to K number of topics in $\mathbb{Z}$ in T consecutive time intervals as a K×T matrix. Then, the similarity of two users u and v, denoted as $d_{\mathbb{U}}(u, v)$, can be defined based on the 2-dimensional cross-correlation of their user-topic contribution time series with no shift $(i = j = 0)$, as follows:

$$d_{\mathbb{U}}(u, v) = \frac{\text{XC}[0, 0](\mathbf{X}_u, \mathbf{X}_v)}{\sqrt{(\mathbf{X}_u \cdot \mathbf{X}_u)(\mathbf{X}_v \cdot \mathbf{X}_v)}} \tag{3.3}$$

where $\mathbf{X}_u$ is the user-topic contribution time series for user u.

We are now able to calculate the similarity between all pairs of users and group similar users that share similar temporal exposition towards similar topics of interest.

3.3.3 User Community

We identify user communities through graph-based partitioning heuristics. We represent users and their pairwise similarity through a weighted undirected graph. Precisely, let $G_{\mathbb{U}} = (\mathbb{V}, \mathbb{E}, g)$ be a weighted user graph in time period T such that $\mathbb{V} = \mathbb{U}$, $\mathbb{E} = \{e_{u,v} : \forall u, v \in \mathbb{U}\}$ and the weight function $g : \mathbb{E} \rightarrow \mathfrak{R}$ is defined as $g(e_{u,v}) = d_{\mathbb{U}}(u, v)$. After constructing the user graph $G_{\mathbb{U}}$ for a given time period T, it is possible to employ a graph partitioning heuristic to extract partitions of users that form latent communities. As in graph-based topic detection, we leverage the Louvain Method (LM). The Louvain Method is suitable for its following characteristics: *i*) this algorithm can be applied to weighted graphs, *ii*) it does not require *a priori* knowledge of the number of partitions when running the algorithm, and *iii*) it is computationally very efficient when applied to large and dense graphs [90]. While modularity maximization is NP-hard, the complexity of LM's greedy implementation is $O(nlogn)$, where n is the number of vertices [15, 90]. Here, the output is a set of induced subgraphs of $G_{\mathbb{U}}$ representing temporal user communities $\mathbb{P}$ that consist of like-minded users who have contributed to the same topics with the same temporal behaviour and contribution degrees.

3.4 Summary

In this chapter, we have proposed an approach to detect communities of like-minded users who share topics of interest with similar temporal behaviour. The approach addresses research questions **RQ1**, i.e., whether the consideration of time plays a role in the quality of the identified communities and **RQ2**, i.e., whether temporal content-based user community detection methods show better performance compared to link-based methods. We model the contribution of each user towards topics using multivariate time series and apply 2-dimensional cross-correlation on all pairs of such time series to find similar users in topics of interest and temporal behaviour. We employ Louvain clustering, a heuristic graph partitioning algorithm based on modularity optimization, to extract our final user communities. To find topics from the social network, we used state-of-the-art topic detection methods with different approaches, as alternatives, in order to show that our approach and its contribution are independent of topic detection algorithms. We used one graph-based and two probabilistic LDA and ToT methods.

3.5 Related Publications

- Hossein Fani, Fattane Zarrinkalam, Ebrahim Bagheri, and WeichangDu, "Time-sensitive topic-based communities on twitter", *Advances in Artificial Intelligence - 29th Canadian Conference on Artificial Intelligence,Canadian AI 2016, Victoria, BC, Canada, May 31 - June 3,*

2016. Proceedings, 2016, pp. 192-204

- Hossein Fani, Ebrahim Bagheri, Fattane Zarrinkalam, Xin Zhao, and Weichang Du. "Finding Diachronic Like-Minded Users." *Computational Intelligence* 34.1 (2018): 124-144.

Chapter 4

Neural User Embeddings

The multivariate user time series method which has been proposed in the previous chapter falls short due to sparsity in the topic space. Indeed, users are interested in very few topics instead of all identified topics. The cross-correlation measure calculates the overall similarity of a given pair of time series based on all topic-time entries including zero entries. Therefore, for instance, a user who has similar temporal interest towards the topic *'War in Afghanistan'* as the other two users `@joe` and `@john` but contributes to no other topics would not join `@joe`, `@john` in the same community. Moreover, this approach is not able to seamlessly incorporate social structure into temporal and topical analysis (links jointly with content).

While we follow the same assumption about temporality in like-minded user community detection, we introduce an alternative time-aware topic-driven method to address these two shortcomings. Inspired by Mikolov et

al.'s word2vec in computational linguistics [69], we proposed distributional representation of users (user embeddings).

Basically, the premise of user embeddings is that similar users should have similar embeddings (or equivalently close points in embedding space). As the main objective of our work is to identify like-minded user communities whose members exhibit temporally similar behaviour toward similar topics of interest, we would like to embed those users who are interested in similar topics at a certain point in time close to each other, and distant from those who have similar interest towards the same topics but in different time intervals. Our proposed temporal topic-driven user embedding model represents a step forward in this respect compared to our proposed method that is based on time series analysis.

We build documents whose elements are users, not words. We extend the concept of co-occurrence of words in documents to users such that two users co-occur if they show the same interest toward the same topics in similar time intervals. The key contribution of this method is to learn user vector representations from users' topics of interest with the expectation that temporally like-minded users end up closer to each other in the vector space. We hypothesize that an appropriate embedding method would bring significant performance into our main downstream task of like-minded user community detection compared to the state of the art. To build user embeddings, we first formally formulate what we mean by a like-minded pair of users. Then, we propose an embedding method which preserves pairwise like-minded prox-

imity of the users through maximizing the likelihood that two like-minded users stay close to each other in vector space.

Similar to the multivariate user time series method, the proposed neural user embeddings are to address research questions **RQ1**, i.e., whether the consideration of time plays a role in the quality of the identified communities and **RQ2**, i.e., whether temporal content-based user community detection methods show better performance compared to link-based methods.

Our work in this chapter moves beyond the proposed neural user embeddings and is extended to interpolating users' social network connections (links) as well as users' interests over time (temporal content). Earlier *non*-temporal user community detection methods have already shown improvement when incorporating social network structure (links) with topics of interest (content) compared to those in which links and content are used separately [93, 113]. However, to the best of our knowledge, existing *temporal* user community detection methods are only content-based and hence do not study the effect of social network structure and temporal evolution of user content simultaneously. Our experiments show that while social network structure is not a discriminative enough feature on its own for identifying high quality user communities, it does improve the quality of the identified user communities when effectively interpolated with temporal contents. It is worth noting that the social network structure is assumed to be static and remains stable over time in our work. The main reason for this assumption is that the social network structure has a significantly lower pace of change

compared to how fast content is generated over time and distributed across the social network [74].

In order to simultaneously consider users' temporal content and their social network structure when identifying user communities, we embed both users' temporal interests and their social network structure into a dense vector representation using neural embedding mechanisms. The user embeddings, which are derived from two different information sources (modalities), i.e., *i*) temporal content-based embeddings based on users' topics of interest over time, and *ii*) network embeddings based on social network neighborhoods, are linearly interpolated to build a single final *multimodal* user embedding. The linear interpolation of two user embeddings at the embeddings level allows us to investigate how and to what extent users' dynamic topics of interest and/or users' social network structure contribute to the quality of the inferred user communities. We perform experiments on Twitter data and evaluate our work in two application scenarios: news recommendation and user prediction, to explore the impact of the different user embeddings and their interpolation.

Beyond **RQ1** and **RQ2**, this chapter also addressed **RQ3**, i.e., whether link-based and temporal content-based community detection methods have synergistic effect on each other.

In summary, the main contributions of this chapter are as follows:

1. We propose a community detection method that considers users' topical interests and their temporal evolution in tandem by learning neural user

representations, which embeds users in an embedding space where those users who have similar inclination towards similar topics in similar time intervals will be embedded close to each other.

2. We employ neural graph embedding techniques to embed information from users' social network structure into user representations.

3. We build a single set of multimodal embeddings from embeddings of temporal social content and social network structure through their linear interpolation in order to elucidate the contribution of users' temporal content on the one hand, and social network structure, on the other hand, for finding user communities.

4. We identify temporal content-based user communities which are topically, temporally and structurally cohesive, based on our multimodal user embeddings.

4.1 Approach Overview

Having formally laid out the problem in Section 3.1, i.e., given a set of users $\mathbb{U}$, partitioning of $\mathbb{U}$, denoted by $\mathbb{P}$, into non-overlapping subsets $\mathbb{C}$ is desired in which each $u \in \mathbb{U}$ is only a member of one subset such that $\mathbb{P} = \{\mathbb{C} : \mathbb{C} \subseteq \mathbb{U}, |\mathbb{C}| > 1\}$; $\forall \mathbb{C}_i, \mathbb{C}_{j \neq i} \in \mathbb{P} : \mathbb{C}_i \cap \mathbb{C}_j = \varnothing$ and members of each $\mathbb{C}_i$ in $\mathbb{P}$ show highly similar temporal disposition with regards to active topics on the social network and high dissimilarity with members of any other $\mathbb{C}_{j \neq i} \in \mathbb{P}$, we seek

Algorithm 1 Overview of the proposed approach to find user communities

Inputs:

$\mathbb{U}$, the set of users;
$\mathcal{D} = (\mathbb{U}, \mathbb{M}, \mathrm{T})$, temporal social content;
$\mathcal{G} = (\mathbb{U}, \mathbb{A})$, the social network;

Output:

$\mathbb{P} = \{\mathbb{C} : \mathbb{C} \subseteq \mathbb{U}, |\mathbb{C}| > 1\}$ such that $\forall \mathbb{C}_i, \mathbb{C}_{j \neq i} \in \mathbb{P} : \mathbb{C}_i \cap \mathbb{C}_j = \varnothing$

1: **parallel_exec:** //User representation learning - parallel execution
2: $\quad \mathbf{W}_{\mathcal{D}} = f(\mathcal{D})$; //Temporal content-based user embeddings, § 4.2.
3: $\quad \mathbf{W}_{\mathcal{G}} = g(\mathcal{G})$; //Link-based user embeddings, § 4.3.
4: $\mathbf{W} = h(\mathbf{W}_{\mathcal{D}}, \mathbf{W}_{\mathcal{G}})\{\textbf{return } \alpha\mathbf{W}_{\mathcal{D}} + (1-\alpha)\mathbf{W}_{\mathcal{G}};\}$; //Interpolation, § 4.4.
5: $\mathbb{P} = \mathrm{Cluster}(\mathbb{U}, \mathbf{W})$ //User community detection, § 4.5.

to find $\mathbb{P}$ through three pipelined phases: 1) temporal content-based and topological user representation learning (Sections 4.2 and 4.3 respectively), 2) interpolation of user embeddings (Section 4.4), and 3) user community detection (Section 4.5). Foremost, we provide an overview of this process after which the details of each step will be presented.

The overview of the approach discussed in this chapter to find user communities is outlined in Algorithm 1. We define temporal social content as $\mathcal{D} = (\mathbb{U}, \mathbb{M}, \mathrm{T})$ where $\mathbb{U}$ is the user set, $\mathbb{M}$ is the textual user-generated content corpus (e.g., tweets), and T is the time period broken down into time intervals. We define the social network structure as a directed graph $\mathcal{G} = (\mathbb{U}, \mathbb{A})$ whose vertices are users in $\mathbb{U}$ and edges are ordered pairs of user elements such as $(u, v) \in \mathbb{A}$ indicating a social tie from u to v (e.g., u is following v).

Our proposed approach consists of creating user representations from two

different information sources (modalities), i.e. 1) temporal content-based embeddings from temporal social content $\mathcal{D} = (\mathbb{U}, \mathbb{M}, \mathrm{T})$, and 2) link-based embeddings from the social network structure $\mathcal{G} = (\mathbb{U}, \mathbb{A})$. On Line 2 of Algorithm 1, we learn user vector representations $\mathbf{W}_{\mathcal{D}}$ from users' content with the expectation that temporally like-minded users end up closer to each other in the vector space. To build this type of user embeddings, we first formally formulate what we mean by a like-minded pair of users with respect to social content only. Then, we propose a representation learning method, which preserves pairwise proximity of the users through maximizing the likelihood that two like-minded users stay close to each other in vector space. Likewise, on Line 3, we learn user vector representations $\mathbf{W}_{\mathcal{G}}$ but from users' social network neighborhood with the assumption that similar users are those that are densely connected to each other due to homophily. We use unsupervised random-walk based graph representation learning to learn user representations such that geometric relationships in the learned vector space reflect the structure of the original social network. Learning vector representations from temporal social content and social network structure are independent and could be run in parallel (Line 1). These monomodal user representations are then linearly interpolated into a single consolidated multimodal representation on Line 4 tailored for the task of user community detection on Line 5.

4.2 Temporal Content-based User Embeddings

In order to learn temporal content-based neural embeddings ($\mathbf{W}_{\mathcal{D}}$) for social network users, we consider social content to be in the form of a triple $\mathcal{D} = (\mathbb{U}, \mathbb{M}, \mathrm{T})$ where $\mathbb{U}$ is the set of users, $\mathbb{M}$ is the collection of content generated by $\mathbb{U}$ and T is the number of consecutive time intervals. We identify a set of topics $\mathbb{Z}$ from $\mathbb{M}$ over the T time intervals as we did in Section 3.2 and build user-topic contribution time series for each user accordingly as in Section 3.3.1. The user-topic contribution time series of each user $u \in \mathbb{U}$ towards a set of identified topics $\mathbb{Z}$ over time intervals $1 \leq t \leq \mathrm{T}$ is a K-variate time series $\mathbf{X}_u = (\mathbf{x}_{u,1}, \mathbf{x}_{u,2}, ..., \mathbf{x}_{u,\mathrm{T}})$ where $\mathbf{x}_{u,t}$ is a vector of elements $\mathbf{x}_{uz,t} \in \Re^{[0,1]}$ showing the degree of interest for the user u towards the topic z at time interval t. The stacking of all users' topic preference time series will generate a cuboid $\mathbf{X} = \{x_{uz,t} : u \in \mathbb{U}, z \in \mathbb{Z}, 1 \leq t \leq \mathrm{T}\}$.

4.2.1 Temporal Context Model

One key novelty of this work is to learn user vector representations from users' content with the expectation that temporally like-minded users end up closer to each other in the vector space. To build user embeddings, we first formally formulate what we mean by a like-minded pair of users with respect to social content only within time. Then, we propose an embedding method which preserves pairwise like-minded proximity of the users through maximizing the likelihood that two like-minded users stay close to each other

in vector space.

The premise of our approach is that the more two users share common interests in similar time intervals, the more similar these users would be and hence the likelihood of these users being in the same community should increase. As an example, let us consider the same three users that were introduced in Figures 1.1 and 3.1 earlier. Figure 4.1 shows a subset of the topic preference time series of these three users for a 10 day time period for a limited set of topics. An interesting observation is that while the visualization of the users' topic preference time series based on a heatmap in Figure 3.1 showed us that users `@joe` and `@john` share similar temporal interests, which is different from `@mary`, it becomes clear that the actual degree of interest is not within the same range. For instance, even for the two users who are considered to be quite similar, their degrees of interest for Topic 44 are 0.35 and 0.14, respectively, which are quite different. This shows that it would be quite difficult to identify users that not only have similar temporal trends but also similar degrees of interest. For this reason, we relax the similarity condition to allow for cells with similarity values within a range to be considered to be similar. The softened condition of similarity is referred to as the condition of homogeneity. For the sake of clarifying the concept of condition of homogeneity, let us assume that any degree of interest below 0.1 is insignificant and can be ignored (shown in grey in Figure 4.1). Assuming the condition of homogeneity considers values above 0.1 to be similar, users `@joe` and `@john` will now share four regions of similarity in Figure 4.1. This

would not be possible without this relaxed condition. On the other hand, `@mary` still maintains its difference with the other two users with only one and zero regions of interests with the other two users. Based on the condition of homogeneity, we now consider `@joe` and `@john` to be similar as they share the many similar regions and `@mary` to be distant from them.

The condition of homogeneity and the number of shared regions between users allows us to formally define an objective function for learning user embeddings. Our objective function will endeavor to place those users who share many regions of similarity close to each other and far away from those users who do not share any regions of similarity with them. Expressed more formally, the shared regions between two users act as a context for the users when they are embedded into a neural embedding space. For instance, the four shared regions for `@joe` and `@john` act as context for each of the users and allows our embedding model to learn similar representations for these two users. In the following, we will propose a deterministic method for finding shared regions between any two users, which will be later used as context for learning user embedding representations. We first define the shared regions as follows:

Definition 1. ***Region of Like-mindedness.*** *A three-dimensional subspace of* $\boldsymbol{X}$*, such as* R*, is defined to be a region of like-mindedness iff (1) all the values in this subspace are equal with respect to a certain condition of homogeneity* c*; notationally,* $\forall x, x' \in R; c(x) = c(x')$ *and (2) it is maximal such that there exists no other regions of like-mindedness such as* R' *such*

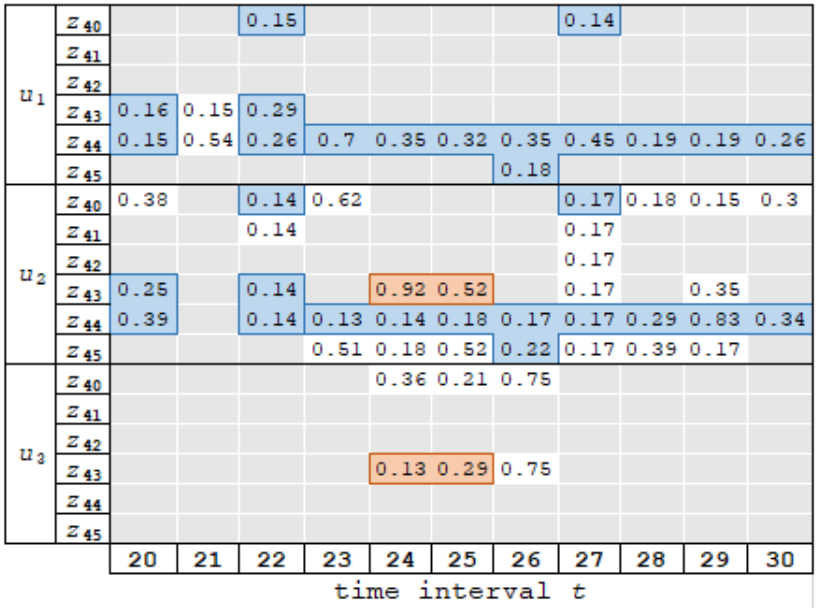

Figure 4.1: Topic preference time series for three sample Twitter users in Figure 3.1 with $\{u_1 = \texttt{@joe}, u_2 = \texttt{@john}, u_3 = \texttt{@mary}\} \times \{z_{40} \,..\, z_{45}\} \times t \in \{20 \,..\, 30\}$. The values are *unnormalized* probabilities for every topic in each document, most of which are equal to the smoothing parameter alpha ($\alpha = \frac{5.0}{|\mathbb{Z}|}$) in the LDA topic modeling method. Also, the values are rounded to two digit precision.

that R is subsumed by R'. The set of all regions of like-mindedness is called $\mathcal{R}$.

We adopt a similar strategy to [118] to find the set of all regions of like-mindedness $\mathcal{R}$ in $\mathbf{X}$. First, we find $\mathcal{R}$ in user and topic dimensions at each time interval t. The output is two-dimensional (2-d) regions indexed by time interval $1 \leq t \leq \mathrm{T}$, i.e., $\mathcal{R}_t$. Then, we merge $\mathcal{R}_t$ of different time intervals to build the required $\mathcal{R}$. The details are as follows:

Finding $\mathcal{R}$ for time interval t ($\mathcal{R}_t$). The process for finding $\mathcal{R}_t$ is

dependent on $\mathbf{X}$ and the condition of homogeneity denoted by c. We let $x_{uz_i,t}$ be the extent of u's interest in z_i and define $\mathbb{U}_{z_iz_j,t}(c)$ to be the set of all those users who are interested in both topics z_i and z_j given c. In our definition, $\mathbb{U}_{z_iz_j,t}(c)$ is considered to be maximal if it is not possible to include an additional user while maintaining c. Based on $\mathcal{U}_t = \{\mathbb{U}_{z_iz_j,t}(c) : z_i, z_j \in \mathbb{Z}\}$, we form a multigraph $\mathrm{G}_t = (\mathbb{Z}, \mathcal{U}_t)$ whose nodes are the set of topics and for each $\mathbb{U}_{z_iz_j,t}(c) \in \mathcal{U}_t$ a directed edge connecting z_i to z_j is added to G_t, which is labeled with the set of users in $\mathbb{U}_{z_iz_j,t}(c)$.

In Figure 4.2, we clarify how the multigraph would look like by visualizing it for time interval 22 for the three users introduced earlier. We assume two alternatives for the condition of homogeneity, *i*) regions that have a value above 0.1 will be considered to be similar, and *ii*) regions that have a value above 0.1 and the differences of values fall in the range [0, 0.1) will be considered to be similar. The multigraphs are shown in Figure 4.2(a) and 4.2(b) respectively.

Once the multigraph has been constructed for time interval t (G_t), we perform a depth first search traversal on G_t in order to find $\mathcal{R}_t$, a process which has been outlined in Algorithm 2. We initially commence the process by considering all of the users with an empty set of topics ($r = \mathbb{U} \times \varnothing$; all users $\mathbb{U}$). The algorithm gradually considers each topic and incrementally adds it to the set. In each recursive stage, we have a candidate denoted as $r = \mathrm{A} \times \mathrm{B}$ and a set of yet-to-be-processed topics C. The candidate will be added to $\mathcal{R}_t$ if it satisfies the condition of homogeneity and is not already

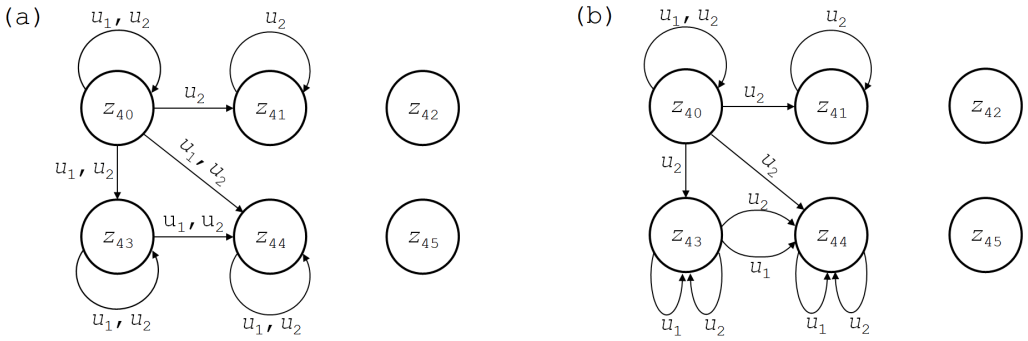

Figure 4.2: The Multigraph constructed from the three users introduced in Figure 1.1 in time interval t_{22} when the condition of homogeneity c is (a) a value above 0.1, and (b) the difference of values above 0.1 falls in the range of [0, 0.1).

subsumed by another region. Since in graph G_t, we only create a region of like-mindedness based on a topic (loop) or a pair of topics (directed edge), we need to check condition c as we traverse a DFS path over the directed edges of the graph in order to extend the region of like-mindedness to include more topics and users. Further, we remove all other regions that are subsumed by r when r is added to $\mathcal{R}_t$ (Lines 2 to 4). Once r is added, we now expand its topic set to include one of the remaining topics that have not been considered yet as long as there is a directed edge between a topic in r and the new topic in G_t. The algorithm is recursively called on the new candidate that includes a new topic (Lines 5 to 12).

For the sake of further clarification, let us review the process proposed in Algorithm 2 for the multigraph depicted in Figure 4.2(a). The algorithm starts by initializing r to consist of all the three of the users but an empty set of considered topics and a complete set of unexplored top-

Algorithm 2 Finding regions of like-mindedness for time interval t ($\mathcal{R}_t$)

Inputs:
c, homogeneity condition;
G_t, multigraph at time interval t;
$\mathbb{U}$, set of users;
$\mathbb{Z}$, set of topics of interest;
Output:
$\mathcal{R}_t$, set of regions of like-mindedness for time interval t
Initialization:
$\mathcal{R}_t = \varnothing$;
find_r_t($r = \mathbb{U} \times \varnothing, \mathrm{C} = [z_1, z_1, z_2, z_2, ..., z_{|\mathbb{Z}|}, z_{|\mathbb{Z}|}]$);

```
1: procedure find_r_t(r = A × B, C)
2:     if (r ⊨ c) ∧ (∄r′ ∈ R_t : r ⊂ r′) then
3:         ∀r″ ∈ R_t if r″ ⊂ r then R_t ← R_t \ r″
4:         R_t ← R_t ∪ r
5:     for all z_j ∈ Z do
6:         A ← r.A; B ← r.B ∪ z_j; C ← C \ z_j
7:         if r.B = ∅ then find_r_t(A × B, C)
8:         else
9:             for all z_i ∈ r.B do
10:                for all (z_i → z_j) ∈ U_t do
11:                    A ← r.A ∩ U_{z_i z_j, t}
12:                    find_r_t(A × B, C)
```

ics ($r = \{u_1, u_2, u_3\} \times \varnothing, \mathrm{C} = [z_{40}, z_{40}, z_{41}, z_{41}, ..., z_{45}, z_{45}]$). The algorithm then selects the first topic (Topic 40) by removing it from C and adding it to the empty set of topics in r (Line 7). Given the current state of r ($\{u_1, u_2, u_3\} \times \{z_{40}\}$) does not satisfy the condition for homogeneity, we select the next topic, which is again Topic 40 given the directed looping edge. The new r ($\{u_1, u_2\} \times \{z_{40}\}$) now satisfies the condition of homogeneity and is hence added to $\mathcal{R}_t$ (Line 4). The subsequent step is to consider Topic 41 because there is a direct edge from Topic 40 to Topic 41. Based on this transition, the new r will be $\{u_2\} \times \{z_{40}, z_{41}\}$, which produces a new element in $\mathcal{R}_t$.

Finding regions of like-mindedness ($\mathcal{R}$). Algorithm 2 identifies $\mathcal{R}_t$ separately for each of the time intervals; however, we will need to identify $\mathcal{R}$ across the whole time period that spans all of the individual time intervals. We adopt a similar strategy for expanding the individual $\mathcal{R}_t$s into $\mathcal{R}$ as explained in Algorithm 3. We build a multigraph G which consists of the time intervals as its nodes and edges representing transitions between time intervals such as i and j only when $\{r.\text{A} \cap r'.\text{A}\} \times \{r.\text{B} \cap r'.\text{B}\} \times \{i, j\}$ satisfies c given two regions $r \in \mathcal{R}_i$ and $r' \in \mathcal{R}_j$.

Algorithm 3 Finding regions of like-mindedness ($\mathcal{R}$)

Inputs:
c, homogeneity condition;
$\mathbb{U}$, set of users;
$\mathbb{Z}$, set of topics of interest;
G, multigraph for the whole time intervals;
$\mathcal{R}_t$ for each time interval $1 \leq t \leq \text{T}$;
Output:
$\mathcal{R}$, set of regions of like-mindedness for the whole time intervals
Initialization:
$\mathcal{R} = \varnothing$;
find_r(R = $\mathbb{U} \times \mathbb{Z} \times \varnothing$, D=[1, 1, 2, 2, ..., T, T]);

```
1: procedure find_r(R = A × B × C, D)
2:     if (R ⊨ c) ∧ (∄R′ ∈ R : R ⊂ R′ then
3:         ∀R″ ∈ R if R″ ⊂ R then R ← R \ R″
4:         R ← R ∪ R
5:     for all j ∈ D do
6:         A ← R.A; B ← R.B; C ← R.C ∪ j; D ← D \ j
7:         if R.C = ∅ then
8:             find_r(A × B × C, D)
9:         else
10:            for all i ∈ R.C do
11:                for all (i → j) ∈ G do
12:                    //r ∈ R_i, r′ ∈ R_j : {r.A ∩ r′.A} × {r.B ∩ r′.B} × {i, j}
13:                    A ← R.A ∩ {r.A ∩ r′.A}
14:                    B ← R.B ∩ {r.B ∩ r′.B}
15:                    find_r(A × B × C, D)
```

Algorithm 3 produces $\mathrm{R} = \mathrm{A} \times \mathrm{B} \times \mathrm{C} \in \mathcal{R}$ where A is a set of users who have the similar interests towards topics in B in time intervals in C based on a defined condition of homogeneity. In essence, this provides us with information on which users, when and how, expressed similar preferences towards topics of the social network. This is valuable for determining which users are similar to each other across different time intervals and topics. Those users who are placed together in the same R can be considered to be more similar to each other compared to those users who are not in the same R. We consider regions of like-mindedness such as R to serve as context for each user. Based on such context, we would like to learn user embeddings that maximize the likelihood of users who have been seen together in the same Rs to be close to each other in the embedding space and those who are not seen together to be embedded far apart from each other. Let us first discuss the time complexity of finding regions of like-mindedness.

Time complexity analysis. In each time interval t, it takes $O(|\mathbb{U}| \times |\mathbb{Z}|^2)$ to calculate $\mathbb{U}_{z_i z_j, t}(c)$ for all pairs of z_i and $z_j \in \mathbb{Z}$ and build the multi-graph G_t considering the fact that testing the condition of homogeneity can be done in $O(1)$. Furthermore, performing depth-first-search (DFS) on the graph to find regions of like-mindedness $\mathcal{R}_t$ takes $O(|\mathbb{U}|^{|\mathbb{Z}|})$ in the worst case, which happens when there exists an edge between each pair of z_i and z_j associated with $\mathbb{U}_{z_i z_j, t}(c)$ containing only one user. The analysis of the time complexity for finding $\mathcal{R}$ is similar but in the context of the number of time intervals and the size of $\mathcal{R}_t$ for each time interval. Here, for each pair of time

intervals i and j, and a pair of $\mathcal{R}_i$ and $\mathcal{R}_j$, we test the condition of homogeneity which takes $O(|r| \times \mathrm{T}^2)$ plus a final DFS in $O(|r|^{\mathrm{T}})$ where $|r|$ is the number of all $\mathcal{R}_t$. As seen, the most expensive parts are the DFS traversal on the multigraphs in the first and second steps which highly depend on the condition for homogeneity c.

We would like to note that the proposed method is efficient in practice because of the following considerations:

1. In the real world, users are only interested in a limited set of topics in each time interval and over the whole time period. For this reason, users' topic preference time series are quite sparse with many topics not even examined or relevant for each user. Therefore, the number of edges in the multigraphs is quite small. Recall that one of the major components of the time complexity of the method was due to the DFS traversal, which will be quite small given the sparsity of the multigraphs in practice.

2. In addition, the depth of the DFS traversal is quite shallow given the fact that the number of users is far larger than the number of topics and time intervals. When compared to the number of users, the number of topics and time intervals can be considered to be constant values.

3. Algorithms 2 and 3 can be easily parallelized across different time intervals.

4.2.2 Temporal Content-based User Vector Representation

We approach the problem of learning user representations as a maximum likelihood (ML) problem through which similar users to a given user are identified based on the user's context. We define the context for each user to consist of all those users who have been observed with this user in similar regions of like-mindedness ($\mathcal{R}$). As such, the more two users are seen in each other's contexts, the more likely it would be for them to be similar to each other. We adopt the continuous bag-of-word (CBOW) model from [69] to learn user representations.

Definition 2. ***Temporal Content-based User Embedding Objective.*** *Given the set of all regions of like-mindedness $\mathcal{R}$, the embedding function $f : \mathbb{U} \rightarrow \mathfrak{R}^d$ maps each user $u \in \mathbb{U}$ onto a d-dimensional real space $[0, 1]^d; d \ll |\mathbb{U}|$, such that the following objective is optimized:*

$$\arg\max_{f} \sum_{\mathrm{R} \in \mathcal{R}, u \in \mathrm{R}} \log \Pr(u | \mathrm{R} \setminus u) \tag{4.1}$$

In order to make the optimization tractable, we assume conditional independence for observing users in a region of like-mindedness. So,

$$\Pr(u | \mathrm{R} \setminus u) = \prod_{v \in \mathrm{R} \setminus u} \Pr(u | v) \tag{4.2}$$

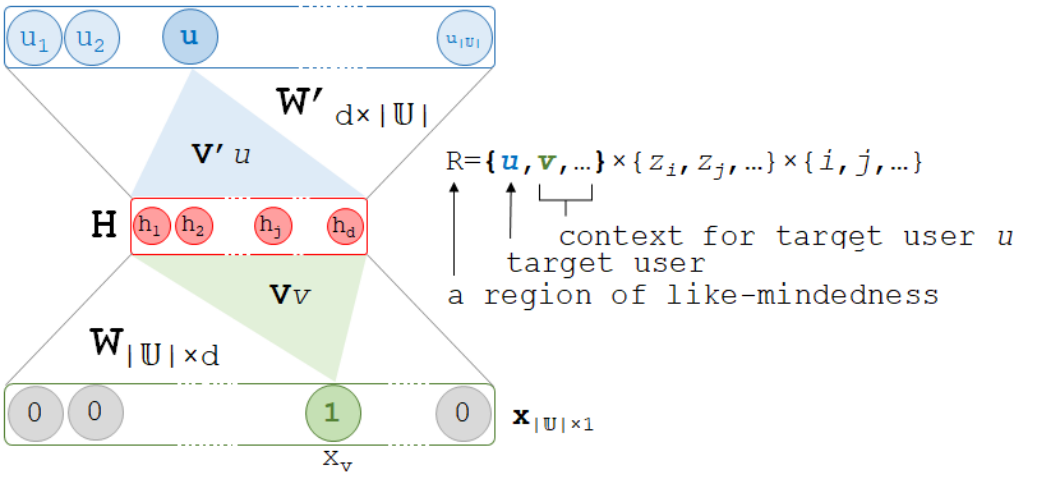

Figure 4.3: The neural network architecture to learn temporal content-based user vector representations.

We adopt the architecture shown in Figure 4.3 to learn user representations. It should be noted that the size of the hidden layer (d) will be the size of the user representation vectors. Furthermore, given the model learns to predict a user given its context, the size of the input and output layers is equivalent to the number of users. We use a one-hot encoding representation to refer to users in the input ($\mathbf{I}$) and output layers. The structure of the hidden layer neurons is linear $\mathbf{H} = \mathbf{W}_{\mathcal{D}}^{\top}\mathbf{I}$ where $\mathbf{W}_{\mathcal{D}}$ has a size of $|\mathbb{U}| \times d$ and is the input to the hidden layer. Similarly, the weights between the nodes in the hidden and output layers are denoted by $\mathbf{W}'_{\mathcal{D}}$ of size $d \times |\mathbb{U}|$. Also, we refer to a user v's corresponding row in $\mathbf{W}_{\mathcal{D}}$ as $\mathbf{V}_v$ The network performs user prediction given its context through a softmax function by approximating the likelihood of observing the target user v given some other user u observed together in at least one region of like-mindedness. This conditional

probability is defined as follows:

$$\Pr(u|v) = \frac{\exp(\mathbf{V}'^{\top}_{u}\mathbf{H})}{\sum_{w\in\mathbb{U}}\exp(\mathbf{V}'^{\top}_{w}\mathbf{H})} = \frac{\exp(\mathbf{V}'^{\top}_{u}\mathbf{V}_{v})}{\sum_{w\in\mathbb{U}}\exp(\mathbf{V}'^{\top}_{w}\mathbf{V}_{v})} \tag{4.3}$$

Given the conditional independence assumption in Equation (4.2) and the above conditional probability in Equation (4.3), we can simplify Equation (4.1) as:

$$\arg\max_{f} \sum_{\mathrm{R}\in\mathcal{R},u\in\mathrm{R}} \Big[\sum_{v\in\mathrm{R}\setminus u} \big[(\mathbf{V}'^{\top}_{u}\mathbf{V}_{v}) - \log\sum_{w\in\mathbb{U}}\exp(\mathbf{V}'^{\top}_{w}\mathbf{V}_{v})\big]\Big] \tag{4.4}$$

However, this formulation is computationally intractable as its time complexity is proportional to the size of $\mathbb{U}$. Morin and Bengio [73] have proposed hierarchical softmax to approximate the full softmax efficiently in practice. Accordingly, instead of a matrix, the hidden layer to output layer connection is a binary Huffman tree whose leaves are users. For each user u, there is a path $u_1, u_2, ..., u_{h(u)}$ of height $h(u)$ from the root, u_1, to her respective leaf, $u_{h(u)}$. This choice leads to speedup from $O(|\mathbb{U}|)$ to $O(log|\mathbb{U}|)$. Hierarchical softmax defines $\Pr(u|v)$ as follows:

$$\Pr(u|v) = \prod_{i=1}^{h(u)-1} s((-1)^{(u_{i+1}\neq child(u_i))} \times \mathbf{V}'^{\top}_{u_i}\mathbf{V}_{v}) \tag{4.5}$$

where $s(x)$ is the sigmoid function. $\mathbf{V}'^{\top}_{u_i}\mathbf{V}_{v}$ shows the similarity between the vector representation of user v and the internal user u_i. At each internal

user u_i, if we choose the left (right) child as the correct u_{i+1} in the path from the root to the user's leaf, we have the probability $s((-1)^0 \times x) = s(x)$, else the right (left) child would result in $s((-1)^1 \times x) = s(-x)$ such that $s(x) + s(-x) = 1$. The intuition is that the more an output user u is similar with the ancestors of input user v, the higher the probability would be that they are the same.

Our neural network is trained using stochastic gradient descent and updates $\mathbf{W}_{\mathcal{D}}$ and $\mathbf{W}'_{\mathcal{D}}$ gradually via backpropagation. After the training converges, a pair of like-minded users $u, v \in \mathbb{U}$ will have highly similar vector representations, denoted by $\mathbf{V}_u$ and $\mathbf{V}_v$ in $\mathbf{W}_{\mathcal{D}}$ with respect to the temporal social content $\mathcal{D} = (\mathbb{U}, \mathbb{M}, \mathrm{T})$.

The next step of our work is to learn vector representations of users with respect to social network structure $\mathcal{G} = (\mathbb{U}, \mathbb{A})$, denoted by $\mathbf{W}_{\mathcal{G}}$. More specifically, we are interested in providing a concrete implementation for $g(\mathcal{G})$ on Line 3 of Algorithm 1.

4.3 Link-based User Embeddings

Given a social network structure in the form of a double $\mathcal{G} = (\mathbb{U}, \mathbb{A})$ where $\mathbb{U}$ is the set of users and $\mathbb{A}$ is the connections between the users, our objective in this section is to learn neural user representations based on the global position of a user in $\mathcal{G}$ and the structure of her local neighborhood. We employ an unsupervised representation learning method to encode this information

into a low-dimensional dense feature vector in latent space such that the geometric relations in this latent space correspond to social connections (e.g., link or path) in $\mathcal{G}$. Specifically, user embeddings are inferred by maximizing the probability of observing subsequent users in random walks of the graph conditioned on the source user. We formulate user embeddings learnt from the social network structure in a unified framework as follows.

4.3.1 Neighborhood Context Model

Based on the *homophily* principle, similar users tend to form ties in a social network [68]. As such, groups of densely connected users could be a sign of a user community. In the context of the social network structure, users would be considered to belong to similar communities if they share similar neighborhoods and, as such, are to be placed close to each other in the embedding space. The shared neighborhood, hence, presents a context with respect to the social network structure as opposed to the regions of like-mindedness in the temporal context model (Section 4.2.1) or co-occurrence context in word embeddings. There are different strategies for building a neighborhood for a user. For instance, depth-first-search (DFS) and breadth-first-search (BFS) are two immediate, yet extremely biased ways to generate different samples of neighborhoods for a user. BFS favours *structural equivalence*, that is, those users who share similar structural roles such as hubs and are not necessarily connected and could be anywhere in the network, should be embedded closely together. Being more community aware, DFS in contrast, respects ho-

mophily and leads to similar (close) embeddings for densely connected users. In practice, online social networks exhibit mixture behaviours through which some parts show homophily while the other parts reflect structural equivalence. For this reason, stochastic sampling methods, such as random walk, have been introduced to randomly sample different neighborhoods of the same source user. Random walks are also computationally efficient in terms of both space and time [43]. As a result, we form a network neighborhood of a user based on random walks in this work, which is formally defined as follows:

Definition 3. ***Network Neighborhood.*** *The network neighborhood of a given user $u \in \mathbb{U}$, denoted by $\mathbb{N}_u$, is a set of random walks of length l rooted at u on a possibly infinite social network structure $\mathcal{G} = (\mathbb{U}, \mathbb{A})$ generated by a stochastic process with random variables $[x_{1:l}]$ such that $x_1 = u$ and x_l is a user chosen from the neighbors of x_{l-1} according to a probability distribution $Pr(x_l = w | x_{l-1} = v)$ if $(v, w) \in \mathbb{A}$ and 0 otherwise. The set of network neighborhoods for all users is denoted by $\mathcal{N}$.*

While graph embedding methods such as DeepWalk [80] use a pure (unbiased) random walk based on the uniform distribution, other methods [43] introduce parametric biased random walk to trade off between breadth-first or depth-first searches to preserve community structure as well as structural equivalence between users. For instance, the work in [43] proposes a second order random walk with two parameters p (return parameter) and q (in-out

parameter) in $\Pr(x_l = w|x_{l-1} = v)$ to bias the walk as follows:

$$\Pr(x_l = w|x_{l-1} = v) = \begin{cases} 1/p & \text{if } d(x_{l-2}, v) = 0 \\ 1 & \text{if } d(x_{l-2}, v) = 1 \\ 1/q & \text{if } d(x_{l-2}, v) = 2 \end{cases} \tag{4.6}$$

where $d(.,.)$ denotes distance of the shortest path between users in an unweighted graph. While higher p values favour exploration and avoid revisiting already seen users, higher q allows the search to obtain a local view and approximate BFS behaviour. Unbiased random walks can be seen as a special case when $p = q = 1$.

4.3.2 Link-based User Vector Representation

Once network neighborhoods for all users have been obtained, we learn a user vector representation for each user by optimizing the conditional probability of observing users in the same walk as her. The process is similar to Section 4.2.2 as network neighborhoods can be seen as similar to regions of like-mindedness. To infer the user embeddings, we optimize the following embedding function:

Definition 4. ***Link-based User Embedding Objective.*** *Given the set of network neighborhoods* $\mathcal{N} = \bigcup_{u \in \mathbb{U}} \mathbb{N}_u$*, the embedding function* $g : \mathbb{U} \rightarrow \mathfrak{R}^d$ *maps each user* $v \in \mathbb{U}$ *onto a d-dimensional real space* $[0, 1]^d; d \ll |\mathbb{U}|$*, such that the following objective function is optimized, assuming conditional*

independence:

$$\begin{aligned}\arg\max_{g} \sum_{\mathbb{N}_v \in \mathcal{N}} \log \Pr(\mathbb{N}_v \setminus v | v) &= \arg\max_{g} \sum_{\mathbb{N}_v \in \mathcal{N}} \log(\prod_{u \in \mathbb{N}_v \setminus v} \Pr(u|v)) \\ &= \arg\max_{g} \sum_{\mathbb{N}_v \in \mathcal{N}} \sum_{u \in \mathbb{N}_v \setminus v} \Pr(u|v)\end{aligned} \tag{4.7}$$

We use the same neural architecture as shown in Figure 4.3 but here, given user v, we predict observing users such as u from v's neighborhood, adopting the skip-gram model from [69]. The hidden layer $\mathbf{H}$ is of size d, the input to hidden layer connections is represented by matrix $\mathbf{W}_{\mathcal{G}}$ of size $|\mathbb{U}| \times d$ with each row representing a vector for each user. The input layer $\mathbf{I}$ is a one-hot encoded vector and the hidden layer's neurons are all linear such that $\mathbf{H} = \mathbf{W}^{\top}\mathbf{I}$. Given a user v in the input layer, $\mathbf{H}$ is the transpose of v's corresponding row in $\mathbf{W}_{\mathcal{G}}$, denoted as $\mathbf{V}_v$. In the same way, the connections from the hidden layer to the output layer can be described by matrix $\mathbf{W}'_{\mathcal{G}}$ of size $d \times |\mathbb{U}|$. The softmax function approximates the probability of observing user u taken from $\mathbb{N}_v$ from the same random walk, i.e.,

$$\Pr(u|v) = \frac{\exp(\mathbf{V}'^{\top}_u \mathbf{H})}{\sum_{w \in \mathbb{U}} \exp(\mathbf{V}'^{\top}_w \mathbf{H})} = \frac{\exp(\mathbf{V}'^{\top}_u \mathbf{V}_v)}{\sum_{w \in \mathbb{U}} \exp(\mathbf{V}'^{\top}_w \mathbf{V}_v)} \tag{4.8}$$

where $\mathbf{V}'_u$ is u's corresponding column of matrix $\mathbf{W}'_{\mathcal{G}}$. However, calculating the normalization factor in the denominator is not feasible. Hierarchical softmax and negative sampling are two promising alternatives to accelerate

the computation. Stochastic gradient descent is used to train the neural network and the derivatives are estimated using backpropagation. Users' vector representations with respect to social network structure $\mathcal{G} = (\mathbb{U}, \mathbb{A})$ are vectors of $\mathbf{W}_{\mathcal{G}}$.

4.4 Embeddings Interpolation

Having learnt two different user vector representations of users from the temporal social content $\mathcal{D} = (\mathbb{U}, \mathbb{M}, \mathrm{T})$ and the social network structure $\mathcal{G} = (\mathbb{U}, \mathbb{A})$, denoted by $\mathbf{W}_{\mathcal{D}}$ and $\mathbf{W}_{\mathcal{G}}$, respectively, the next step is to integrate them into a single vector representation, denoted as $\mathbf{W}$, by an interpolation function $h(\mathbf{W}_{\mathcal{D}}, \mathbf{W}_{\mathcal{G}})$ defined on Line 4 of Algorithm 1. We adopt a linear weighting mechanism to interpolate the embeddings mined from the social network structure and temporal social content. Formally,

$$h(\mathbf{W}_{\mathcal{D}}, \mathbf{W}_{\mathcal{G}}) = \alpha \mathbf{W}_{\mathcal{D}} + (1 - \alpha)\mathbf{W}_{\mathcal{G}} \tag{4.9}$$

where α denotes a weighting coefficient to interpolate between temporal content and social network structure in the final user vector representation. For instance, if $\alpha = 0$, the interpolated embeddings lead to the conventional link-based user community detection on the one extreme. On the other extreme, it will solely rely on temporal content if $\alpha = 1$ and becomes a pure temporal content-based method. The effect of embedding interpolation to the overall

performance of user community detection is evaluated by choosing $\alpha \in \mathfrak{R}^{[0,1]}$. Although simple, linear weighting is uninformed, easy to implement, interpretable, and could achieve competitive performance across a wide span of different data types and domains [45, 8, 7]

4.5 Community Detection

Given the interpolated user vector representation $\mathbf{W} = h(\mathbf{W}_{\mathcal{D}}, \mathbf{W}_{\mathcal{G}})$, we identify communities of users through graph-based partitioning heuristics. We represent users and their pairwise distances through a weighted undirected graph. Precisely, let $\mathrm{G} = (\mathbb{U}, \mathbb{E}, w)$ be a weighted user graph such that $\mathbb{E} = \{e_{u,v} : \forall u, v \in \mathbb{U}\}$ and the weight function $w : \mathbb{E} \rightarrow \mathfrak{R}^{[0,1]}$ defined as $w(e_{u,v})$ be the dot-product, or angle, between u and v's embeddings in $\mathbf{W}$. It is possible to employ a graph partitioning heuristic to extract clusters of users that form latent communities. We leverage the Louvain Method (LM) [15] as it *i*) can be applied to weighted graphs, *ii*) does not require *a priori* knowledge of the number of partitions, and *iii*) has an efficient linear time complexity for the problem of graph partitioning. As a result of the application of LM, a set of subgraphs such as $\mathrm{G}[\mathbb{C}]$ are induced where the edges in each subgraph have both ends in the same subgraph. The collection of these subgraphs form the set of user communities $\mathbb{P}$ desired in the problem definition presented in Section 3.1.

4.6 Summary

In this chapter, we have proposed an approach to detect communities through multimodal feature learning (embeddings) of users from their *i*) *temporal* content, *ii*) social network neighborhood. With respect to the temporal content, we model the users' temporal contribution towards topics of interest by introducing the notion of regions of like-mindedness between users. These regions cover users who share not only similar topical interests but also similar temporal behaviour. Given the regions of like-mindedness as context, we train a neural network such that the probability of a user in a region is maximized given other users in the same region (Section 4.2). With regard to the social network neighborhood, we learn user embeddings based on their social network connections (links) through neural graph embeddings (Section 4.3). We then interpolate temporal content-based embeddings with social link-based embeddings to capture both sources of information for representing users (Section 4.4). The approach addresses research questions **RQ1**, i.e., whether the consideration of time plays a role in the quality of the identified communities and **RQ2**, i.e., whether temporal content-based user community detection methods show better performance compared to link-based methods, as well as **RQ3**, i.e., whether link-based and temporal content-based community detection methods have synergistic effect on each other.

4.7 Related Publications

- Hossein Fani, Ebrahim Bagheri and Weichang Du; "Temporally Like-minded User Community Identification through Neural Embeddings." *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017.*

- Hossein Fani, Eric Jiang, Ebrahim Bagheri, Feras Al-Obeidat, Weichang Du and Mehdi Kargar; "User Community Detection via Embedding of Social Network Structure and Temporal Content." *In Information Processing and Management (IP&M)*, 2019, (In press).

Chapter 5

User Community Prediction

Not only the detection of user communities up until the current point in time is of high importance, but also their analysis in future time intervals can be of interest due to their wide range of applications such as personalized recommendations and marketing campaigns. Temporal content-based user community detection methods like the proposed methods in the previous chapters incorporate temporal aspects of users' content and stress that users of the same community would ideally show similar interest patterns for similar topics over time. However, users' temporal content is only used for pairwise user similarity calculation to build content-based user communities as opposed to user similarity *prediction*. As a result, they have limited applicability for identifying user communities of the *future*. In other words, unlike temporal content-based user community detection methods, which employ users' temporal and topical interests for calculating pairwise user similarities

in order to identify user communities up until *now*, our work in this chapter employs such information for predicting user communities in the *future*, which is a step forward compared to the state of the art.

In this chapter, we propose two methods to predict content-based user communities in the future: Granger regression (G-regression) and temporal latent space modeling. In Granger regression, we propose to consider both the temporal evolution of users' interests as well as a stricter form of inter-user influence through the notion of *causal dependency*. We employ Granger causality to determine the degree of inter-user influence that can be used to identify which users play influential roles in the behavioural evolution of one or more other users. Based on Granger causality, we identify a causing user c to influence the affected user e if and when the past observations of c lead to a more accurate prediction of the behaviour of e above and beyond the information contained in past observation of e alone. Although the proposed G-regression method shows promising results to identify user communities in the future, it requires building predictive models on a per user basis and, hence, is practically prohibitive.

In the second method, we propose a temporal latent space model for user community *prediction* in social networks, whose goal is to predict *future* emerging user communities based on past history of users' topics of interest. Our model assumes that each user lies within an unobserved latent space, and similar users in the latent space representation are more likely to be members of the same user community. The model allows each user to adjust

her location in the latent space as her topics of interest evolve over time. In this method, although we use temporal information to predict future users' topical interests similar to the proposed G-regression method, we train only one model for all users and, thus, significantly reduce computational cost compared to G-regression techniques.

These two proposed methods are intended to address **RQ4**, i.e., whether it is possible to predict future-yet-unobserved content-based communities on social networks.

5.1 Granger regression (G-regression)

Users' topics of interest show the dynamic evolution of user behaviour in online social networks, whose effective prediction can be utilized for the task of community prediction. Common time series models, which use observations from past time intervals to predict the users' future topics of interest, have an independence assumption that users' behaviour is considered to evolve independently from that of the other users. These methods overlook the explicit or implicit social interactions that are inherent to social networks. On the other hand, time-aware collaborative filtering approaches such as TimeSVD++ [57] and recurrent recommender network (RRN) [106] propose a valuable step forward by integrating the individual and collective perspectives of the users in addition to their temporal evolution patterns (non-stationarity) under the traditional collaborative filtering framework.

Successful as they are, these approaches, however, do not consider strict inter-user dependence (social influence) and only benefit from users' behavioural correlation to make predictions. In contrast, social-aware recommender systems such as TrustSVD [46] and SocialPMF [53] have already been proposed to address this issue but they in turn overlook temporal evolution.

We propose to consider both the temporal evolution of users' interests as well as a stricter form of inter-user influence through the notion of *causal dependency*. We employ Granger causality [42] to determine the degree of inter-user influence that can be used to identify which users play influential roles in the behavioural evolution of one or more other users. Based on Granger causality, we identify a causing user c to influence the affected user e if and when the past observations of c lead to a more accurate prediction of the behaviour of e above and beyond the information contained in past observation of e alone. This leads to a weighted *directed* network of users, denoted by the *influence network*, in which the edges depict the influence direction of its adjacent users. We use the influence network to perform interest prediction. Specifically, given a topic of interest z and a user e, we find e's influential neighbor(s) from the influence network such as c and build a vector autoregression model (VAR) based on e and c's user-topic contribution time series to predict e's degree of interest toward topic z in the future. Last, a weighted *undirected* graph is formed over the users and their pairwise similarity on predicted degrees of interest in the future, on which the Louvain method [15] is applied to find user communities in the future.

In summary, our G-regression approach consists of users' topic contribution detection, users' influencers identification, users' future topics of interest prediction, and user community detection in the future. In the following, we describe the details of each step.

5.1.1 User Topic Contribution Detection

To capture users' topic contribution, we build user-topic contribution time series representing users' interests towards topics over time similar to Section 3.3.1. Given a set of topics $\mathbb{Z}$ within T time steps (e.g. days) extracted by a topic detection method and a set of users $\mathbb{U}$, the user-topic contribution time series of user e is expressed as $\mathbf{X}_e = (\mathbf{x}_{e,1}, \mathbf{x}_{e,2}, ..., \mathbf{x}_{e,\mathrm{T}})$. At each time interval t, $\mathbf{x}_{e,t}$ is a vector whose elements $x_{ez,t} \in \Re^{[0,1]}$ show the degree of interest for the user e towards the topic z. Assuming there are K topics detected, $\mathbf{x}_{e,t}$ is a K-tuple vector and the user-topic contribution time series will be a K-variate time series.

We identify influence relations between users on a per topic basis. We break down the K-variate user-topic contribution time series into K *univariate* time series $\mathbf{X}_{ez} = (x_{ez,1}, x_{ez,2}, ..., x_{ez,\mathrm{T}})$ indicating the contribution by user e for topic z within time period T.

The main objective of the proposed G-regression method is to accurately predict $x_{ez,\mathrm{T}+1}; \forall z \in \mathbb{Z}, \forall e \in \mathbb{U}$.

5.1.2 User Influence Identification

We leverage the influence between users when delivering their topics of interest over time to predict users' topics of interest. The influence that one user might exert on the other is identified through Granger causality [42]. Granger causality has been perceived as a predictive notion of causality between time series [13] and as such in our case, it can be applied to the user-topic contribution time series. In the bivariate case, user e, the effect, is said to be influenced (Granger-caused) by another user c, the cause, with respect to topic z, if and only if, regressing on past values of both e and c's topic contribution time series is statistically significantly more accurate than doing so with past values of e alone. Formally, let $\mathbf{X}_{ez} = (x_{ez,1}, x_{ez,2}, ..., x_{ez,\mathrm{T}})$ and $\mathbf{X}_{cz} = (x_{cz,1}, x_{cz,2}, ..., x_{cz,\mathrm{T}})$ be two stationary topic contribution time series of user e and c with respect to topic z, and let the two regression models be:

$$\begin{aligned} \mathrm{H}_1 &: x_{ez,\mathrm{T}} = \sum_{l=1}^{\mathrm{L}} a_l x_{ez,t-l} + \sum_{l=1}^{\mathrm{L}} b_l x_{cz,t-l} + \epsilon_1 \\ \mathrm{H}_0 &: x_{ez,\mathrm{T}} = \sum_{l=1}^{\mathrm{L}} a_l x_{ez,t-l} + \epsilon_2 \end{aligned} \tag{5.1}$$

where L is the maximal time lag, a_l and b_l are the regression variable coefficients, and ϵ_1 and ϵ_2 are the residual terms, which are independent and identically distributed (i.i.d) according to a standard Gaussian $\mathcal{N}(0, \sigma^2)$. If H_1 is a significantly better model than H_0 (i.e., provides more precise pre-

dictions), we conclude that $\mathbf{X}_{cz}$ Granger-causes $\mathbf{X}_{ez}$; notationally $c \xrightarrow{\mathrm{G}}_z e$. Among other techniques, the significance level can be tested using the F statistic by the Granger-Sargent test [42], defined as follows:

$$F = \frac{(rss_{\epsilon_2} - rss_{\epsilon_1})/\mathrm{L}}{(rss_{\epsilon_1})/(\mathrm{T} - 2\mathrm{L})} \sim F(\mathrm{L}, \mathrm{T} - 2\mathrm{L}) \tag{5.2}$$

where rss_{ϵ_2} is the restricted residual sum of squares under H_0, rss_{ϵ_1} is the unrestricted residual sum of squares under H_1, T is the number of time steps, and F follows the F-distribution. We reject the null hypothesis that c does not Granger-cause e if the above calculated F is greater than the critical value of the F-distribution for some desired false-rejection probability, e.g., 0.05.

Based on Granger causality between all users for all topics, i.e., $c \xrightarrow{G}_z e; \forall e, c \in \mathbb{U}, \forall z \in \mathbb{Z}$, it is possible to find causal dependency between pairs of users in order to identify influencers. The set of influencers for a user form its influence network.

5.1.3 User Future Interest Prediction

We use the estimated vector autoregression (VAR) model to do one-step-ahead prediction of users' topics of interest at time step T+1, although can be generalized to make predictions for any time period after T. Given a user e and a topic z, we build a VAR model whose variables are e's topic preference time series and her influencer network, identified by Granger causality, up to

time step $t = \mathrm{T}$. Formally,

$$\mathbf{Y}_{z,\mathrm{T}+1} = b + \sum_{l=1}^{\mathrm{L}} A_l \mathbf{Y}_{z,t-l+1} + \epsilon \tag{5.3}$$

where $\mathbf{Y}_{z,t}$ is a vector whose first element is equal to e's degree of interest toward topic z at time step t; notationally $\mathbf{Y}^{(1)}_{ez,t} = x_{ez,t}$. The other elements belong to e's influencers such as c, i.e., $Y^{(i)}_{z,t} = x_{cz,t}; i > 1$. Here, b is a vector of constants (intercepts), A_l is a time-invariant matrix of coefficients and ϵ is a vector of error terms. After model estimation (training) to learn b, A_l and ϵ, the predicted degree of interest for user e towards topic z at time step T+1, denoted as $\hat{x}_{ez,\mathrm{T}+1}$, will be $Y^{(1)}_{z,\mathrm{T}+1}$.

Overall, $|\mathbb{U}| \times (|\mathbb{Z}| \times |\mathbb{U}| + |\mathbb{Z}|); |\mathbb{U}| >> |\mathbb{Z}|$, VAR models should be trained for pairwise Granger causality tests and one-time-step-ahead predictions. While the time complexity of our method is a quadratic function of the number of users, its parallel implementation is able to reduce the complexity to linear complexity, with $|\mathbb{U}|$ users in parallel with each other.

5.1.4 User Community Detection in the Future

The main goal is to predict the user communities whose user members share similar temporal expositions toward similar topics of interest at future time step T+1. To do so, we build a weighted undirected graph $G = (\mathbb{U}, \mathbb{E}, w)$ whose nodes are users and edges are weighted by function w based on the topical similarity between two users at time interval T+1. Based

on our G-regression method, this depends only on the predicted degree of interest for the users towards topics at time step T+1. Given $\hat{x}_{ez,\mathrm{T+1}}$, the predicted degree of interest for user e towards topic z at time step T+1, we build a predicted user-topic contribution vector at time T+1 as $\hat{\mathbf{X}}_{e,\mathrm{T+1}} = (\hat{x}_{ez^1,\mathrm{T+1}}, \hat{x}_{ez^2,\mathrm{T+1}}, ..., \hat{x}_{ez^\mathrm{K},\mathrm{T+1}})$ where $z^i \in \mathbb{Z}$ and $1 \leq i \leq \mathrm{K} = |\mathbb{Z}|$ and the weight function w is defined to be the dot-product, or angle, between e and c's predicted user-topic contribution vector at time T+1, i.e., $w(e,c) = \hat{\mathbf{X}}_{e,\mathrm{T+1}} \hat{\mathbf{X}}^{\intercal}_{c,\mathrm{T+1}}$.

It is possible to employ a graph partitioning heuristic to extract clusters of users that form latent communities. We leverage the Louvain Method (LM) [15] as in Section 3.3.3 and Section 4.5. As a result of the application of LM, a set of subgraphs such as $\mathrm{G}[\mathbb{C}]$ are induced where the edges in each subgraph have both ends in the same subgraph. The collection of these subgraphs form the set of user communities in future time interval T+1.

5.2 Temporal Latent Space Modeling

The proposed G-regression method requires building predictive models on a per-user per-topic basis and, hence, is practically prohibitive. Alternatively, in this section, we propose temporal latent space modeling to predict content-based user communities in the future where only one model is trained for all users and computational cost is significantly reduced. Plus, it exhibits a stronger predictive power compared to the proposed G-regression method.

Latent space modeling [95] has been successfully employed for link prediction in graphs where, given the observed links in the graph, the location of each node in a latent space is learned such that the closer two nodes are in that space, the higher the probability of a link between them would be. In other words, similarity in latent space translates into links in graph space. Latent space modeling inherently preserves homophily where links between nodes are considered clues for similarity and, so, densely connected groups of nodes imply communities. Different approaches based on matrix factorization and deep neural networks [112, 111, 114] have been proposed to learn the latent space representation of the network structure. However, these studies are concerned with static graphs, where the latent representations of the users are fixed. Such works undermine the fact that the latent space needs to evolve over time and, hence, fall short when identifying user communities of the future.

Temporal tensor-factorization approaches [34], or temporal latent space models [120], however, go beyond static networks and assume that the network is dynamic and changes with time. Such models endeavor to learn low-rank latent space representations for dynamic link prediction based on the intuition that nodes can move in the latent space over time. While suitable for predicting links in a social network structure, dynamic link prediction models are inherently deficient when the communities need to take users' content similarity into account, i.e., identify *content-based* user communities in the future which is the main goal of this chapter.

Our approach consists of three subsequent phases: temporal graph identification, temporal latent space inference, and community prediction in the future. In the following, we lay out the details of each step.

5.2.1 Temporal Graph Identification

Given a set of topics $\mathbb{Z}$ from a social network within T time intervals extracted by a topic detection method and a set of users $\mathbb{U}$, we build the user-topic contribution time series of user $u \in \mathbb{U}$ towards topic set $\mathbb{Z}$ within time period T as $\mathbf{X}_u = (\mathbf{x}_{u,1}, x_{u,2}, ..., \mathbf{x}_{u,\mathrm{T}})$ according to Section 3.3.1.

We let temporal graph $G_t = (\mathbb{U}, \mathbb{E}_t, w)$ represent the content-based similarity between the users of the social network whose nodes are users in $\mathbb{U}$ and let $\mathbb{E}_t$ be the set of weighted undirected edges whose weights are based on a similarity function w, which is defined as the cosine similarity of topic preference vectors of the users at time interval t, i.e., $\forall u, v \in \mathbb{U} : w(u, v : t) = \frac{\mathbf{x}_{u,t} \cdot \mathbf{x}_{v,t}}{|\mathbf{x}_{u,t}||\mathbf{x}_{v,t}|}$.

Given $(G_1, G_2, ..., G_\mathrm{T})$, we aim to accurately predict a set of induced subgraphs in $G_{\mathrm{T}+1}$ to form content-based user communities at time interval T+1. The following proposed method can be generalized to make predictions for any time period after T though.

5.2.2 Temporal Latent Space Inference

Within time period T, the stream of graphs $(G_1, G_2, ..., G_{\mathrm{T}})$ could be considered as a dynamic graph $\mathscr{G}$ which is evolving over time. We map each user u at time interval t to a low-rank d-dimensional latent space, denoted by $\mathbf{y}_{u,t}$, while imposing the following assumptions: *i*) users change their latent representations over time, *ii*) two users that are close to each other in $\mathscr{G}$ remain close in latent space, *iii*) two users who are close in latent space share similar topics of interest with each other.

Formally, given a dynamic network $\mathscr{G}$, we find a d-dimensional latent space representation for $\forall u \in \mathbb{U}$ for time interval $1 \leq t \leq \mathrm{T}$ that minimizes the quadratic loss with temporal regularization:

$$\begin{aligned} \arg\min \Big[\sum_{t=1}^{\mathrm{T}} \sum_{u,v \in \mathbb{U}} |w(u,v:t) - \mathbf{y}_{ut}\mathbf{y}_{vt}^{\top}|_F^2 \\ + \lambda \sum_{t=1}^{\mathrm{T}} \sum_{u \in \mathbb{U}} (1 - \mathbf{y}_{ut}\mathbf{y}_{u(t-1)}^{\top}) \Big] \\ \forall u \in \mathbb{U}; \mathbf{y}_{ut} \geq 0, \mathbf{y}_{ut}\mathbf{y}_{ut}^{\top} = 1 \end{aligned} \tag{5.4}$$

where $w(u, v : t)$ is the similarity score for a pair of users u and v in G_t, $\mathbf{y}_{ut}$ is the d-dimensional latent representation for u at time interval t, λ is a regularization parameter, and the term $(1 - \mathbf{y}_{ut}\mathbf{y}_{u(t-1)}^{\top})$ penalizes user u for an immediate large change in its location in latent space. Our model maps each user to a point in a unit hypersphere rather than simplex, because sphere modeling gives a clearer boundary between similar users and dissimilar users

when mapping all user pairs into the latent space.

Optimizing Eq. 5.4 is expensive in terms of both time complexity and storage as it requires all graphs in $\mathscr{G}$ to jointly update all temporal latent representations for users in all time intervals. To optimize Eq. 5.4, we use the *local* block coordinate gradient descent (bc-gd) algorithm [120], in which inference happens sequentially. Specifically, we optimize users' latent representation locally by minimizing the following objective function at each time interval t:

$$\arg\min \sum_{u,v\in\mathbb{U}} (w(u,v:t) - \mathbf{y}_{ut}\mathbf{y}_{vt}^{\top})^2 + \sum_{u\in\mathbb{U}} (1 - \mathbf{y}_{ut}\mathbf{y}_{u(t-1)}^{\top}) \tag{5.5}$$

The local bc-gd algorithm infers users' latent representation from a single graph snapshot G_t and prior initialization from $\mathbf{y}_{u(t-1)}$. The algorithm iteratively updates $\mathbf{y}_{ut}$ until it converges and then moves to the computation of temporal latent space in the next time interval $t+1$. This local sequential update schema greatly reduces the computational cost in practice.

5.2.3 User Community Detection in the Future

The main goal of our work is to predict the user communities whose user members share similar temporal expositions toward similar topics of interest in the future graph $G_{\mathrm{T}+1}$. To do so, we first need to estimate the future graph $G_{\mathrm{T}+1}$. Based on our model, the topical similarity between two users depends only on their latent representations. In other words, the more two latent

representations for a pair of users are close, the more similar the users are in terms of topics of interest. As a result, given $\forall u, v \in \mathbb{U} : \mathbf{y}_{u(\mathrm{T}+1)}$ and $\mathbf{y}_{v(\mathrm{T}+1)}$, we are able to predict future graph $G_{\mathrm{T}+1} = (\mathbb{U}, \mathbb{E}_{\mathrm{T}+1}, w)$ assuming $w(u, v : \mathrm{T}+1) = \mathbf{y}_{u(\mathrm{T}+1)}\mathbf{y}^{\top}_{v(\mathrm{T}+1)}$. However, $\mathbf{y}_{u(\mathrm{T}+1)}$ and $\mathbf{y}_{v(\mathrm{T}+1)}$ are not available and have to be approximated based on temporal latent representations up until time interval T. We assume a user's latent representation at time T to be the proxy of her latent representation at time T+1, as suggested by Zhu et al. [120], i.e.,

$$\mathbf{y}_{u(\mathrm{T}+1)} \simeq \mathbf{y}_{u\mathrm{T}} \tag{5.6}$$

$$w(u, v : \mathrm{T}+1) = \mathbf{y}_{u\mathrm{T}}\mathbf{y}^{\top}_{v\mathrm{T}} \tag{5.7}$$

where $\mathbf{y}_{u\mathrm{T}}$ encapsulates the latent representations of all graph snapshots from 1 up until T-1.

Now, given $G_{\mathrm{T}+1}$, we employ a graph partitioning heuristic to extract clusters of users that form our final user communities in the future. We leverage the Louvain method as it is a linear heuristic for the problem of graph partitioning based on modularity optimization. Louvain can be applied to weighted graphs, does not require *a priori* knowledge about the number of communities, and is computationally efficient on large graphs [91]. The application of Louvain on $G_{\mathrm{T}+1}$ produces a set of induced subgraphs such as $G_{\mathrm{T}+1}[\mathbb{C}]$ whose vertex set $\mathbb{C} \subset \mathbb{U}$ and edge set consists of all of the edges in $\mathbb{E}_{\mathrm{T}+1}$ that have both endpoints in $\mathbb{C}$. Subgraphs with $|\mathbb{C}| \geq 2$ form instances

of user communities.

5.3 Summary

In this chapter, two methods, namely Granger regression (G-regression) and temporal latent space modeling, have been proposed to address **RQ4**, i.e., whether it is possible to predict future-yet-unobserved content-based communities on social networks. Specifically, given a sequence of users' contributions towards a set of topics from time interval 1 to T, the objective is to predict topical user communities in a future time interval T + 1. While the G-regression method exhibits promising user communities in the future, its running complexity is intractable in practice. Temporal latent space modeling, however, is not only computationally efficient but also could outperform the G-regression method.

5.4 Related Publications

- Negar Arabzadeh, Hossein Fani, Fattaneh Zarrinkalam, Ahmed Navivala and Ebrahim Bagheri; "Causal Dependencies for Future Interest Prediction on Twitter." *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018.*

- (Submitted) Hossein Fani, Ebrahim Bagheri and Weichang Du; "Tem-

poral Latent Space Modeling for Community Prediction." *In the 42nd European Conference on Information Retrieval*, 2020.

Chapter 6

Evaluation

In this thesis, we seek to answer four research questions that would provide insight into the role of temporal analysis on the quality of the identified user communities in online social networks. The four research questions (RQ) have been formulated as follows:

- **RQ1.** Does the consideration of temporal evolution of users' topics of interest lead to higher quality communities compared to when time is overlooked?
- **RQ2.** Do temporal content-based methods lead to higher quality communities compared to link-based methods?
- **RQ3.** Do temporal content-based and link-based methods have synergistic impact on each other and reinforce the quality of the identified communities when applied in tandem?

- **RQ4.** Is it possible to predict future-yet-unobserved content-based communities on social networks?

As such, four approaches have been proposed to address the proposed research questions:

- Detecting user communities based on users' temporal social content using multivariate time series analysis to address **RQ1** and **RQ2**;
- Detecting user communities based on users' social network structure and temporal social content using neural embedding to address **RQ1**, **RQ2**, and **RQ3**;
- Predicting user communities in future yet-to-be-observed time intervals using Granger regression method to address **RQ4**; and
- Predicting user communities in future yet-to-be-observed time intervals using temporal latent space modeling to address **RQ4**.

In this chapter, we describe our experiments to evaluate the proposed approaches in terms of the dataset, experimental setup, evaluation methodology, gold standard, and metrics.

6.1 Dataset

The testbed to evaluate the proposed approaches includes a publicly available Twitter dataset. The dataset is collected and published by Abel et al. [2].

It consists of approximately 3 milion tweets in English posted by 135,731 unique users between November 1 and December 31, 2010. In addition to its text, each tweet includes a user id and a timestamp. Additionally, we collected the followership networks of the users using the Twitter API. The whole two-month time period is sampled on a daily basis, i.e., T = 61 days. Figure 6.1 depicts the overall and temporal distributions of different types of tweets and Figure 6.2 depicts the number of tweets per user in this dataset. Twitter suffers from participation inequality, where a minority of users usually contribute the most while the others just free-ride. The statistics show that only 15% of the users contribute more than 16 tweets within the whole two-month period. There are 1% of users who actively participate by posting at least a tweet. There are also other datasets available for user community detection evaluation such as the ones released by Liang et al. [64] and Yang et al. [110]. However, they include no tweets but the ids due to the Twitter's Developer Policy. Indeed, the actual textual body of tweets should be retrieved from Twitter API which is costly in terms of time.

6.2 Finding Topics of Interest ($\mathbb{Z}$)

Topic detection is the initial step to all proposed methods in this thesis. Applying topic modeling methods such as LDA or ToT to extract topics from tweets suffers from the sparsity problem [84, 97, 21] because they are designed for regular documents and not short, noisy and informal texts like tweets.

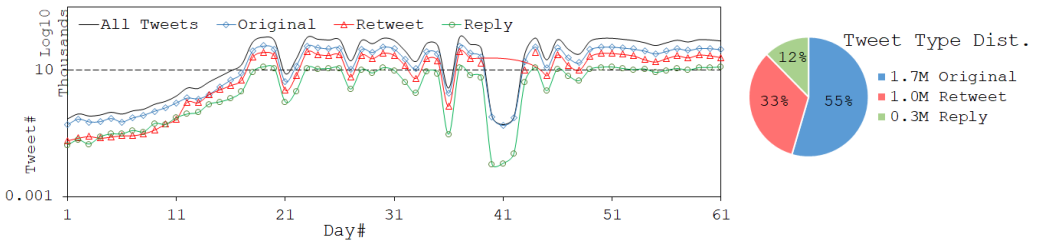

Figure 6.1: Temporal distribution of different types of tweet from November till end of December 2010 in Abel et al. [2]'s dataset .

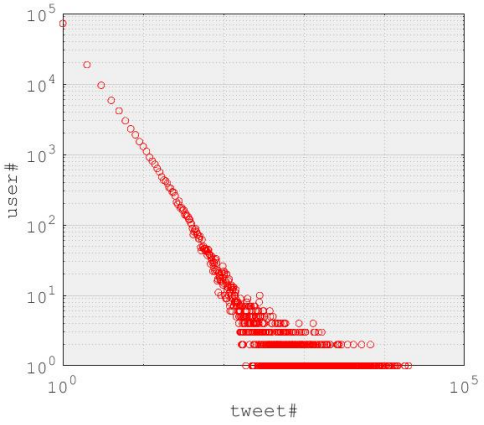

Figure 6.2: The user distribution by the number of tweets from November till end of December 2010 in Abel et al. [2]'s dataset.

As suggested in [97, 37, 98], to obtain better topics from Twitter without modifying the standard topic detection methods, we annotate each tweet with concepts defined in Wikipedia using an existing semantic annotator. We see each concept as a term in the set $\mathbb{W}$. For instance, for a tweet such as *'NATO Leaders Seek Time on Afghan Exit Strategy - http://nyti.ms/cMMDuR'*, a semantic annotator such as TAGME [37] is able to identify and extract several Wikipedia entities, namely '`NATO`'[1], '`Afghanistan`', and '`Exit_Strategy`'.

[1] `en.wikipedia.org/wiki/NATO`

Using entities instead of words can lead to the reduction of noisy content within the topic detection process, because each entity implicitly represents a collection of topical terms which are collectively more meaningful than a single word or a group of less coherent words [81]. We annotated the text of each tweet with Wikipedia entities using the TAGME RESTful API[2], which resulted in 350,731 unique entities.

In order to find topics $\mathbb{Z}$ in our dataset, we apply topic detection methods, described in Section 3.2, on the set of concepts extracted from the tweets. The graph-based approach for topic detection (GbT) identifies topics by grouping a set of concepts that exhibit similar co-occurrence patterns over time. Given that our Twitter dataset consists of tweets from a two-month period, we compute the pairwise similarities between daily (L=61 days) concept signals. Due to the large number of identified concepts (350K), it is expensive to measure pairwise similarity through cross-correlation between all pairs of concepts. However, a large number of signals are trivial and not informative. We screen out the trivial concepts as suggested in [36, 104]. Filtering the trivial concepts significantly reduces the number of signals down to 782 and makes the computation of concept similarities practically feasible. The remaining concepts are then clustered using the Louvain Method to form topics. We were able to find K=47 topics, which served as our topic set $\mathbb{Z}_{\text{GbT}}$.

We also use LDA and ToT to discover topics. LDA-based approaches to

[2] `services.d4science.org/web/tagme/documentation`

topic detection need *a priori* knowledge of the number of topics, contrary to GbT. Therefore, we have opted to select the topic set size for LDA and ToT based on the number of topics detected by GbT. We aggregate daily tweets of each user to form a single document. Then, we apply LDA and ToT on the constructed documents to find topics, $\mathbb{Z}_{\text{LDA}}$ and $\mathbb{Z}_{\text{ToT}}$, respectively. We have used MALLET[3] for LDA and an open-source implementation available on GitHub[4] for ToT.

Given the three extracted topic sets $\mathbb{Z}_{\text{GbT}}$, $\mathbb{Z}_{\text{LDA}}$, and $\mathbb{Z}_{\text{ToT}}$, we are interested in determining whether or not our temporal approaches can provide a more accurate representation of user communities compared to the *non*-temporal and the state-of-the-art temporal approaches. It should be noted that the main goal of our experiments is to determine the role and impact of temporality when building user communities. Therefore, we intentionally keep the parameters for topic detection methods constant (e.g. the number of topics in LDA and ToT) so as to avoid any unintended effects on the results and keep the scope of the experiments unchanged.

6.3 User Community Detection Evaluation

Contrary to small real-world social networks or synthetic ones, true gold standard user communities are not available in most cases for real world applications [19]. As such, well-defined quality measures such as Rand index,

[3] `http://mallet.cs.umass.edu/topics.php`
[4] `http://github.com/ahmaurya/topics_over_time`

Jaccard index, or normalized mutual information (NMI) that require comparison to the gold standard cannot be used for evaluation. On the other hand and in the absence of a gold standard, quality functions such as modularity [40] are not helpful either since they are based on the explicit links between the users (structural). In our approach and the baselines, the links between the users are inferred through a learning process and are not always explicit. For instance, a near perfect method may result in a low modularity because graph edges are sparse and do not form densely connected user sets. Conversely, a weak method may connect topically dissimilar users together forming communities of users that do not share similar interests but result in a high modularity. So, the communities that achieve high structural quality in an inferred similarity graph are not necessarily optimal [72].

Fortunately, the performance of community detection methods can be measured through observations made at the application level, as suggested in [19, 72]. In these evaluation strategies, a user community detection method is considered to have better quality *iff* its output communities improve an underlying application such as retweet prediction [117], timestamp prediction [101], news recommendation [2] and user prediction. We deploy news recommendation and user prediction. By using these applications, we explore whether and which community detection method is able to provide stronger performance compared to the other state of the art community detection techniques and hence systematically answer the four research questions.

To this end, we curate a gold standard dataset, which consists of the set

of news articles that have been mentioned in the users' tweets. The reason we collect such a gold standard is because it can be safely assumed that users would only post links to news articles if they are interested in the topic of that news. Given our work is based on entity mentions in tweets, we also semantically annotate the news articles that have been collected in the gold standard dataset. Each entry in the gold standard can be viewed as a triple (u, a, t), which refers to user u posting article a at time interval t. Formally, the gold standard is defined as $\mathbb{G} = \{(u, a, t) : u \in \mathbb{U}, a \in \mathbb{D}, 1 \leq t \leq \mathrm{T} = 61\}$ where $\mathbb{U}$ and $\mathbb{A}$ are the set of users and news articles, respectively. In our experiments, the gold standard consisted of 25,756 triples derived from 3,468 articles shared by 1,922 users. To avoid leakage, tweets which include a URL in the gold standard have been removed from the training set. It is worth noting that almost half of the tweets in our dataset include at least one URL, precisely 1,437,713 out of 2,948,742 tweets with 787,680 unique URLs, among which we could only crawl 3,468 news articles to build the gold standard. This leads to removing 13,742 tweets and left 2,935,000 tweets for training purposes.

6.3.1 News Recommendation

Our first set of experiments rely on the assumption that an accurate clustering of users into communities would place those users who have similar topical interest evolution over time next to each other in the same community. As such, recommending news articles to the users of the same community

should be possible and effective due to the similarity between user interests. Based on the gold standard, an effective recommendation for a user would be one that has been observed as one of the triples $(u, a, t) \in \mathbb{G}$. In order to make recommendations based on the identified communities, we perform the following two steps:

1. We consider each identified community separately in every time interval t $(1 \leq t \leq \mathrm{T} = 61)$ and compute the selected community's overall topic of interest at that time. The overall topics of interest for a community are calculated by using the sum of topic preference time series of all users in that community. More formally, they are computed as $\sum_{u \in \mathbb{C}} x_{uz,t}$. All news articles in the gold standard are ranked descendingly based on their cosine similarity to the overall topic of interest for the community in each time interval.

2. Each user member of a community is recommended the ranked list of news articles that are assigned to the community.

The news recommendation application will perform best when the users that are placed within the same community exhibit the same temporal topical interests and hence are interested in similar news articles at each time interval; therefore, it is a suitable extrinsic evaluation method to measure how well the community detection method has been able to effectively partition users into different communities based on their temporal interests.

Metrics. We evaluate the ranked list of news articles for recommen-

dation by standard information retrieval metrics: Precision at rank k (P_k), Mean Reciprocal Rank (MRR), and Success at rank k (S_k). P_k is the proportion of relevant news articles in the top-k recommended items:

$$\mathrm{P}_k = \frac{1}{|\mathbb{U}|} \sum_{u \in \mathbb{U}} \frac{tp_u}{k} \tag{6.1}$$

where tp_u (true positive) is the number of relevant news articles for user u in her top-k rank list of recommendations. MRR is the inverse of the first position that a correct item occurs within the ranked list,

$$\mathrm{MRR} = \frac{1}{|\mathbb{U}|} \sum_{u \in \mathbb{U}} \frac{1}{rank_u} \tag{6.2}$$

where $rank_u$ refers to the rank position of the first relevant news article for the user u. S_k shows the probability that at least one correct item occurs within the top-k items of the ranked list:

$$\mathrm{S}_k = \frac{1}{|\mathbb{U}|} \sum_{u \in \mathbb{U}} (rank_u \leq k) \tag{6.3}$$

In case $k = 1$, S_1 would be equal to P_1.

6.3.2 User Prediction

We perform a second set of experiments based on the user prediction application. Given the gold standard $\mathbb{G}$ and the user communities $\mathbb{P}$, this time the

goal is to predict which users posted a news article a at time interval t. To do so, we find the closest community to the news article in terms of topics of interest at time interval t. This is done based on the cosine similarity of the community's overall topics of interest at time t and news article a. Then, the members of the community would constitute the predicted users. The logic behind why this approach helps us qualify the output communities of the different approaches is the same as the news recommender application. However, while the performance of the news application is evaluated based on information retrieval metrics, the user prediction application is evaluated based on *classification* metrics.

Metrics. We adopt three standard classification metrics, i.e., Precision, Recall, and F-measure, to report user prediction performance. Precision is the probability that a predicted poster of a news article is the actual poster of the article:

$$\text{Precision} = \frac{tp}{(tp + fp)} \tag{6.4}$$

where tp is the true positive count, i.e., the number of users correctly assigned to the news article and fp is the false positive count, i.e., the number of users assigned incorrectly. Recall, or hit rate, is the probability that a true poster of a news article has been correctly assigned to the posted news article:

$$\text{Recall} = \frac{tp}{(tp + fn)} \tag{6.5}$$

where fn is the false negative count, i.e., the number of actual posters that

have not been assigned to their posted news articles. F-measure is the harmonic mean of Recall and Precision and is defined as:

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6.6}$$

6.3.3 Baselines

In our work and in order to answer research questions **RQ1**, **RQ2**, and **RQ3** with respect to the user community detection, we systematically compare the following baseline methods. For all comparisons, the paired statistical significance has been carried out using Student's t-test at the 5% significance level, unless otherwise explicitly stated.

6.3.3.1 CD

This is a *non*-temporal content-based community detection baseline. We build non-temporal content-based communities over the set of users. We project daily user-topic contribution time series of each user to the topic space by aggregating the values over the whole time period. Then, we calculate the topic-based similarity of users based on the cosine similarity of their corresponding topic vectors. Finally, we create a weighted graph over the users and their pairwise similarity and apply Louvain in Pajek[5] to find communities. Pajek is a social network analysis software that facilitates quantitative or qualitative analysis of social networks through either numerical or

[5] `vlado.fmf.uni-lj.si/pub/networks/Pajek/`

visual representation. Pajek's key strengths are its impressive visualization capabilities and its ability to handle very large networks comprising several million vertices.

6.3.3.2 TCD-Timeseries

Proposed in Chapter 3, this is a temporal content-based approach which models users' contributions toward the topics of interest within time through a multivariate time series. The approach uses two dimensional cross correlation to measure the similarity of a pair of users' time series. We use the implementation in MATLAB for calculating time series cross-correlation[6]. Finally, we use Louvain in Pajek for its community detection step.

6.3.3.3 TCD-Embedding

Proposed in Chapter 4, this is a temporal content-based method based on temporal user embeddings that does not consider the network structure proposed in Section 4.2. This baseline could be considered as a variation of user embedding interpolation in Equation 4.9 where $\alpha = 1$ to filter out link-based embeddings.

We adopt the implementation of triCluster[7] [118] to find the regions of like-mindedness in users' topic contribution time series. The condition for homogeneity c is set based on two alternatives: c_1; the regions are considered homogeneous if the difference of their values falls in the range [0, 0.1), and

[6] www.mathworks.com/help/signal/ref/xcorr2.html

[7] `www.cs.rpi.edu/ zaki/software/TriCluster.tar.gz`

c_2; the regions are considered homogeneous if their values are greater than the threshold 0.1 (LDA's alpha prior). We proceed to extend the CBOW architecture in Gensim[8] to learn user vector representations. The training phase uses a learning rate of 0.025 and in each epoch we decrease it by 0.002 for 200 epochs. The window size for the representation learning process is set to 2. We perform the experiments on different vector sizes of $d = \{100, 200, ..., 500\}$.

6.3.3.4 GrosToT

This is a temporal content-based generative process for topics and communities proposed by Hu et al. [50]. The number of topics is set to $\mathbb{Z} = 50$ and we perform experiments on increasing numbers of communities for C={5, 10, ..., 30} untill we see no performance gain. The number of iterations is set to 1,000. This method is a mixture model in which all users are members of all communities with a probability distribution. In our comparison, we only consider the community with the highest probability as the user's community.

6.3.3.5 Link-CD

This is a link-based method based on link-based user vector representations, proposed in Section 4.3, which does not consider user content. This baseline could be considered as a variation of user vector interpolation in Equation 4.9

[8] `radimrehurek.com/gensim/models/word2vec.html`

where $\alpha = 0$ to filter out temporal content-based embeddings.

In order to infer the user embeddings from the social network structure $\mathcal{G} = (\mathbb{U}, \mathbb{A})$ whose vertices are users $\mathbb{U}$ and edges are ordered pairs of user elements such as $(u, v) \in \mathbb{A}$, we use the formulation presented in Section 4.3 owing to its scalability ($O(|\mathbb{U}|)$) and unsupervised representation learning as opposed to more sophisticated neural-based graph embedding techniques such as deep neural graph representations (DNGR) [18] and structural deep network embeddings (SDNE) [99] with higher time complexity ($O(|\mathbb{U}|^2)$ and $O(|\mathbb{U}||\mathbb{A}|)$, respectively). Graph convolutional networks (GCN) [56] with running complexity of $O(|\mathbb{A}|)$ and its variations [108] are the state of the art in inductive tasks, i.e., they are able to generalize previously unseen users, which is crucial in evolving social networks. In our work, however, we assume that the social network structure $\mathcal{G}$ remains stationary and, hence, employing GCN-based methods does not add much value to our experiments.

We created 10 random walks of length $l \in \{40, 80\}$ for each user and the window size for the training process is set to $\{5, 10\}$ while the learning rate and the number of epoches are set to 0.002 and 200, respectively. The return (p) and in-out (q) parameters are set to a default value 1.

6.3.3.6 TCD(α)-Embedding

Proposed in Chapter 4, this baseline interpolates temporal content-based embeddings with the link-based ones based on Equation 4.4, proposed in Section 4.4, where $\alpha \in \Re^{[0,1]}$.

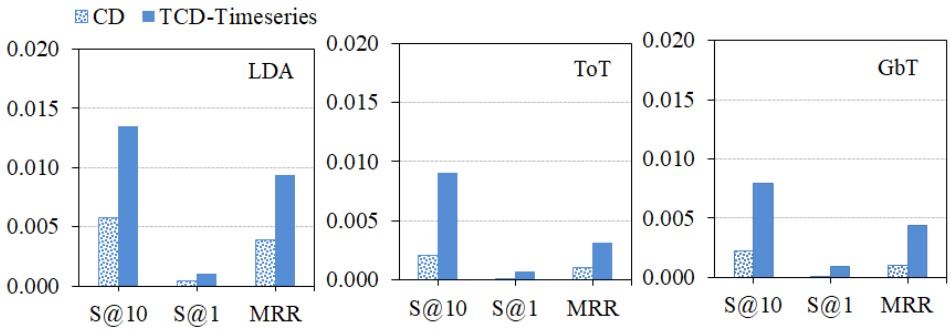

Figure 6.3: The performance of the proposed multivariate user time series method (TCD-Timeseries) and non-temporal community detection method (CD) in the context of the news recommendation application using different topic detection methods, LDA [14], ToT [101], and GbT. The ranking metrics and their amplitude are shown in horizontal and vertical axes respectively.

6.3.4 RQ1: TCD-Timeseries vs. CD

We begin by considering research question **RQ1**, i.e., whether the consideration of temporal evolution of users' topics of interest lead to higher quality communities compared to when time is overlooked. Here, we compare our time series approach, proposed in Chapter 3, against non-temporal baseline (CD).

Figure 6.3 summarizes the performance of the proposed time series approach against non-temporal baselines in terms of MRR, S@1 and S@10 in the context of a news recommendation application as explained in Section 6.3.1. As shown, the time series-based community detection (TCD-Timeseries) method along with different topic detection methods, GbT, LDA, and ToT outperform the non-temporal counterparts in all metrics. This means that incorporating temporal aspects of users' interests for extracting like-minded

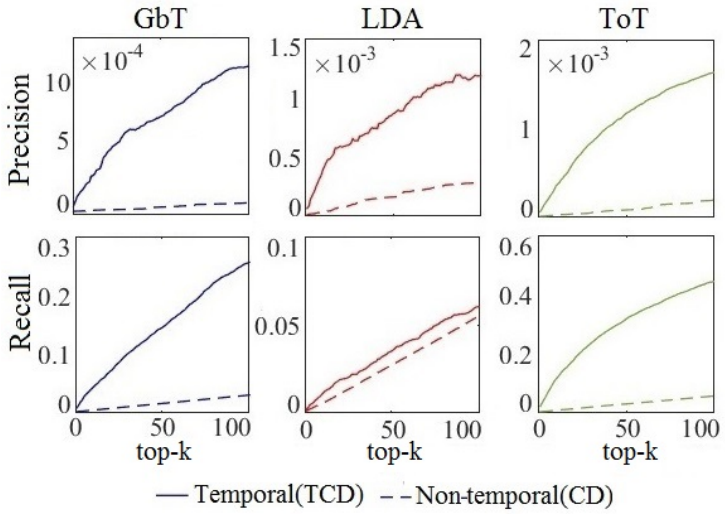

Figure 6.4: The performance of the proposed multivariate user time series method (TCD-Timeseries) and non-temporal community detection method (CD) in the context of the user prediction application using different topic detection methods, LDA [14], ToT [101], and GbT. The x axis shows the number of top communities which are selected for the task of user prediction.

communities leads to more cohesive communities that consequently results in better news recommendations. This characteristic allows us to make recommendations to users that are topically relevant and made at the appropriate time.

We evaluate and compare the quality of user predictions made based on the time series-based approach and non-temporal baseline. The results are presented in Figure 6.4 for top-k; $1 \leq k \leq 100$.

As seen, methods which use time series-based temporal community detection (TCD-Timeseries) unanimously outperform the ones which use non-temporal community detection (CD) in terms of precision and recall. This

reinforces our hypothesis that the communities that are built using the time series-based approach are topically and temporally coherent such that when the user who mentions a certain news article needs to be identified at time interval t, both the content and time of the news can be taken into consideration to make more accurate predictions. On the other hand, while non-temporal communities do consider the topic of the news article, they fail to take time into account. This will result in many false positives when predicting the user because while a user may have had interest in a certain topic in previous time intervals, she might have lost interest later and therefore naturally be much less likely to post about that topic as time passes.

For instance, let us consider the three sample Twitter users again. All users `@joe`, `@john` and `@mary` are interested in the 'War in Afghanistan' topic (z_{44} in Figure 3.1) but with a one month time difference. As was observed, `@joe`, `@john` show their interest in the topic in November whereas `@mary` did so in December. Now, if a news article has been observed on Twitter talking about 'War in Afghanistan' on December 17, it is very likely that `@mary` is the user who is posting this news as opposed to the other two users. The same logic applies if the same news article has been seen but on November 25. This time the likelihood of `@joe` or `@john` posting this news is much higher. As it turns out in our experiments, the non-temporal community detection methods were not able to make a distinction between the three users and would hence predict all users to be the posters in both cases.

In summary, in response to the **RQ1**, i.e., whether the consideration of

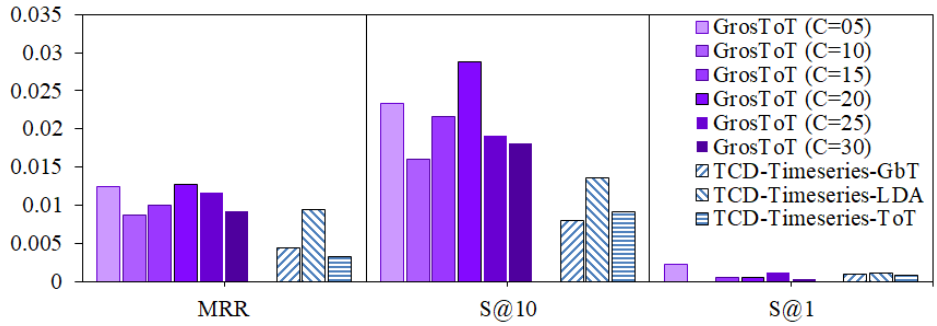

Figure 6.5: The performance of the proposed multivariate user time series method (TCD-Timeseries) using different topic detection methods, LDA [14], ToT [101], and GbT and the state of the art (GrosToT [50]) in the context of the news recommendation application. The ranking metrics and their amplitude are shown in horizontal and vertical axes respectively.

time plays a role in the quality of the identified communities or not, the comparative result shows that considering users' temporal behaviour is an influential contributor to the identification of high quality user communities in the context of news recommendation and user prediction.

6.3.5 RQ1: TCD-Timeseries vs. State of the Art

To compare the time series-based approach with the state of the art, we run GrosToT [50] on the ground truth. Figure 6.5 depicts the performance of GrosToT as the number of communities changes compared with the time series approach. As shown, LDA variant of TCD-Timeseries achieve competitive performance compared with GrosToT but not statistically significant where TCD-GbT and TCD-LDA show poorer performance on MRR and S@10, respectively. The reason for this better performance by GrosToT

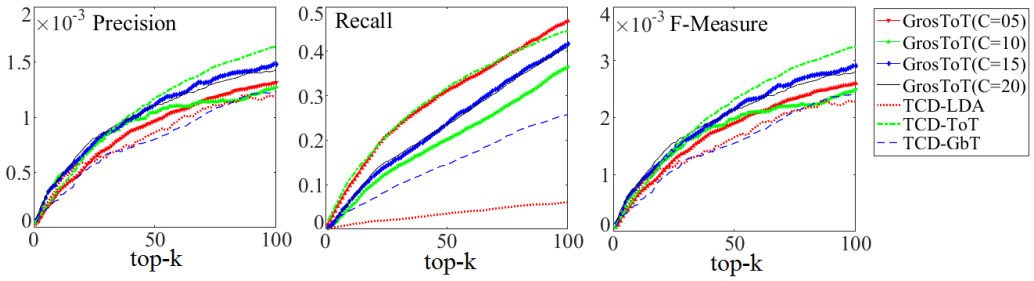

Figure 6.6: The performance of the proposed multivariate user time series method (TCD-Timeseries) using different topic detection methods, LDA [14], ToT [101], and GbT and the state of the art (GrosToT [50]) in the context of the user prediction application. The ranking metrics and their amplitude are shown in x and y axes respectively.

could be the fact that the time series representation of the users suffers from sparsity and is not able to capture both topical and temporal disposition of users more effectively.

We have evaluated the performance of the state-of-the-art competitor, i.e., GrosToT, in the task of user prediction as well. The results are presented in Figure 6.6 in terms of precision, recall, and f-measure. As observed, GrosToT with different numbers of communities does not show a coherent performance. While GrosToT with five communities (C = 5) shows better performance in terms of recall compared with the other GrosToT variations, GrosToT with C = 15 and C = 20 communities show higher precision. When comparing GrosToT with the best performing variation of the proposed time series approach, i.e., TCD-Timeseries with ToT, one can make two observations:

1. TCD-Timeseries and GrosToT(C = 5) show competitive performance

in terms of recall. However, it should be noted that higher recall values for GrosToT(C = 5) are expected given the fact that a lower number of communities will essentially group users in fewer clusters hence producing higher recall. However, when looking at the precision for GrosToT(C = 5), it can be seen that this higher recall has come at the cost of a much lower precision compared with TCD-Timeseries.

2. In terms of precision, both TCD-Timeseries and GrosToT(C = 20) show competitive performance. GrosToT(C = 20) performs slightly better for top-k; $k \leq 35$, whereas TCD-Timeseries shows slightly better results for $k > 35$. Overall, when considering f-measure, GrosToT(C = 20) and TCD-Timeseries show very competitive performance while TCD-Timeseries outperforms for top-k; $k > 35$.

In summary, our time series approach and the state of the art (GrosToT) in temporal user community detection show competitive performance in the context of news recommendation and user prediction with respect to the **RQ1**. While our time series approach could not outperform the state of the art, which we attribute to sparsity in topic space in user-topic contribution time series, our second alternative approach, i.e., user neural embedding, is able to outperform the state of the art as shown in the next section. LDA has been selected as the topic detection method in all the following experiments since its use with our time series method outperforms the other variations, i.e., using GbT and ToT as underlying topic detection methods.

6.3.6 RQ1: TCD-Embedding vs. State of the Art

As we show in Figure 4.1, it is very unlikely that two users have the exact same probability value for a given topic in the same time interval. As such, we have introduced the condition of homogeneity to relax the condition for matching users with each other in a given time interval over some topics. One option (c_1) for defining the condition of homogeneity is to allow for slight variations between topic contributions by different users. For instance, we could allow the difference to be in a certain range, e.g., [0, 0.1), which is the strategy that we have adopted in the previous set of experiments reported earlier. It is alternatively possible to define the condition of homogeneity (c_2) in a way that two values would be considered similar if they both have a value greater than a given threshold, e.g., LDA's alpha prior 0.1 , as shown in Figure 4.2(a). This way, we are treating values as binary values; hence, a pair of users with a degree of interest towards a given topic at a same time interval are considered like-minded only if both users have a value greater than the threshold. We have additionally performed experiments with this alternative condition of homogeneity, denoted by the '-b' suffix in the figures.

In order to explore whether the explicit embedding of time within users' vector representations lead to higher quality communities compared to when time is incorporated into a generative process, we compare the temporal content-based baselines, namely GrosToT [50] and TCD-Embedding in which only temporal user vector representations have been utilized, in Figure 6.7. As shown, TCD-Embedding achieves better performance compared with the

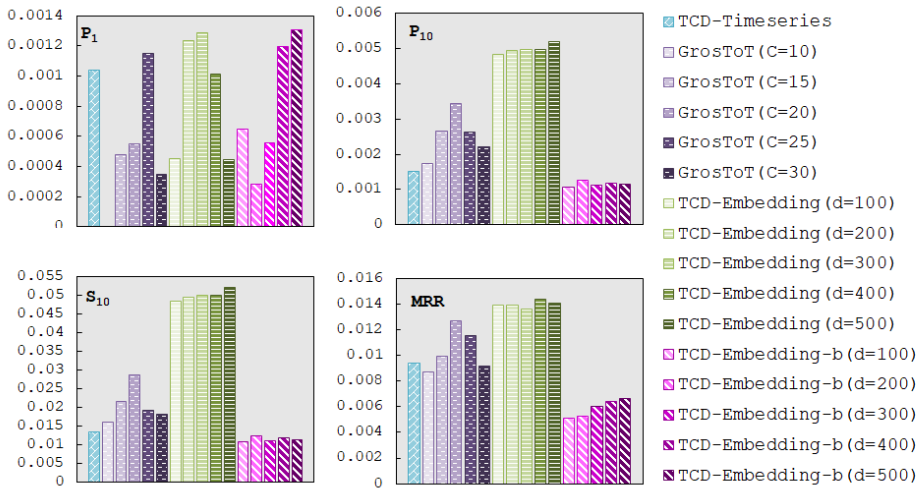

Figure 6.7: The performance of the proposed neural user embedding method (TCD-Embedding) under two alternative conditions of homogeneity using LDA [14] compared to the state of the art (GrosToT [50]) in the context of the news recommendation application. The vertical axis show the amplitude of the ranking metrics.

temporal approach proposed by Hu et al. (GrosToT) for different dimension sizes. Specifically, the result shows that TCD-Embedding with $d = 300$ is the best and GrosToT is the runner up. We attribute the better performance of TCD-Embedding to the fact that the embedding function preserves both topical and temporal proximity of users more effectively and, consequently, the extracted user communities capture temporal content-based similarity of users more coherently than the other two baselines. This demonstrates the effectiveness of explicitly embedding time into user vector representations. Based on the results in Figure 6.7, we conclude that the explicit embedding of time in user vector representations leads to higher quality user communi-

ties compared to when time is incorporated as a component in a generative process.

Comparing the two alternative conditions of homogeneity, we observe that c_1 outperforms c_2. As seen in Figure 6.7, TCD-Embedding-b is not even able to outperform GrosToT baseline. The reason might be the fact that considering two users who have a value more than a threshold (0.1, LDA's alpha prior) to be like-minded regardless of their degrees of interest has a confounding effect on the user embeddings. That is, dissimilar users end up with close embeddings and finally become members of the same communities.

Similar to the news recommendation task, we seek to evaluate the performance of our embedding method (TCD-Embedding) vs. State of the art (GrosToT) but in the context of the user prediction application. We summarize the performance of user prediction for temporal baselines in terms of classification metrics in Figure 6.8. As shown, TCD-Embedding outperforms other baselines in all metrics (except for $d = 100$). This reinforces the fact that when time is explicitly embedded in the user representations that it will lead to higher quality communities compared to representations that incorporate time within a generative process. Contrary to the news recommendation application where TCD-Embedding did not outperform GrosToT baseline, in the user prediction application it performs better. In summary, in user prediction, our method is able to outperform the state of the art regardless of the condition of homogeneity.

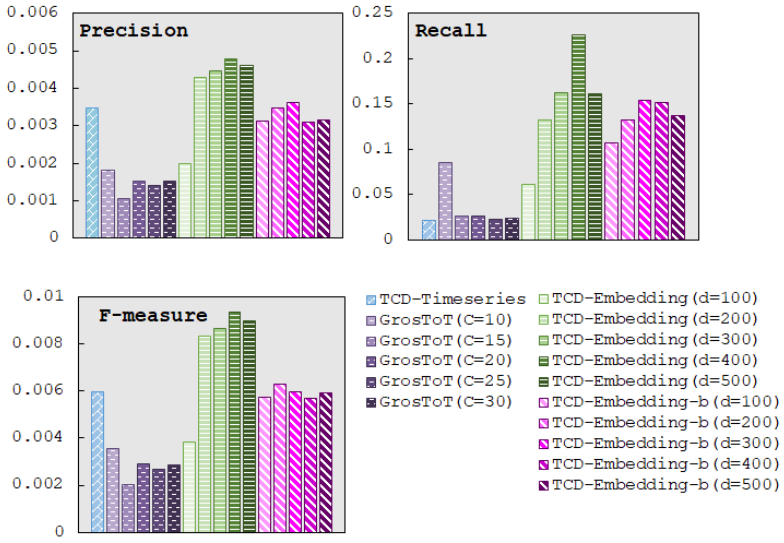

Figure 6.8: The performance of the proposed neural user embedding method (TCD-Embedding) under two alternative conditions of homogeneity using LDA [14] compared to the state of the art (GrosToT [50]) in the context of the user prediction application. The vertical axis show the amplitude of the classification metrics.

6.3.7 RQ2: TCD-Embedding vs. Link-CD

In order to answer research question **RQ2**, i.e., whether temporal content-based user community detection methods show better performance compared to link-based methods, we compare the quality of the output communities in Figure 6.9. As seen, linked-based methods (Link-CD) show their best performance with $d = 300$ and a random walk length $l = 80$ but still perform worse than the poorest version of TCD-Embedding with $d = 100$. As an example, all the variations of Link-CD produce *zero* in terms of P_1. This points to the fact that link-based methods produce lower quality communities compared to temporal content-based counterparts.

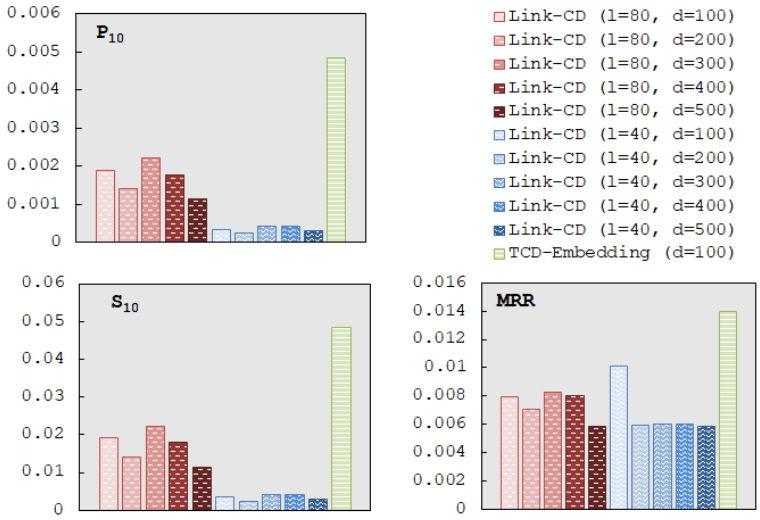

Figure 6.9: The performance of link-based community detection baseline (Link-CD) vs. worst case of TCD-Embedding ($d = 100$) in terms of ranking metrics in the context of the news recommendation application. The vertical axis shows the amplitude of the metrics. All Link-CD variations had a performance of *zero* in terms of P_1.

Likewise, Figure 6.10 shows that the temporal content-based user community detection methods outperform link-based methods in the user prediction application. Specifically, the best link-based baseline (Link-CD with $d = 300$ and random walk length $l = 80$) performs worse than the poorest version of LDA-TCD with $d = 100$. This reinforces our findings in the news prediction application that link-based methods produce lower quality communities compared to content-based baselines.

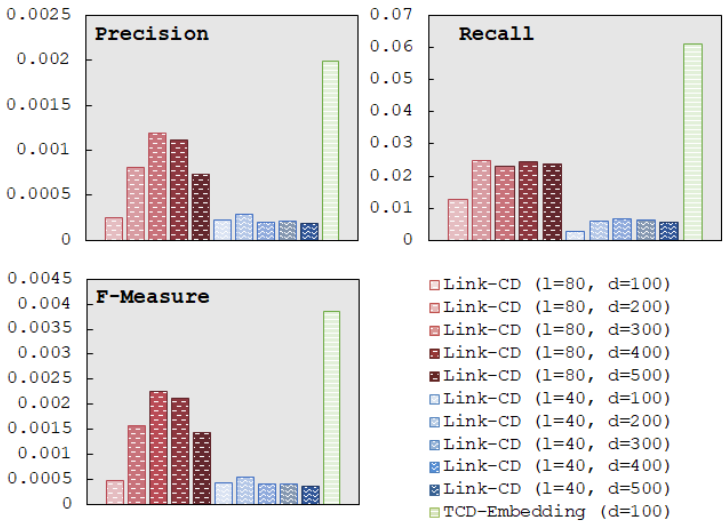

Figure 6.10: The performance of link-based community detection baseline (Link-CD) vs. worst case of TCD-Embedding ($d = 100$) in terms of classification metrics in the context of the user prediction application. The vertical axis shows the amplitude of the metrics.

6.3.8 RQ3: TCD(α)-Embedding vs. TCD-Embedding

In order to answer research question **RQ3**, i.e., whether link-based and temporal content-based community detection methods have a synergistic effect on each other, we use TCD(α)-Embedding in which the user vector representations from temporal social content are interpolated with link-based ones. As both types of user representation yield best results for user communities at $d = 300$, i.e., TCD-Embedding ($d = 300$) and Link-CD ($l = 80, d = 300$), we investigate the effect of social structure in temporal user community detection only for user vector representations of size $d = 300$ in TCD(α)-Embedding. Figure 6.11 shows the results for decreasing values of α in order to show

the impact of link-based methods on improving the quality of content-based methods. As shown, we start with $\alpha = 1$ where there is no link-based user vector representation involved and the representation is essentially equivalent to TCD-Embedding. As we gradually put more weight on the link-based user vector representation, the results improve up to an extremum, which happens at $\alpha = 0.6$. This demonstrates the fact that the link-based user representation is helping with user community detection and identifying user relationships that cannot be otherwise derived based solely on user content. However, the impact of link-based user embeddings needs to be controlled as the increase in the weight of the link-based user representation beyond $\alpha = 0.6$ leads to declining community quality.

In order to answer research question **RQ3** with regards to the synergistic impact of temporal content-based and link-based user embeddings in the user prediction application, similar to the new prediction application, we employ a TCD(α)-Embedding baseline with an embedding dimension size of $d = 300$. Figure 6.12 shows the results for decreasing values of α. The left corner of each diagram in Figure 6.12 represents the performance of TCD-Embedding due to $\alpha = 1$ and as such no link-based user vector representation is involved. As seen, the gradual increase in the weight of the link-based user representation leads to improved performance up to $\alpha = 0.5$ and 0.6 for $l = 80$ and $l = 40$, respectively. However, we observe declining performance as α decreases untill the end when TCD(α)-Embedding becomes a pure Link-CD method at $\alpha = 0$. This demonstrates the fact that while link-based user

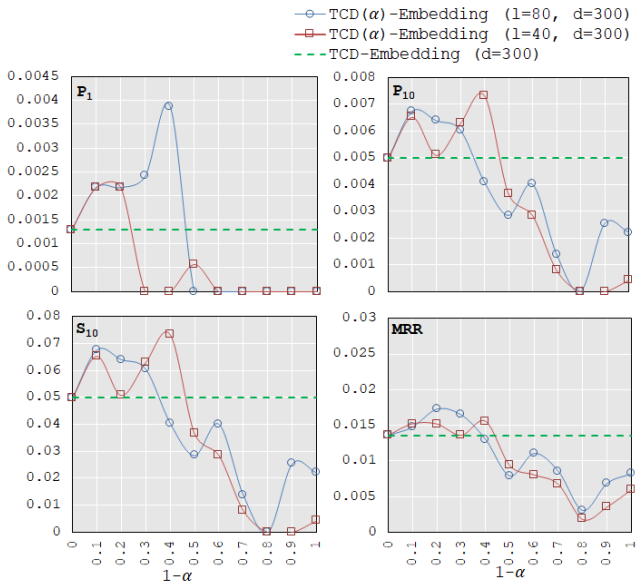

Figure 6.11: The performance of user communities through linear interpolation of temporal content-based and link-based user vector representations of size $d = 300$ in the context of news recommendation. The vertical axis shows the amplitude of the ranking metrics.

representations alone do not produce high quality user communities, they can help improve the performance of content-based methods if interpolated effectively.

6.4 RQ4: User Community Prediction Evaluation

Similar to user community detection, we evaluate the proposed methods for user community prediction at the application level. In this evaluation strategy, a community prediction method is considered better *iff* its output

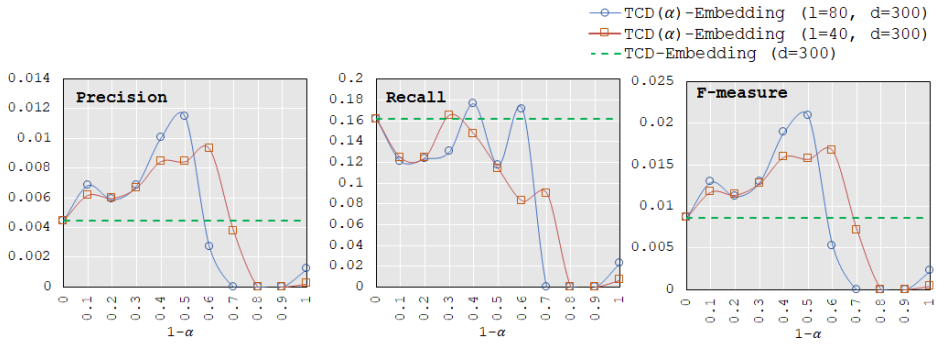

Figure 6.12: The quality of the identified user communities as a results of the linear interpolation of link-based and temporal content-based user vector representation in TCD(α)-Embedding in the context of the user prediction application. The vertical axis shows the amplitude of the classification metrics.

communities in future time intervals improve an underlying application. We deploy two applications, namely news recommendation, and user prediction. By using these applications, we explore whether our proposed method is able to provide stronger performance compared to the state of the art. We build the gold standard from a set of news articles whose URLs have been posted by user u at time T+1. We see each entry as a triple (u, a, T+1) consisting of the news article a, user u, and the time interval T+1 to form our gold standard.

We compare the quality of user communities predicted by our proposed method against the baselines in the context of news recommendation and user prediction applications.

News Recommendation. To evaluate user communities in the future in the context of the news recommender application, we recommend news arti-

cles in two steps:

1. For each community $\mathcal{C}$, we recommend news articles in a ranked list based on the similarity of the article a and the community's overall topic preference vector at time T+1. The overall topic preference vector for a community is the sum over all users' topic preference vectors belonging to the community, i.e., $\sum_{u \in \mathcal{C}} \mathbf{x}_{u(\mathrm{T}+1)}$.

2. We recommend the news article a to a user $u \in \mathcal{C}$ based on the same ranked list as her community's list. A true community is one whose members are interested in the same topics of interest in the future. As a result, at time T+1, a news article is about the same topics of interest as the community's overall interests *iff* all the members post about the same or similar news articles.

We evaluate the recommended list of news articles using standard retrieval metrics such as MRR, nDCG@5, and nDCG@10.

User Prediction. The other application with which we evaluate our approach is the user prediction application. Here, given the user communities of the future, the goal is to predict which users posted a news article a at time T+1. To do so, we consider members of the *closest* community to a news article in terms of topics of interest at time T+1 to be the potential posters. We use precision (PR), recall (REC), and F-measure (F1) to report user prediction performance.

6.4.1 Baselines

In order to answer the last research question, **RQ4**, we compare the following baseline methods from four categories of temporal user community detection (GrosToT [50] and TCD-Embedding), temporal collaborative filtering (TimeSVD++ [57] and RRN [106]), regression (VAR and G-regression), and temporal latent space modeling (Chimera [5] and our proposed model).

6.4.1.1 GrosToT [50]

This the same GrosToT baseline proposed by Hu et al. [50] in user community detection but trained on users' temporal content up until the last day but one, i.e., T=60. The final communities are then used as an estimate for communities on day 61.

6.4.1.2 TCD-Embedding (d=300)

This is our neural user embedding method, which is proposed to identify temporal content-based user communities, at its best parameter settings with embedding dimension size d=300 as shown in Section 6.3.6. We trained this baseline on users' temporal content up until the day before the last in our dataset, i.e., T=60. The final communities are then used as an estimate for communities on day 61.

6.4.1.3 TimeSVD++ [57]

Temporal collaborative filtering methods are able to predict users' topics of interest in future and, hence, can be used for the task of content-based community prediction, among which we choose **TimeSVD++** as a collaborative filtering baseline. TimeSVD++ [57] is the temporal extension to SVD++. The implementation in librec[9] was used in our experiments. We performed a grid search over the bin size in {1,2,4,8,16,32,64} and factors size in {10,20,40,80} to select the best settings. Other settings were left to a default value, i.e., learning rate=0.01 and regularization $\lambda = 0.1$.

6.4.1.4 Recurrent Recommender Networks (RRN) [106]

This is a temporal collaborative filtering approach based on recurrent neural nets. We performed grid search over bin size in {1,2,4,8,16,32,64} and users and topics' dynamic states size in {10,20,40,80}. Other hyperparameters were set to a default: single-layer LSTM with 40 hidden neurons and embeddings size of 40. The implementation is kindly provided by its authors.

6.4.1.5 G-regression

Proposed in Chapter 5, this is the Granger regression method which finds the top influencers for each user with respect to each topic up until day T=60 using Granger causality. Then, a user's degree of interest toward each topic on day T+1=61 is predicted using both the user and her influencers'

[9] `www.librec.net`

temporal content in a *bivariate* var model. We performed experiments on an increasing number of influencers $k \in \{1, 2, 5, 10, 20\}$. To find communities in the future, a weighted graph is formed over the users and their pairwise similarity on predicted degrees of interest on day 61, on which Louvain is applied.

Our users' topic preference time series satisfy the Granger causality assumption of stationarity as they passed different stationarity tests, namely the Phillips-Perron [82], Dickey-Fuller [29] and KPSS [58]. The significance level and the maximum number of lags were set to 0.05 and 2, respectively. The bayesian information criterion (bic) was used to find optimal lag. We used the first 4 values of the users' topic preference time series as the pre-samples to initiate the var models estimation.

6.4.1.6 Vector Autoregression (VAR)

This approach is the vector autoregression (VAR) method to predict users' future topics of interests in a *univariate* var model. A user's degree of interest toward each topic on day T+1=61 is predicted using only the user's temporal content. To find communities on day 61, we took a similar approach as in the G-regression baseline.

6.4.1.7 Temporal Latent Space Modeling

Proposed in Chapter 5, this is the temporal latent space modeling method. We adopt the sequential (local) version of block coordinate gradient descent

proposed in [120]. By setting the temporal smoothness (regularization) parameter $\lambda = 0.01$, we performed experiments on increasing numbers of dimensions $d \in \{10, 20, ..., 100\}$ for learning temporal latent representation of users in 1,000 iterations. We apply Louvain on the estimated user graph at time T+1=61 using Pajek to identify user communities in the future.

6.4.1.8 Chimera [5]

To the best of our knowlegde, this baseline is the most related and recent baseline to our proposed temporal latent space model. Appel et al. use *shared* matrix factorization to embed social network dynamics and temporal content in a shared feature space followed by a traditional clustering technique, such as k-means, to identify user communities. We performed experiments on increasing numbers of communities for C=$\{5, 10, 20, 30\}$ and varying embedding dimensions d=$\{5, 10, 20, 30\}$.

6.4.2 Results

Foremost, we explored the impact of the size of the influencer network on the predictive power of our G-regression method and then compared it to other baselines. We evaluated the performance of G-regression for varying number of top-influencers $k \in \{1, 2, 5, 10, 20\}$ that were used in the influence network using rating metrics including mean absolute error (MAE) and root-mean-squared error (RMSE), and ranking metrics including nDCG, MAP, and P@5. As seen in Figure 6.13, the accuracy of our approach did not show

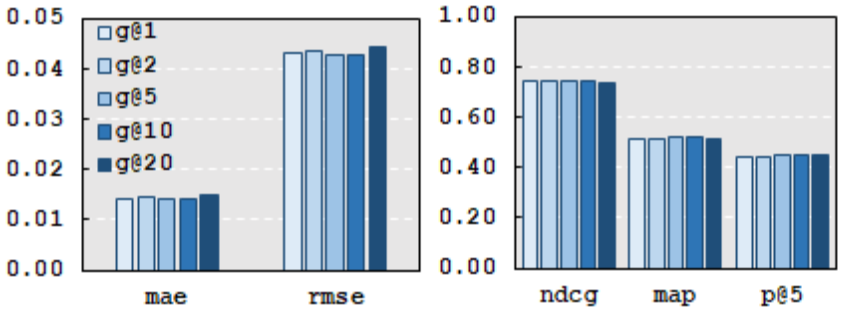

Figure 6.13: The performance of proposed G-regression method with varying number of influencers (g@k) in terms of prediction error (the lower the better) and ranking metrics (the higher the better). The vertical axis shows the amplitude of the metrics.

statistically significant improvement or deterioration on any of the metrics for different number of influencers (g@k). This can be due to two factors: *i*) the VAR model structure which is used to incorporate the k influencers' topic preference time series selects the top (k=1) influencer's topic preference time series as its salient component for all baselines $k > 1$, and *ii*) the pairwise (bivariate) Granger causality test could potentially lead to misleading influencers as mentioned by Ding et al. [70] for cases when more than one causes are considered. For instance, let c, m, and e be three users where $c \xrightarrow{G}_z m$ and $m \xrightarrow{G}_z e$. Pairwise Granger analysis would yield $c \xrightarrow{G}_z e$ and not be able to distinguish whether the causality between e and c is direct or mediated by m. As such, we conclude that only considering each user's top influencer is sufficient to accurately predict the user's future interests; therefore, without loss of generality, we compare our proposed approach with the baselines based on g@1.

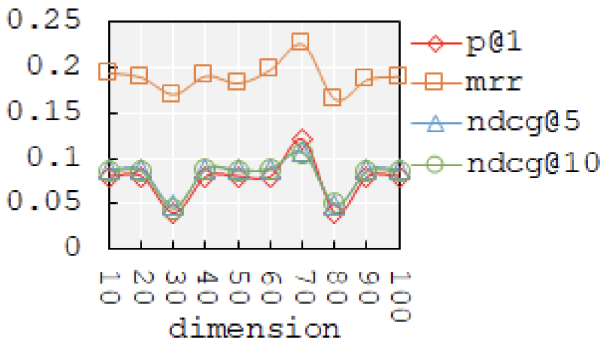

Figure 6.14: The impact of dimension size on the proposed temporal latent space modeling method in the context of news recommendation application. The vertical axis shows the amplitude of the ranking metrics.

Next, we analyze the effect of dimension d in our temporal latent space inference algorithm. We vary d from 10 to 100 and report the performance in Figure 6.14. As seen, the overall trend indicates that the recommendation performance in terms of all ranking metrics increases with number of dimensions up to an extremum at $d = 70$.

We compared our proposed G-regression and temporal latent space method at its best setting (d=70) against the baselines at their best settings in Table 6.1. As shown, our proposed temporal latent space method outperforms other baselines in terms of all ranking metrics in the context of news recommendation. We attribute the accuracy of the temporal latent space modeling approach to the fact that it directly models and leverages the impact of users' pairwise similarity over their topics of interest within the time dimension, i.e., sequence of similarity graphs, which has been overlooked in all of the other baselines. G-regression is a predictive model based on inter-user topical in-

fluence within the time dimension and was able to be the next in terms of MRR. GrosToT, which is neither a predictive model nor aware of temporal similarity among users, however, had inefficient performance. It is worth noting that due to capturing sequences of inter-user similarities indirectly through collaborative filtering, RRN was able to become the runner-up in terms of nDCG@5 and nDCG@10. It is worth noting that the underlying news recommendation system we used for the evaluation might not be capable of being useful in practice as the results are very low due to the applied naive recommendation algorithm. However, the main goal herein is not to propose a state of the art method for news recommendation but to show the performance gain of the proposed methods compared to the baselines given an existing recommendation system.

The other application with which we evaluate our approach is the user prediction application. We use precision (PR), recall (REC), and f-measure (F1) to report user prediction performance. We further compare our method at its best which happens to be at d=80, against the baselines at their best setting in Table 6.1. In terms of precision, our proposed methods, G-regression and temporal latent space modeling, were able to outperform other baselines. In terms of *recall*; however, the baselines could achieve higher performance and our methods were not as strong. The reason for such high recall for the baselines is the fact that the baseline methods cluster users into very few, yet large user communities, as seen in Figure 6.15. For instance, RRN was able to excel in recall due to its low number of communities. In an

Table 6.1: The performance comparison of the proposed G-regression and temporal latent space modeling vs. the state-of-the-art baselines for community prediction in the context of news recommendation and user prediction applications in terms of ranking and classification metrics respectively.

	News Recommendation			User Prediction		
	MRR	nDCG@5	nDCG@10	PR	REC	F1
Temporal Latent Space Modeling	**0.2254**	**0.1080**	**0.1050**	**0.0123**	0.0352	**0.0148**
Chimera (C=20, d=30) [5]	0.1760	0.0558	0.0545	0.0065	0.0938	0.0105
GrosToT(C=20) [50]	0.1733	0.0560	0.0490	0.0070	0.1358	0.0130
TCD-Embedding (d=300)	0.0652	0.0400	0.0400	0.0070	0.1364	0.0132
VAR	0.1540	0.0480	0.0490	0.0012	0.1481	0.0025
G-regression	**0.1816**	0.0480	0.0490	**0.0087**	0.4423	**0.0169**
RRN [106]	0.1732	**0.0731**	**0.0795**	0.0040	**0.7407**	0.0079
TimeSVD++ [57]	0.1412	0.0577	0.0641	0.0026	**0.6574**	0.0052

extreme, if a method only identifies one community that includes all of the users, recall would be 1. As such the lower the number of the communities is, the higher the recall would be. However, this comes at the cost of precision. Overall, the f-measure metric points to higher quality communities identified based on our proposed work. This reinforces the fact that when users' pairwise similarity with respect to the topics of interest over time are explicitly embedded in a sequence of graphs, it will lead to higher quality user communities in the future. Further, Figure 6.15 shows that unlike some of the baselines where the majority of the users are placed in few communities and the other communities only have a few members, our approaches have been able to proportionally distribute users across different communities.

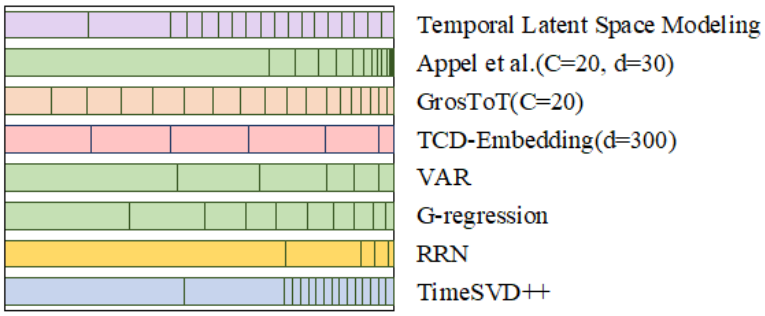

Figure 6.15: User distribution in communities. Our temporal latent space method leads to a higher number of communities with a proportional distribution of users in the communities while the baseline methods like RRN have a higher skewness. Disproportionate distribution of users in communities can lead to poor application-level performance.

6.5 Summary

6.5.1 Findings

Based on our experiments on the news recommendation and user prediction tasks, we can summarize our findings with regards to the four research questions as follows:

1. We find that the consideration of temporal evolution of user-generated content is key in finding effective user communities. Our observations show that the incorporation of time in the user representations leads to higher quality user communities compared to when time is not considered.

2. Further, we find that the neural embedding of time into the user representation leads to higher quality communities compared to when time

is included as a part of a generative process.

3. We observed that the communities identified through link-based methods are poorer compared to when temporal content-based methods are employed.

4. We find that while link-based methods show poorer performance compared to temporal content-based methods, they can still have synergistic impact on the performance of temporal content-based methods. In other words, the interpolation of link-based and temporal content-based methods lead to higher quality user communities.

5. Finally, we find that it is possible to predict future-yet-unobserved content-based communities on social networks through proposed temporal latent space modeling.

In summary, we conclude that when embeddings learnt based on temporal content-based methods are interpolated with the embeddings learnt from link-based community detection methods, they result in the highest quality communities as shown within the context of news recommendation and user prediction tasks. The findings have been evaluated from both the perspective of information retrieval and classification metrics. With respect to community prediciton task, our work is among the first to explore the idea of predicting *topical* user communities on social networks. Our experiments show that our temporal latent space approach is able to predict communities

of like-minded users with respect to topics of interest in future yet-to-be-observed time interval and outperform the state of the art.

6.6 Related Publications

- Hossein Fani, Masoud Bashari, Fattane Zarrinkalam, Ebrahim Bagheri and Feras Al-Obeidat; "Stopword Detection for Streaming Content." *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings.*

Chapter 7

Conclusions

This chapter concludes the thesis and provides an outlook on future work. First, a summary is presented and concluding remarks are made for the thesis. Next, open research problems for future work are described.

7.1 Concluding Remarks

Identifying and extracting user communities is an important step towards understanding social network dynamics from a macro perspective. For this reason, the work in this thesis explored various aspects related to the identification of user communities.

This thesis first proposed two approaches to detect communities of like-minded users who share topics of interest with similar temporal behaviour. Then, it put forward two approaches to predict the same type of user com-

munities but in a future yet-to-be-observed time interval. In the following, a number of findings that can be concluded from the work done in this thesis are summarized.

- To date, user community detection methods employ either explicit links between users (link analysis), or users' topics of interest in posted content (content analysis), or in tandem. Little work has considered temporal evolution when identifying user communities in a way to group together those users who share not only similar topical interests but also similar temporal behaviour towards their topics of interest. In this thesis, we identified user communities through two alternative representations for users. First, we modeled the contribution of each user towards topics using multivariate time series and applied 2-dimensional cross-correlation on all pairs of such time series to find similar users in topics of interest and temporal behaviour. We employed the Louvain clustering, a heuristic graph partitioning algorithm based on modularity optimization, to create our final user communities. To find topics from the social network, we used state-of-the-art topic detection methods with different approaches, as alternatives, in order to show that our approach and its contribution were independent of topic detection algorithms. We used one graph-based and two probabilistic LDA and ToT methods. We examined our approach on two application scenarios: news recommendation and user prediction. According to our results, our temporal topic-based community detection method based on multi-

variate user time series was able to effectively identify user communities that were formed around temporally similar behaviour towards shared topics and show competitive performance compared to the state of the art. However, this method suffered from sparsity in multivariate user time series. Further, it was not possible to seamlessly augment social links as an additional information source. As such, we proposed a neural embeddings method to model users based on their temporal content similarity. Inspired by Mikolov et al.'s word2vec [69] in computational linguistics, while we followed the same underlying premise about temporality in like-minded user community detection, we introduced a time-aware topic-driven distributional representation (embeddings) of users. We applied cosine similarity on all pairs of user embeddings to find topically and temporally similar users. We employed the Louvain clustering to create our final user communities. This approach not only was able to outperform the state of the art by addressing the sparsity problem, but was also able to support social links as well. In summary, we found that (1) methods that consider temporal evolution of content, our proposed methods in particular, showed better performance compared to their non-temporal counter-parts; (2) content-based methods produced higher quality communities compared to link-based methods.

- Futher, we learnt user embeddings based on their social network connections (links) through neural graph embeddings. We systematically interpolated temporal content-based embeddings and social link-based

embeddings to capture both social network connections and temporal content evolution for representing users. We evaluated the quality of each embedding type in isolation and also when interpolated together and demonstrated their performance on a same testbed, i.e., under the two application scenarios, namely news recommendation and user prediction. We found that (3) while link-based methods are weaker than content-based methods, their interpolation with content-based methods leads to improved quality of the identified communities.

- Performing community prediction in the future can be quite challenging, especially on social networks such as Twitter, due to the rapid changes in community topics and evolution of user interactions. In this context, temporal collaborative filtering methods, which benefit from similar user behavioural patterns over time to predict how a user's interests might evolve in the future, can be employed to perform user community prediction. In this thesis, we proposed that instead of considering the whole user base within a collaborative filtering framework, it is possible to much more accurately predict such user communities by only considering the behavioural patterns of the most influential user related to the user of interest. We modeled influence as a form of causal dependency between users. To this end, we employed the concept of *Granger causality* to identify causal dependencies. While our experiments showed that the consideration of only one causally dependent user leads to much more accurate prediction of users' future interests,

the running time complexity was prohibitive. Therefore, we proposed a temporal latent space model for user community prediction in social networks, whose goal was to predict future emerging user communities based on past history of users topics of interest. Our model assumed that each user lies within an unobserved latent space, and similar users in the latent space representation are more likely to be members of the same user community. The model allowed each user to adjust its location in the latent space as her topics of interest evolve over time. We found that (4) this method not only is computationally tractable, but also outperforms existing approaches.

7.2 Future Work

In this section, open research problems for future work are described.

7.2.1 User Community Detection

The work presented in this thesis for detecting diachronically like-minded user communities can be improved from a number of aspects with additional research. In the following, some of these aspects are discussed and directions for future research are provided:

- In all our experiments, the time period was broken into daily intervals. This was a fair choice for our dataset because it consists of tweets from a two-month period. It would be, however, interesting to study the

effect of other time interval spans, e.g., weekly vs monthly, in order to construct the user-topic contribution time series for the user representation. Possible future directions of our work would be exploring the impact of time interval size on the quality of the derived communities and additionally finding ways in which the optimal time interval length can be learnt through hyper-parameter search techniques.

- We used news recommendation and user prediction applications to evaluate the proposed methods. However, other application scenarios have been proposed for extrinsic evaluation when there is no labeled dataset such as retweet prediction [117] or timestamp prediction [101]. Examining the performance of the proposed methods in such underlying applications would be a potential future direction.

- To produce the output user communities, we employed the Louvain clustering method, which was an effective and efficient partitioning clustering method. It would be interesting to extend our approach beyond disjoint user communities and study overlapping user communities by using overlapping clustering techniques.

- In our multimodal neural user embedding approach, we interpolated temporal content and social network structure at the user vector representation level for the task of temporal user community detection to explore the synergy between social links and temporal contents. This inherently limited the vectors for both types of representation to have

the same embedding size. One possible future direction would be to explore temporal content-based and link-based user vectors at the score level, i.e., the final similarity scores of temporal content-based user vector representations could be interpolated with the similarity scores of link-based user vectors. This way, the embedding size of the information sources becomes irrelevant.

- We interpolated temporal content-based and link-based user vector representations through a weighted linear function in order to systematically investigate the synergy between links and temporal content. Another direction for our future research is to learn the embedding interpolation function through joint representation instead. Hence, the synergy effect would not be limited to linear functions and might lead to a greater impact.

7.2.2 User Community Prediction

Our work is among the first to explore the idea of predicting topical user communities on social networks. Our work has limitations in both proposed methods, namely Granger regression and temporal latent space modeling, though. In the following some of these limitations are discussed and directions for future research are provided:

- Communities encounter various evolution stages such as birth, death, growth, shrinking, splitting and/or merging. In other words, two com-

munities combine into a bigger community or another community may be divided into several smaller communities. However, in this thesis we disregarded information about evolution patterns to detect future user communities which otherwise could greatly improve the accuracy of the community prediction task.

- In Granger regression, we used the pairwise bivariate Granger causality test in our influence identification step. However, while our empirical experiments showed promising performance on this basis, some researchers have also pointed out that causal influences might not be highly accurate [70]. Conditional Granger causality has been proposed to alleviate this problem [70].

- In Granger regression, the original formulation of Granger causality is linear. However, extensions to nonlinear cases now exist. Ancona et al. [4] propose to use a radial basis function to perform a global nonlinear regression. Such extensions would potentially improve the Granger regression performance and are worth further exploration in future work.

- With respect to temporal latent space modeling, our method penalized significant changes in positions of users' latent representation in the latent space, which may not be warranted in some circumstances like bursty topics. Our work can be generalized to such cases based on intuitions from Deng et al. [26].

- Although our temporal latent space modeling used simple, yet effective approximations, the latent representations at time T+1 can be formulated as $\mathbf{y}_{u(\mathrm{T}+1)} = f(\mathbf{y}_{u1},\ ..\ \mathbf{y}_{ut},\ ..\ \mathbf{y}_{u\mathrm{T}})$ and more sophisticated estimations such as the nonparametric method suggested by Sarkar et al. [94] can be used in the future.

Bibliography

[1] H. A. Abdelbary, A. M. ElKorany, and R. Bahgat, *Utilizing deep learning for content-based community detection*, 2014 Science and Information Conference, Aug 2014, pp. 777–784.

[2] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao, *Analyzing user modeling on Twitter for personalized news recommendations*, User Modeling, Adaption and Personalization - 19th International Conference, UMAP 2011, Girona, Spain, July 11-15, 2011. Proceedings, 2011, pp. 1–12.

[3] Mohammad Akbari and Tat-Seng Chua, *Leveraging behavioral factorization and prior knowledge for community discovery and profiling*, Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017, 2017, pp. 71–79.

[4] Nicola Ancona, Daniele Marinazzo, and Sebastiano Stramaglia, *Radial basis function approach to nonlinear granger causality of time series*,

Physical Review E **70** (2004), no. 5, 056221.

[5] Ana Paula Appel, Renato L. F. Cunha, Charu C. Aggarwal, and Marcela Megumi Terakado, *Temporally evolving community detection and prediction in content-centric networks*, Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part II, 2018, pp. 3–18.

[6] Negar Arabzadeh, Hossein Fani, Fattane Zarrinkalam, Ahmed Navivala, and Ebrahim Bagheri, *Causal dependencies for future interest prediction on Twitter*, Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, 2018, pp. 1511–1514.

[7] Ognjen Arandjelovic, *Weighted linear fusion of multimodal data: A reasonable baseline?*, Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016, 2016, pp. 851–857.

[8] Ognjen Arandjelovic and Roberto Cipolla, *A new look at filtering techniques for illumination invariance in automatic face recognition*, Seventh IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2006), 10-12 April 2006, Southampton, UK, 2006, pp. 449–454.

[9] Albert-László Barabási and Réka Albert, *Emergence of scaling in random networks*, Science **286** (1999), no. 5439, 509–512.

[10] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco, *Efficient methods for influence-based network-oblivious community detection*, ACM TIST **8** (2017), no. 2, 32:1–32:31.

[11] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio A. F. Almeida, *Characterizing user behavior in online social networks*, Proceedings of the 9th ACM SIGCOMM Internet Measurement Conference, IMC 2009, Chicago, Illinois, USA, November 4-6, 2009, 2009, pp. 49–62.

[12] Adrian Benton, Raman Arora, and Mark Dredze, *Learning multiview embeddings of Twitter users*, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers, 2016.

[13] Prasanta Bhattacharya and Rishabh Mehrotra, *The information network: Exploiting causal dependencies in online information seeking*, Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval, CHIIR 2016, Carrboro, North Carolina, USA, March 13-17, 2016, 2016, pp. 223–232.

[14] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, *Latent dirichlet allocation*, Journal of Machine Learning Research **3** (2003), 993–1022.

[15] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, *Fast unfolding of communities in large networks*, Journal of Statistical Mechanics: Theory and Experiment **2008** (2008), no. 10, P10008.

[16] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner, *On modularity clustering*, IEEE Trans. Knowl. Data Eng. **20** (2008), no. 2, 172–188.

[17] Jinxin Cao, Hongcui Wang, Di Jin, and Jianwu Dang, *Combination of links and node contents for community discovery using a graph regularization approach*, Future Generation Comp. Syst. **91** (2019), 361–370.

[18] Shaosheng Cao, Wei Lu, and Qiongkai Xu, *Deep neural networks for learning graph representations*, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA., 2016, pp. 1145–1152.

[19] Tanmoy Chakraborty, Zhe Cui, and Noseong Park, *Metadata vs. ground-truth: A myth behind the evolution of community detection methods*, Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018, 2018, pp. 45–46.

[20] Mingming Chen, Konstantin Kuzmin, and Boleslaw K. Szymanski, *Community detection via maximization of modularity and its variants*, IEEE Trans. Comput. Social Systems **1** (2014), no. 1, 46–65.

[21] Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo, *BTM: topic modeling over short texts*, IEEE Trans. Knowl. Data Eng. **26** (2014), no. 12, 2928–2941.

[22] Evangelos Christou, *Social media monitoring: A practical case example of city destinations*, Social Media in Travel, Tourism and Hospitality, Routledge, 2016, pp. 315–334.

[23] Rob Claxton, Jonathan Reades, and B Anderson, *On the value of digital traces for commercial strategy and public policy: Telecommunications data as a case study*, pp. 105–112, World Economic Forum, 2012 (English).

[24] Alessio Conte, Tiziano De Matteis, Daniele De Sensi, Roberto Grossi, Andrea Marino, and Luca Versari, *D2K: scalable community detection in massive networks via small-diameter k-plexes*, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018, 2018, pp. 1272–1281.

[25] Allison Davis, Burleigh B Gardner, Mary R Gardner, and W Lloyd Warner, *Deep south: A sociological anthropological study of caste and class*, University of Chicago Press, 1941.

[26] Dingxiong Deng, Cyrus Shahabi, Ugur Demiryurek, Linhong Zhu, Rose Yu, and Yan Liu, *Latent space model for road networks to predict time-varying traffic*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 1525–1534.

[27] Qiao Deng, Zhoujun Li, Xiaoming Zhang, and Jiali Xia, *Interaction-based social relationship type identification in microblog*, Behavior and Social Computing, International Workshop on Behavior and Social Informatics, BSI 2013, Gold Coast, QLD, Australia, April 14-17, 2013 and International Workshop on Behavior and Social Informatics and Computing, BSIC 2013, Beijing, China, August 3-9, 2013, Revised Selected Papers, 2013, pp. 151–164.

[28] Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis, *Weighted graph cuts without eigenvectors A multilevel approach*, IEEE Trans. Pattern Anal. Mach. Intell. **29** (2007), no. 11, 1944–1957.

[29] David A. Dickey, *Dickey-fuller tests*, International Encyclopedia of Statistical Science, 2011, pp. 385–388.

[30] Christopher P. Diehl, Galileo Namata, and Lise Getoor, *Relationship identification for social network discovery*, Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada, 2007, pp. 546–552.

[31] Laura Dietz, Steffen Bickel, and Tobias Scheffer, *Unsupervised prediction of citation influences*, Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007, 2007, pp. 233–240.

[32] Ying Ding, *Community detection: Topological vs. topical*, J. Informetrics **5** (2011), no. 4, 498–514.

[33] Claire Donnat, Marinka Zitnik, David Hallac, and Jure Leskovec, *Learning structural node embeddings via diffusion wavelets*, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018, 2018, pp. 1320–1329.

[34] Daniel M. Dunlavy, Tamara G. Kolda, and Evrim Acar, *Temporal link prediction using matrix and tensor factorizations*, TKDD **5** (2011), no. 2, 10:1–10:27.

[35] Elena Erosheva, Stephen Fienberg, and John Lafferty, *Mixed-membership models of scientific publications*, Proceedings of the National Academy of Sciences **101** (2004), no. suppl 1, 5220–5227.

[36] Hossein Fani, Masoud Bashari, Fattane Zarrinkalam, Ebrahim Bagheri, and Feras Al-Obeidat, *Stopword detection for streaming content*, Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings, 2018, pp. 737–743.

[37] Paolo Ferragina and Ugo Scaiella, *Fast and accurate annotation of short texts with wikipedia pages*, IEEE Software **29** (2012), no. 1, 70–75.

[38] Santo Fortunato, *Community detection in graphs*, Physics Reports **486** (2010), no. 3, 75 – 174.

[39] Hongchang Gao and Heng Huang, *Self-paced network embedding*, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018, 2018, pp. 1406–1415.

[40] M. Girvan and M. E. J. Newman, *Community structure in social and biological networks*, Proceedings of the National Academy of Sciences **99** (2002), no. 12, 7821–7826.

[41] Michelle Girvan and Mark EJ Newman, *Community structure in social and biological networks*, Proceedings of the National Academy of Sciences **99** (2002), no. 12, 7821–7826.

[42] C. W. J. Granger, *Investigating causal relations by econometric models and cross-spectral methods*, Econometrica **37** (1969), no. 3, 424–438.

[43] Aditya Grover and Jure Leskovec, *node2vec: Scalable feature learning for networks*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 855–864.

[44] Amit Gruber, Michal Rosen-Zvi, and Yair Weiss, *Latent topic models for hypertext*, UAI 2008, Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, Helsinki, Finland, July 9-12, 2008, 2008, pp. 230–239.

[45] Yu Guan, Xingjie Wei, Chang-Tsun Li, and Yosi Keller, *People identification and tracking through fusion of facial and gait features*, Biometric Authentication - First International Workshop, BIOMET 2014, Sofia, Bulgaria, June 23-24, 2014. Revised Selected Papers, 2014, pp. 209–221.

[46] Guibing Guo, Jie Zhang, and Neil Yorke-Smith, *Trustsvd: Collaborative filtering with both the explicit and implicit influence of user trust and of item ratings*, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA., 2015, pp. 123–129.

[47] Dongxiao He, Dayou Liu, Di Jin, and Weixiong Zhang, *A stochastic model for detecting heterogeneous link communities in complex networks*, Proceedings of the Twenty-Ninth AAAI Conference on Artificial

Intelligence, January 25-30, 2015, Austin, Texas, USA., 2015, pp. 130–136.

[48] G.C. Homans, *The human group*, International Library of Sociology, Taylor & Francis, 2013.

[49] Liangjie Hong and Brian D. Davison, *Empirical study of topic modeling in Twitter*, Proceedings of the First Workshop on Social Media Analytics (New York, NY, USA), SOMA '10, ACM, 2010, pp. 80–88.

[50] Zhiting Hu, Junjie Yao, and Bin Cui, *User group oriented temporal dynamics exploration*, Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada., 2014, pp. 66–72.

[51] Zhiting Hu, Junjie Yao, Bin Cui, and Eric P. Xing, *Community level diffusion extraction*, Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015, 2015, pp. 1555–1569.

[52] Yin Huang, Han Dong, Yelena Yesha, and Shujia Zhou, *A scalable system for community discovery in Twitter during hurricane sandy*, 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGrid 2014, Chicago, IL, USA, May 26-29, 2014, 2014, pp. 893–899.

[53] Mohsen Jamali and Martin Ester, *A matrix factorization technique with trust propagation for recommendation in social networks*, Proceedings of the 2010 ACM Conference on Recommender Systems, RecSys 2010, Barcelona, Spain, September 26-30, 2010, 2010, pp. 135–142.

[54] Di Jin, Zheng Chen, Dongxiao He, and Weixiong Zhang, *Modeling with node degree preservation can accurately find communities*, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA., 2015, pp. 160–167.

[55] Kazi Lutful Kabir, Liban Hassan, Zahra Rajabi, Nasrin Akhter, and Amarda Shehu, *Graph-based community detection for decoy selection in template-free protein structure prediction*, Molecules (Basel, Switzerland) **24** (2019), 854.

[56] Thomas N. Kipf and Max Welling, *Semi-supervised classification with graph convolutional networks*, International Conference on Learning Representations ICLR, 2017.

[57] Yehuda Koren, *Collaborative filtering with temporal dynamics*, Commun. ACM **53** (2010), no. 4, 89–97.

[58] Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, and Yongcheol Shin, *Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?*, Journal of econometrics **54** (1992), no. 1-3, 159–178.

[59] Vincent Labatut, Nicolas Dugué, and Anthony Perez, *Identifying the community roles of social capitalists in the Twitter network*, 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2014, Beijing, China, August 17-20, 2014, 2014, pp. 371–374.

[60] Andrea Lancichinetti and Santo Fortunato, *Community detection algorithms: a comparative analysis*, Physical review E **80** (2009), no. 5, 056117.

[61] Frank Thomson Leighton and Satish Rao, *Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms*, J. ACM **46** (1999), no. 6, 787–832.

[62] Jure Leskovec, Kevin J. Lang, and Michael W. Mahoney, *Empirical comparison of algorithms for network community detection*, Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010, 2010, pp. 631–640.

[63] Chunshan Li, William K. Cheung, Yunming Ye, Xiaofeng Zhang, Dianhui Chu, and Xin Li, *The author-topic-community model for author interest profiling and community discovery*, Knowl. Inf. Syst. **44** (2015), no. 2, 359–383.

[64] Shangsong Liang, Zhaochun Ren, Yukun Zhao, Jun Ma, Emine Yilmaz, and Maarten de Rijke, *Inferring dynamic user interests in streams of short texts for user clustering*, ACM Trans. Inf. Syst. **36** (2017), no. 1, 10:1–10:37.

[65] Hongtao Liu, Hui Chen, Mao Lin, and Yu Wu, *Community detection based on topic distance in social tagging networks*, TELKOMNIKA Indonesian Journal of Electrical Engineering **12** (2014), no. 5, 4038–4049.

[66] Jie Liu, Zhicheng He, Lai Wei, and Yalou Huang, *Content to node: Self-translation network embedding*, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018, 2018, pp. 1794–1802.

[67] Pablo Mateos, *Demographic, ethnic, and socioeconomic community structure in social networks*, Encyclopedia of Social Network Analysis and Mining, 2nd Edition, 2018.

[68] Miller McPherson, Lynn Smith-Lovin, and James M Cook, *Birds of a feather: Homophily in social networks*, Annual Review of Sociology **27** (2001), no. 1, 415–444.

[69] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean, *Distributed representations of words and phrases and their compositionality*, Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems

2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States., 2013, pp. 3111–3119.

[70] Ding Mingzhou, Chen Yonghong, and Bressler Steven L., *Granger causality: Basic theory and application to neuroscience*, pp. 437–460, Wiley-Blackwell, 2006.

[71] Aalaa Mojahed, Joao H. Bettencourt-Silva, Wenjia Wang, and Beatriz de la Iglesia, *Applying clustering analysis to heterogeneous data using similarity matrix fusion (SMF)*, Machine Learning and Data Mining in Pattern Recognition - 11th International Conference, MLDM 2015, Hamburg, Germany, July 20-21, 2015, Proceedings, 2015, pp. 251–265.

[72] Farnaz Moradi, Tomas Olovsson, and Philippas Tsigas, *An evaluation of community detection algorithms on large-scale email traffic*, Experimental Algorithms - 11th International Symposium, SEA 2012, Bordeaux, France, June 7-9, 2012. Proceedings, 2012, pp. 283–294.

[73] Frederic Morin and Yoshua Bengio, *Hierarchical probabilistic neural network language model*, Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, AISTATS 2005, Bridgetown, Barbados, January 6-8, 2005, 2005.

[74] Seth A. Myers and Jure Leskovec, *The bursty dynamics of the Twitter information network*, Proceedings of the 23rd International Conference

on World Wide Web (New York, NY, USA), WWW '14, ACM, 2014, pp. 913–924.

[75] Ramesh Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen, *Joint latent topic models for text and citations*, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008, 2008, pp. 542–550.

[76] Nagarajan Natarajan, Prithviraj Sen, and Vineet Chaoji, *Community detection in content-sharing social networks*, Advances in Social Networks Analysis and Mining 2013, ASONAM '13, Niagara, ON, Canada - August 25 - 29, 2013, 2013, pp. 82–89.

[77] M. E. J. Newman, *Modularity and community structure in networks*, Proceedings of the National Academy of Sciences **103** (2006), no. 23, 8577–8582.

[78] ———, *Spectral methods for community detection and graph partitioning*, Physical Review E **88** (2013), no. 4, 042822.

[79] Dunlu Peng, Xie Lei, and Ting Huang, *DICH: A framework for discovering implicit communities hidden in tweets*, World Wide Web **18** (2015), no. 4, 795–818.

[80] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena, *Deepwalk: online learning of social representations*, The 20th ACM SIGKDD Interna-

tional Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014, 2014, pp. 701–710.

[81] Georgios Petkos, Symeon Papadopoulos, Luca Maria Aiello, Ryan Skraba, and Yiannis Kompatsiaris, *A soft frequent pattern mining approach for textual topic detection*, 4th International Conference on Web Intelligence, Mining and Semantics (WIMS 14), WIMS '14, Thessaloniki, Greece, June 2-4, 2014, 2014, pp. 25:1–25:10.

[82] Peter C. B. Phillips and Pierre Perron, *Testing for a unit root in time series regression*, Biometrika **75** (1988), no. 2, 335–346.

[83] Ioannis Psorakis, Stephen Roberts, Mark Ebden, and Ben Sheldon, *Overlapping community detection using bayesian non-negative matrix factorization*, Physical Review E **83** (2011), no. 6, 066114.

[84] Jipeng Qiang, Ping Chen, Tong Wang, and Xindong Wu, *Topic modeling over short texts by incorporating word embeddings*, Advances in Knowledge Discovery and Data Mining - 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part II, 2017, pp. 363–374.

[85] Dimitrios Rafailidis, Pavlos Kefalas, and Yannis Manolopoulos, *Preference dynamics with multimodal user-item interactions in social media recommendation*, Expert Syst. Appl. **74** (2017), 11–18.

[86] Stuart A. Rice, *The identification of blocs in small political bodies*, The American Political Science Review **21** (1927), no. 3, 619–627.

[87] Yossi Richter, Elad Yom-Tov, and Noam Slonim, *Predicting customer churn in mobile networks through analysis of social groups*, Proceedings of the SIAM International Conference on Data Mining, SDM 2010, April 29 - May 1, 2010, Columbus, Ohio, USA, 2010, pp. 732–741.

[88] Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth, *The author-topic model for authors and documents*, UAI '04, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence, Banff, Canada, July 7-11, 2004, 2004, pp. 487–494.

[89] Giulio Rossetti and Rémy Cazabet, *Community discovery in dynamic networks: A survey*, ACM Comput. Surv. **51** (2018), no. 2, 35:1–35:37.

[90] Randolf Rotta and Andreas Noack, *Multilevel local search algorithms for modularity clustering*, J. Exp. Algorithmics **16** (2011), 2.3:2.1–2.3:2.27.

[91] Randolf Rotta and Andreas Noack, *Multilevel local search algorithms for modularity clustering*, ACM Journal of Experimental Algorithmics **16** (2011).

[92] Yiye Ruan, David Fuhry, and Srinivasan Parthasarathy, *Efficient community detection in large networks using content and links*, 22nd In-

ternational World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, 2013, pp. 1089–1098.

[93] Mrinmaya Sachan, Danish Contractor, Tanveer A. Faruquie, and L. Venkata Subramaniam, *Using content and interactions for discovering communities in social networks*, Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012, 2012, pp. 331–340.

[94] Purnamrita Sarkar, Deepayan Chakrabarti, and Michael I. Jordan, *Nonparametric link prediction in dynamic networks*, Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012, 2012.

[95] Purnamrita Sarkar and Andrew W. Moore, *Dynamic social network analysis using latent space models*, Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada], 2005, pp. 1145–1152.

[96] Tom AB Snijders and Alessandro Lomi, *Beyond homophily: Incorporating actor variables in statistical network models*, Network Science **7** (2019), no. 1, 1–19.

[97] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas, *Short text classification in Twitter to improve*

information filtering, Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010, 2010, pp. 841–842.

[98] Andrea Varga, Amparo Elizabeth Cano Basave, Matthew Rowe, Fabio Ciravegna, and Yulan He, *Linked knowledge sources for topic classification of microposts: A semantic graph-based approach*, J. Web Sem. **26** (2014), 36–57.

[99] Daixin Wang, Peng Cui, and Wenwu Zhu, *Structural deep network embedding*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 1225–1234.

[100] Meng Wang, Chaokun Wang, Jeffrey Xu Yu, and Jun Zhang, *Community detection in social networks: An in-depth benchmarking study with a procedure-oriented framework*, PVLDB **8** (2015), no. 10, 998–1009.

[101] Xuerui Wang and Andrew McCallum, *Topics over time: a non-markov continuous-time model of topical trends*, Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006, 2006, pp. 424–433.

[102] Duncan J Watts and Steven H Strogatz, *Collective dynamics of small-worldnetworks*, nature **393** (1998), no. 6684, 440.

[103] Robert S Weiss and Eugene Jacobson, *A method for the analysis of the structure of complex organizations*, American Sociological Review **20** (1955), no. 6, 661–668.

[104] Jianshu Weng and Bu-Sung Lee, *Event detection in Twitter*, Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011, 2011.

[105] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He, *TwitterRank: finding topic-sensitive influential twitterers*, Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010, 2010, pp. 261–270.

[106] Chao-Yuan Wu, Amr Ahmed, Alex Beutel, Alexander J. Smola, and How Jing, *Recurrent recommender networks*, Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017, 2017, pp. 495–503.

[107] Jibing Wu, Lianfei Yu, Qun Zhang, Peiteng Shi, Lihua Liu, Su Deng, and Hongbin Huang, *Multityped community discovery in time-evolving heterogeneous information networks based on tensor decomposition*, Complexity **2018** (2018), 9653404:1–9653404:16.

[108] Yuexin Wu, Hanxiao Liu, and Yiming Yang, *Graph convolutional matrix completion for bipartite edge prediction*, Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2018, Volume 1: KDIR, Seville, Spain, September 18-20, 2018., 2018, pp. 49–58.

[109] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y. Chang, *Network representation learning with rich text information*, Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015, 2015, pp. 2111–2117.

[110] Jaewon Yang and Jure Leskovec, *Patterns of temporal variation in online media*, Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011, 2011, pp. 177–186.

[111] Liang Yang, Xiaochun Cao, Dongxiao He, Chuan Wang, Xiao Wang, and Weixiong Zhang, *Modularity based community detection with deep learning*, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016, 2016, pp. 2252–2258.

[112] Liang Yang, Xiaochun Cao, Di Jin, Xiao Wang, and Dan Meng, *A unified semi-supervised community detection framework using latent space*

graph regularization, IEEE Trans. Cybernetics **45** (2015), no. 11, 2585–2598.

[113] Tianbao Yang, Rong Jin, Yun Chi, and Shenghuo Zhu, *Combining link and content for community detection: a discriminative approach*, Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009, 2009, pp. 927–936.

[114] Fanghua Ye, Chuan Chen, and Zibin Zheng, *Deep autoencoder-like nonnegative matrix factorization for community detection*, Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, 2018, pp. 1393–1402.

[115] Zhijun Yin, Liangliang Cao, Quanquan Gu, and Jiawei Han, *Latent community topic analysis: Integration of community discovery with topic modeling*, ACM TIST **3** (2012), no. 4, 63:1–63:21.

[116] Fattane Zarrinkalam, Hossein Fani, Ebrahim Bagheri, Mohsen Kahani, and Weichang Du, *Semantics-enabled user interest detection from Twitter*, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015, Singapore, December 6-9, 2015 - Volume I, 2015, pp. 469–476.

[117] Fattane Zarrinkalam, Mohsen Kahani, and Ebrahim Bagheri, *Mining user interests over active topics on social networks*, Inf. Process. Manage. **54** (2018), no. 2, 339–357.

[118] Lizhuang Zhao and Mohammed Javeed Zaki, *Tricluster: An effective algorithm for mining coherent clusters in 3d microarray data*, Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, USA, June 14-16, 2005, 2005, pp. 694–705.

[119] Ding Zhou, Eren Manavoglu, Jia Li, C. Lee Giles, and Hongyuan Zha, *Probabilistic models for discovering e-communities*, Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006, 2006, pp. 173–182.

[120] Linhong Zhu, Dong Guo, Junming Yin, Greg Ver Steeg, and Aram Galstyan, *Scalable temporal latent space inference for link prediction in dynamic social networks*, IEEE Trans. Knowl. Data Eng. **28** (2016), no. 10, 2765–2777.

[121] Shenghuo Zhu, Kai Yu, Yun Chi, and Yihong Gong, *Combining content and link for classification using matrix factorization*, SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007, 2007, pp. 487–494.

Vita

Candidate's full name: Hossein Fani
University attended (with dates and degrees obtained):

Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran
Master of Computer Science, 2009

Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran
Bachelor of Computer Science, 2005

Publications:

Journal Publications:
Hossein Fani, Ebrahim Bagheri, Fattane Zarrinkalam, Xin Zhao and Weichang Du; "Finding Diachronic Like-Minded Users." *Computational Intelligence* 34(1): 124-144 (2018)

Hossein Fani, Eric Jiang, Ebrahim Bagheri, Feras Al-Obeidat, Weichang Du and Mehdi Kargar; "User Community Detection via Embedding of Social Network Structure and Temporal Content." *Information Processing and Management*, 2019. (to appear)

Yue Feng, Fattane Zarrinkalam, Ebrahim Bagheri, **Hossein Fani** and Feras Al-Obeidat; "Entity linking of tweets based on dominant entity candidates." *Social Network Analysis and Mining* 8(1): 46:1-46:16 (2018)

Conference Presentations:

Hossein Fani, Amin Mirlohi, Hawre Hosseini, Reiner Herpers; "Swim Stroke Analytic: Front Crawl Pulling Pose Classification." *In IEEE International Conference on Image Processing (ICIP'18)*, 2018.

Negar Arabzadeh, **Hossein Fani**, Fattane Zarrinkalam, Ahmed Navivala, Ebrahim Bagheri; "Causal Dependencies for Future Interest Prediction on Twitter." *In Conference on Information and Knowledge Management, (CIKM'18)*, 2018.

Maryam Khodabakhsh, **Hossein Fani**, Fattane Zarrinkalam, Ebrahim Bagheri; "Predicting Personal Life Events from Streaming Social Content." *In Conference on Information and Knowledge Management, (CIKM'18)*, 2018.

Hossein Fani, Masoud Bashari, Fattane Zarrinkalam, Ebrahim Bagheri, and Feras Al-Obeidat; "Stopword Detection for Streaming Content." *In Advances in Information Retrieval: 40th European Conference on IR Research (ECIR'18)*, 2018.

Hossein Fani, Ebrahim Bagheri, and Weichang Du; "Temporally Like-minded User Community Identification through Neural Embeddings." *In the 2017 Conference on Information and Knowledge Management (CIKM'17)*, 2017.

Hossein Fani and Ebrahim Bagheri; "Community detection in social networks." *Encyclopedia with Semantic Computing and Robotic Intelligence*, 01:1630001, 2017.

Fattane Zarrinkalam, **Hossein Fani**, Ebrahim Bagheri and Mohsen Kahani; "Predicting Users Future Interests on Twitter." *In 39th European Conference on Information Retrieval (ECIR'17)*, 2017.

Fattane Zarrinkalam, **Hossein Fani**, Ebrahim Bagheri and Mohsen Kahani; "Inferring Implicit Topical Interests on Twitter." *In 38th European Conference on Information Retrieval (ECIR'16)*, 2016.

Hossein Fani, Fattane Zarrinkalam, Ebrahim Bagheri and Weichang Du; "Time-Sensitive Topic-Based Communities on Twitter." *In Advances in*

Artificial Intelligence - 29th Canadian Conference on Artificial Intelligence, Canadian AI'16, 2016.

Hossein Fani; "Temporal Formation and Evolution of Online Communities." *Doctoral Consortium, The 9th ACM International Conference on Web Search and Data Mining (WSDM'16)*, 2016.

Fattane Zarrinkalam, **Hossein Fani**, Ebrahim Bagheri, Mohsen Kahani and Weichang Du; "Semantics-enabled User Interest Detection from Twitter." *In ACM/IEEE/WIC Web Intelligence, IEEE*, 2015.

Yue Feng, **Hossein Fani**, Ebrahim Bagheri and Jelena Jovanovic; "Lexical Semantic Relatedness for Twitter Analytics." *In IEEE International Conference on Tools with Artificial Intelligence (ICTAI'15)*, 2015.

Hossein Fani and Ebrahim Bagheri; "An Ontology for Describing Security Events." *In The 27th International Conference on Software Engineering and Knowledge Engineering, SEKE'15*, 2015.