**Resource: A Framework for Community Prediction**

**Soroush Ziaeinejad**  
University of Windsor, Canada  
ziaeines@uwindsor.ca

**Saeed Samet**  
University of Windsor, Canada  
ssamet@uwindsor.ca

**Hossein Fani**  
University of Windsor, Canada  
hfani@uwindsor.ca

---

**ABSTRACT**

Online user communities exhibit distinct temporal dynamics in response to popular topics or breaking events. Despite abundant community detection libraries, there is yet to be one that provides access to the possible user communities in future time intervals. To bridge this gap, we contribute SEERa, an open-source end-to-end community prediction framework to identify future user communities in a text streaming social network. SEERa incorporates state-of-the-art temporal graph neural networks to model inter-user topical affinities at each time interval via streams of temporal graphs and learns temporal vector representations for users. This all takes place while users’ topics of interest and hence their inter-user topical affinities are changing over time. SEERa predicts yet-to-be-seen user communities based on the final positions of users’ vectors in the latent space. Notably, our framework serves as a one-stop-shop to future user communities for Social Information Retrieval and Social Recommendation systems. While there are strong research papers on the community prediction problem, SEERa is the first framework to be publicly released for this purpose.

---

**1 INTRODUCTION**

Social networks are the prominent medium for communication and social interaction wherein communities of like-minded users who are interested in similar topics emerge due to homophily [24]. Identifying user communities finds immediate application in scalable item recommendation and marketing campaigns [5, 31]. Traditionally, user community detection methods were based on explicit links, e.g., followership. However, explicit links conflate topical similarity among users when connections are missing or due to kinship [4, 33]. To this end, latent community detection methods have been proposed based on users’ textual posts and their topics of interest. Whether to find explicit or latent communities, existing methods primarily focus on identifying communities at the current point in time and fall short of extracting communities that will emerge in the future, except for Fani et al. [13] and Chimera [3] where temporal latent space embedding and shared matrix factorization are proposed respectively to predict the possible user communities in future yet-to-be-observed time intervals. In fact, online user communities are temporal since users enjoy varying behavior in response to popular topics or breaking events [11, 32, 35].

Meanwhile, myriad of temporal graph neural networks like [9, 26, 28] have been proposed for a range of social network tasks such as link prediction that can be employed for community prediction owing to their capability to model and learn from graph-structured data. Also, a dizzying array of topic modeling libraries [6, 18, 27, 30] and an increasing research in Social Information Retrieval (Social IR) and Social Recommendation that integrates social network contexts to enhance rankings and recommendations [20, 21, 36, 37] have been introduced to date. However, there is yet to be one end-to-end (no human in the loop) case-specific framework that unifies topic modeling approaches with temporal graph embedding methods for the task of topical community prediction in social networks.

To establish a unified software platform for the task of topical community prediction, we contribute SEERa, the first open-source python-based framework for identifying the possible user communities in future yet-to-be-observed time intervals. SEERa models inter-user topical affinities in each time interval in user graphs whose nodes and edges are users and their topical affinities, respectively. Given the stream of graphs, SEERa applies a range of state-of-the-art temporal latent space embedding and graph neural networks to capture the users’ temporal topics of interest and their inter-user affinities from the past up until now. Temporal embedding methods allow users to change their positions in latent space as their topics of interest evolve over time: the more similar temporal and topical behavior the users exhibit, the closer their final latent positions would be, upon which SEERa can accurately estimate user communities in the future whose members will almost surely share similar topics. Out of the box, SEERa includes a news article recommendation module to exhibit how future community prediction is effective in applications in which an accurate estimation of users’ interests in the future is desired. Key contributions of SEERa include:

1. **SEERa** predicts future user communities and can be configured for any online social platform with streaming content.
2. **SEERa** is designed with extensibility in mind. While already hosts a wide variety of methods in each layer of its pipeline, it facilitates the addition of new topic modeling methods, temporal graph embeddings, graph clustering methods, and application-level use cases;
3. **SEERa** facilitates reproducibility and repeatability of the research work on future user community prediction on a shared underlying application. Also, it helps practitioners with cross-comparing different topic modeling algorithms with temporal graph embedding methods and pick the most suitable one for their application. As a part of its release, it includes a news article recommendation application. Indeed, community-level...
The primary purpose of the data access layer (dal) is to (i) load raw texts posted by users within a time period \(T\) (load_tweets()), and (ii) group them into coherent contexts, namely documents, where a user describes one or a few topics of her interest (data_preparation()). A document is either (i) a single post of a user (time=False), or (ii) can be enriched by all posts of a user in each time interval (time=True), known as pooling [25], which is based on a prior assumption that whether users follow different topics of interest in each of their posts or remain consistent to a single or a few topics of interest within a time interval. Figure 4 shows that document’s tokens are either raw words (tagme=False) or semantic entities (tagme=True). The choice of semantic entities is encouraged if posts are informal, short, and noisy (e.g., tweets) as opposed to long formal documents (e.g., weblogs). SEERa employs TagMe [15], a fast and accurate semantic annotator for short texts that uses Wikipedia as its background knowledge to disambiguate and identify semantic entities while filtering stopwords and uninformative tokens.

Once documents are formed, they are led to the topic modeling layer (tml) for users’ topics of interest detection.

### 2.2 Topic Modeling Layer (tml)

This layer identifies topics of interest from user’s documents using latent Dirichlet allocation (LDA) [7]. It can also be easily extended to any other topic modeling method such as TwitterLDA [2] for short streaming content. This layer returns a list of topics as probability distributions over all unique tokens, and documents as distributions over topics. To this end, SEERa benefits from gensim [27] (python) to compute topic models.
Table 1: Dominant topics on Twitter for Nov. 1 – Dec. 15, 2010.

<table>
<thead>
<tr>
<th>topic1</th>
<th>topic2</th>
<th>topic3</th>
</tr>
</thead>
<tbody>
<tr>
<td>papers 0.017</td>
<td>new_york 0.006</td>
<td>aung_san_su_1 0.018</td>
</tr>
<tr>
<td>secret 0.017</td>
<td>live 0.006</td>
<td>myanmar 0.014</td>
</tr>
<tr>
<td>usaid 0.017</td>
<td>cigarette 0.005</td>
<td>activist 0.011</td>
</tr>
</tbody>
</table>

Figure 5: User Modeling Layer (uml).

as well as mallet[23] (java libraries). The quality of output topics under different numbers of topics (ntopics) and other settings can be quantitatively tuned by coherence [10]. Table 1 shows dominant topics from Nov. 1, until Dec. 15, 2010, when ‘Aung San Suu Kyi’, a political activist, was among the users’ topics of interest.

2.3 User Modeling Layer (uml)

Given the user’s documents as distributions over identified topics from (uml), from those documents at time interval t, SEERA calculates pairwise inter-user topical similarities at t and connects users who share strong similarities in their topics of interest in the graph g_t, as shown in Figure 5. The final output of this layer is the stream of graphs for the entire time period [g_1]^T. The time interval can be adjusted for, e.g., daily, weekly, biweekly, or monthly. We underscore that graphs are generated independently at this layer, and temporal relations between graphs are overlooked. SEERA employs dynamic graph embedding methods to address this issue in the next layer.

2.4 Graph Embedding Layer (gel)

Given [g_t]^T, the graph embedding layer (gel) applies temporal embedding methods to embed users (nodes) into a low d-dimensional vectors [v_t]^T as they pass through time while capturing temporal and topical affinities from their initial position v_1 to a final position v_T, as shown Figure 6. Users’ vectors at the final time interval, v_T, depend on preceding [v_1...v_(T-1)] via observation of [g_1...g_(T-1)] which is in contrast with static models that obtain v_T based solely on g_T. Our framework employs state-of-the-art temporal graph neural network methods including Dynamic Auto-Encoder, Dynamic RNN that uses LSTM, and Dynamic AERN which uses AE followed by LSTM network [16], and temporal latent space modeling based on non-negative matrix factorization [17].

2.5 Community Prediction Layer (cpl)

SEERA uses users’ vectors at the final time interval, v_T, as the proxy to predict the topical similarities of users in future user time intervals, i.e., graph g_T^>T, based on pairwise cosine similarity (v_r \cdot v_j) of the users r and j. We are, additionally, developing more advanced prediction methods such as multivariate regression on v_T to predict v_T^>T as a better proxy for g_T^>T. Finally, a clustering method on g_T^>T unfolds densely connected subgraphs as the future user communities. As shown in Figure 7, while we use Louvain [8],

the canonical method in community detection, this layer can be seamlessly extended to any clustering methods [22, 29]. We are currently extending SEERAs to clustering methods to be applied directly on users’ d-dimensional vectors at the final time interval v_T to obtain the future user communities, dispensing the intermediate graph representation g_T^>T.

To quantitatively evaluate the quality of the predicted future user communities, our framework includes two methods based on the availability of golden standard: (1) intrinsic evaluation using well-known metrics such as rand index and mutual information, and (2) extrinsic evaluation, that is, to what extent an extra knowledge about users’ membership to future communities will provide synergy to the efficacy and efficiency of an underlying application.

2.6 Application Layer (apl): News Article Recommendation

Future community prediction helps applications with prior insight into the future and to excel at effectiveness and efficiency. Out of the box, we implemented an important use case in Social Recommendation as suggested in [1, 14]: news article recommendations. Future user communities can signal a recommender system to target users based on topics of the communities to which they belong. Therefore, community prediction not only increases the efficiency by targeting communities, as opposed to individual users, but also improves the efficacy of the recommendations via prior knowledge about topics of interest in the future. For instance, in Figure 7(b), g_4 is the predicted graph in the future and u_2 and u_3 are members of the same community. They will receive the same set of news articles whose topics are different from the news articles that u_1 receives.

To this end, we define community-level topics of interest as the sum of all its members’ predicted topics of interest in the future. Top-k news articles are recommended to each community (i.e., all of its members) according to cosine similarity scores between the news articles’ topics and community-level topics. The recommendation table is then evaluated against users’ mentions of news articles in their posts. We use pytrec_eval [34] to report information retrieval metrics such as MRR and nDCG.
The lda_model is saved.

Data Reading:
Start and end dates: from (1-11-2010) to (15-11-2010)
#Tweets: 9414
#Users: 895
#Documents: 1 day

Topic Modeling:
#Topics: 25
The lda_model is saved.

User Modeling:
Shape of users_topic matrix: (15, 895, 25)
15 graphs are created for 15 days with 895 users.

Graph Embedding:
Embedding Dimension: 40
Method: Node2Vec
Embedded Matrix: (895, 40) is saved.

Graph Clustering:
method: Louvain
#Nodes: 895 / #Edges: 20960

Clustering output: 24 communities
C0: 88 users / Favorite topic: T09 (94.6%)
C1: 67 users / Favorite topic: T09 (94.6%)
C2: 66 users / Favorite topic: T10 (89.3%)
C3: 45 users / Favorite topic: T11 (86.6%)
C4: 41 users / Favorite topic: T21 (82.9%)
C5: 5 users / Favorite topic: T19 (100%)

Table 2: SEERA’s benchmark on a Twitter dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>News Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRR</td>
</tr>
<tr>
<td>Community Prediction</td>
<td></td>
</tr>
<tr>
<td>Fani et al. [13]</td>
<td>0.225</td>
</tr>
<tr>
<td>Appel et al. [3]</td>
<td>0.176</td>
</tr>
<tr>
<td>Temporal Community Detection</td>
<td></td>
</tr>
<tr>
<td>Hu et al. [19]</td>
<td>0.173</td>
</tr>
<tr>
<td>Fani et al. [12]</td>
<td>0.065</td>
</tr>
</tbody>
</table>

Table 2 shows SEERA’s main benchmark results on a Twitter dataset of 2M tweets authored by 135K users within 59+1 days and the performance of predicted future communities on the 60th day for 50 topics of interest. Complete results have been reported publicly in the codebase. We are testing scalability of SEERA on a large-scale dataset including 300M tweets authored by 3.5M users within 6 months and the result will be released once our experiments have been accomplished.

3 QUICK START

SEERA can be obtained by:
git clone https://github.com/fani-lab/seera.git
SEERA has hyperparameters at each layer that should be set at ./src/params.py. Figure 8 shows the summary of parameters. For instance, number of topics or embedding dimension can be adjusted in ml and gel sections, respectively. The entry point to the framework is ./src/main.py that executes the pipeline until the final delivery of future user communities followed by their evaluation on news article recommendation. Figure 9 display an example output of our framework. From the figure, we ran SEERA on a sample dataset from Nov. 1 to Nov. 15, 2010 (today) to predict future communities on Nov. 15+1 (tomorrow). SEERA found 25 communities from which cluster C0 have 80 users, most of them (88.7%) interested in topic T14, '2010s_Haiti_cholera_outbreak'. Should a news recommender system recommend news articles to users tomorrow, it would select news articles about topic T14 for community C0, expecting a high click-through rate. Due to the SEERA’s documentation and ease of use, researchers and practitioners familiar with python programming can run, modify, and use SEERA in different applications.

4 CONCLUDING REMARKS

In this paper, we propose SEERA, an open-source end-to-end framework, to identify future user communities in a text streaming social network. SEERA is to address the emerging need in Social IR and Social Recommendation systems for future communities. SEERA is designed with extensibility in mind. While hosts a wide variety of methods in each layer of its pipeline, it easily accommodates new methods as well as application-level use cases; (2) can be easily configured for any online social platform with streaming content; (3) improves the efficiency and efficacy of the underlying time-sensitive applications using the predicted future user communities; and (4) benefits the information retrieval and recommender systems community with reproducibility and repeatability of the research work on community prediction.