
This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/4477/

Link to published version:

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.
Correlated Community Estimation Models Over a Set of Names

Suresh Veluru∗, Yogachandran Rahulamathavan∗, Suresh Manandhar†, and Muttukrishnan Rajarajan∗

∗Information Security Group, School of Engineering and Mathematical Sciences, City University London, London, EC1V 0HB, United Kingdom.
†Department of Computer Science, University of York, Heslington, York, YO10 5GH, United Kingdom.
E-mail: {Suresh.Veluru.1, Yogachandran.Rahulamathavan.1, R.Muttukrishnan}@city.ac.uk

Abstract—Generally surnames (family name) or forenames are evolved over generations which can be used to understand population origins, migration, identity, social norms and cultural customs. These forenames or surnames may have hidden structure associated with them called communities. Each community might have strong correlation among several forenames and surnames. In addition, the correlation might be across communities of forenames or surnames. Popular statistical generative model such as Latent Dirichlet Allocation (LDA) has been developed to find topics in a corpus of documents. However, the LDA model can be proposed to identify hidden communities in names data set. This paper proposes several variants of latent Dirichlet allocation models to capture correlation between surnames and forenames within the communities and across the communities over a set of names collected at different locations. Initially, we propose surname correlated LDA model and forename correlated LDA model. These models identify communities in surnames or forenames and extract corresponding correlated forenames or surnames in each community respectively. Later, we propose surname community correlated LDA model and forename community correlated LDA model. These models estimate correlation among each surname community to the communities of forenames and vice versa respectively. We experiment for India and United Kingdom names data sets and conclusions are drawn.

Keywords—Latent Dirichlet Allocation; Communities; Probabilistic Generative Models; Bayesian Statistics; Correlation;

I. INTRODUCTION

Due to rapid growth of digital data, knowledge discovery and data mining have great potential which would turn data into useful information and knowledge. Text mining (sometimes called ‘mining from text documents’) is to extract knowledge from a set of text documents [16]. Names analysis is popular in geography [15] which relay on the fact that family names (surnames) or names represent ethnic, geographic, cultural and genetic structures in human populations. However, these methods in geography use elementary statistical approaches to analyse names data set. Many advanced statistical methods have limited applications in names analysis.

Knowledge discovery in names data set involves identifying relationship among group of people (surnames) or identifying communities in names data set. It is a well known fact that people migrate from one location to other due to job prospects, economic prosperity, political unrest, etc. However, the surnames of migrants retain semantic similarity to surnames of the people at their original locations. In order to address this issue of identifying semantic surnames, Veluru et al. [26] [25] recently applied statistical methods such as vector space model and latent semantic indexing (LSI) in names data set. Further, email address categorization has been performed based on semantics of surnames. The generative probabilistic model can be applied to identify hidden communities in a names data set.

Generative probabilistic model such as Latent Dirichlet Allocation (LDA) becomes attractive and powerful in natural language processing for topic modelling [12]. It works on discrete data of words in a corpus of documents and overcomes the limitations of LSI and probabilistic LSI (pLSI). It assumes document contains “bag-of-words” which means the order of words in the document can be neglected and also assume that the order of documents can be neglected [12]. This is called exchangeability assumption in the language of probability. de Finette [3] established a classic theorem that states any collection of exchangeable random variables has a representation as a mixture distribution. Hence, LDA model estimates statistical inferences of topics via mixing distribution in a collection of documents.

A names data set contains a set of names collected at several locations in a country which does not depend on order of names collected at each location or order of locations in the data set. The assumption of exchangeability in a names data set is obvious since the order of names in each location and the order of locations can be neglected. Hence, LDA can be applied on names data set that identifies hidden structure associated in it called communities. However, name consists of forename and surname. It is possible that several surnames highly correlate to several forenames. For example, surname Smith is highly correlated to forename David in British community. Hence communities can be estimated either on surnames or forenames and corresponding correlated forenames or surnames can be extracted respectively.

Indeed, several forenames correlate across communities of surnames and viceversa. For example, Sarah and John correlate across many surname communities in United Kingdom. The challenge is to find correlation across communities of surnames or forenames. For example, especially with cross cultural marriages, it may be possible that a community of forenames share high likely with certain communities of surnames and less likely with some other communities of surnames.

This paper proposes several variants of LDA models to address above issues. Initially surname correlated LDA model...
and forename correlated LDA model are proposed. These two models find communities in surnames and forename respectively and extracts corresponding forenames and surnames in each community respectively. Later, we propose surname community correlated LDA model and forename community correlated LDA model. The surname community correlated LDA finds communities in surnames and extracts correlated forename communities for each surname community. Similarly, the forename community correlated LDA model finds communities in forenames and extracts correlated surname communities for each forename community.

This paper is organized as follows. Section II sets out the related work. Section III describes proposed models called correlated community estimation models. Section IV presents the experimental results and finally Section V presents conclusion and future work of the paper.

II. RELATED WORK

This section describes related work for surname analysis. Many surname analysis techniques have been developed in geography such as identifying spatial concentration of surnames [5], migrant surname analysis [19], uncertainty in the analysis of ethnicity classification [22], and ethnicity and population structure analysis [21]. However, statistical analysis that measures the degree of similarity between surname mixes has been developed by comparing relative frequencies of surnames at different locations such as isonymy [18] and Lasker distance [24]. These measures are complementary measures such that the inverse natural logarithm of the isonymy creates a more intuitive measure called Lasker distance. These are applicable to study inbreeding between marital partners or social groups, but do not explicitly address the semantic similarity between surnames. Hence, an advanced statistical analysis method has been developed for email address categorization based on semantics of surnames [26].

E-mail address categorization based on semantics of surnames has two phase [26]. In the first phase, the semantics of surnames are identified by representing a set of names at different locations using a vector space model followed by latent semantics of surnames. Further, clustering of surnames is done using average-link clustering method. In the second phase, suffix tree is constructed for an e-mail address which has been used to identify if any surname present in the email address as substring. If surname is present as substring in the email address then the email address is categorize into the cluster of surname. However, LDA and variants of LDA models have been proposed in text analysis which have not been developed in names analysis. Several variants of LDA have been developed which incorporates meta-data information in generative models that are classified into downstream models and upstream models [8]. Downstream models use standard document-topic distribution and incorporates metadata-topic distribution in parallel to the standard topic-word distribution [6], [13], [29], [30], [9]. However, upstream models replace document-topic distribution with metadata-topic distribution which incorporate additional information and use standard topic-word distribution without any change [2], [20], [8], [7], [10], [11]. Several other variants of LDA models have been developed for several applications such as topic modelling beyond bag-of-words [28], finding scientific topics [27], entity resolution [4] [17] [1], community identification in social networks [14], dynamic models for time series [31], and tag recommendation [23]. However, these variants of LDA models cannot be applied directly to identify communities and their correlation in name data set which is described in the following section.

III. CORRELATED COMMUNITY ESTIMATION MODELS

This section describes the proposed models over names data set. Initially, subsection III-A describes LDA model to estimate communities in either forenames or surnames. Subsection III-B and III-C propose community correlated estimation models and community-community correlated estimation models respectively.

A. Community Estimation Model

This subsection describes the use of LDA for community estimation.

Consider the location space of a region or a country consisting of a set of locations where each location has a bag of names. Let a name can be represented as \( f(i), s(i) > \) where \( f(i) \) be forename and \( s(i) \) be surname of name \( i \). Let there be \( L = \{ l_1, l_2, \ldots, l_m \} \) locations and let be a location has \( N \) names. Let \( W_s \) and \( W_f \) be set of unique surnames and forenames respectively.

LDA is a generative probabilistic model that can be applied to estimate communities over a set of names where names could be either surnames or forenames. Without loss of generality, let us formulate community estimation model in surnames. Consider a community characterized by a distribution over surnames and a location contains a random mixtures of communities. Let \( \phi(W_s) \) or \( \phi(W_f) \) denotes multinomial distributions of communities over the set of surnames \( W_s \) or a set of forenames \( W_f \) respectively. Let \( \theta^{L,J} \) denotes a random mixtures of communities over a set of locations. In statistical theory, if a location contains surnames as a random mixtures over latent \( K \) communities then the probability of \( i^{th} \) surname \( s_i \) in a given location as

\[
P(s(i)) = \sum_{j=1}^{K} P(s(i)|z(s(i)) = j)P(z(s(i)) = j)
\]  

where \( z(s(i)) \) denotes community assignment for surname \( s(i) \), \( P(s(i)|z(s(i)) = j) \) is the probability the surname \( s(i) \) given community \( j \), and \( P(z(s(i)) = j) \) is the probability of choosing a community \( j \) in the current location. Hence \( P(s(i)|z(s(i)) = j) \) is \( \phi^{s(i)} \) and \( P(z(s(i)) = j) = \theta^{l,J} \).

Each community estimation model using LDA works as follows. Location contains a distribution over communities that can be modelled using a Dirichlet distribution \( \theta^{L,J} \) with hyper-parameter \( \alpha \). Surnames in each location \( l_c \) are generated by picking community \( j \) from distribution \( \theta^{l_c} \) and picking a surname \( s_i \) from the community \( j \) according to \( P(s(i)|z(s(i)) = j) = \phi^{s(i)} \) generated from a Dirichlet distribution with hyper-parameter \( \beta \). Here \( \alpha \) and \( \beta \) specify the priori on \( \theta^{L,J} \) and \( \phi^{W_s} \) respectively. Each hyper parameter has single value which is assumed to be symmetric Dirichlet prior.
The complete LDA model for community estimation over surnames data set is given by

\[
s(i) | z_{s(i)}, \phi^{(z_{s(i)})} \sim \text{Discrete}(\phi_{w_s}) \\
\phi^{(W_s)} \sim \text{Dirichlet}(\beta) \\
z_{s(i)} | \theta^{(i)} \sim \text{Discrete}(\theta^{(i)}) \\
\theta^{(L)} \sim \text{Dirichlet}(\alpha)
\]

Now, estimating \( \theta^{(L)} \) and \( \phi^{(W_s)} \) establishes distributions of communities over a set of location L and distributions of communities over surnames \( W_s \). The goal is to estimate \( \theta^{(L)} \) and \( \phi^{(W_s)} \) by maximizing posterior distribution over community assignments to surnames using Bayesian statistics as given by (2)

\[
P(z_{s(i)} | s(i)) = \frac{P(s(i) | z_{s(i)}).P(z_{s(i)})}{\sum_{z_{s(i)}} P(s(i) | z_{s(i)}).P(z_{s(i)})}
\]

where \( \phi^{(W_s)} \) and \( \theta^{(L)} \) are multinomial Dirichlet distributions with priori \( \alpha \) and \( \beta \) are given by (3) and (4) respectively

\[
P(s(i) | z_{s(i)}) = \left( \frac{\Gamma([W_s | \beta])}{\Gamma(\beta)^{|W_s|}} \right) K \prod_{j=1}^{K} \frac{\Pi_{s(i)} \Gamma(n_{s(i)} + \beta)}{\Gamma(n_{s(i)} + |W_s| \beta)}
\]

\[
P(z_{s(i)}) = \left( \frac{\Gamma(K \alpha)}{\Gamma(\alpha)^{|K|}} \right) L \prod_{i=1}^{L} \frac{\Pi_{f} \Gamma(n_{f(i)} + \alpha)}{\Gamma(n_{f(i)} + K \alpha)}
\]

Here \( n_{s(i)}^{(j)} \) is number of times surname \( s(i) \) belongs to community \( j \), \( n_{s(i)}^{(j)} \) is number of times all surnames belong to community \( j \), \( n_{s(i)}^{(l)} \) is number of times any surname from location \( l \) belongs to community \( j \), and \( n_{s(i)}^{(l)} \) is number of times all surnames present in location \( l \). Also, \( \Gamma(\_\_) \) is the gamma function and \(||\_\|\) is the size of the set.

\[
\begin{array}{c|c|c|c}
\alpha & z_{s(i)} \text{ or } z_{f(i)} & L & K \\
\hline
\theta^{(i)} & s(i) \text{ or } f(i) & N & \phi^{(W_s)} & \beta
\end{array}
\]

Fig. 1. The Graphical representation of community estimation model.

The graphical representation of community estimation model using LDA is given in Figure 1. Each node is a random variable which is labelled according to its role in the generative process. Slashed nodes are observed variables. The rectangular "plate" denotes replication.

Unfortunately, the distribution given in (2) cannot be computed directly since the sum in the denominator does not factorize. In this paper, we follow [27] and apply Gibbs sampling to estimate the distribution in (2).

Gibbs sampling applies Markov chain Monte Carlo (MCMC) in which the next state is reached by sequentially sampling all variables from their distribution when conditioned on the current values of all other variables and data. Hence, it converges to the posterior distribution on \( z_{s(i)} \) or \( z_{f(i)} \)

summing out to \( \theta^{(L)} \) and \( \phi^{(W_s)} \) using standard Dirichlet Integrals as given in (5).

\[
P(z_{s(i)} = j | z_{s(-i)}, s(-i)) \propto \frac{n_{s(i)}^{(j)} + \beta}{|W_s| \beta} \frac{n_{s(i)}^{(l)} + \alpha}{n_{s(i)}^{(l)} + K \alpha} 
\]

Note that \( n_{s(i)}^{(l)} \) indicates the count that does not include the current assignment of \( z_{s(i)} \). That is \( n_{s(i)}^{(l)} = n_{s(i)}^{(j)} - 1 \).

It can be observed from the posterior probability in (5) is proportionate to multiplication of the probability of surname \( s(i) \) which belongs to community \( j \) and the probability of community \( j \) in location \( l \). Hence, the distributions \( \theta^{(L)} \) and \( \phi^{(W_s)} \) can be estimated as given in equations (6) and (7) respectively.

\[
\hat{\theta}_{j}^{(s(i))} = \frac{n_{s(i)}^{(j)} + \alpha}{n_{s(i)}^{(j)} + K \alpha}
\]

\[
\hat{\phi}_{j}^{(s(i))} = \frac{n_{s(i)}^{(j)} + \beta}{n_{s(i)}^{(j)} + |W_s| \beta}
\]

Similarly, community estimation model using LDA can be performed over a set of forenames. However, these models do not estimate correlation between forenames or surnames within communities or across communities.

**B. Community Correlated Estimation Models**

This subsection proposes community correlated LDA models that jointly identify correlated surnames and forenames within each community. For example, *surname correlated LDA model* proposes to find communities in surnames and extracts corresponding correlated forenames.

If a location contains a random mixtures of \( K \) communities then the probability of \( i^{th} \) correlated forename \( f^{*}(i) \) corresponding to the surname \( s(i) \) in a given location as

\[
P(f^{*}(i)) = \sum_{j=1}^{K} P(f(i) | z_{s(i)} = j)P(z_{s(i)} = j)
\]

where \( P(f^{*}(i) | z_{s(i)} = j) \) is the probability of correlated forename \( f^{*}(i) \) corresponding to community assignment of surname \( z_{s(i)} \) from which surname \( s(i) \) was drawn. Hence, a new distribution \( \phi_{f^{*}(i)}^{(W_r)} \) can be obtained which represents communities in forenames that correlate with surnames under the community \( j \). Hence \( P(f^{*}(i) | z_{s(i)} = j) \) is \( \phi_{f^{*}(i)}^{(W_r)}(j) \).

The complete *surname correlated LDA model* is given by

\[
s(i) | z_{s(i)}, \phi^{(z_{s(i)})} \sim \text{Discrete}(\phi^{(z_{s(i)})}) \\
\phi^{(W_s)} \sim \text{Dirichlet}(\beta) \\
z_{s(i)} | \theta^{(i)} \sim \text{Discrete}(\theta^{(i)}) \\
\theta^{(L)} \sim \text{Dirichlet}(\alpha) \\
f^{*}(i) | z_{s(i)}, \phi^{(z_{f(i)})} \sim \text{Discrete}(\phi^{(z_{f(i)})}) \\
\phi^{(W_f)} \sim \text{Dirichlet}(\beta)
\]
Surname communities in the United Kingdom

Sarah

For example, forenames can form hidden communities in the correlated forenames. Correlated forenames across several surname communities and surnames can be seen that there might be many common sections.

\[ P(f^*(i) | z_{s(i)}) = \left( \frac{\Gamma(|W_f|, \beta_1)}{\Gamma(\beta_1)^{|W_f|}} \right)^K \prod_{j=1}^{K} \frac{\Pi_{f(i)} \Gamma(n_{f(i)}^{(j)} + \beta_1)}{\Gamma(n_{f(i)}^{(j)} + |W_f|, \beta_1)} \]  

(9)

Here \( n_{f(i)}^{(j)} \) is number of times forename \( f(i) \) belongs to community \( j \). However, the estimation of \( \phi(W_s) \) corresponds to probability given in (9) can be obtained by

\[ \hat{\phi}_j^{f(i)} = \frac{n_{f(i)} + \beta_1}{n_{f(i)} + |W_f|, \beta_1} \]  

(10)

Similarly, the surname correlated LDA model can be estimated which gives correlated surnames for each community of forenames. However, these models do not infer correlation between communities of forenames and communities of surnames. The following subsection proposes the community-community correlated estimation models.

C. Community-Community Correlated Estimation models

This subsection proposes community-community correlated estimation models. We propose two models which are surname community correlated LDA model and forename community correlated LDA model.

The surname community correlated LDA model initially estimates communities over surnames as explained in subsection III-A. Let \( K \) be number of communities obtained in surnames. It can be seen that there might be many common correlated forenames across several surname communities and thus can form hidden communities in the correlated forenames. For example, forenames Sarah and Paul shared across many surname communities in United Kingdom.

The surname community correlated LDA model chooses proportions of several surname communities that correlate with each surname community whereas earlier models choose proportions of communities in a location. Hence the surname community correlated LDA model can find forename communities such that the distribution of forenames in each forename community is based on correlation of forename community to several surname communities.

The surname community correlated estimation model using LDA works as follows. Surname communities correlate with several forename communities. If a surname community correlates with a random mixture of \( K \) forename communities then the probability of \( i^{th} \) forename \( f(i) \) that correlates with a surname community as

\[ P(f(i) | z_{s(i)}) = \sum_{j=1}^{K} P(f(i) | z_{f(i)} = j) P(z_{f(i)} = j | z_{s(i)}) \]  

(11)

where \( z_{f(i)} \) denotes latent forename-community assignment \( j \) from which \( i^{th} \) forename \( f(i) \) was drawn. \( P(f(i) | z_{f(i)} = j) \) is the probability the correlated forename \( f(i) \) under the forename-community \( j \), and \( P(z_{f(i)} = j | z_{s(i)}) \) is the probability of choosing forename-community \( j \) that correlates to a surname-community \( z_{s(i)} \). The idea behind this model is that the forenames that correlated with each surname-community are generated by picking the forename-community \( j \) from distribution \( \Lambda(K) \) and picking a forename from the forename-community \( j \) according to \( \hat{P}(f(i) | z_{f(i)} = j) \). Hence a new multinomial distribution \( \Lambda(K) \) with a Dirichlet prior \( \gamma \) represents proportions of several forename-communities shared over surname-communities and \( \phi(W_s) \) denotes a multinomial distribution of communities over a set of forenames with a Dirichlet prior \( \beta_1 \). Note that \( \gamma \) and \( \beta_1 \) are symmetric Dirichlet priori can take scalar values.

The complete surname community correlated LDA model is given by

\[ s(i) | z_{s(i)}, \phi^{(z_{s(i)})}, \phi(W_s) \sim \text{Discrete}(\phi^{(z_{s(i)})}) \]

\[ \phi^{(W_s)} \sim \text{Dirichlet}(\beta) \]
Fig. 4. Surname Community Correlated LDA model

\[
z_{s(i)} | \theta^{(l)} \sim \text{Discrete}(\theta^{(l)})
\]
\[
\theta^{(L)} \sim \text{Dirichlet}(\alpha)
\]
\[
f(i) | z_{f(i)}, \phi(z_{f(i)}) \sim \text{Discrete}(\phi(z_{f(i)}))
\]
\[
\phi(W_i) \sim \text{Dirichlet}(\beta_1)
\]
\[
z_{f(i)} | \Lambda^{(k1)} \sim \text{Discrete}(\Lambda^{(k1)})
\]
\[
\Lambda^{(K)} | z_{s(i)} \sim \text{Dirichlet}(\gamma)
\]

The posterior distribution on \(z_{s(i)}\), the distributions of \(\theta^{(L)}\), and the distributions of \(\phi(W_i)\) are given by equations (2), (4), and (3) respectively. However, the community-community correlated estimation model can be performed to estimate \(\Lambda^{(K)}\) and \(\phi(W_i)\) by maximizing community assignments to forenames using (12)

\[
P(z_{f(i)} | f(i)) = \frac{P(f(i)|z_{f(i)}).P(z_{f(i)}|z_{s(i)})}{\sum_{z_{f(i)}} P(f(i)|z_{f(i)}).P(z_{f(i)}|z_{s(i)})}
\]

where \(\phi(W_i)\) and \(\Lambda^{(K)}\) are multinomial Dirichlet distributions with priori \(\beta_1\) and \(\gamma\) are given by (13) and (14) respectively

\[
P(f(i)|z_{f(i)}) = \frac{\Gamma(|W_i|\beta_1)}{\Gamma(\beta_1)^{|W_i|}} \prod_{j=1}^{K_i} \frac{\Pi_{f(i)} \Gamma(n_{(j)}^{f(i)} + \beta_1)}{\Gamma(n_{(j)}^{f(i)} + |W_i|\beta_1)}
\]

\[
P(z_{f(i)}|z_{s(i)}) = \frac{\Gamma(K_1\gamma)}{\Gamma(\gamma)^{K_1}} \prod_{k=1}^{K} \Gamma(n_{(k)}^{f(k_i)} + \gamma) / \Gamma(n_{(k)}^{f(k_i)} + K_1\gamma)
\]

Here \(n_{(j)}^{f(i)}\) is number of times forename \(f(i)\) belongs to community \(j\), \(n_{(j)}^{f(k)}\) is number of times all forenames belong to community \(j\), \(n_{(j)}^{f(k_i)}\) is number of times forename \(i\) that correlates with surname-community \(k_i\) belongs to community \(j\), and \(n_{(k)}^{f(k_i)}\) is number of times all forenames that correlate with surname-community \(k_i\). Also, \(\Gamma(.)\) is the standard gamma function and \(|.|\) is the size of the set.

We will use Gibbs sampling to find the posterior distribution on \(z_{f(i)}\), which integrating out to \(\Lambda^{(K)}\) and \(\phi(W_i)\) using standard Dirichlet Integrals as given in equation (15).

\[
P(z_{f(i)} = j | z_{f(-i)}, f(-i)) \propto \frac{n_{(j)}^{f(i)} + \beta_1}{n_{(j)}^{f(-i)} + |W_i|\beta_1} \cdot \frac{n_{(i)}^{f(k)} + \gamma}{n_{(i)}^{f(k_i)} + K_1\gamma}
\]

Note that \(n_{(i)}^{f(k_i)}\) indicates the count that does not include the current assignment of \(z_{f(i)}\). That is \(n_{(i)}^{f(k_i)} = n_{(j)}^{f(k)} - 1\).

It can be observed from the posterior probability in (15) is proportionate to multiplication of the probability of forename \(f(i)\) which belongs to community \(j\) and the probability of forename-community \(j\) correlates with surname-community \(i\).

Hence, the distributions \(\Lambda^{(K)}\) and \(\phi(W_i)\) can be estimated as given in equations (16) and (17) respectively.

\[
\hat{\Lambda}_j^{(k)} = \frac{n_{(j)}^{f(k_i)} + \gamma}{n_{(j)}^{f(k_i)} + K_1\gamma}
\]

\[
\hat{\phi}_j^{f(i)} = \frac{n_{(j)}^{f(i)} + \beta_1}{n_{(j)}^{f(i)} + |W_i|\beta_1}
\]

Similarly forename community correlated LDA model estimates communities in forenames and introduces an additional multinomial distribution that captures correlation among communities in surnames over communities of forenames. The graphical

IV. EXPERIMENTAL RESULTS

This section describes experimental results. We have two countries names data set, viz., United Kingdom (UK) and India. United Kingdom corpus has 0.924 million names collected over 115 locations in United Kingdom. India corpus has 17.4 million names collected over 277 locations which covered 28 provinces and 6 union territories. Names in 100 random
locations chosen as train data set and names in 15 remaining locations chosen as test data set for United Kingdom. Similarly, names in 250 random locations chosen as train data set and names in 27 random locations chosen as test data set for India. Test data set consists of held-out names from several locations that evaluates the estimated model from training set.

Experiments are carried out using Gibbs sampler to estimate communities and their correlations in UK and India names data set. The number of communities are chosen from 15,20,25,30. The hyper-parameters such as \( \alpha, \beta, \beta_1, \) and \( \gamma \) are symmetric Dirichlet prior and each hyper parameter is chosen single value which is 0.1. Gibbs sampling runs up to 1000 iterations.

### A. Community Correlated Estimation Models

This subsection presents the result of surname correlated LDA model and forename correlated LDA model. The results of surname correlated LDA model present estimated communities in surnames and correlated forenames in each surname communities. The results of forename correlated LDA model present estimated communities in forenames and correlated surnames in each forename communities.

1) UK names data set: Table I shows the results of surname correlated LDA model. Communities 5, 10, 24, and 25 have top 10 most likely surnames and their correlated top 10 most likely forenames for UK. Surnames belong to community 5 and 25 are British or European and the correlated forenames are also British or European. Surnames in community 10 seem to be Indian or Pakistani surnames and surnames mo-hamad, muhammad, and ali are seem to be correlated Indian or Pakistani forenames and also with some other correlated British forenames. Similarly, surnames in community 24 seem to be Chines along with some correlated Chines and British forenames. Some British forenames appear across many surname communities.

<table>
<thead>
<tr>
<th>Surname</th>
<th>PROB.</th>
</tr>
</thead>
<tbody>
<tr>
<td>matos</td>
<td>0.000436</td>
</tr>
<tr>
<td>duff</td>
<td>0.000374</td>
</tr>
<tr>
<td>neves</td>
<td>0.000374</td>
</tr>
<tr>
<td>fevrier</td>
<td>0.000374</td>
</tr>
<tr>
<td>molina</td>
<td>0.000374</td>
</tr>
<tr>
<td>wallace</td>
<td>0.000313</td>
</tr>
<tr>
<td>roos</td>
<td>0.000313</td>
</tr>
<tr>
<td>springham</td>
<td>0.000313</td>
</tr>
<tr>
<td>asare</td>
<td>0.000313</td>
</tr>
<tr>
<td>decarvalho</td>
<td>0.000313</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Forename</th>
<th>PROB.</th>
</tr>
</thead>
<tbody>
<tr>
<td>john</td>
<td>0.005376</td>
</tr>
<tr>
<td>michael</td>
<td>0.005353</td>
</tr>
<tr>
<td>maria</td>
<td>0.003351</td>
</tr>
<tr>
<td>paul</td>
<td>0.001348</td>
</tr>
<tr>
<td>david</td>
<td>0.003047</td>
</tr>
<tr>
<td>peter</td>
<td>0.002946</td>
</tr>
<tr>
<td>andrew</td>
<td>0.002238</td>
</tr>
<tr>
<td>sarah</td>
<td>0.002136</td>
</tr>
</tbody>
</table>

Table II shows the results of forename correlated LDA model. Communities 4, 7, 20, and 25 have top 10 most likely forenames and their correlated top 10 most likely surnames for UK. Forenames in community 4, 7, 20, and 25 are British, Pakistani, Indian, and Chines and correlated surnames seem to be from same communities and however, there are some

<table>
<thead>
<tr>
<th>Forename</th>
<th>PROB.</th>
</tr>
</thead>
<tbody>
<tr>
<td>smith</td>
<td>0.009057</td>
</tr>
<tr>
<td>brown</td>
<td>0.006248</td>
</tr>
<tr>
<td>wilson</td>
<td>0.004873</td>
</tr>
<tr>
<td>stewart</td>
<td>0.004834</td>
</tr>
<tr>
<td>thomson</td>
<td>0.003626</td>
</tr>
<tr>
<td>murray</td>
<td>0.004166</td>
</tr>
<tr>
<td>scott</td>
<td>0.003832</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Surname</th>
<th>PROB.</th>
</tr>
</thead>
<tbody>
<tr>
<td>khan</td>
<td>0.017986</td>
</tr>
<tr>
<td>hassan</td>
<td>0.015627</td>
</tr>
<tr>
<td>ali</td>
<td>0.012530</td>
</tr>
<tr>
<td>ahmed</td>
<td>0.013584</td>
</tr>
<tr>
<td>patel</td>
<td>0.009099</td>
</tr>
<tr>
<td>akhtar</td>
<td>0.009047</td>
</tr>
<tr>
<td>mahmood</td>
<td>0.004551</td>
</tr>
<tr>
<td>begum</td>
<td>0.004461</td>
</tr>
<tr>
<td>iqbal</td>
<td>0.003831</td>
</tr>
<tr>
<td>singh</td>
<td>0.003966</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Forename</th>
<th>PROB.</th>
</tr>
</thead>
<tbody>
<tr>
<td>scott</td>
<td>0.004998</td>
</tr>
<tr>
<td>murray</td>
<td>0.004998</td>
</tr>
<tr>
<td>asaf</td>
<td>0.004998</td>
</tr>
<tr>
<td>rahman</td>
<td>0.004998</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Forename</th>
<th>PROB.</th>
</tr>
</thead>
<tbody>
<tr>
<td>john</td>
<td>0.005376</td>
</tr>
<tr>
<td>michael</td>
<td>0.005353</td>
</tr>
<tr>
<td>maria</td>
<td>0.003351</td>
</tr>
<tr>
<td>paul</td>
<td>0.001348</td>
</tr>
<tr>
<td>david</td>
<td>0.003047</td>
</tr>
<tr>
<td>peter</td>
<td>0.002946</td>
</tr>
<tr>
<td>andrew</td>
<td>0.002238</td>
</tr>
<tr>
<td>sarah</td>
<td>0.002136</td>
</tr>
</tbody>
</table>
common surnames between *Pakistani* and *Indian*. Also, there are some common surnames in *British* and *Chines*.

2) *India names data set*: Table III shows the results of *surname correlated LDA model*. Communities 8, 10, 12, and 25 have top 10 most likely surnames and their correlated top 10 most likely forenames for India names data set. Surnames in community 8, community 10, community 12, and community 25 are belong to *Bengali*, *Orrisa*, *Marathi*, and *Assami* surnames. The correlated forenames correspond to each surname community are presented which share some forenames across two or more community groups. For example, forenames *abhijit*, *amit*, and *sanjay* share across *Bangali* and *Assami* surname communities. Forenames *sanjay* and *manoj* share across *Orrisa* and *Assami* communities.

Table IV shows the results of *forename correlated LDA model*. Communities 9, 16, 28, and 25 have top 10 most likely forenames and their correlated top 10 most likely surnames for India names data set. The distributions of surname communities can be interpreted easily in Indian names data set whereas the distributions of forename communities are hard to interpret since forenames can share across many surname communities in Indian names data set. However, it is clear from the Table IV that the correlated surnames of each forename community appear in same surname community in Table III. Hence, the *forename correlated LDA model* clearly finding communities in forenames and correlated surnames.

### Table III. Surname Correlated Estimation Model for India Data Set

<table>
<thead>
<tr>
<th>Community 8</th>
<th>Community 10</th>
<th>Community 12</th>
<th>Community 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surname</td>
<td>PROB.</td>
<td>Surname</td>
<td>PROB.</td>
</tr>
<tr>
<td>das</td>
<td>0.047381</td>
<td>sahoo</td>
<td>0.030447</td>
</tr>
<tr>
<td>gosht</td>
<td>0.036121</td>
<td>mohanty</td>
<td>0.026827</td>
</tr>
<tr>
<td>roy</td>
<td>0.030650</td>
<td>mishra</td>
<td>0.025976</td>
</tr>
<tr>
<td>banerjee</td>
<td>0.024349</td>
<td>das</td>
<td>0.021119</td>
</tr>
<tr>
<td>chakraborty</td>
<td>0.024280</td>
<td>nayak</td>
<td>0.020382</td>
</tr>
<tr>
<td>mukherjee</td>
<td>0.024237</td>
<td>behera</td>
<td>0.019462</td>
</tr>
<tr>
<td>saha</td>
<td>0.020182</td>
<td>panda</td>
<td>0.019417</td>
</tr>
<tr>
<td>sarkar</td>
<td>0.018965</td>
<td>dash</td>
<td>0.017908</td>
</tr>
<tr>
<td>dutta</td>
<td>0.018494</td>
<td>sahu</td>
<td>0.016069</td>
</tr>
<tr>
<td>chatterjee</td>
<td>0.018213</td>
<td>mahapatra</td>
<td>0.014038</td>
</tr>
</tbody>
</table>

### Table IV. Forename Correlated Estimation Model for India Data Set

<table>
<thead>
<tr>
<th>Community 9</th>
<th>Community 16</th>
<th>Community 18</th>
<th>Community 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forename</td>
<td>PROB.</td>
<td>Forename</td>
<td>PROB.</td>
</tr>
<tr>
<td>amit</td>
<td>0.006203</td>
<td>amit</td>
<td>0.017785</td>
</tr>
<tr>
<td>abhijit</td>
<td>0.004291</td>
<td>rajul</td>
<td>0.000500</td>
</tr>
<tr>
<td>sandeep</td>
<td>0.003975</td>
<td>deepak</td>
<td>0.001347</td>
</tr>
<tr>
<td>ashish</td>
<td>0.003703</td>
<td>sanyan</td>
<td>0.000987</td>
</tr>
<tr>
<td>sanyan</td>
<td>0.003453</td>
<td>subrata</td>
<td>0.000972</td>
</tr>
<tr>
<td>subrata</td>
<td>0.003447</td>
<td>sanyan</td>
<td>0.000987</td>
</tr>
<tr>
<td>parthi</td>
<td>0.003381</td>
<td>santooshkumar</td>
<td>0.000972</td>
</tr>
<tr>
<td>kaushek</td>
<td>0.003284</td>
<td>ashok</td>
<td>0.000963</td>
</tr>
<tr>
<td>kaushek</td>
<td>0.003284</td>
<td>manorajan</td>
<td>0.000943</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Surname</th>
<th>PROB.</th>
<th>Surname</th>
<th>PROB.</th>
<th>Surname</th>
<th>PROB.</th>
<th>Surname</th>
<th>PROB.</th>
</tr>
</thead>
<tbody>
<tr>
<td>kumar</td>
<td>0.053129</td>
<td>patel</td>
<td>0.008968</td>
<td>abhijit</td>
<td>0.004777</td>
<td>amit</td>
<td>0.007978</td>
</tr>
<tr>
<td>singh</td>
<td>0.042147</td>
<td>jignesh</td>
<td>0.004512</td>
<td>arindram</td>
<td>0.004504</td>
<td>arindram</td>
<td>0.005050</td>
</tr>
<tr>
<td>sharma</td>
<td>0.030981</td>
<td>chigag</td>
<td>0.000630</td>
<td>hardik</td>
<td>0.001133</td>
<td>subrata</td>
<td>0.000404</td>
</tr>
<tr>
<td>gupta</td>
<td>0.020334</td>
<td>chigag</td>
<td>0.000630</td>
<td>hardik</td>
<td>0.001133</td>
<td>subrata</td>
<td>0.000404</td>
</tr>
<tr>
<td>Jain</td>
<td>0.014350</td>
<td>mehta</td>
<td>0.007534</td>
<td>bhavesh</td>
<td>0.000303</td>
<td>sanyan</td>
<td>0.000393</td>
</tr>
<tr>
<td>Shah</td>
<td>0.009734</td>
<td>bhatt</td>
<td>0.007292</td>
<td>hiren</td>
<td>0.002983</td>
<td>sanyan</td>
<td>0.000393</td>
</tr>
<tr>
<td>Mishra</td>
<td>0.008965</td>
<td>desai</td>
<td>0.006895</td>
<td>prajapati</td>
<td>0.006267</td>
<td>sanyan</td>
<td>0.000393</td>
</tr>
<tr>
<td>Yadav</td>
<td>0.007624</td>
<td>pandya</td>
<td>0.006218</td>
<td>dutta</td>
<td>0.013930</td>
<td>sanyan</td>
<td>0.000393</td>
</tr>
<tr>
<td>Verma</td>
<td>0.007109</td>
<td>panchal</td>
<td>0.006112</td>
<td>chatterjee</td>
<td>0.015577</td>
<td>sanyan</td>
<td>0.000393</td>
</tr>
<tr>
<td>Agarwal</td>
<td>0.007090</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hence the *forename correlated LDA model* clearly finding communities in forenames and correlated surnames.
Fig. 6. Forename Community Correlated LDA Model for UK Names Data Set

Fig. 7. Surname Community Correlated Estimation Model for India Names Data Set
1) UK names data set: Figure 6 shows the results of forename community correlated LDA model. Forename communities (FCommunity) 10, 16, and 17 seem to be Indian, Chines, and British surnames respectively. It is observed that surname community 1 correlates with all three communities. However, it highly correlates to British forename community (0.8071) whereas surname community 6 and 4 are only correlate to British community. Surname community 2 contains Indian or Pakistani surnames which highly correlates to Indian forenames with probability 0.7611. However, surname community 7 contains Chines surnames which are highly correlates to Chines forenames with probability 0.6044.

2) India names data set: Figure 7 shows the results of surname community correlated LDA model. Surname communities (SCommunity) 1, 11, and 13 have top 10 most likely surnames and their correlated top 3 forename communities where each forename community represents top 10 most likely surnames of that community for UK names data set. Surnames in community 1, 11, and 13 are seem to be Assam, Bengali, and Marathi communities in India. It is observed that forename community 7 shares all these three surname communities and however it highly correlates to Assam surname community with probability 0.4346. Also, forename community 2 highly correlates to Bengali community with probability 0.7864 and forename community 4 highly correlates to Marathi surname community with probability 0.8071.

C. Performance Comparison

A held out test data set is used to compare the performance of the proposed models for Indian and UK names data set. Perplexity is a standard measure to compare the performance of a probabilistic model. A lower perplexity score indicates better generalization performance. The perplexity is defined as exponential of negative normalized predictive likelihood of test data under the model. Figure 8 represents perplexity comparison for probabilistic models against number of communities for UK and India data sets. Perplexity can also be used to study the strengths of communities under different scenarios. For India names data set, surname LDA model has less perplexity than forename LDA model which means surnames capture better communities than forenames whereas UK names data set, forenames capture better communities than surnames.

The perplexity of the surname correlated LDA model has almost same performance as forename LDA since the correlated forenames probabilities are used to calculate perplexity, but it extracts additional information such as correlated forenames of each surname community for both the names data set. Similarly, the proposed forename correlated LDA model has almost same performance as surname LDA, but it extracts correlated surnames for both the names data sets. However, the proposed surname community correlated LDA and forename community correlated LDA have improved performance compared to all other methods. These models also provide interaction among communities of surnames and communities of forenames.

For Indian names data set, the forename community correlated LDA model performs better than surname community correlated LDA model whereas for UK names data set, the surname community correlated LDA model performs better than forename community correlated LDA model. It means surname capture good communities which interact across several forename communities in India whereas forename capture good communities which interact across several surname communities in UK. Intuitively, surname communities in India and forename communities in UK are well established in all the proposed methods. However, the performance of community-community correlated LDA models are better than all other models. The average perplexity of surname community correlated LDA model and forename correlated LDA model are 3.01944 and 7.6573 for UK names data set whereas these values are 9.66520 and 4.8298 for India names data set respectively.

V. CONCLUSION AND FUTURE WORK

This paper used the probabilistic generative model such as LDA to find communities over a set of names collected
at different locations. In addition, this paper proposed several variants of LDA models to capture correlation among surnames and forenames within the communities and across the communities. Initially, this paper presented surname correlated LDA model and forename correlated LDA model. These models find communities in surnames or forenames and extracts correlated forenames or surnames respectively. The performance of surname correlated LDA model or forename correlated LDA model is similar to performance of LDA model to find communities in surnames or forenames independently. However, these proposed models extract correlated forenames or surnames. Later, this paper proposed surname community correlated LDA model and forename community correlated LDA model. These models establish interaction among communities of surnames and communities of forenames. These two models have lower perplexity compared to all other related models which means the performance of these two models are better than all other related models in this paper. The experiments for proposed models are conducted against number of communities for India and UK names data set. It has been observed that surnames form good communities in India names data set and forenames form good communities in UK names data set. This paper assumes the number of communities are known in advance. In future work, we will propose to derive optimal number of clusters from the given names data set.

**ACKNOWLEDGEMENTS**

This work was supported through the EPSRC Grant: “The Uncertainty of Identity: Linking Spatiotemporal Information Between Virtual and Real Worlds” (EP/J005266/1).

**REFERENCES**


