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.
E-mail:{Suresh.Veluru.1, Yogachandran.Rahulamathavan.1, R.Muttukrishnan}@city.ac.uk
†Department of Computer Science, University of York, Heslington, York, YO10 5GH, United Kingdom
E-mail: suresh.manandhar@york.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 correlated 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 each location using a vector space model followed by latent semantic analysis. 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 $\theta^{(L)}$ 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

$$P(s(i)) = \sum_{j=1}^{K} P(s(i)|z_{s(i)} = j)P(z_{s(i)} = j) \tag{1}$$

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 $\phi_j^{s(i)}$ and $P(z_{s(i)} = j)$ is θ_j^l.

Each community estimation model using LDA works as follows. Location contains a distribution over communities that can be modelled using a Dirichlet distribution $\theta^{(L)}$ with hyper-parameter α. Surnames in each location l_i are generated by picking community j from distribution $\theta^{(l_i)}$ and picking a surname s_i from the community j according to $P(s(i)|z_{s(i)} = j) = \phi_j^{(s(i))}$ generated from a Dirichlet distribution with hyper-parameter β. Here α and β specify the priori on $\theta^{(L)}$ 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

$$\begin{aligned} s(i)|z_{s(i)}, \phi^{(z_{s(i)})} &\sim \text{Discrete}(\phi^{(z_{s(i)})}) \\ \phi^{(W_s)} &\sim \text{Dirichlet}(\beta) \\ z_{s(i)}|\theta^{(l_i)} &\sim \text{Discrete}(\theta^{(l_i)}) \\ \theta^{(L)} &\sim \text{Dirichlet}(\alpha) \end{aligned}$$

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)

$$P(z_{s(i)}|s(i)) = \frac{P(s(i)|z_{s(i)}).P(z_{s(i)})}{\sum_{z_{s(i)}} P(s(i)|z_{s(i)}).P(z_{s(i)})} \tag{2}$$

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

$$P(s(i)|z_{s(i)}) = \left(\frac{\Gamma(|W_s|\beta)}{\Gamma(\beta)^{|W_s|}}\right)^K \prod_{j=1}^{K} \frac{\Pi_{s(i)}\Gamma(n_{(j)}^{s(i)} + \beta)}{\Gamma(n_{(j)}^{s(.)} + |W_s|\beta)} \tag{3}$$

$$P(z_{s_i}) = \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^{|L|} \prod_{l_i=1}^{|L|} \frac{\Pi_j\Gamma(n_{(j)}^{s(l_i)} + \alpha)}{\Gamma(n_{(.)}^{s(l_i)} + K\alpha)} \tag{4}$$

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

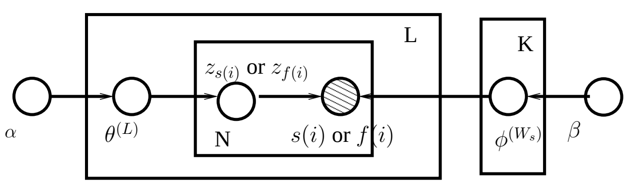

Fig. 1. The Graphical representation of community estimation model.

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).

$$P(z_{s(i)} = j|z_{s(-i)}, s(-i)) \propto \frac{n_{(-i,j)}^{s(i)} + \beta}{n_{(-i,j)}^{s(.)} + |W_s|\beta} . \frac{n_{(-i,j)}^{s(l_i)} + \alpha}{n_{(-i,.)}^{s(l_i)} + K.\alpha} \tag{5}$$

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

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_i. Hence, the distributions $\theta^{(L)}$ and $\phi^{(W_s)}$ can be estimated as given in equations (6) and (7) respectively.

$$\hat{\theta}_j^{s(l_i)} = \frac{n_{(j)}^{s(l_i)} + \alpha}{n_{(.)}^{s(l_i)} + K.\alpha} \tag{6}$$

$$\hat{\phi}_j^{s(i)} = \frac{n_{(j)}^{s(i)} + \beta}{n_{(j)}^{s(.)} + |W_s|\beta} \tag{7}$$

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

$$P(f^*(i)) = \sum_{j=1}^{K} P(f(i)|z_{s(i)} = j)P(z_{s(i)} = j) \tag{8}$$

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 $\phi_j^{(W_f)}$ 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 $\phi_j^{f(i)}$.

The complete *surname correlated LDA model* is given by

$$\begin{aligned} s(i)|z_{s(i)}, \phi^{(z_{s(i)})} &\sim \text{Discrete}(\phi^{(z_{s(i)})}) \\ \phi^{(W_s)} &\sim \text{Dirichlet}(\beta) \\ z_{s(i)}|\theta^{(l_i)} &\sim \text{Discrete}(\theta^{(l_i)}) \\ \theta^{(L)} &\sim \text{Dirichlet}(\alpha) \\ f^*(i)|z_{s(i)}, \phi^{(z_{f(i)})} &\sim \text{Discrete}(\phi^{(z_{f(i)})}) \\ \phi^{(W_f)} &\sim \text{Dirichlet}(\beta_1) \end{aligned}$$

The posterior distribution on $z_{s(i)}$, the distributions of $\theta^{(L)}$, and the distributions of $\phi^{(W_s)}$ are given by equations (2), (4),

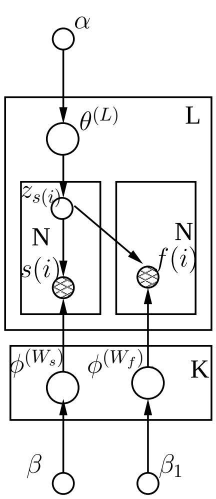

Fig. 2. The graphical representation of surname Correlated LDA model

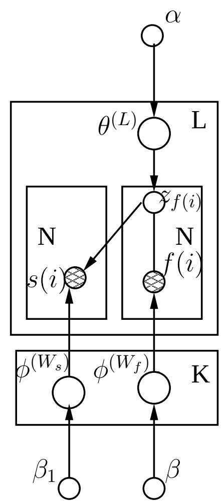

Fig. 3. The graphical representation of forename Correlated LDA Model

and (3) respectively. The new distribution $\phi^{(W_s)}$ can be obtained as given in (9) which is a multinomial Dirichlet distribution with a symmetric prior β_1 that corresponds to distribution of communities over correlated forenames.

$$P(f^*(i)|z_{s(i)}) = \left(\frac{\Gamma(|W_f|.\beta_1)}{\Gamma(\beta_1)^{|W_f|}}\right)^K \prod_{j=1}^{K} \frac{\Pi_{f(i)}\Gamma(n_{(j)}^{f(i)} + \beta_1)}{\Gamma(n_{(j)}^{f(.)} + |W_f|.\beta_1)} \tag{9}$$

Here $n_{(j)}^{f(i)}$ is number of times forename $f(i)$ belongs to community j, $n_j^{f(.)}$ is number of times all forenames belong to community j. However, the estimation of $\phi^{(W_s)}$ corresponds to probability given in (9) can be obtained by (10)

$$\hat{\phi}_j^{f(i)} = \frac{n_{(j)}^{f(i)} + \beta_1}{n_{(j)}^{f(.)} + |W_f|.\beta_1} \tag{10}$$

Similarly *forename 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 forename 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 correlate with a random mixture of K_1 forename communities then the probability of i^{th} forename $f(i)$ that correlates with a surname community as

$$P(f(i)|z_{s(i)}) = \sum_{j=1}^{K_1} P(f(i)|z_{f(i)} = j)P(z_{f(i)} = j|z_{s(i)}) \tag{11}$$

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 $P(f(i)|z_{f(i)} = j)$. Hence a new multinomial distribution $\Lambda^{(K)}$ with a Dirichlet prior γ represents proportions of several forename-communities shared over surname-communities and $\phi^{(W_f)}$ denotes a multinomial distribution of communities over a set of forenames with a Dirichlet prior β_1. Note that γ and β_1 are symmetric Dirichlet priori can take scalar values.

The complete *surname community correlated LDA model* is given by

$$\begin{aligned} s(i)|z_{s(i)}, \phi^{(z_{s(i)})} &\sim \text{Discrete}(\phi^{(z_{s(i)})}) \\ \phi^{(W_s)} &\sim \text{Dirichlet}(\beta) \end{aligned}$$

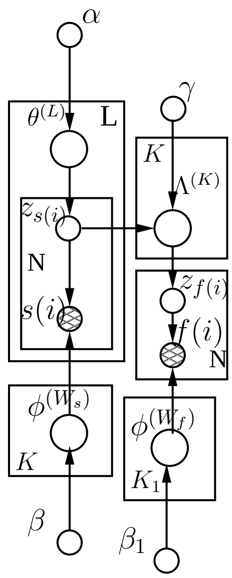

Fig. 4. Surname Community Correlated LDA model

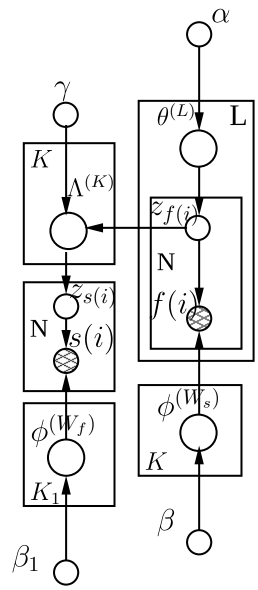

Fig. 5. Forename Community Correlated LDA Model

$$
\begin{aligned}
z_{s(i)}|\theta^{(l_i)} &\sim \text{Discrete}(\theta^{(l_i)}) \\
\theta^{(L)} &\sim \text{Dirichlet}(\alpha) \\
f(i)|z_{f(i)}, \phi^{(z_{f(i)})} &\sim \text{Discrete}(\phi^{(z_{f(i)})}) \\
\phi^{(W_f)} &\sim \text{Dirichlet}(\beta_1) \\
z_{f(i)}|\Lambda^{(k_i)} &\sim \text{Discrete}(\Lambda^{(k_i)}) \\
\Lambda^{(K)}|z_{s(i)} &\sim \text{Dirichlet}(\gamma)
\end{aligned}
$$

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

$$P(z_{f(i)}|f(i)) = \frac{P(f(i)|z_{f(i)}).P(z_{f(i)}|z_{s(i)})}{\sum_{z_{f(i)}} P(f(i)|z_{f(i)}).P(z_{f(i)}|z_{s(i)})} \quad (12)$$

where $\phi^{(W_f)}$ and $\Lambda^{(K)}$ are multinomial Dirichlet distributions with priori β_1 and γ are given by (13) and (14) respectively

$$P(f(i)|z_{f(i)}) = \left(\frac{\Gamma(|W_f|.\beta_1)}{\Gamma(\beta_1)^{|W_f|}}\right)^{K_1} \prod_{j=1}^{K_1} \frac{\Pi_{f(i)}\Gamma(n_{(j)}^{f(i)} + \beta_1)}{\Gamma(n_{(j)}^{f(.)} + |W_f|.\beta_1)} \quad (13)$$

$$P(z_{f(i)}|z_{s(i)}) = \left(\frac{\Gamma(K_1\gamma)}{\Gamma(\gamma)^{K_1}}\right)^{K} \prod_{k_i=1}^{K} \frac{\Pi_j\Gamma(n_{(j)}^{f(k_i)} + \gamma)}{\Gamma(n_{(.)}^{f(k_i)} + K_1\gamma)} \quad (14)$$

Here $n_{(j)}^{f(i)}$ is number of times forename $f(i)$ belongs to community j, $n_j^{f(.)}$ 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 $n_{(.)}^{f(k_i)}$ 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 $\phi^{(W_f)}$ using standard Dirichlet Integrals as given in equation (15).

$$P(z_{f(i)} = j|z_{f(-i)}, f(-i)) \propto \frac{n_{(-i,j)}^{f(i)} + \beta_1}{n_{(-i,j)}^{f(.)} + |W_f|\beta_1} . \frac{n_{(-i,j)}^{f(k_i)} + \gamma}{n_{(-i,.)}^{f(k_i)} + K_1.\gamma} \quad (15)$$

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

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 $\phi^{(W_f)}$ can be estimated as given in equations (16) and (17) respectively.

$$\hat{\Lambda}_j^{f(k_i)} = \frac{n_{(-i,j)}^{f(k_i)} + \gamma}{n_{(-i,.)}^{f(k_i)} + K_1.\gamma} \quad (16)$$

$$\hat{\phi}_j^{f(i)} = \frac{n_{(-i,j)}^{f(i)} + \beta_1}{n_{(-i,j)}^{f(.)} + |W_f|\beta_1} \quad (17)$$

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

TABLE I. SURNAME CORRELATED ESTIMATION MODEL FOR UNITED KINGDOM DATA SET

Community 5		Community 10		Community 24		Community 25	
Surname	**PROB.**	**Surname**	**PROB.**	**Surname**	**PROB.**	**Surname**	**PROB.**
matos	0.000436	patel	0.02416	derrick	0.000951	smith	0.008004
duff	0.000374	khan	0.01912	luo	0.000840	wilson	0.007720
neves	0.000374	ali	0.01254	zhao	0.000784	brown	0.007701
fevrier	0.000374	ahmed	0.01200	trott	0.000562	stewart	0.006923
molina	0.000374	hussain	0.00935	jhon	0.000562	campbell	0.006895
wallace	0.000313	singh	0.00909	billy	0.000562	anderson	0.006639
roos	0.000313	shah	0.00860	rosa	0.000562	robertson	0.006354
springham	0.000313	begum	0.00738	venn	0.000562	thomson	0.005842
asare	0.000313	kaur	0.00415	skuse	0.000562	murray	0.005691
decarvalho	0.000313	rahman	0.00321	whyet	0.000562	scott	0.004998
Forename	**PROB.**	**Forename**	**PROB.**	**Forename**	**PROB.**	**Forename**	**PROB.**
john	0.005376	john	0.00591	david	0.006408	john	0.021779
michael	0.003553	mohammed	0.00589	paul	0.005975	david	0.018649
anna	0.003553	david	0.00528	susan	0.005802	james	0.016307
maria	0.003351	michael	0.00396	sarah	0.005629	robert	0.010693
paul	0.003148	richard	0.00375	helen	0.005543	margaret	0.010663
david	0.003047	paul	0.00345	john	0.005111	paul	0.009936
peter	0.002946	sarah	0.00340	emma	0.004332	william	0.009744
james	0.002541	muhammad	0.00314	karen	0.004332	andrew	0.009088
andrew	0.002338	ali	0.00308	xheng	0.003554	michael	0.008694
sarah	0.002136	susan	0.00294	cheng	0.003467	elizabeth	0.008310

TABLE II. FORNAME CORRELATED ESTIMATION MODEL FOR UNITED KINGDOM DATA SET

Community 4		Community 7		Community 20		Community 25	
Forename	**PROB.**	**Forename**	**PROB.**	**Forename**	**PROB.**	**Forename**	**PROB.**
james	0.027032	mohammed	0.021833	mandeep	0.004284	eugen	0.002144
john	0.026830	imran	0.006522	gurpreet	0.002254	rice	0.001576
william	0.021617	abdul	0.004561	jas	0.002040	dren	0.001434
margaret	0.020741	muhammad	0.003928	hardeep	0.002040	yu	0.001292
david	0.015551	mohammad	0.003739	harjinder	0.002040	hil	0.001292
elizabeth	0.014764	salma	0.003485	kamaljit	0.002040	smith	0.001150
robert	0.013438	shabana	0.003485	amandeep	0.001720	feng	0.001150
mary	0.012764	usman	0.002790	manjit	0.001613	robin	0.001150
fiona	0.012068	asif	0.002726	sandeep	0.001506	srun	0.001008
thomas	0.010315	saima	0.002600	harpreet	0.001506	hai	0.000866
Surnames	**PROB.**	**Surname**	**PROB.**	**Surname**	**PROB.**	**Surname**	**PROB.**
smith	0.009057	khan	0.017968	singh	0.008697	jones	0.002907
brown	0.006248	hussain	0.015627	kaur	0.004890	smith	0.002164
campbell	0.005285	ali	0.012250	khan	0.002543	brown	0.001941
wilson	0.004873	ahmed	0.010584	patel	0.002353	li	0.001792
stewart	0.004834	patel	0.009909	hussain	0.002036	luo	0.001271
robertson	0.004421	akhtar	0.005047	ali	0.001782	david	0.001197
thomson	0.004362	mahmood	0.004551	ahmed	0.001782	dong	0.001048
anderson	0.004342	begum	0.004461	begum	0.001528	liu	0.000899
murray	0.004166	iqbal	0.003831	gill	0.001401	zhao	0.000825
scott	0.003832	singh	0.003066	sandhu	0.001211	ma	0.000751

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 α, β, β_1, and γ are symmetric Dirichlet priori and each hyper parameter is chosen single value which is 0.1. Gibbs sampling runs over 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 forenames *mohamad, 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 II shows the result 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 III. SURNAME CORRELATED ESTIMATION MODEL FOR INDIA DATA SET

Community 8	
Surname	**PROB.**
das	0.047381
ghosh	0.036121
roy	0.030650
banerjee	0.024349
chakraborty	0.024280
mukherjee	0.024237
saha	0.020182
sarkar	0.018965
dutta	0.018494
chatterjee	0.018213
Forename	**PROB.**
amit	0.006203
abhijit	0.004291
arindam	0.003975
anirban	0.003703
abhishek	0.003668
suman	0.003501
sanjay	0.003453
subrata	0.003447
partha	0.003381
kaushik	0.003284

Community 10	
Surname	**PROB.**
sahoo	0.030447
mohanty	0.026827
mishra	0.025976
das	0.021119
nayak	0.020382
behera	0.019462
panda	0.019417
dash	0.017908
sahu	0.016069
mohapatra	0.014038
Forename	**PROB.**
santosh	0.001891
manoj	0.001464
deepak	0.001347
manas	0.001289
biswajit	0.001133
rajesh	0.001031
sanjay	0.000987
santoshkumar	0.000972
ashok	0.000963
manoranjan	0.000943

Community 12	
Surname	**PROB.**
patil	0.032645
kulkarni	0.021224
joshi	0.012290
jadhav	0.011181
shinde	0.009799
pawar	0.009491
deshpande	0.009260
deshmukh	0.008316
gaikwad	0.006371
shaikh	0.006037
Forename	**PROB.**
sachin	0.007677
rahul	0.005960
amit	0.005789
prashant	0.005357
amol	0.004734
sandeep	0.004635
nitin	0.004469
santosh	0.004446
nilesh	0.004393
yogesh	0.004157

Community 25	
Surname	**PROB.**
gogoi	0.002211
saikia	0.001997
baruah	0.001807
borah	0.001712
deka	0.001190
kalita	0.001118
hazarika	0.001071
bora	0.000857
sarmah	0.000738
phukan	0.000643
Forename	**PROB.**
lal	0.000244
ram	0.000157
rajesh	0.000107
amit	0.000088
pankaj	0.000082
pranjal	0.000082
sanjay	0.000075
manoj	0.000063
anil	0.000063
abhijit	0.000063

TABLE IV. FORNAME CORRELATED ESTIMATION MODEL FOR INDIA DATA SET

Community 9	
Forename	**PROB.**
amit	0.017765
rahul	0.010283
sandeep	0.009322
ashish	0.009181
deepak	0.009120
manish	0.008506
abhishek	0.008176
sanjay	0.007081
rajesh	0.007064
gaurav	0.006828
Surnames	**PROB.**
kumar	0.053129
singh	0.042147
sharma	0.030981
gupta	0.020334
jain	0.014350
shah	0.009734
mishra	0.008965
yadav	0.007624
verma	0.007109
agarwal	0.007090

Community 16	
Forename	**PROB.**
patel	0.008988
jignesh	0.004512
chirag	0.003621
hardik	0.003603
dhaval	0.003272
bhavesh	0.003032
hiren	0.002983
mehul	0.002811
nirav	0.002512
kalpesh	0.002340
Surname	**PROB.**
patel	0.064114
shah	0.034320
parmar	0.008046
joshi	0.007688
mehta	0.007534
bhatt	0.007292
desai	0.006895
prajapati	0.006267
pandya	0.006218
panchal	0.006112

Community 18	
Forename	**PROB.**
abhijit	0.004777
arindam	0.004504
biswajit	0.004138
subrata	0.004004
anirban	0.003980
partha	0.003934
suman	0.003932
sourav	0.003696
sandip	0.003603
kaushik	0.003564
Surname	**PROB.**
das	0.043842
ghosh	0.031245
roy	0.025780
banerjee	0.020718
chakraborty	0.020569
mukherjee	0.020309
saha	0.017109
sarkar	0.016326
dutta	0.015930
chatterjee	0.015577

Community 25	
Forename	**PROB.**
amol	0.007978
nilesh	0.00500
sachin	0.004320
ashwini	0.004234
ganesh	0.003819
sagar	0.003708
swapnil	0.003667
amruta	0.003222
snehal	0.003169
rupali	0.003166
Surname	**PROB.**
patil	0.024298
kulkarni	0.017574
joshi	0.010838
jadhav	0.008986
shinde	0.007711
deshpande	0.007711
pawar	0.007203
deshmukh	0.006019
shaikh	0.005634
singh	0.005482

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*, *amith*, and *sanjay* share across *Bangali* and *Assami* surname communities. Forenames *sanjay* and *manoj* share across *Orrisa* and *Assami* communities.

Table IV shows the result 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.

B. Community-Community Correlated Estimation Models

This subsection presents the result of *surname community correlated LDA model* and *forename community correlated LDA model*. In the subsection IV-A, it has been observed that there could be several forenames or surnames that correlate across several surname communities or forename communities. It is important to establish communities in forenames or surnames that correlate with given communities of surnames or forenames. Hence the results of *community-community correlated estimation models* provide interaction between surname communities and forename communities. The results of *surname community correlated LDA model* provide top 3 most likely correlated forename communities of each surname community. Similarly, The results of *forename community correlated LDA model* provide top 3 most likely correlated surname communities of each forename community.

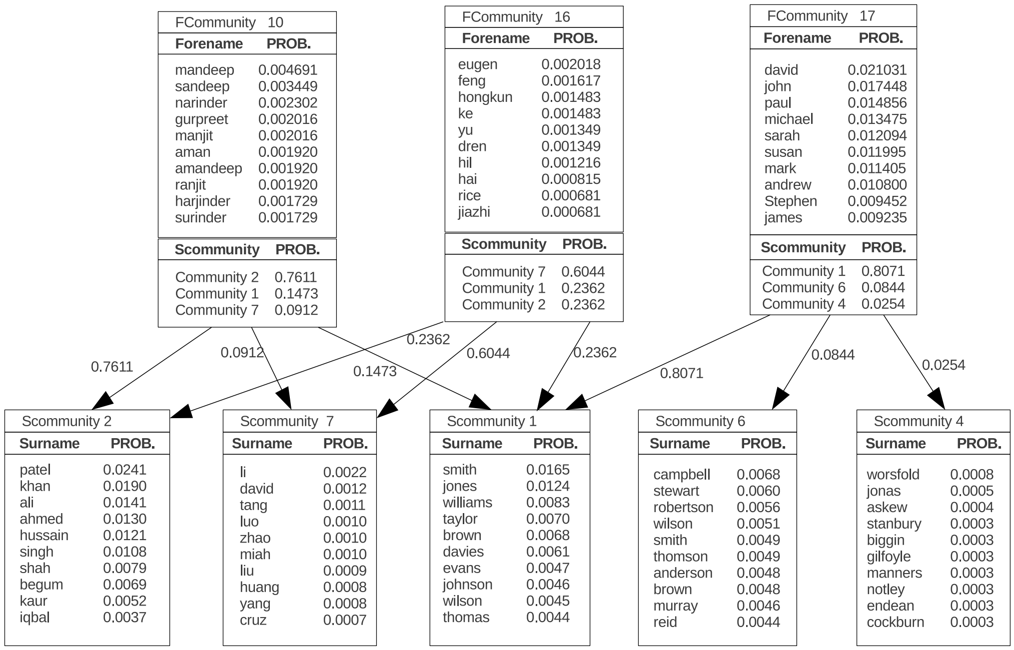

Fig. 6. Forename Community Correlated LDA Model for UK Names Data Set

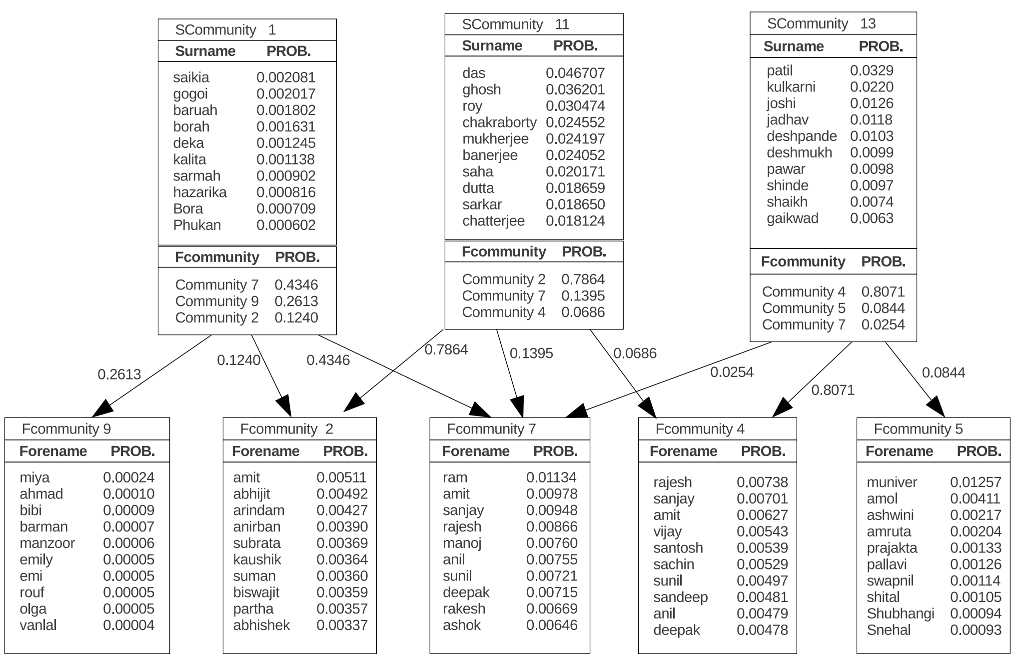

Fig. 7. Surname Community Correlated Estimation Model for India Names Data Set

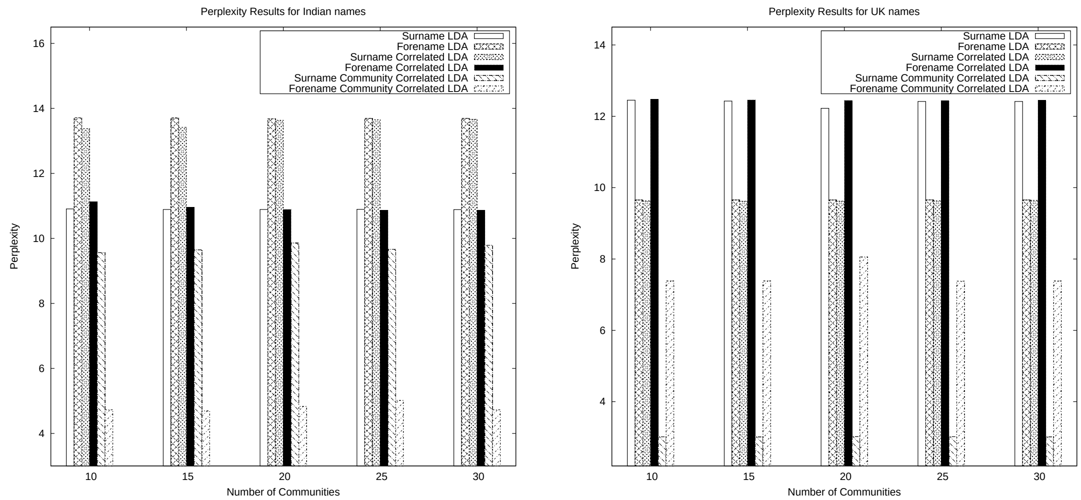

Fig. 8. Perplexity values for India and UK names data set for different probabilistic models

1) UK names data set: Figure 6 shows the results of *forename community correlated LDA model*. Forename communities (FCommunity) 10, 16, and 17 have top 10 most likely forenames and their correlated top 3 surname communities where each surname community represents top 10 most likely surnames of that community for UK names data set. Forenames in community 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 foreanme community represents top 10 most likely forenames of that community for India 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 forenames and communities of surnames. 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

[1] A. M. Dai and A. J. Storkey. The grouped author-topic model for unsupervised entity resolution. In *Proc. of the 21th international conference on Artificial neural networks*, pages 241–249, 2011.

[2] A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30(1):249–272, 2007.

[3] B. de Finetti. *Theory of probability*. John Wiley & Sons Ltd., 1975.

[4] Indrajit Bhattacharya and Lise Getoor. A latent dirichlet allocation model for entity resolution. Technical report, University of Maryland, College Park, MD, USA, 2005.

[5] James A. Cheshire and Paul A. Longley. Identifying spatial concentrations of surnames. *International Journal of Geographical Information Science*, DOI:10.1080/13658816.2011.591291:1–17, 2011.

[6] D. M. Blei and M. I. Jordan. Modelling annotated data. In *Proc. of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, 2003.

[7] D. Mimmo and A. McCallum. Expertise modeling for matching papers with reviewers. In *Proc. of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 500–509, 2007.

[8] D. Mimmo and A. McCallum. Topic models conditioned on arbitrary features with dirichlet- multinomial regression. In *Proc. of UAI*, 2008.

[9] D. Newman, C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *Proc. of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 680–686, 2006.

[10] D. Ramage, C. D. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. In *Proc. of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–465, 2011.

[11] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 1, pages 248–256, 2009.

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

[13] E. Erosheva, S. Fienberg and J. Lafferty. Mixed-membership models of scientific publications. In *Proc. of the National Academy of Sciences of the United States of America*, volume 101, pages 5220–5227, 2004.

[14] H. Zhang, B. Qiu, C. L. Giles, H. C. Foley, and J. Yen. An lda-based community structure discovery approach for large-scale social networks. In *Proc. of the 20th international joint conference on Artifical intelligence (IJCAI'07)*, pages 200–207. IEEE Intelligence and Security Informatics, 2007.

[15] J. Burt, G. Barder, and D. Rigby. *Elementary statistics for geographers*. Guilford Press, 2009.

[16] K. Aas and L. Eikvil. Text categorisation: A survey. In *Technical Report 941*. Norwegian Computing Center, 1999.

[17] L. Shu, B. Long, and W. Meng. A latent topic model for complete entity resolution. In *Proc. of the 2009 IEEE International Conference on Data Engineering*, pages 880–891, 2009.

[18] G. Lasker. Using surnames to analyse population structure. *Naming, Society and Regional Identity*, pages 3–24, 2002.

[19] Paul A. Longley, James A. Cheshire, and P Mateos. Creating a regional geography of britain through the spatial analysis of surnames. *Geoforum*, doi:10.1016, 2011.

[20] M. Steyvers, P. Smyth, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proc. of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315, 2004.

[21] P Mateos, Paul A. Longley, and David O'Sullivan. Ethnicity and population structure in personal naming networks. *PLoS ONE*, 6(9):1–12, 2011.

[22] P Mateos, A Singleton, and P A Longley. Uncertainty in the analysis of ethnicity classifications: Issues of extent and aggregation of ethnic groups. *Journal of Ethnic and Migration Studies*, 35(9):1437–1460, 2009.

[23] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. Latent dirichlet allocation for tag recommendation. In *Proc. of the third ACM conference on Recommender systems (RecSys '09)*, pages 61–68, 2009.

[24] A Rodriguez-Larralde, A. Pavesi, G. Siri, and I. Barrai. Isonamy and the genetic structure of sicily. *Journal of Biosocial Science*, 26:9–24, 1994.

[25] Suresh Veluru, R Yogachandran, and M Rajarajan. Surname identification and correction in a corpus of forename surname dataset. In *Proc. of the UK Workshop on Computational Intelligence 2012 (UKCI 2012)*, 2012.

[26] Suresh Veluru, R Yogachandran, P. Viswanath, P Longley, and M Rajarajan. E-mail address categorization based on semantics of surnames. In *Proc. of the IEEE Symposium on Computational Intelligence and Data Mining (ICDM)*, pages 222–229, 2013.

[27] T. L. Griffiths and M. Steyvers. Finding scientific topics. In *Proc. of the National Academy of Sciences of the United States of America*, volume 101, pages 5228–5235, 2004.

[28] H. M. Wallach. Topic modeling : Beyond bag-of-words. In *Proc. of the 23rd International Conference on Machine Learning*, pages 977–983, 2006.

[29] X. Wang and A. McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *Proc. of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, 2006.

[30] X. Wang, N. Mohanty and A. McCallum. Group and topic discovery from relations and their atrributes. In *Proc. of the 3rd international workshop on Link discovery*, pages 28–35, 2005.

[31] X. Wei, J. Sun, and X. Wang. Dynamic mixture models for multiple time series. In *Proc. of the 20th international joint conference on Artifical intelligence (IJCAI'07)*, pages 2909–2914. Norwegian Computing Center, 2007.