# ErLinkTopic: A GENERATIVE PROBABILISTIC FRAMEWORK FOR ANALYZING REGIONAL COMMUNITIES IN SOCIAL NETWORKS

**Tran Van Canh** [(1)], **Michael Gertz** [(2)], **and Dang Hong Linh** [(1)]
[1] *Institute of Engineering and Technology, Vinh University, Vietnam*
[2]*Institute of Computer Science, Heidelberg University, Germany*
Received on 5/4/2019, accepted for publication on 22/6/2019

**Abstract:** Understanding how communities evolve over time have become a hot topic in the field of social network analysis due to the wide range of its applications. In this context, several approaches have been introduced to capture changes in the community members. Our claim is that a community is characterized by not only the identity of users but complex features such as the topics of interest, and the regional and geographic characteristics. Studying changes in such features of communities also provides informative findings for related applications. This leads to the main goal of the study in this paper, which is to capture the evolution of complex features describing communities. Particularly, we introduce a probabilistic framework called $ErLinkTopic$ model. The model is able to extract regional $LinkTopic$ [1] communities and to capture gradual changes in three features describing each community, i.e., community members, the prominence of topics describing communities, and terms describing such topics. It further supports the study of regional and geographic characteristics of communities as well as changes in such features. Experimental evaluations have been conducted using $Twitter$ data to evaluate the model in terms of its effectiveness and efficiency in extracting communities and capturing changes in the features describing each community.

## 1 Introduction

Several models and algorithms have been developed for extracting communities in social networks. Typical approaches rely on the link structure of users, which is presented as a graph. This leads to the application of different graph clustering algorithms to detect such link-based communities, e.g., [2]-[4]. Recent studies, however, pay more attention to finding topical communities. By this, topical analysis is applied to the messages of users to derive topics indicating their interests. The extracted topics are used as another feature, besides the link structures to identify relationships between users. The key idea is that by leveraging more common features of users one can discover more meaningful communities. That is, users in a community exhibit both structural and hidden semantic links to each others. The main approach to extracting communities based on this idea is to develop a probabilistic model simulating a process of generating the observed features of users from hidden

[1)] Email: canhtv@vinhuni.edu.vn (T. V. Canh) communities. In the proposed models, e.g., [5]-[7], the two important features, namely the contextual links of users and the regional aspect of communities, have been either neglected or paid only very little attention to. In [1], the authors developed a novel probabilistic model *rLinkTopic* to add these features into account. However, *rLinkTopic* does not cover the dynamic of communities. Nevertheless, communities in a social network evolve over time due to several reasons. A user is interested in the topics of a community and joins as a new member while some users might leave the community. The happening of social events, e.g., an election, and other phenomena also lead to the evolution of communities. Such an evolution is implied by changes in the features describing a community. These include, for example, users in the community, topics of the community, and geographic locations of the users. Given that a community is characterized by even more features, analyzing its evolution thus is a challenging task. This is because one has to have a complex model that is able to discover communities and to capture changes in as many features describing a community as possible. To date, existing approaches for the analysis of evolving communities attempt to study changes with respect to one feature, which are the community members [8]-[11]. The concept of *evolution* is therefore defined only in the context of the user population of a community over time. Because of this, no information is obtained with respect to how other features of the community evolve. From an application perspective, one is usually interested not only in the dynamics of users, e.g., which users are in a community at what time, but also in other features that describe the community over time. These observations motivate our study and development of a comprehensive framework that takes more features of interest into account to study the evolution of communities in social networks. Particularly, in this paper, we introduce a probabilistic model called *ErLinkTopic* that is an extension of the *rLinkTopic* model developed in [1] for extracting regional *LinkTopic* communities and analyzing their complex evolution. By stating complex evolution, we are particularly interested in changes in the features describing a community as formalized in the *rLinkTopic* model. These include (1) the community membership of users in a community; (2) topic proportion of a community; and (3) terms occurring in a community topic. Also, because information about geographic locations is associated with users' postings, the model further supports the study of changes in the regional and geographic characteristics of communities. The paper is organized as follows. Section 2 gives an overview of the background and related work for this paper. Section 3 presents the underlying data model and introduces notations used to present the *ErLinkTopic* model. In Section 4, we first describe how *rLinkTopic* is extended to build *ErLinkTopic* that can discover communities and, at the same time, capture their evolution (Section 4.1). We then give detailed steps to derive a Gibbs sampling algorithm to compute the posterior distribution of the *ErLinkTopic* model (Section 4.2). The results of our experimental evaluations using *Twitter* data are presented in Section 5 before we conclude the paper in Section 6.

# 2 Background and the rLinkTopic Model

## 2.1 Study of Evolving Communities

In addition to extracting static communities, e.g., [1], [3], [7], [12]-[15], several models have been introduced to study the evolution of communities regarding changes in the community members over time. Three main approaches have been applied, namely snapshot community matching, evolutionary clustering, and probabilistic models.

The MONIC framework for finding and monitoring cluster transactions was proposed in [16]. The authors consider the number of common objects (users) between two clusters (community structures) at two consecutive snapshots as a measure to decide whether a cluster has transited to or evolved from another. Based on this measure, five events called *becomes, splits, merges, disappears, and appears* that might happen to a community during two consecutive snapshots are defined. Sitaram Asur et al. [8] developed a similar framework to study community evolution. By matching snapshot communities, the authors formalized five temporal events that are identically interpreted as those in MONIC. Other measures called *stability, sociability, popularity, and influence* to study the behavior of users in a network were defined in this framework also. Palla et al. [17], [18] introduced a *Clique Percolation Model* and proposed a method to capture the evolution of communities between two consecutive snapshots by creating a union graph and matching community structures found in this graph with community structures found at the two snapshots. Studies based on the evolutionary clustering approach build *unified* models to find *temporal smooth* evolving communities. The main idea of this approach is that the objective function employed in graph partitioning algorithms consists of two components, the *history quality* and the *snapshot quality.* The snapshot quality measures how accurate the resulting clusters capture the structure of the network at the current snapshot, while the history quality measures how consistent the resulting clusters are, with respect to the clusters discovered at the previous snapshot. Algorithms are designed to find a partition that is trade-off to these two quality components. The first study in this direction was introduced by Chakrabarti et al. [9]. In their work, the $k$-means and hierarchical clustering algorithms were extended to produce evolving clusters. Lin et al. [10], [19] developed a FacetNet framework, which is based on non-negative matrix factorization [20] to approximate the structure of a snapshot. The snapshot quality and history quality are computed using Kullback Leibler divergence distance. Evolving communities are identified by optimizing the clustering solution with respect to both the snapshot quality and the history quality. The authors of FacetNet also introduced a similar framework called MetaFac that employs metagraph factorization to extract communities in dynamic and rich media networks [11]. Other studies on the evolutionary clustering approach employed spectral clustering methods. Examples include the studies by Chi et al. [21], [22].

The probabilistic modeling approaches extract communities from each snapshot and make prediction about the evolution of communities using Bayesian prediction strategy. A probabilistic model is developed to discover communities in each snapshot, which is basically similar to the idea applied to extract static communities. However, to capture the evolution of communities, the community membership of users at the previous snapshot is used as a prior knowledge for computing such a membership at the current snapshot. Communities gradually evolve over time, which is indicated by changes in the membership of users in communities discovered over snapshots [23], [24].

### 2.2 The rLinkTopic Model

Although geographic and regional aspects of communities find many practical applications, e.g., in social studies and marketing, to date, existing approaches to community detection have paid little attention to these features when analyzing social network data. To address these shortcomings, in [1], the authors introduced the concept of regional link-topic communities and proposed a novel probabilistic model called *rLinkTopic* for extracting such communities. The model jointly considers the spatio-temporal proximity of users in terms of the messages they post over time, together with contextual links and message topics to determine communities. Each community derived by *rLinkTopic* is not only described by a mixture of topics but also by its regional properties. It is noted that, in the *rLinkTopic* model, a social network is formalized as a sequence of snapshots. The model relies on the occurrences of users in each snapshot to identify users who occur in the network within spatio-temporal proximity. This *co-occurrence* feature together with the contextual links and the topics of user postings are employed to extract communities. By this, the temporal order of the occurrences of users, i.e., the order of snapshots, is not important and is discarded in the *rLinkTopic* model. Our aim in this paper is to take advantage of the *rLinkTopic* model to extract communities; and, at the same time, to capture community evolution. For the latter aspect, the temporal order is crucial, because it is used to explain the evolution of the characteristics of a community over time.

## 3 Data Model and Notations

This section describes the data model underlying our framework and introduced notations used throughout this paper. We model a social network as a sequence of sliding windows, each of which consists of a number of consecutive snapshots. The general idea is that communities are extracted within each sliding window, i.e., the temporal order of the snapshots in a sliding window is discarded. Information about the community structures obtained from the current sliding window then is employed to derive communities at the next sliding window. Adopting the data model introduced in the *rLinkTopic* model [1], the concept of sliding windows is formalized as follows.

**Definition 3.1** (Network Sliding Window)**.** Given a social network $SN = \{sn_1, sn_2, ..., sn_T\}$ and a time span $\triangle t = [t_s, t_e]$, a sliding window $\mathcal{W}_t$ of size $\triangle t$ is a sequence of consecutive snapshots $\mathcal{W}_t = \{sn_{t_s}, ..., sn_{t_e}\}$.

Having the sliding window defined, a social network is now considered a sequence of sliding windows, i.e., $SN = \{\mathcal{W}_1, \mathcal{W}_2, ..., \mathcal{W}_T\}$, which is the underlying data model for the *ErLinkTopic* framework presented in the next section. To present the *ErLinkTopic* model, the main notations used in the *rLinkTopic* model [1] are employed and some other notations are introduced, all of which are described in Table 1.

**Tab. 1:** *Notations used in the ErLinkTopic model for extracting regional LinkTopic communities and analyzing their evolution.*

| **Notation** | **Description** |
|---|---|
| $U$ | set of users in social network, $u$ is a user in $U$ |
| $C$ | set of communities, $c$ is a community in $C$ |
| $V$ | vocabulary set, $w$ is a word in $V$ |
| $Z$ | set of community topics, $z$ is a topic in $Z$ |
| $R_{\mathcal{W}_t}$ | set of geographic regions created from snapshots of sliding window $\mathcal{W}_t$ |
| $\theta_t$ | set of community distributions in geographic regions $R_{\mathcal{W}_t}$, i.e., $\theta_t = \{\theta_r\}, r \in R_{\mathcal{W}_t}$ |
| $\phi_t$ | set of user distributions for communities $C$ at window $\mathcal{W}_t$, i.e., $\phi_t = \{\phi_{t;c}\}, c \in C$ |
| $\pi_t$ | set of topic proportions of communities $C$ at window $\mathcal{W}_t$, i.e., $\pi_t = \{\pi_{t;c}\}, c \in C$ |
| $\varphi_t$ | set of term distributions for topics $Z$ at window $\mathcal{W}_t$ , i.e., $\varphi_t = \{\varphi_{t;z}\}, z \in Z$ |
| $\boldsymbol{r}_t$ | region assignments of the occurrences of users at window $\mathcal{W}_t$ |
| $\boldsymbol{c}_t$ | community assignments of the occurrences of users at window $\mathcal{W}_t$ |
| $\boldsymbol{z}_t$ | topic assignments of the messages of users at window $\mathcal{W}_t$ |

# 4 ErLinkTopic Probabilistic Model

This section presents in detail the *ErLinkTopic* model for extracting regional *LinkTopic* communities and analyzing their evolution. In Section 4.1, a discussion explaining how *rLinkTopic* is employed to develop *ErLinkTopic* is given. We present the steps to derive a Gibbs sampling algorithm for the *ErLinkTopic* model in Section 4.2.

## 4.1 rLinkTopic to ErLinkTopic

Typically, a two-step approach is applied to study the evolution of communities. In the first step, communities are extracted independently of the occurrences of users at different time points, e.g., snapshots or sliding windows. In the second step, a matching of the communities obtained from consecutive time points is accomplished. Based on the result of the matching, the evolution of communities is then explained. For example, if the *rLinkTopic* model is employed to study community evolution based on this two-step approach, then one would run the model independently on each sliding window to extract communities. Communities obtained from consecutive sliding windows are then matched to find out their evolution. Almost all of existing studies for the analysis of evolving communities follow this strategy [8], [16], [18]. Even that, this typical approach has two main shortcomings. First, the matching procedure always requires extensive computations and the selection of a matching solution is a subjective task. This issue becomes even harder for our setting, because we aim at studying the evolution of multiple features describing a community. The second weakness affecting the result more is that this approach fails to capture the gradual evolution of communities. It is because communities are independently extracted from different sliding windows and none of the obtained information is employed while deriving new communities. That is, for example, the community structures obtained from the previous sliding window are not used in the extraction of communities at the current sliding window. Obviously, community memberships of a user at the current sliding window should be derived based on the memberships of that user in communities discovered from the previous sliding window. This happens similarly to the evolution of the topic proportion of a community, and the evolution of terms in a topic. To handle these observations, the *ErLinkTopic* model is developed to discover communities over sliding windows in the way that information about the community structures obtained from a sliding window is used for deriving communities at the next window. That is, the community membership of users, the topic proportion of communities, and the distribution of terms in topics obtained from sliding window $\mathcal{W}_{t-1}$ are used as prior knowledge provided to compute the corresponding distributions at sliding window $\mathcal{W}_t$. This is basically done by extending the *rLinkTopic* model. The key idea in the *rLinkTopic* model is that we employ the conjugacy between the *Dirichlet* distribution and the $Multinomial$ distribution to model the features describing a community. Such features include (1) the distribution $\phi_c$ of users, (2) the topic proportion $\pi_c$, (3) the distribution $\varphi_z$ of terms in a topic associated with $c$, and (4) the geographic areas where $c$ is observed, which is characterized by the likelihood of $c$ in regions, denoted $\theta_{r,c}, r \in R$. As a result, the posterior distribution of each of these variables is also a *Dirichlet* distribution. Therefore, it is straightforward to extend the *rLinkTopic* model so that it can be used to discover communities and, at the same time, to capture their gradual evolution. More precisely, the scenario of extracting and capturing the evolution of communities over two sliding windows $\mathcal{W}_{t-1}$ and $\mathcal{W}_t$ is as follows. First, applying the $rLinkTopic$ model to the occurrences of users in the snapshots of $\mathcal{W}_{t-1}$ to extract communities from that sliding window. Each identified community $c$ is characterized by the posterior distributions of the (1) users in $c$, denoted $\phi_{t-1;c}$, (2) topic proportion of $c$, denoted $\pi_{t-1;c}$, (3) terms in topics associated with $c$, denoted $\varphi_{t-1;z}$, $z \in Z$, and (4) locations of $c$, denoted $\theta_{t;r,c}, r \in R_{\mathcal{W}_{t-1}}$, derived at sliding window $\mathcal{W}_{t-1}$. The estimated value of each of these variables except $\theta_t$ is then used as an evidence to compute the corresponding variables at the next step for extracting communities from sliding window $\mathcal{W}_t$. By this, all features describing a community are obtained over time and their changes are gradually captured. Figure 4.1 shows the graphical model representing the generative process of the *ErLinkTopic* model as described. It is a sequence of $rLinkTopic$ models linked to each other. Each block describes the extraction of communities in a sliding window.

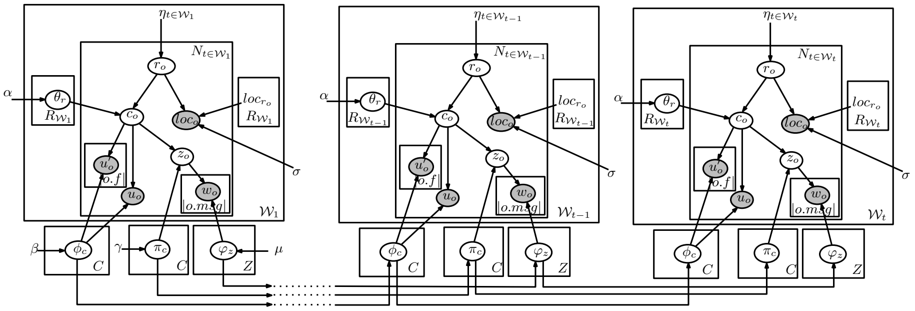


**Fig. 1:** *Graphical model presenting the generative process of the ErLinkTopic model. It consists of a sequence of rLinkTopic models linked to each other.*

## 4.2 Posterior Estimation for ErLinkTopic Model

There are assumptions implicitly employed in the *ErLinkTopic* model shown in Figure 4.1. First, the distributions $\phi_t$ of users in communities, the topic proportions $\pi_t$ of communities, and the distributions $\varphi_t$ of terms in topics at the current sliding window $\mathcal{W}_t$ are conditionally independent of the occurrences of users at the previous sliding window $\mathcal{W}_{t-1}$, given the corresponding distributions obtained from $\mathcal{W}_{t-1}$, i.e., $\phi_{t-1}$, $\pi_{t-1}$, and $\varphi_{t-1}$. Second, the occurrences of users in the snapshots of sliding window $\mathcal{W}_t$ are conditionally independent of all other information, given $\phi_t$, $\pi_t$, $\varphi_t$, and $\theta_t$. Having such assumptions employed, the joint distribution of the *ErLinkTopic* model is represented as follows.

$$\begin{aligned} P(SN,\phi,\theta,\pi,\varphi,\boldsymbol{r},\boldsymbol{c},\boldsymbol{z}|\beta,\gamma,\mu,\alpha,\eta,\sigma) &= P(\mathcal{W}_1,\phi_1,\theta_1,\pi_1,\varphi_1,\boldsymbol{r_1},\boldsymbol{c_1},\boldsymbol{z_1}|\beta,\gamma,\mu,\alpha,\eta,\sigma) \\ &\times \prod_{t=2}^{T} P(\mathcal{W}_t,\phi_t,\theta_t,\pi_t,\varphi_t,\boldsymbol{r_t},\boldsymbol{c_t},\boldsymbol{z_t}|\phi_{t-1},\pi_{t-1},\varphi_{t-1},\alpha,\eta,\sigma) \end{aligned} \tag{1}$$

Based on Eq. 1, the posterior distribution of the model is derived incrementally over sliding windows. Particularly, it is first computed based on the occurrences of users in the snapshots of the first sliding window $\mathcal{W}_1$ and the hyperparamters of the model. This is actually the posterior estimation of the *rLinkTopic* model applied to the snapshots of $\mathcal{W}_1$. For each of the next sliding windows, information about the community structures derived from the previous step, together with the user occurrences in the snapshots of that sliding window are used to extract communities.

The posterior distribution of the model at sliding window $\mathcal{W}_t$ $(t > 1)$ is computed based on the user occurrences in the snapshots of $\mathcal{W}_t$ and the posterior distribution derived from $\mathcal{W}_{t-1}$, which is presented as follows.

$$\begin{aligned} P(\phi_t,\theta_t,\pi_t,\varphi_t,\boldsymbol{r_t},\boldsymbol{c_t},\boldsymbol{z_t} \quad | \quad & \mathcal{W}_t,\phi_{t-1},\pi_{t-1},\varphi_{t-1},\alpha,\eta,\sigma) = \\ & \frac{P(\mathcal{W}_t,\phi_t,\theta_t,\pi_t,\varphi_t,\boldsymbol{r_t},\boldsymbol{c_t},\boldsymbol{z_t}|\phi_{t-1},\pi_{t-1},\varphi_{t-1},\alpha,\eta,\sigma)}{P(\mathcal{W}_t|\phi_{t-1},\pi_{t-1},\varphi_{t-1},\alpha,\eta,\sigma)} \end{aligned} \tag{2}$$

The above posterior distribution is estimated by sampling from the joint distribution of the model applied to the user occurrences in the snapshots of sliding window $\mathcal{W}_t$, given the information derived from the previous sliding window $\mathcal{W}_{t-1}$ and the hyperparameters, which is computed as follows.

$$\begin{aligned} P(\mathcal{W}_t,\phi_t,\theta_t,\pi_t,\varphi_t,\boldsymbol{r_t},\boldsymbol{c_t},\boldsymbol{z_t}|\phi_{t-1},\pi_{t-1},\varphi_{t-1},\alpha,\eta,\sigma) = \prod_{sn_t\in\mathcal{W}_t}\prod_{o\in sn_t} P(r_o|\eta_t)P(loc_o|loc_{r_o},\sigma) \times & \qquad \text{(I)} \\ \prod_{sn_t\in\mathcal{W}_t} P(\theta_t|\alpha) \prod_{o\in sn_t} P(c_o|\theta_{t,r_o}) \times & \qquad \text{(II)} \\ P(\phi_t|\phi_{t-1}) \prod_{sn_t\in\mathcal{W}_t}\prod_{o\in sn_t} P(u_o|\phi_{t,c_o}) \prod_{u'\in o.f} P(u'|\phi_{t,c_o}) \times & \qquad \text{(III)} \\ P(\pi_t|\pi_{t-1}) \prod_{sn_t\in\mathcal{W}_t}\prod_{o\in sn_t} P(z_o|\pi_{t,c_o}) \times & \qquad \text{(IV)} \\ P(\varphi_t|\varphi_{t-1}) \prod_{sn_t\in\mathcal{W}_t}\prod_{o\in sn_t}\prod_{w\in o.msg} P(w|\varphi_{t,z_o}) & \qquad \text{(V)} \end{aligned} \tag{3}$$

**Tab. 2:** *Notations used to present the count variables in the ErLinkTopic model. Each variable is computed based on the user occurrences in the snapshots of one sliding window.*

| Notation | Description |
|---|---|
| $n_c^{(r)}$ | number of occurrences in region $r$ that are assigned to community $c$ |
| $n_u^{(c)}$ | number of occurrences of user $u$ that are assigned to community $c$ |
| $n_{f.u}^{(c)}$ | number of times user $u$ is contextually linked by other users in community $c$ |
| $n_w^{(z)}$ | number of occurrences of term $w$ that are assigned to topic $z$ |
| $n_z^{(c)}$ | number of messages in community $c$ that are assigned to topic $z$ |

Adopting the notations defined in Table 4.2, the above joint distribution is simplified so that the posterior distribution in Eq. 2 is then estimated as follows.

$$P(\phi_t, \theta_t, \pi_t, \varphi_t, \boldsymbol{r_t}, \boldsymbol{c_t}, \boldsymbol{z_t} | \mathcal{W}_t; \phi_{t-1}, \pi_{t-1}, \varphi_{t-1}, \alpha, \eta, \sigma) \propto \prod_{sn_t \in \mathcal{W}_t} \prod_{o \in sn_t} P(r_o|\eta_t) P(loc_o|loc_{r_o}, \sigma) \times$$
$$\prod_{r \in R_{\mathcal{W}_t}} \prod_{c \in C} \theta_{r,c}^{n_c^{(r)} + \alpha_c - 1} \times \prod_{c \in C} \prod_{u \in U} \phi_{t;c,u}^{n_u^{(c)} + n_{f.u}^{(c)} + \phi_{t-1;c,u} - 1} \times$$
$$\prod_{c \in C} \prod_{z \in Z} \pi_{t;c,z}^{n_z^{(c)} + \pi_{t-1;c,z} - 1} \times \prod_{z \in Z} \prod_{w \in V} \varphi_{t;z,w}^{n_w^{(z)} + \varphi_{t-1;z,w} - 1} \tag{4}$$

By integrating out the multinomial parameters $\phi_t$, $\pi_t$, $\varphi_t$, and $\theta_t$, the posterior distribution of the region assignments $\boldsymbol{r_t}$, community assignments $\boldsymbol{c_t}$, and topic assignments $\boldsymbol{z_t}$ of the user occurrences in the snapshots of sliding window $\mathcal{W}_t$ becomes

$$P(\boldsymbol{r_t}, \boldsymbol{c_t}, \boldsymbol{z_t} | \mathcal{W}_t; \phi_{t-1}, \pi_{t-1}, \varphi_{t-1}, \alpha, \eta, \sigma) \propto \underbrace{\prod_{sn_t \in \mathcal{W}_t} \prod_{o \in sn_t} P(r_o|\eta_t) P(loc_o|loc_{r_o}, \sigma) \times}_{(T_1)}$$
$$\underbrace{\prod_{r \in R_{\mathcal{W}_t}} \frac{\prod_{c \in C} \Gamma(n_c^{(r)} + \alpha_c)}{\Gamma(\sum_{c \in C} n_c^{(r)} + \alpha_c)}}_{(T_2)} \times \underbrace{\prod_{c \in C} \frac{\prod_{u \in U} \Gamma(n_u^{(c)} + n_{f.u}^{(c)} + \phi_{t-1;c,u})}{\Gamma(\sum_{u \in U} n_u^{(c)} + n_{f.u}^{(c)} + \phi_{t-1;c,u})}}_{(T_3)} \times$$
$$\underbrace{\prod_{c \in C} \frac{\prod_{z \in Z} \Gamma(n_z^{(c)} + \pi_{t-1;c,z})}{\Gamma(\sum_{z \in Z} n_z^{(c)} + \pi_{t-1;c,z})}}_{(T_4)} \times \underbrace{\prod_{z \in Z} \frac{\prod_{w \in V} \Gamma(n_w^{(z)} + \varphi_{t-1;z,w})}{\Gamma(\sum_{w \in V} n_w^{(z)} + \varphi_{t-1;z,w})}}_{(T_5)}. \tag{5}$$

From Eq. 5, the joint distribution of the region assignment $r_o$, community assignment $c_o$, and topic assignment $z_o$ of occurrence $o$ is obtained as follows.

$$\begin{aligned}
&P(r_o, c_o, z_o|\boldsymbol{r_{t;-o}}, \boldsymbol{c_{t;-o}}, \boldsymbol{z_{t;-o}}, \mathcal{W}_t; \phi_{t-1}, \pi_{t-1}, \varphi_{t-1}, \alpha, \eta, \sigma) = P(r_o|\eta_t)P(loc_o|loc_{r_o}, \sigma)\times \\
&\qquad \frac{n^{(r_o)}_{-o,c_o} + \alpha_{c_o}}{\sum_{c\in C} n^{(r_o)}_{-o,c} + \alpha_c} \times \frac{n^{(c_o)}_{-o,u_o} + n^{(c_o)}_{f.u_o} + \phi_{t-1;c_o,u_o}}{\sum_{u\in U} n^{(c_o)}_{-o,u} + n^{(c_o)}_{f.u} + \phi_{t-1;c_o,u}} \times \\
&\qquad \frac{n^{(c_o)}_{-o,z_o} + \pi_{t-1;c_o,z_o}}{\sum_{z\in Z} n^{(c_o)}_{-o,z} + \pi_{t-1;c_o,z}} \times \frac{\prod_{w\in o.msg}\prod_{i=1}^{n_w.msg}(i-1+n^{(z_o)}_{-w,w} + \varphi_{t-1;z_o,w})}{\prod_{i=1}^{n.msg}(i-1+\sum_{w\in V} n^{(z_o)}_{-w,w} + \varphi_{t-1;z_o,w})} \qquad (6)
\end{aligned}$$

Finally, the sampling rule for each of the assignment variables $r_o$, $c_o$, and $z_o$ is obtained similarly to the corresponding sampling rule in the $rLinkTopic$ model, which is presented as follows.

**1. Sampling rule for region assignment:**

$$\begin{aligned}
P(r_o = r|c_o, z_o, \boldsymbol{r_{-o}}, \boldsymbol{c_{-o}}, \boldsymbol{z_{-o}}, \mathcal{W}_t; \cdot) = &\quad P(r|\eta_t)P(loc_o|loc_r, \sigma) \times \frac{n^{(r)}_{-o,c_o} + \alpha_{c_o}}{\sum_{c\in C} n^{(r)}_{-o,c} + \alpha_c} \\
\propto &\quad exp(-\frac{|loc_o, loc_r|}{\sigma^2}) \times \frac{n^{(r)}_{-o,c_o} + \alpha_{c_o}}{\sum_{c\in C} n^{(r)}_{-o,c} + \alpha_c} \qquad (7)
\end{aligned}$$

**2. Sampling rule for community assignment:**

$$\begin{aligned}
P(c_o = c|r_o, z_o, \boldsymbol{c_{-o}}, \boldsymbol{r_{-o}}, \boldsymbol{z_{-o}}, \mathcal{W}_t; \cdot) \propto &\frac{n^{(c)}_{-o,u_o} + n^{(c)}_{-o,f.u_o} + \phi_{t-1;c,u_o}}{\sum_{u\in U} n^{(c)}_{-o,u} + n^{(c)}_{-o,f.u} + \phi_{t-1;c,u}} \\
&\times \frac{n^{(r_o)}_{-o,c} + \alpha_c}{\sum_{c'\in C} n^{(r_o)}_{-o,c'} + \alpha_{c'}} \times \frac{n^{(c)}_{-o,z_o} + \pi_{t-1;c,z_o}}{\sum_{z\in Z} n^{(c)}_{-o,z} + \pi_{t-1;c,z}} \qquad (8)
\end{aligned}$$

**3. Sampling rule for topic assignment:**

$$\begin{aligned}
P(z_o = z|r_o, c_o, \boldsymbol{r_{-o}}, \boldsymbol{c_{-o}}, \boldsymbol{z_{-o}}, \mathcal{W}_t; \cdot) \propto &\frac{\prod_{w\in o.msg}\prod_{i=1}^{n_w.msg}(i-1+n^{(z)}_{-w,w} + \varphi_{t-1;z_o,w})}{\prod_{i=1}^{n.msg}(i-1+\sum_{w\in V} n^{(z)}_{-w,w} + \varphi_{t-1;z_o,w})} \\
&\times \frac{n^{(c_o)}_{-o,z} + \pi_{t-1;c_o,z}}{\sum_{z'\in Z} n^{(c_o)}_{-o,z'} + \pi_{t-1;c_o,z'}} \qquad (9)
\end{aligned}$$

**Gibbs sampling algorithm.** The Gibbs sampling algorithm for the $ErLinkTopic$ model is shown in Algorithm 1. Input of the algorithm is a sequence of sliding windows $SN = \{\mathcal{W}_1, \mathcal{W}_2, ..., \mathcal{W}_T\}$ and the hyperparameters. Hidden variables are first estimated for the first sliding window $\mathcal{W}_1$ using the $rLinkTopic$ model with the given hyperparameters. From the second sliding window, the $rLinkTopic$ model is employed in the way that the values of $\phi_{t-1}$, $\pi_{t-1}$ and $\varphi_{t-1}$ obtained from the previous sliding window are used as the prior hyperparameters of model. Based on the sequence of each of these variables computed over sliding windows, the evolution of communities regarding the community membership of users, the topic proportion of communities, and the distribution of terms in topics is then analyzed. It is noted that *ErLinkTopic* has the same computational complexity as *rLinkTopic*. For a snapshot $sn_t$ having $|R_t|$ regions, the computation for an occurrence $o$ at a sampling step has complexity $O(|R_t| + |C| + |Z|)$. Therefore, the complexity of the algorithm for a network of $T$ snapshots and with $I$ iterations for sampling will be $O(I \times T \times |sn_t| \times (|R_t| + |C| + |Z|))$.

**Algorithm 1:** Gibbs sampling algorithm for the *ErLinkTopic* probabilistic model.

```
Input:
$SN = \{\mathcal{W}_1, \mathcal{W}_2, ..., \mathcal{W}_T\}$: sequence of network sliding windows
$|C|$: number of communities to be extracted
$|Z|$: number of topics associated with communities
$minRad$: a threshold to determine representative locations of regions
$\sigma$: prior standard deviation for Gaussian
$\alpha, \beta, \gamma, \mu$: Dirichlet hyperparameters
Output:
set of evolving communities characterized by:
(1) $\theta = \{\theta_1, \theta_2, ..., \theta_T\}$: sequence of distributions of communities in regions
(2) $\phi = \{\phi_1, \phi_2, ..., \phi_T\}$: sequence of distributions of users in communities
(3) $\pi = \{\pi_1, \pi_2, ..., \pi_T\}$: sequence of topic proportions of communities
(4) $\varphi = \{\varphi_1, \varphi_2, ..., \varphi_T\}$: sequence of distributions of terms in topics
/* first sliding window */
$\phi_1, \pi_1, \varphi_1, \theta_1 \leftarrow$ rLinkTopic($\mathcal{W}_1, |C|, |Z|, \alpha, \beta, \gamma, \mu, minRad, \sigma$);
/* from second sliding window */
foreach $t = 2..T$ do
    $\phi_t, \pi_t, \varphi_t, \theta_t \leftarrow$ rLinkTopic($\mathcal{W}_t, |C|, |Z|, \alpha, \phi_{t-1}, \pi_{t-1}, \varphi_{t-1}, minRad, \sigma$);
    /* detect changes in community memberships of users */
    detectChangesFrom($\phi_{t-1}, \phi_t$);
    /* detect changes in topic proportions of communities */
    detectChangesFrom($\pi_{t-1}, \pi_t$);
    /* detect changes in topics of communities */
    detectChangesFrom($\varphi_{t-1}, \varphi_t$);
```

**Tab. 3:** *Statistics of Twitter datasets used to evaluate the ErLinkTopic model in extracting regional LinkTopic communities and analyzing their evolution.*

| Dataset | Users/Filtered | Tweets/Filtered | Terms/Filtered | Time |
|---|---|---|---|---|
| **Sub-England** | 1.720.956/18.264 | 13.114.353 /6.572.764 | 2.915.851/15.215 | June 01 - Nov 28 |
| **Sub-US** | 980.924/14.756 | 6.301.435/3.654.000 | 2.135.098/16.260 | June 01 - Nov 28 |

# 5 Experiments

This section presents the experimental results of applying our approach to extracting and analyzing the evolution of regional *LinkTopic* communities in social networks. Particularly, by using *Twitter* data, we show the effectiveness and efficiency of the *ErLinkTopic* model in terms of discovering communities and, at the same time, capturing changes in the features describing communities. Our framework is implemented in Java. All experiments are run on an Intel(R) Core(TM) i7-4770 CPU @ 3.40G with 16GB RAM, running Ubuntu 64bit.

## 5.1 Twitter Datasets

We use two six-month interval Twitter datasets collected from the **EUROPE** and **US** for conducting the experiments. The first subset is called **Sub-England** dataset and the second subset is called **Sub-US** dataset. A filtering step is applied so that users posting less than 180 messages, i.e., on average 1 message a day, and terms occurring less than 360 times, i.e., on average 2 time a day, are removed from the **Sub-US** dataset. Such numbers applied to filter users and terms in the **Sub-England** dataset are 180 and 540, respectively. Relevant statistics of the two datasets before and after filtering users and terms are summarized in Table 11. The main objective of our experiments is to extract communities and capture their evolution from which to study how the features describing a community evolve over time. Besides this, it is also necessary to verify the efficiency of the *ErLinkTopic* model regarding the computational complexity.

## 5.2 Evaluation measures

To study the evolution of features associated with communities, the following notations are introduced, given the parameters $numU, numZ$, and $numV$.

1. $U(c, t, numU)$: set of $numU$ users that have the highest likelihood in community $c$ at sliding window $\mathcal{W}_t$.

2. $Z(c, t, numZ)$: set of $numZ$ topics that have the highest likelihood in community $c$ at $\mathcal{W}_t$.

3. $V(z, t, numV)$: set of $numV$ terms that have the highest likelihood in topic $z$ at $\mathcal{W}_t$.

Based on these notations, the evolution of a community with respect to the community members, community topics, and terms in topics is formalized in the following sections.

**Dynamics of users.** To capture the dynamics of users in community $c$ over two consecutive sliding windows $\mathcal{W}_{t-1}$ and $\mathcal{W}_t$, we introduce a *user dynamic* measure $\partial_\phi(c, t-1, t, numU)$, computed as follows.

$$\partial_\phi(c, t-1, t, numU) = \frac{numU - |U(c, t-1, numU) \cap U(c, t, numU)|}{numU} \in [0, 1] \qquad (10)$$

**Topic-prominence dynamic.** The $\partial_\pi(c, t-1, t, numZ)$ is defined to determine the frequency of updating the prominence of the topics associated with community $c$.

$$\partial_\pi(c, t-1, t, numZ) = \frac{numZ - |Z(c, t-1, numZ) \cap Z(c, t, numZ)|}{numZ} \in [0, 1] \tag{11}$$

**Term dynamic.** Finally, the $\partial_\varphi(z, t-1, t, numV)$ is defined to measure the frequency of changes of terms occurring in a topic $z$.

$$\partial_\varphi(z, t-1, t, numV) = \frac{numV - |V(z, t-1, numV) \cap V(z, t, numV)|}{numV} \in [0, 1] \tag{12}$$

## 5.3 Dynamic Measure Analysis

Based on the results extracted from the three different settings of sliding windows, i.e., 1-week interval, 2-week interval, and 1-month interval, we study the dynamics of communities in terms of changes in (1) the members of each community using the user dynamic measure $\partial_\phi(c, t-1, t, numU)$, (2) the prominence of topics associated with each community using the topic-prominence dynamic measure $\partial_\pi(c, t-1, t, numZ)$, and (3) terms occurring in each community topic using the term dynamic measure $\partial_\varphi(z, t-1, t, numW)$. We visualize the community membership of users in each community and the likelihood of terms in each topic to determine appropriate values for $numU$ and $numW$, respectively. By studying the community membership of users, we find two prevalent points at $numU = 5$ and $numU = 30$ where the likelihood of users in every community strongly decreases. However, the top 5 users in all communities change frequently at every sliding window. We therefore select $numU = 30$ for evaluating the dynamics of users in communities. Applying the same method we determine that a good value for $numW$ is 20.

Finally, we choose $numZ = 5$ for measuring the dynamics of the prominence of community topics. The following findings are obtained from both two datasets.

1. Communities evolve gradually over a short time interval of sliding windows. This evolving trend applies to all three features of interests, i.e., community members, community topics, and terms describing a topic. Changes to these features happen more often when longer time intervals are employed to form a sliding window. This finding confirms that social networks and especially communities in social networks are dynamic structures.

2. Community members evolve faster than community topics, which is indicated by a larger value of $\partial_\phi(c, t-1, t, numU)$ compared to the value of $\partial_\pi(c, t-1, t, numZ)$ or $\partial_\varphi(z, t-1, t, numW)$. This implies that the topics discussed by a community are more stable regarding both the topic prominence and terms describing topics even though users might change topics of interest and leave a community and join other communities more often. The dynamic measures of three communities extracted from the **Sub-US** dataset and five communities extracted from the **Sub-England** dataset are presented in Table 5.3 and Table 5.3, respectively.

**Tab. 4:** *Dynamic measures computed at the first five sliding windows for three selected communities extracted from the* ***Sub-US*** *dataset.*

| **Two selected politics communities:** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Sliding Window** | **1-week interval** | | | **2-week interval** | | | **1-month interval** | | |
| | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ |
| 01 | 0.40 | 0.20 | 0.35 | 0.73 | 0.60 | 0.40 | 0.93 | 0.40 | 0.30 |
| 02 | 0.60 | 0.20 | 0.40 | 0.76 | 0.40 | 0.40 | 0.93 | 0.40 | 0.40 |
| 03 | 0.63 | 0.40 | 0.25 | 0.70 | 0.40 | 0.35 | 0.96 | 0.40 | 0.65 |
| 04 | 0.53 | 0.40 | 0.35 | 0.63 | 0.40 | 0.60 | 0.93 | 0.40 | 0.70 |
| 05 | 0.66 | 0.0 | 0.45 | 0.76 | 0.20 | 0.35 | 0.70 | 0.40 | 0.75 |
| **Average** | **0.56** | **0.24** | **0.36** | **0.71** | **0.40** | **0.41** | **0.89** | **0.40** | **0.56** |
| 01 | 0.56 | 0.20 | 0.20 | 0.76 | 0.40 | 0.30 | 0.86 | 0.40 | 0.55 |
| 02 | 0.76 | 0.20 | 0.30 | 0.70 | 0.20 | 0.25 | 0.96 | 0.40 | 0.68 |
| 03 | 0.70 | 0.20 | 0.20 | 0.73 | 0.20 | 0.10 | 0.96 | 0.40 | 0.60 |
| 04 | 0.66 | 0.0 | 0.15 | 0.66 | 0.40 | 0.15 | 0.86 | 0.60 | 0.72 |
| 05 | 0.56 | 0.0 | 0.20 | 0.63 | 0.30 | 0.30 | 0.90 | 0.60 | 0.62 |
| **Average** | **0.65** | **0.12** | **0.21** | **0.70** | **0.30** | **0.22** | **0.91** | **0.48** | **0.63** |
| **Two selected job communities:** | | | | | | | | | |
| **Sliding Window** | **1-week interval** | | | **2-week interval** | | | **1-month interval** | | |
| | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ |
| 01 | 0.66 | 0.10 | 0.20 | 0.76 | 0.40 | 0.40 | 0.86 | 0.60 | 0.35 |
| 02 | 0.63 | 0.20 | 0.25 | 0.86 | 0.40 | 0.40 | 1.00 | 0.40 | 0.45 |
| 03 | 0.76 | 0.20 | 0.20 | 0.86 | 0.20 | 0.35 | 0.93 | 0.60 | 0.60 |
| 04 | 0.66 | 0.0 | 0.25 | 0.93 | 0.60 | 0.60 | 1.00 | 0.20 | 0.70 |
| 05 | 0.76 | 0.0 | 0.15 | 0.80 | 0.80 | 0.10 | 0.86 | 0.40 | 0.80 |
| **Average** | **0.69** | **0.10** | **0.21** | **0.84** | **0.48** | **0.37** | **0.93** | **0.44** | **0.58** |
| 01 | 0.76 | 0.20 | 0.20 | 0.75 | 0.60 | 0.35 | 0.85 | 0.40 | 0.60 |
| 02 | 0.63 | 0.20 | 0.25 | 0.73 | 0.20 | 0.40 | 0.80 | 0.40 | 0.65 |
| 03 | 0.66 | 0.0 | 0.20 | 0.80 | 0.60 | 0.65 | 0.93 | 0.60 | 0.55 |
| 04 | 0.70 | 0.0 | 0.25 | 0.76 | 0.20 | 0.55 | 0.96 | 0.40 | 0.70 |
| 05 | 0.60 | 0.0 | 0.15 | 0.63 | 0.40 | 0.55 | 0.93 | 0.50 | 0.50 |
| **Average** | **0.67** | **0.08** | **0.21** | **0.73** | **0.40** | **0.50** | **0.89** | **0.46** | **0.60** |
| **Two selected weather community:** | | | | | | | | | |
| **Sliding Window** | **1-week interval** | | | **2-week interval** | | | **1-month interval** | | |
| | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ |
| 01 | 0.63 | 0.30 | 0.25 | 0.63 | 0.60 | 0.40 | 0.90 | 0.40 | 0.40 |
| 02 | 0.70 | 0.0 | 0.45 | 0.70 | 0.60 | 0.45 | 1.00 | 0.20 | 0.70 |
| 03 | 0.66 | 0.0 | 0.50 | 0.76 | 0.20 | 0.50 | 0.93 | 0.60 | 0.75 |
| 04 | 0.66 | 0.0 | 0.40 | 0.86 | 0.80 | 0.55 | 0.96 | 0.0 | 0.70 |
| 05 | 0.76 | 0.0 | 0.30 | 0.66 | 0.60 | 0.45 | 0.93 | 0.60 | 0.70 |
| **Average** | **0.68** | **0.06** | **0.38** | **0.72** | **0.56** | **0.47** | **0.94** | **0.36** | **0.65** |
| 01 | 0.66 | 0.20 | 0.45 | 0.73 | 0.40 | 0.50 | 0.83 | 0.40 | 0.55 |
| 02 | 0.50 | 0.30 | 0.55 | 0.76 | 0.40 | 0.40 | 0.93 | 0.40 | 0.50 |
| 03 | 0.63 | 0.0 | 0.25 | 0.80 | 0.10 | 0.60 | 1.00 | 0.40 | 0.55 |
| 04 | 0.50 | 0.0 | 0.30 | 0.73 | 0.20 | 0.55 | 0.86 | 0.20 | 0.65 |
| 05 | 0.56 | 0.20 | 0.15 | 0.70 | 0.40 | 0.60 | 0.93 | 0.40 | 0.70 |
| **Average** | **0.59** | **0.14** | **0.34** | **0.74** | **0.30** | **0.53** | **0.91** | **0.36** | **0.59** |

**Tab. 5:** *Dynamic measures computed at the first five sliding windows for five selected communities extracted from the* ***Sub-England*** *dataset.*

| **A selected football community:** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Sliding Window** | **1-week interval** | | | **2-week interval** | | | **1-month interval** | | |
| | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ |
| 01 | 0.40 | 0.0 | 0.35 | 0.63 | 0.20 | 0.50 | 0.73 | 0.40 | 0.60 |
| 02 | 0.53 | 0.20 | 0.40 | 0.73 | 0.0 | 0.45 | 0.83 | 0.20 | 0.50 |
| 03 | 0.50 | 0.0 | 0.35 | 0.76 | 0.20 | 0.35 | 0.86 | 0.20 | 0.65 |
| 04 | 0.53 | 0.20 | 0.45 | 0.80 | 0.0 | 0.50 | 0.83 | 0.20 | 0.60 |
| 05 | 0.46 | 0.0 | 0.45 | 0.83 | 0.20 | 0.60 | 0.70 | 0.40 | 0.65 |
| **Average** | **0.48** | **0.08** | **0.40** | **0.75** | **0.12** | **0.48** | **0.79** | **0.28** | **0.60** |
| **A selected social media community:** | | | | | | | | | |
| **Sliding Window** | **1-week interval** | | | **2-week interval** | | | **1-month interval** | | |
| | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ |
| 01 | 0.46 | 0.0 | 0.20 | 0.66 | 0.0 | 0.25 | 0.76 | 0.20 | 0.35 |
| 02 | 0.53 | 0.0 | 0.25 | 0.70 | 0.0 | 0.35 | 0.86 | 0.40 | 0.45 |
| 03 | 0.66 | 0.20 | 0.25 | 0.76 | 0.20 | 0.30 | 0.83 | 0.20 | 0.60 |
| 04 | 0.66 | 0.0 | 0.35 | 0.86 | 0.0 | 0.40 | 0.80 | 0.20 | 0.50 |
| 05 | 0.56 | 0.20 | 0.15 | 0.86 | 0.40 | 0.25 | 0.86 | 0.20 | 0.40 |
| **Average** | **0.57** | **0.08** | **0.24** | **0.76** | **0.12** | **0.31** | **0.82** | **0.24** | **0.46** |
| **A selected weather community:** | | | | | | | | | |
| **Sliding Window** | **1-week interval** | | | **2-week interval** | | | **1-month interval** | | |
| | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ |
| 01 | 0.45 | 0.20 | 0.20 | 0.76 | 0.20 | 0.45 | 0.75 | 0.40 | 0.50 |
| 02 | 0.51 | 0.0 | 0.30 | 0.80 | 0.20 | 0.35 | 0.80 | 0.20 | 0.40 |
| 03 | 0.53 | 0.0 | 0.22 | 0.73 | 0.0 | 0.30 | 0.85 | 0.20 | 0.55 |
| 04 | 0.60 | 0.20 | 0.40 | 0.73 | 0.40 | 0.40 | 0.75 | 0.20 | 0.65 |
| 05 | 0.55 | 0.20 | 0.10 | 0.60 | 0.20 | 0.55 | 0.83 | 0.40 | 0.50 |
| **Average** | **0.53** | **0.12** | **0.24** | **0.72** | **0.20** | **0.41** | **0.80** | **0.32** | **0.52** |
| **A selected food community:** | | | | | | | | | |
| **Sliding Window** | **1-week interval** | | | **2-week interval** | | | **1-month interval** | | |
| | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ |
| 01 | 0.45 | 0.20 | 0.10 | 0.73 | 0.20 | 0.40 | 0.80 | 0.20 | 0.50 |
| 02 | 0.50 | 0.0 | 0.30 | 0.66 | 0.0 | 0.75 | 0.83 | 0.20 | 0.40 |
| 03 | 0.30 | 0.20 | 0.20 | 0.76 | 0.30 | 0.35 | 0.73 | 0.40 | 0.55 |
| 04 | 0.50 | 0.20 | 0.15 | 0.83 | 0.20 | 0.25 | 0.90 | 0.20 | 0.30 |
| 05 | 0.53 | 0.0 | 0.20 | 0.63 | 0.0 | 0.50 | 0.85 | 0.40 | 0.60 |
| **Average** | **0.46** | **0.12** | **0.19** | **0.72** | **0.14** | **0.45** | **0.82** | **0.28** | **0.47** |
| **A selected music and event community:** | | | | | | | | | |
| **Sliding Window** | **1-week interval** | | | **2-week interval** | | | **1-month interval** | | |
| | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ | $\partial_\phi$ | $\partial_\pi$ | $\partial_\varphi$ |
| 01 | 0.30 | 0.0 | 0.20 | 0.63 | 0.0 | 0.25 | 0.72 | 0.20 | 0.40 |
| 02 | 0.40 | 0.20 | 0.30 | 0.73 | 0.20 | 0.45 | 0.80 | 0.20 | 0.60 |
| 03 | 0.45 | 0.0 | 0.32 | 0.76 | 0.20 | 0.80 | 0.65 | 0.20 | 0.55 |
| 04 | 0.41 | 0.0 | 0.20 | 0.80 | 0.0 | 0.35 | 0.85 | 0.40 | 0.45 |
| 05 | 0.50 | 0.20 | 0.35 | 0.73 | 0.40 | 0.50 | 0.80 | 0.40 | 0.40 |
| **Average** | **0.41** | **0.08** | **0.27** | **0.73** | **0.16** | **0.47** | **0.76** | **0.28** | **0.48** |

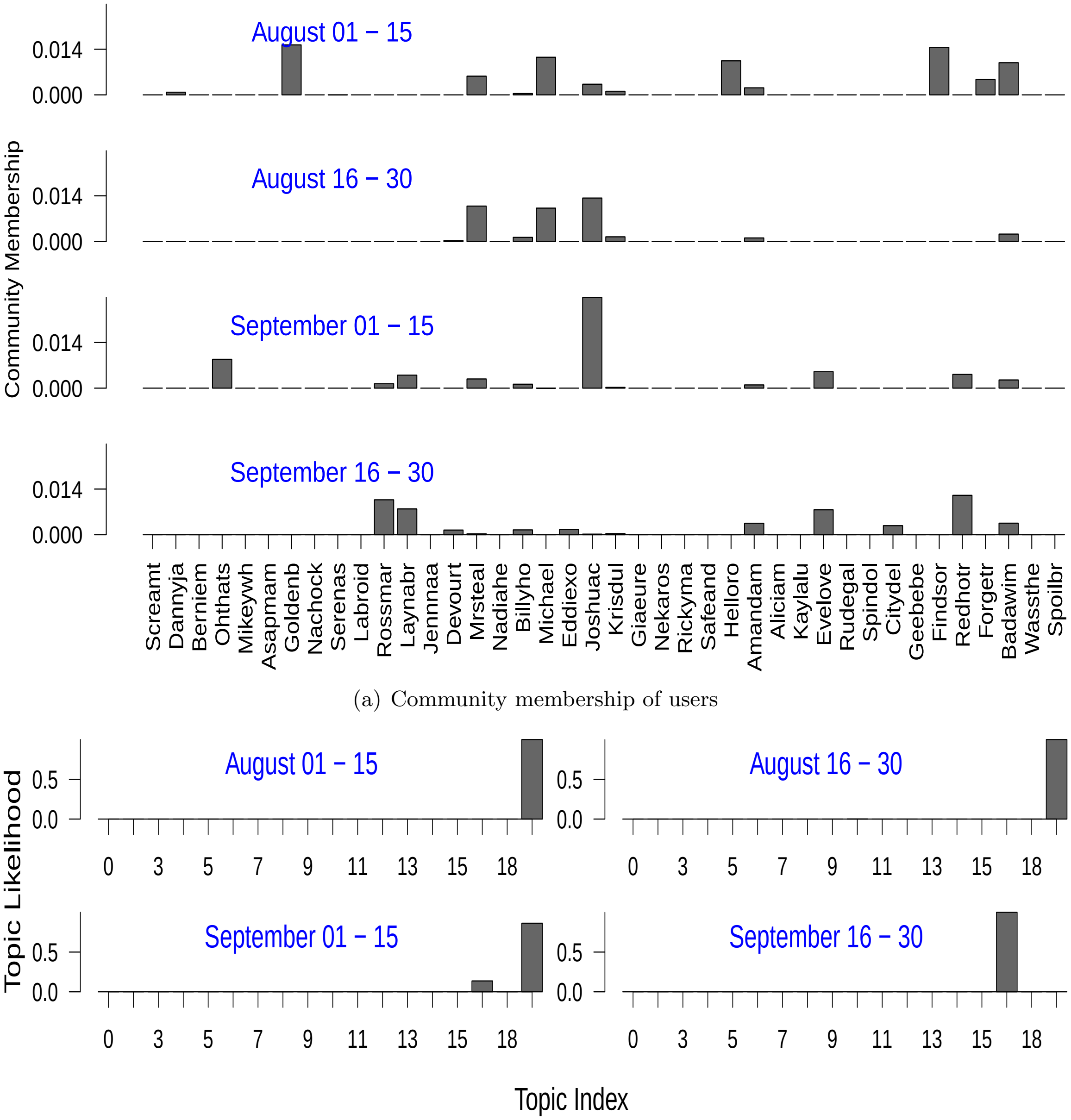
## 5.4 Evolving Communities

Example communities extracted from the **Sub-US** dataset are presented in this section to demonstrate the effectiveness of the *ErLinkTopic* model in extracting evolving communities. For this purpose, topics associated with communities extracted by the model are first manually classified into the groups *politics, jobs, social activities, weather, music and social events, social media, social networks, sports,* and *general.* A topic is labeled as *general* if terms occurring in that topic are about different subjects making it unclear for a classification. We manually label each community based on the prominence of topics associated with it. Generally, each community is associated with at most two topics at a time point. The evolution of each community is characterized by changes in the community membership of users, the prominence of topics, and the likelihood of terms in each topic as well. Evolving phenomena that are observed from communities extracted from our datasets include the stability, generalization, specification, and shifting of the prominence of topics associated with a community; the growth and shrinkage of community members; and the stability of terms describing topics. In our experiments, we rarely find the stability of community members, especially when a sliding window of more than 2-week interval is applied. This indicates that users in social networks in general and particularly *Twitter* users are dynamic in terms of posting messages associated with contextual links of different topics reflecting their complex life and changing geographic locations over time.

As an example, we find an interesting trend from the **Sub-US** dataset that communities characterized by a *job* topic tend to shift their interest to politics before the election in the US in 2012. Figure 5.4 shows an example. At first, this community is associated with a topic described by terms about jobs (the topic indexed 19) during August 2012. The shifting of topics happens at the beginning of September 2012, where the likelihood of the topic described by terms about politics (the topic indexed 16) increases. By the end of September 2012, the community is characterized by only the *politics* topic.

## 5.5 Evaluation of Runtime

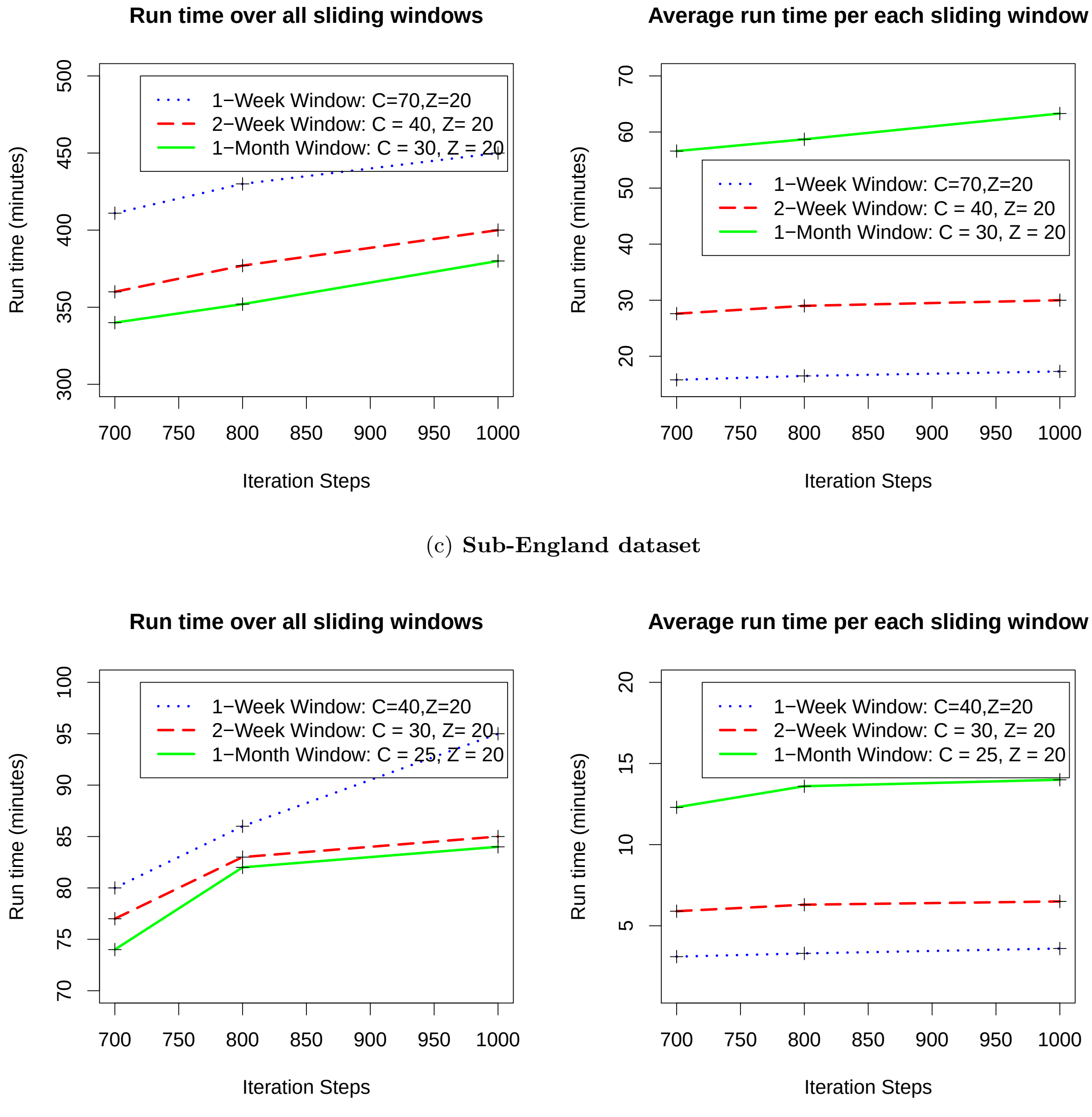
This section discusses the running time of the *ErLinkTopic* algorithm applied to the datasets used in the experiments presented. Particularly, for each time interval of sliding windows, we measure the running time of the algorithm using three different settings of the number of iterations for sampling. In the first setting, the model is run with 820 steps for the *Burn-In* stage and 180 steps for collecting assignment samples and updating multinomial parameters. The results (i.e., the communities, topics, and their evolution) presented in this paper are derived from this configuration. In the second setting, 700 steps for the *Burn-In* stage and 100 steps for collecting assignment samples and updating multinomial parameters are employed. Such steps of iterations for the last setting are 600 and 100, respectively. The results show that for each dataset the model takes almost the same time when it is run with different time intervals of sliding windows, given that the same number of communities $|C|$ and number of topics $|Z|$ are assigned to the model. Also, the running time of the algorithm increases linearly to the number of iterations and the number of communities applied. Details of the evaluations are summarized in Figure 5.4.



(a) Community membership of users

(b) Prominence of topics associated with the community

**Fig. 3:** *The evolution of community members and the shifting of the prominence of a topic about jobs (indexed 19) to a topic about politics (indexed 16) of a community discovered from the* ***Sub-US*** *dataset.*



(c) **Sub-England dataset**

(d) **Sub-US dataset**

**Fig. 4:** *Running time of the ErLinkTopic algorithm applied to the* ***Sub-England*** *dataset (c) and* ***Sub-US*** *dataset (d). Three time intervals (1 week, 2 weeks, and 1 month) are employed to create sliding windows. For each time interval, three settings of the number of iterations (700, 800, and 1000) are used in the ErLinkTopic algorithm.*

## 6 Conclusion

We have presented a probabilistic model called *ErLinkTopic* to analyze regional link-topic communities. Important features that have not been considered in existing studies, i.e., capturing and analyzing the evolution of community attributes, are addressed in our framework. There are aspects in the proposed framework that we would like to study in order to improve the model. First, in this framework, regions are derived from the density of geographic locations of users within each snapshot. This implies an assumption that regions might change over time. Because of this, the model ignores the evolution of the community distribution in each region. There should be an improvement for the model in a way that it is able to capture region evolution as well. Second, due to the lack of ground truth in real-world datasets, evaluating the results of extracting feature-based communities and analyzing their evolution is a challenging task. Finally, in our framework, we assume there are no changes in the number of communities $|C|$ and the number of topics $|Z|$ across time. It should be more appropriate if a *Dirichlet* process is employed so that these constraints are relaxed.

## REFERENCES

[1] Canh T. V., Gertz M., “rlinktopic: A probabilistic model for discovering regional linktopic communities,” In *ASONAM 2014*, eds. Wu X., Ester M., Xu G., IEEE Computer Society, 2014, pp. 24-26.

[2] Kernighan, B.W., Lin S.. “An Efficient Heuristic Procedure for Partitioning Graphs”, *The Bell system technical journal,* **49**(1), pp. 291-307, 1970.

[3] Newman M. E. J., Girvan M., “Finding and evaluating community structure in networks”, *Pattern Recognition Letters,* **69**(5), pp. 413-421, 2004.

[4] Ruan J., Zhang W., “An efficient spectral algorithm for network community discovery and its applications to biological and social networks,” In *Proceedings of the 2007, Seventh IEEE International Conference on Data Mining. ICDM ’07, Washington, DC, USA, IEEE Computer Society,* 2007, pp. 643-648.

[5] Pathak A. B. N., Erickson K., “Social topic models for community extraction,” In *The 2nd SNA-KDD Workshop ’08 (SNA-KDD’08), Las Vegas, Nevada, USA*, 2008.

[6] Sachan M., Contractor D., Faruquie T. A., Subramaniam L. V., “Using content and interactions for discovering communities in social networks,” In *Proceedings of the 21st International Conference on World Wide Web. WWW ’12, New York, NY, USA, ACM*, 2012, pp. 331-340.

[7] Zheng G., Guo J., Yang L., Xu S., Bao S., Su Z., Han D., Yu Y., “Mining topics on participations for community discovery,” In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. SIGIR ’11*, New York, NY, USA, ACM, 2011, pp. 445-454.

[8] Asur S., Parthasarathy S., Ucar D., “An event-based framework for characterizing the evolutionary behavior of interaction graphs,” In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, ACM*, 2007, pp. 913-921.

[9] Chakrabarti D., Kumar R., Tomkins A., “Evolutionary clustering,” In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge discovery and Data Mining, KDD ’06, New York, USA, ACM*, 2006, pp. 554-560.

[10] Lin Y. R., Chi Y., Zhu S., Sundaram H., Tseng B. L, “Analyzing communities and their evolutions in dynamic social networks,” *ACM Trans. Knowl. Discov. Data*, **3**(2) 8:1–8:31, 2009.

[11] Lin Y. R., Sun J., Sundaram H., Kelliher A., Castro P., Konuru R., “Community discovery via metagraph factorization,” *ACM Trans. Knowl. Discov. Data*, **5**(3), 17:1–17:44, 2011.

[12] Costa G., Ortale R., “A bayesian hierarchical approach for exploratory analysis of communities and roles in social networks,” In *ASONAM*, IEEE Computer Society, 2012, pp. 194-201.

[13] Natarajan N., Sen P., Chaoji V., “Community detection in content-sharing social networks”, In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ASONAM ’13, New York, NY, USA, ACM*,2013, pp. 82–89.

[14] Zeng Z., Wu B., “Detecting probabilistic community with topic modeling on sampling subgraphs,” In *ASONAM, IEEE Computer Society*, 2012, pp. 623-630.

[15] Zhou D., Manavoglu E., Li J., Giles, C.L., Zha, H., “Probabilistic models for discovering e-communities”, In *Proceedings of the 15th International Conference on World Wide Web. WWW ’06, New York, NY, USA, ACM*, 2006, pp. 173-182.

[16] Spiliopoulou M., Ntoutsi I., Theodoridis, Y., Schult, R. “Monic: modeling and monitoring cluster transitions,” In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge discovery and Data Mining. KDD ’06, New York, NY, USA, ACM*, 2006, pp. 706-711.

[17] Palla G., Derúnyi I., Farkas I., Vicsek T., “Uncovering the overlapping community structure of complex networks in nature and society,” *Nature*, **435**(7043), pp. 814-818, 2005.

[18] Palla G., lászló Barabási A., Vicsek T., Hungary B., “Quantifying social group evolution,” *Nature*, **446**, 2007.

[19] Lin Y. R., Chi Y., Zhu S., Sundaram H., Tseng,B. L., “Facetnet: a framework for analyzing communities and their evolutions in dynamic networks,” In: *Proceedings of the 17th International Conference on World Wide Web. WWW ’08, New York, NY, USA, ACM*, 2008, pp. 685-694.

[20] Dhillon I. S., Sra S., “Generalized nonnegative matrix approximations with Bregman

divergences," In *Neural Information Proc. Systems*, pp. 283–290, 2005.

[21] Chi Y., Song X., Zhou D., Hino K., Tseng B. L., "Evolutionary spectral clustering by incorporating temporal smoothness," In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge discovery and Data Mining. KDD '07, New York, NY, USA, ACM*, 2007, pp. 153-162.

[22] Chi Y., Song X., Zhou D., Hino K., Tseng B. L., "On evolutionary spectral clustering," *ACM Trans. Knowl. Discov. Data*, **3**(4), 17:1–17:30, 2009.

[23] Hofman J.M., Wiggins C.H., "A bayesian approach to network modularity," *Physical Review Letters*, **100**(25), pp. 1–4, 2007.

[24] Yang T., Chi Y., Zhu S., Gong Y., Jin R., "Detecting communities and their evolutions in dynamic social networks-a bayesian approach," *Machine Learning*, **82**, pp. 157–189, 2001. DOI: 10.1007/s10994-010-5214-7.

## TÓM TẮT

## MÔ HÌNH SINH XÁC SUẤT PHÁT HIỆN VÀ HỖ TRỢ PHÂN TÍCH NHÓM CỘNG ĐỒNG TRÊN MẠNG XÃ HỘI

Bài báo này giới thiệu mô hình xác xuất sinh dữ liệu có khả năng học cấu trúc và hỗ trợ phân tích sự phát triển của các nhóm cộng đồng trên mạng xã hội được xác định dựa trên các tiêu chí về vùng không gian địa lý (region), chủ đề quan tâm (topic), và tương tác (interaction). Chúng tôi trình bày chi tiết mô hình sinh xác suất (generative model) *ErLinkTopic* từ việc mở rộng mô hình *rLinkTopic* [1] và thuật toán Gibbs sampling tương ứng. Kết quả đánh giá thuật toán bằng việc sử dụng dữ liệu từ mạng xã hội Twitter cho thấy các kết quả khá thú vị khẳng định tính khả thi của thuật toán.