The Annals of Applied Statistics
2018, Vol. 12, No. 2, 815–845
https://doi.org/10.1214/18-AOAS1164

LOGLINEAR MODEL SELECTION AND HUMAN MOBILITY[1]

BY ADRIAN DOBRA AND REZA MOHAMMADI

University of Washington and University of Amsterdam

Methods for selecting loglinear models were among Steve Fienberg's research interests since the start of his long and fruitful career. After we dwell upon the string of papers focusing on loglinear models that can be partly attributed to Steve's contributions and influential ideas, we develop a new algorithm for selecting graphical loglinear models that is suitable for analyzing hyper-sparse contingency tables. We show how multi-way contingency tables can be used to represent patterns of human mobility. We analyze a dataset of geolocated tweets from South Africa that comprises 46 million latitude/longitude locations of 476,601 Twitter users that is summarized as a contingency table with 214 variables.

1. Introduction. Steve Fienberg was one of the founders of modern multivariate categorical data analysis. In two of the books he wrote early in his career [Bishop, Fienberg and Holland (1975), Fienberg (1980)] he laid out key notation, definitions, modeling techniques, and also open research directions for building approaches for analyzing contingency tables. More than forty years ago, he argued that interactions of various orders among categorical variables are of great interest—a fact that is now recognized in the literature from several fields (e.g., biology, social sciences, public health, transportation research). Hierarchical loglinear models that represent log expected cell counts as sums of main effects of variables cross-classified in a table, and interactions of two, three or more of these variables are well suited to capture complex multivariate patterns of dependencies. The selection of the interaction structure in hierarchical loglinear models was a problem Steve discussed in considerable length in Bishop, Fienberg and Holland [(1975), Chapter 9], Fienberg [(1980), Chapter 4], and also in several papers he subsequently published later on in his career.

Fienberg (1970) laid out one of the first strategies for hierarchical loglinear model determination which is based on partitioning the Pearson or the likelihood-ratio goodness-of-fit statistics into several additive parts. Steve's approach starts with a hierarchy of models, and a significance level. Interactions are sequentially added or deleted based on a series of tests that correspond with the partitioned

Received November 2017; revised March 2018.

[1]Supported in part by the National Science Foundation Grant DMS/MPS-1737746 to University of Washington.

Key words and phrases. Contingency tables, model selection, human mobility, graphical models, Bayesian structural learning, birth–death processes, pseudo-likelihood.

components of the most complex models. The model search stops when the difference between consecutive models is significant. Steve properly recognized that a good model building strategy must walk the fine line between goodness-of-fit and parsimony, that is, including more interactions to obtain a better fit of the data, and leaving fewer interactions in the model to create simpler representations of the association structure. However, this early method for loglinear model selection can compare only models that are nested (i.e., a simpler model is obtained from a more complex one by deleting interactions), and can be successfully used for datasets that involve no more than 5 variables.

Due in part to Steve's early contributions and ideas, several approaches to selection of loglinear models have started to emerge [Edwards and Havránek (1985), Agresti (1990), Whittaker (1990)], but these methods turned out to be quite ineffective even for contingency tables with 7 variables. One bottleneck is due to the exponential increase in the number of possible hierarchical loglinear models: while there are 7580 models with 5 variables, there are about 5.6×10^{22} models with 8 variables [Dellaportas and Forster (1999)]. Moreover, contingency tables that involve a large number of variables are sparse and their nonzero counts are imbalanced. That is, almost all the counts in large tables are zero; most of their positive counts are small (1, 2 or 3), and there are always a few counts that are quite large. Sparsity and imbalance give rise to severe difficulties when performing model selection due to the invalidation of the asymptotic approximations to the null distribution of the generalized likelihood-ratio test statistic, or the nonexistence of the maximum likelihood estimates [Fienberg and Rinaldo (2007, 2012)].

The Bayesian paradigm avoids some of these issues through the specification of prior distributions for model parameters [Clyde and George (2004)]. Dellaportas and Forster (1999) represents a key contribution that proposed a Markov chain Monte Carlo (MCMC) algorithm to identify loglinear models with high posterior probability. Other notable papers develop various stochastic search schemes for discrete data [Madigan and Raftery (1994), Madigan and York (1995, 1997), Tarantola (2004), Dellaportas and Tarantola (2005), Dobra and Massam (2010)]. These methods are known to work well for datasets with no more than 8 variables. Another approach for Bayesian model selection in contingency tables is called copula Gaussian graphical models [Dobra and Lenkoski (2011)], and it has successfully been used to analyze a 16-dimensional table. More recently, ultra-sparse high-dimensional contingency tables have been analyzed using probabilistic tensor factorizations induced through a Dirichlet process (DP) mixture model of product multinomial distributions [Dunson and Xing (2009), Canale and Dunson (2011), Bhattacharya and Dunson (2012), Kunihama and Dunson (2013)]. These papers present simulation studies and real-world data examples that involve up to 50 categorical variables.

Penalized likelihood methods for categorical data have focused on Markov random fields for binary variables [Höfling and Tibshirani (2009), Ravikumar, Wainwright and Lafferty (2010)]. Wainwright and Jordan (2008) show that higher-order interactions and variables with three or more categories can be modeled by

introducing additional binary variables in the model specification. Such claims have never been tested on known examples; from a theoretical perspective, there is no proof that the extension of the work of Höfling and Tibshirani (2009) or Ravikumar, Wainwright and Lafferty (2010) to general multi-way tables preserves the hierarchical structure of loglinear parameters, or yields consistent parameter and model estimates. The group lasso estimator for loglinear models [Nardi and Rinaldo (2012)], despite having desirable theoretical properties, does not provide guarantees that the hierarchical structure of interaction terms is preserved.

In this paper we introduce a Bayesian framework for loglinear model determination that is suitable for the analysis of a contingency table with 214 variables. Our method determines graphical loglinear models that are a special type of hierarchical loglinear models [Whittaker (1990), Lauritzen (1996)]. Our key application comes from human mobility. In this context, multivariate categorical data capture the movement of individuals across multiple geographical areas irrespective of the order in which these areas were visited, or of their spatial proximity. The goal of our analysis will be to identify graphs that have vertices associated with each area in the corresponding graphical loglinear models. The complete subgraphs of these graphs define interaction terms of joint presence and absence patterns from two, three or several areas. A missing edge between two areas means that, conditional on presence or absence in the rest of the areas, the presence or absence of a random individual in the first area is independent of the presence or absence of the same individual in the second area.

The structure of the paper is as follows. In Section 2 we discuss the relevance of massive unsolicited geolocated data for human mobility research, and in Section 3 we explain the role of loglinear models in modeling human movement. In Section 4 we describe our collection process of a geolocated Twitter dataset from South Africa; these data are subsequently transformed in the 214 dimensional contingency table we analyze in Section 7. Our modeling framework is presented in Section 5. In Section 6 we provide information about the efficiency of our proposed method in a simulation study. In Section 8 we give some concluding comments.

2. Research on human mobility. Human mobility, or movement over short or long distances for short or long periods of time, is an important yet understudied phenomenon in the social and demographic sciences. Migration processes represent a special case of human mobility that involve movements over longer periods of time and over longer distances. The impact of migration on human well-being, macro-social, political, and economic organization is a hot topic in the current literature [Donato (1993), Durand et al. (1996), Harris and Todaro (1970), Massey (1990), Massey et al. (1993, 2010), Massey and Espinosa (1997), Stark and Bloom (1985), Stark and Taylor (1985), Taylor (1987), Todaro (1969), Todaro and Maruszko (1987), VanWey (2005), Williams (2009)]. Similar advances in understanding human mobility have been hindered by difficulties in recording and

measuring how humans move on a minute and detailed scale. A notable exception is the relatively rich literature focusing on urban mobility and transportation studies. But much of this literature relies on travel surveys which are expensive to collect, have small sample sizes and limited spatial and temporal scales, are updated infrequently, and suffer from recall bias [Calabrese et al. (2013), Stopher and Greaves (2007), Wolf, Oliveira and Thompson (2003)]. Until recently, studies of mobility could not benefit from large scale data to widely address how differentials in mobility influence other outcomes. This is quite problematic given that mobility is likely a fundamental factor in behavior and macro-level social change, with potential associations with key issues that face human societies today, including spread of infectious diseases, responses to armed conflict and natural disasters, health behaviors and outcomes, economic, social, and political well-being, and migration.

Massive unsolicited geolocated data from mobile phones have recently become available for the study of human mobility. Such data are continuously collected by social media websites, search engines, and wireless-service providers [Becker et al. (2013)]. Every time a person makes a voice call, sends a text message, goes online or posts through a social media service from their mobile phone, a record is generated with information about the time and day, duration and type of communication, as well as positional information. This could be the exact latitude and longitude of the mobile phone, or an identifier of the cellular tower that handled the request. The approximate spatiotemporal trajectory of a mobile phone and its user can be reconstructed by linking the records associated with that phone. This exciting new type of data holds immense promise for studying human behavior with precision and accuracy on a vast scale never before possible with surveys or other data collection techniques [Tatem (2014), Dobra, Williams and Eagle (2015), Williams et al. (2015)].

User communications and check-ins through social media platforms such as Twitter generate publicly-available world-wide databases of human activity that can be readily accessed online free of charge. Recent evidence suggests that Twitter is a reliable source for examining human mobility patterns whose quality is comparable at the ecological level with mobile phone call records [Jurdak et al. (2015)]. The dual cultural role of Twitter as both a microblog and a social network is evidenced by the Library of Congress' decision to store a permanent, daily updated archive of the site from its first tweet. Social media offers location sharing services whose growing popularity generate digital traces that can be located in space and time. Each day, Twitter records 7 million tweets with explicit geolocation (latitude and longitude) information from mobile devices with GPS sensors [Neubauer et al. (2015)] that represent about 1.6% of the total number of tweets [Leetaru et al. (2013)]. The geographic information from geolocated tweets (geotweets) reveals the locations of human settlements and transportation networks [Leetaru et al. (2013)]. As the number of smartphone users continues to rise around the world, especially in low income countries, the potential of geolocated social

media data to improve our knowledge of human geography will constantly grow. These are the data we collect and analyze in this paper—see Sections 4 and 7.

3. Modeling human mobility. The majority of the literature on human mobility is concerned with Lèvy flights models and with Markov process models. Let us assume that traveling patterns are observed with respect to p distinct areas or locations $\{1, 2, \ldots, p\}$. Denote by N_{ij} the number of individuals that traveled from location i to location j in a given time interval, and by P_{ij} the probability that a random individual will travel to location j given that they are currently at location i. A class of stochastic process models called Lèvy flights [Brockmann, Hufnagel and Geisel (2006)] is one of the most popular ways of modeling human mobility, or to model its limits [Gonzalez, Hidalgo and Barabasi (2008)]. This model represents the probability of traveling a distance d as a power law: $\mathsf{P}(d) \propto d^{-(1+\beta)}$, where $\beta < 2$ is a diffusion parameter. The Lèvy flight model says that traveling a shorter distance is more likely than traveling longer distances, but long-distance travel can still occur even if it is rare. While this assumption is reasonable, the model implies that P_{ij} depends exclusively on d_{ij}—the distance between locations i and j. This represents a serious limitation since it implies that traveling to destinations that are located at the same distance from an origin is equally likely. A more recent contribution [Guerzhoy and Hertzmann (2014)] builds on multiplicative factor models from social network analysis [Hoff (2008)] to improve the Lèvy flights model which lacks the ability to quantify the desirability of certain travel locations. They propose a model in which $\mathsf{P}_{ij} \propto \exp(f(d_{ij}, \tau) + \mathbf{u}_i^T \mathbf{v}_j)$, where $f(d_{ij}, \tau)$ is a function of distance d_{ij}, of general parameter τ, and $\mathbf{u}_i, \mathbf{v}_j \in \mathbb{R}^q$ are location-specific latent factors. In particular, $\mathbf{u}_i^T \mathbf{v}_j$ represents the affinity of locations i and j. Inference for this latent factor model is performed based on its log-likelihood that is proportional to $-\sum_{i,j} N_{ij} \log P_{ij}$.

Both the Lèvy flights models [Brockmann, Hufnagel and Geisel (2006)] and the multiplicative latent factor models [Guerzhoy and Hertzmann (2014)] are based on the crude assumption that human travel can be seen as a Markov process in which the probability of traveling to a location depends only on the origin of the trip's segment, and does not depend on previous locations visited. However, individuals are likely to travel repeatedly across multiple locations in a given period of time. Markov process models break mobility trajectories that involve multiple locations into pairs of consecutive locations, and, by doing so, lose key dependencies that are induced by multiple locations being visited by the same individuals in the reference time frame.

Loglinear models also have a long tradition in the human mobility literature, specifically, to estimate flows of migration by origin, destination, age, sex, and other categorical sociodemographic variables such as economic activity group [Raymer, Abel and Smith (2007), Smith, Raymer and Giulietti (2010), Raymer et al. (2013)]. Migration flows are represented as origin-destination migration flow

tables. These are square tables in which the rows and columns correspond with places, regions, aggregation of places or countries of interest. The (i, j) cell contains a count of the number of individuals that left from region i and moved to region j over the course of a specified time frame. The inclusion of other categorical variables lead to higher-dimensional migration flow tables. Modeling these tables involves spatial interaction loglinear models of the form [Raymer, Abel and Smith (2007)]

$$\log(\lambda_{ijk}) = \log(\alpha_i) + \log(\beta_j) + \log(m_{ijk}),$$

where λ_{ijk} is the expected migration flow from origin i to destination j for a combination of levels k of one, two or more additional categorical variables, and m_{ijk} is auxiliary information on the migration flow. The characteristics of the origin i and the destination j are represented through the parameters α_i and β_j. However, migration flow tables cannot capture the movement of those individuals that live in more than three regions during the time frame of observation. An example individual that left from region 1 to move to region 2, then moved again to region 3, would contribute with a count of 1 to the $(1, 2)$ and $(2, 3)$ cells of the resulting migration flow table. But, the link between these two counts will be lost. For this reason, loglinear models that estimate migration flows suffer from the same shortcoming as Markov process models.

4. Description of the geolocated Twitter data. In this article we analyze a large-scale database of geolocated tweets from South Africa. This sub-Saharan country has been selected due to its high rates of internal and external migration caused by violent internal conflicts, war, political and economical instability, poverty, racial discrimination. Statistics South Africa reports that, in October 2016, 3.5 million travelers passed through South Africa's ports of entry. They were made up of 925,796 South African residents and 2.6 million foreign travelers. In this country, human mobility is known to be one of the major contributors to the spread of infectious diseases (HIV, tuberculosis, malaria) [Tatem (2014), Dobra et al. (2017)].

Our geotweets database was put together in a two step process. First, geolocated tweets posted in South Africa between September 2011 and September 2016 have been obtained directly from Twitter through GNIP, a reseller of social data owned by Twitter, as part of a no-cost collaborative research agreement between the University of Washington and Twitter. A geotweet is classified to have been posted inside a country based on a country code field derived by GNIP from the latitude and longitude of the tweet. Second, we used the Twitter REST APIs [Twitter, Inc. (2017)] to obtain geolocated tweets of the 476,601 users whose geotweets have been captured in the first step. The REST APIs allow access to up to 3200 most recent geotweets in each user's timeline irrespective of the time when they have been posted, or the location they have been posted from. For this purpose, we used a customized version of the `smappR` R package [SMaPP (2017)]. The second data

collection step took place continuously between January and December 2016. During this period, the most recent geotweets of each of the 476,601 users have been retrieved at least twice per month.

The total number of unique geotweets acquired in both steps is 46,210,370. The actual tweets have been discarded after we extracted tuples of the form `<user key, time of the posting, latitude, longitude, ...>` from the rich content of each tweet. To assure privacy protection, each Twitter user is identified by a randomly generated key which replaces their Twitter identifier. Additional filtering steps were performed to eliminate any nonhuman activity (e.g., Twitter bots) or any geotweets with coding errors. We emphasize that this database comprises only public information which can be viewed online, and replicated using the APIs provided by Twitter or downloaded directly from a third party provider of social media data such as GNIP.

For each of the 476,601 users, we estimated their country of residence as follows. We estimated the amount of time a user spent in a country they visited as the cumulative periods of time between consecutive geotweets posted in that country. A user's country of residence was defined as the country with the largest amount of time spent among all the countries this user tweeted from. Our method for identifying the users' country of residence has certain limitations. First, it is possible that a user could choose to post geotweets only when they are away from their country of residence. Second, it is also possible that our two step process of collecting geotweets might have missed relevant time intervals in which a user tweeted from their country of residence. However, after carefully examining the spatial patterns of geotweets with respect to the estimated countries of residence, we are confident that our method of determination worked fine for a large percentage of users. Based on this procedure, we classified 41,049 (8.62%) of the 476,601 users as visitors of South Africa, and the rest as locals, that is, individuals that most likely see South Africa as their home country.

We subsequently mapped the geotweets into the 213 municipalities of South Africa—see Figure 1. This allowed us to determine, for each user, the municipalities they were present and absent during the five years data collection time frame. Here we assume that absence from a municipality is implied by the user not posting any geotweets within its boundaries. These presence/absence patterns together with the Local (yes/no) variable define a 214 dimensional binary contingency table. This table is hyper-sparse: only 55,015 cells contain positive counts (the base 10 logarithm of the percentage of nonzero counts is -132.813). Among the 5015 nonzero counts, there are 46,175 (83.93%) counts of 1, 3439 (6.25%) counts of 2, 1411 (2.56%) counts of 3, 747 (1.36%) counts of 4, and 476 (0.87%) counts of 5. The top five largest counts are 58,929, 42,781, 28,731, 28,197, and 22,313, and represent the number of users that were locals to South Africa and posted geotweets only from one of following five metropolitan municipalities: Johannesburg (JHB, Gauteng), Cape Town (CPT, Western Cape), Tshwane (TSH, Gauteng), eThekwini (ETH, KwaZulu-Natal), and Ekurhuleni (EKU, Gauteng),

FIG. 1. *Administrative divisions of South Africa*: *nine provinces divided into* 52 *districts* (*dashed lines*) *that are further divided into* 213 *municipalities* (*dotted lines*). *The locations of the main cities with more than* 1 *million inhabitants* (*Johannesburg*, *Pretoria*, *Soweto*, *Cape Town*, *Port Elizabeth and Durban*) *are also shown*.

respectively—see Figure 1. The sixth largest count is 9568, and represents the number of users that were locals to South Africa, and posted geotweets from two metropolitan municipalities, Johannesburg (JHB) and Ekurhuleni (EKU). The seventh largest count count is 8464, and represents the number of users that were visitors (non-locals) to South Africa, and posted geotweets only from Johannesburg (JHB). In the next section we present our framework for determining the multivariate patterns of interactions among these 214 binary variables.

5. Bayesian structural learning in graphical loglinear models. An undirected interaction graph $G = (V, E)$ ($V = \{1, \ldots, p\}$ are vertices, and $E \subset V \times V$ are edges) is defined for a hierarchical loglinear model $\mathcal{H}$ that involves p categorical variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ as follows. A vertex $i \in V$ of G corresponds with variable X_i. An edge $e = (i, j)$ appears in G if and only if the variables X_i and X_j appear together in an interaction term of $\mathcal{H}$. Model $\mathcal{H}$ is graphical if the subsets of V that are the vertices of the complete subgraphs of G that are maximal with respect to inclusion, are also maximal interaction terms in $\mathcal{H}$ [Lauritzen (1996)]. In this case, the absence of an edge between vertices i and j in G means that X_i and X_j are conditional independent given the remaining variables $X_{V\setminus\{i,j\}}$. For this reason, the interaction graph G of a graphical loglinear model is called a conditional independence graph. This graph also has a predictive interpretation. Denote by $\mathsf{nbd}_G(i) = \{j \in V : (i, j) \in E\}$ the neighbors of vertex i in G. Then X_i is conditionally independent of $X_{V\setminus(\mathsf{nbd}_G(i)\cup\{i\})}$ given $X_{\mathsf{nbd}_G(i)}$ which implies

that, given G, a mean squared optimal prediction of X_i can be made from the neighboring variables $X_{\mathsf{nbd}_G(i)}$.

We focus on the structural learning problem [Jones et al. (2005), Drton and Maathuis (2017)] which aims to estimate the structure of G (i.e., which edges are present or absent in E) from the available data $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})$. In a Bayesian framework, we explore the posterior distribution of G conditional on the data $\mathbf{x}$, that is,

$$\mathsf{P}(G \mid \mathbf{x}) = \frac{\mathsf{P}(G)\mathsf{P}(\mathbf{x} \mid G)}{\sum_{G \in \mathcal{G}_p} \mathsf{P}(G)\mathsf{P}(\mathbf{x} \mid G)}, \tag{5.1}$$

where $\mathsf{P}(G)$ is a prior distribution on the space $\mathcal{G}_p$ of undirected graphs with p vertices, and $\mathsf{P}(\mathbf{x} \mid G)$ is the marginal likelihood of the data conditional on G [Jones et al. (2005)]. Identifying the graphs with the largest posterior probability (5.1) is a complex problem because $2^{\binom{p}{2}}$, the number of undirected graphs in $\mathcal{G}_p$, becomes large very fast as p increases. For example, for $p = 20$, the number of undirected graphs in $\mathcal{G}_p$ exceeds 10^{70}. In this paper we introduce a computationally efficient search algorithm that takes advantage of parallelizable local computations at the vertex level that moves fast towards regions with high posterior probabilities (5.1).

5.1. *Bayesian structural learning via birth–death processes.* To efficiently explore the graph space $\mathcal{G}_p$, Mohammadi and Wit (2015) developed the birth–death Markov chain Monte Carlo (BDMCMC) algorithm. This is a trans-dimensional MCMC algorithm, and represents an alternative to the well known reversible jump MCMC algorithm [Green (1995)]. The version of BDMCMC presented in Mohammadi and Wit (2015) was developed specifically for Gaussian graphical models. In this section we give a general formulation for sampling from any distributions on a space of graphs $\mathcal{G}_p$.

The BDMCMC algorithm is based on a continuous time birth–death Markov process [Preston (1975)]. Its underlying sampling scheme traverses $\mathcal{G}_p$ by adding and removing edges corresponding to the birth and death events. Given that the process is at state $G = (V, E)$, we define the birth and death events as independent Poisson processes as follows:

Birth event—each edge $e \in \overline{E}$ where $\overline{E} = \{e \in V \times V : e \notin E\}$, is born independently of other edges that do not belong to G as a Poisson process with rate $B_e(G)$. If the birth of edge e occurs, the process jumps to $G^{+e} = (V, E \cup \{e\})$ which is a graph with one edge more than G.

Death event—each edge $e \in E$ dies independently of other edges that belong to G as a Poisson process with rate $D_e(G)$. If the death of edge e occurs, the process jumps to $G^{-e} = (V, E \setminus \{e\})$ which is a graph with one edge less than G.

This birth–death Markov process is a jump process with intensity $\alpha(G) = \sum_{e \in \overline{E}} B_e(G) + \sum_{e \in E} D_e(G)$. Its waiting time to the next jump follows an exponential distribution with expectation $1/\alpha(G)$. The birth and death probabilities

are

$$\mathsf{P}(\text{birth of edge } e) \propto B_e(G) \qquad \text{for } e \in \overline{E}, \tag{5.2}$$

$$\mathsf{P}(\text{death of edge } e) \propto D_e(G) \qquad \text{for } e \in E. \tag{5.3}$$

The following theorem provides sufficient conditions on the birth and death rates to guarantee that the corresponding process on $\mathcal{G}_p$ has stationary distribution (5.1).

THEOREM 5.1. *The birth–death process defined by the birth and death probabilities* (5.2) *and* (5.3) *has the stationary distribution* $\mathsf{P}(G \mid \mathbf{x})$ *given in* (5.1), *if the following detailed balance condition is satisfied*:

$$B_e(G)\mathsf{P}(G \mid \mathbf{x}) = D_e\big(G^{+e}\big)\mathsf{P}\big(G^{+e} \mid \mathbf{x}\big), \tag{5.4}$$

where $e \in \overline{E}$, $G = (V, E)$, *and* $G^{+e} = (V, E \cup \{e\})$.

PROOF. See Section 1 in the Supplementary Material [Dobra and Mohammadi (2018)]. □

Based on Theorem 5.1, we define the birth and death rates of the BDMCMC algorithm as a function of the ratio of the corresponding posterior probabilities to optimize the convergence speed:

$$B_e(G) = \min\left\{\frac{\mathsf{P}(G^{+e} \mid \mathbf{x})}{\mathsf{P}(G \mid \mathbf{x})}, 1\right\} \qquad \text{for each } e \in \overline{E},$$

$$D_e(G) = \min\left\{\frac{\mathsf{P}(G^{-e} \mid \mathbf{x})}{\mathsf{P}(G \mid \mathbf{x})}, 1\right\} \qquad \text{for each } e \in E.$$

We show the birth and death rates as follows:

$$R_e(G) = \min\left\{\frac{\mathsf{P}(G^{*} \mid \mathbf{x})}{\mathsf{P}(G \mid \mathbf{x})}, 1\right\} \qquad \text{for each } e \in \{E \cup \overline{E}\}, \tag{5.5}$$

where for the birth of edge e we take $G^* = (V, E \cup \{e\})$, and for the death of edge e we take $G^* = (V, E \setminus \{e\})$.

Algorithm 1 provides the pseudo-code for the BDMCMC algorithm which samples from the posterior distribution (5.1) on $\mathcal{G}_p$ by using the above birth–death mechanism. In Section 5.3 we explain how to efficiently compute the ratio of posterior probabilities in the birth and death rates (5.5) for multivariate discrete data by using the marginal pseudo-likelihood approach [Pensar et al. (2017)].

5.2. *Posterior estimation via sampling in continuous time.* Figure 2 illustrates how the output of Algorithm 1 can be used to estimate posterior quantities of interest. The output consists of a set of sampled graphs, a set of waiting times $\{W_1, W_2, \ldots\}$, and a set of jumping times $\{t_1, t_2, \ldots\}$. Based on the Rao–Blackwellized estimator [Cappé, Robert and Rydén (2003)], the estimated posterior probability of each sampled graph is proportional to the expectation of length

Algorithm 1 . BDMCMC algorithm for undirected graphical models

Input: A graph $G = (V, E)$ with p nodes and data $\mathbf{x}$
  for N iterations **do**
    for all the possible edges in parallel **do**
      Calculate the birth and death rates in (5.5)
    end for
    Calculate the waiting time for G by $W(G) = \frac{1}{\sum_{e\in\overline{E}} B_e(G)+\sum_{e\in E} D_e(G)}$
    Update G based on birth/death probabilities in (5.2) and (5.3)
  end for
Output: Samples from the posterior distribution (5.1).

of the holding time in that graph which is estimated as the sum of the waiting times in that graph. The posterior inclusion probability of an edge $e \in V \times V$ is estimated by

$$\widehat{\mathsf{P}}(\text{edge } e \mid \mathbf{x}) = \frac{\sum_{t=1}^{N} \mathsf{I}(e \in G^{(t)}) W(G^{(t)})}{\sum_{t=1}^{N} W(G^{(t)})}, \tag{5.6}$$

where N denotes the number of iterations, $\mathsf{I}(e \in G^{(t)})$ denotes an indicator function: $\mathsf{I}(e \in G^{(t)}) = 1$ if $e \in G^{(t)}$, and 0 otherwise.

5.3. *Birth and death rates with the marginal pseudo-likelihood.* We assume that the observed random variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ are categorical, with each variable X_i taking values in a discrete set $\mathcal{X}_i = \{1, 2, \ldots, r_i\}$. The determination of the birth and death rates (5.5) involves the marginal likelihood conditional

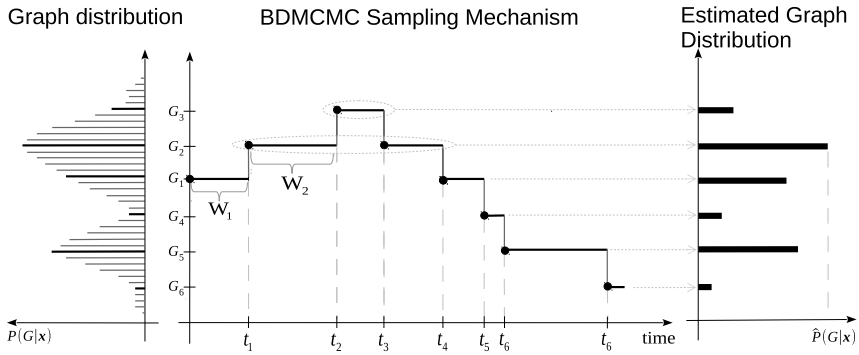


FIG. 2. *The left and right panels show the true and estimated posterior distribution* (5.1) *on the space the graphs. The middle panel shows an example output from an application of Algorithm* 1 *where* $\{W_1, W_2, \ldots\}$ *denote waiting times, and* $\{t_1, t_2, \ldots\}$ *denote jumping times.*

on a graph $G \in \mathcal{G}_p$:

$$\mathsf{P}(\mathbf{x} \mid G) = \int_{\Theta_G} \mathsf{P}(\mathbf{x} \mid \theta_G, G)\mathsf{P}(\theta_G \mid G)\, \mathrm{d}\theta_G, \tag{5.7}$$

where $\theta_G \in \Theta_G$ are the parameters of a multivariate model associated with G, $\mathsf{P}(\theta_G \mid G)$ is prior for θ_G, and $\mathsf{P}(\mathbf{x} \mid \theta_G, G)$ is the full likelihood function. However, the exact calculation of the marginal likelihood $\mathsf{P}(\mathbf{x} \mid G)$ is possible only for decomposable graphs G which represent a small fraction of the graphs in $\mathcal{G}_p$ [Massam, Liu and Dobra (2009)]. Numerical approximations for the marginal likelihood for arbitrary undirected graphs have been developed [Dobra and Massam (2010)], but their application is computationally expensive for datasets that involve $p \geq 20$ variables. This high computational effort renders them inapplicable for the Twitter mobility data described in Section 4 with $p = 214$ observed variables.

A computationally cheaper alternative comes from approximating the full likelihood $\mathsf{P}(\mathbf{x} \mid \theta_G, G)$ with the pseudo-likelihood [Besag (1975, 1977)] which is the product of the full conditionals of the random variables $\mathbf{X}$ given their neighbors in G:

$$\mathsf{P}_{\mathrm{pl}}\big(\mathbf{x} \mid \theta_G^{\mathrm{pl}}, G\big) = \prod_{d=1}^{n} \prod_{i=1}^{p} \mathsf{P}\big(X_i = x_i^{(d)} \mid \mathbf{X}_{\mathsf{nbd}_G(i)} = \mathbf{x}_{\mathsf{nbd}_G(i)}^{(d)}, \theta_{i,G}^{\mathrm{pl}}\big). \tag{5.8}$$

We denote $\mathcal{X}_A = \times_{j \in A} \mathcal{X}_j$ for $A \subseteq \{1, \ldots, p\}$, $\theta_{i,\cdot l} = \{\theta_{i,kl} : k \in \mathcal{X}_i\}$, and $\theta_{i,G}^{\mathrm{pl}} = \{\theta_{i,\cdot l} : l \in \mathcal{X}_{\mathsf{nbd}_G(i)}\}$. In (5.8), $\theta_G^{\mathrm{pl}} = \times_{i=1}^{p} \theta_{i,G}^{\mathrm{pl}} \in \Theta_G^{\mathrm{pl}}$ are the set of parameters of the full conditionals

$$\mathsf{P}(X_i = k \mid \mathbf{X}_{\mathsf{nbd}_G(i)} = l) = \theta_{i,kl} \qquad \text{for } i = 1, \ldots, p,$$

where $k \in \mathcal{X}_i$, $l \in \mathcal{X}_{\mathsf{nbd}_G(i)}$. Thus, the pseudo-likelihood (5.8) can be written as

$$\mathsf{P}_{\mathrm{pl}}\big(\mathbf{x} \mid \theta_G^{\mathrm{pl}}, G\big) = \prod_{i=1}^{p} \prod_{k \in \mathcal{X}_i} \prod_{l \in \mathcal{X}_{\mathsf{nbd}_G(i)}} \theta_{i,kl}^{n_{i,kl}}, \tag{5.9}$$

where $n_{i,kl}$ represents the number of samples $x^{(d)}$, $d = 1, 2, \ldots, n$, such that $x_i^{(d)} = k$ and $x_{\mathsf{nbd}_G(i)}^{(d)} = l$.

For computational convenience, we assume that the set of parameters $\theta_{i,G}^{\mathrm{pl}}$ and $\theta_{i',G}^{\mathrm{pl}}$ associated with the full conditionals of X_i and $X_{i'}$, $i \neq i'$ are independent. This assumption is certainly not consistent with the assumption that the full conditionals are derived from the same full joint distribution of $\mathbf{X}$. Nevertheless, the approximation of the full likelihood with the pseudo-likelihood (5.7) is based on the same premise [Besag (1975, 1977)]. We also assume that, within the same full conditional associated with the variable X_i, the parameters $\theta_{i,\cdot l}$ and $\theta_{i,\cdot l'}$ associated with the different levels l and l' of the variables $\mathbf{X}_{\mathsf{nbd}_G(i)}$ are independent

[Pensar et al. (2017)]. We impose a prior for θ_G^{pl} that factorizes according to these two assumptions:

$$\mathsf{P}(\theta_G^{\mathrm{pl}}) = \prod_{i=1}^{p} \mathsf{P}(\theta_{i,G}) = \prod_{i=1}^{p} \prod_{l \in \mathcal{X}_{\mathsf{nbd}_G(i)}} \mathsf{P}(\theta_{i,\cdot l}). \tag{5.10}$$

Furthermore, we choose a Dirichlet prior on the conditional probabilities of X_i at level $l \in \mathcal{X}_{\mathsf{nbd}_G(i)}$ of $\mathbf{X}_{\mathsf{nbd}_G(i)}$:

$$\theta_{i,\cdot l} \sim \mathsf{Dir}(\alpha_{i,1l}, \ldots, \alpha_{i,r_i l}). \tag{5.11}$$

From (5.9), (5.10), and (5.11), it follows that the marginal pseudo-likelihood is [Pensar et al. (2017)]

$$\mathsf{P}_{\mathrm{pl}}(\mathbf{x} \mid G) = \prod_{i=1}^{p} \mathsf{P}(\mathbf{x}_i \mid \mathbf{x}_{\mathsf{nbd}_G(i)}), \tag{5.12}$$

with

$$\mathsf{P}(\mathbf{x}_i \mid \mathbf{x}_{\mathsf{nbd}_G(i)}) = \prod_{l \in \mathcal{X}_{\mathsf{nbd}_G(i)}} \frac{\Gamma(\alpha_{i,\cdot l})}{\Gamma(\alpha_{i,\cdot l} + n_{i,\cdot l})} \prod_{k \in \mathcal{X}_i} \frac{\Gamma(\alpha_{i,kl} + n_{i,kl})}{\Gamma(\alpha_{i,kl})}, \tag{5.13}$$

where $\alpha_{i,\cdot l} = \sum_{k \in \mathcal{X}_i} \alpha_{i,kl}$ and $n_{i,\cdot l} = \sum_{k \in \mathcal{X}_i} n_{i,kl}$.

A prior on the space of graphs $\mathcal{G}_p$ that encourages sparsity by penalizing for the inclusion of additional edges in the graph $G = (V, E)$ is [Jones et al. (2005)]

$$\mathsf{P}(G) \propto \left(\frac{\beta}{1-\beta}\right)^{|E|} = \left(\prod_{i=1}^{p} \left(\frac{\beta}{1-\beta}\right)^{|\mathsf{nbd}_G(i)|}\right)^{1/2}, \tag{5.14}$$

where $\beta \in (0, 1)$ is set to a small value, for example, $\beta = 1/\binom{p}{2}$. While other priors on $\mathcal{G}_p$ are available [Dobra, Lenkoski and Rodriguez (2011)], the prior (5.14) can be decomposed as the product of independent priors for the p full conditionals given G such that the probability of inclusion of a vertex in each of these conditionals is equal with β as shown in (5.14).

The marginal posterior distribution on $\mathcal{G}_p$ based on the marginal pseudo-likelihood (5.12) and the prior on $\mathcal{G}_p$ (5.14) is

$$\mathsf{P}_{\mathrm{pl}}(G \mid \mathbf{x}) \propto \mathsf{P}_{\mathrm{pl}}(\mathbf{x} \mid G)\mathsf{P}(G) = \prod_{i=1}^{p} \mathsf{P}(\mathbf{x}_i \mid \mathbf{x}_{\mathsf{nbd}_G(i)}) \left(\frac{\beta}{1-\beta}\right)^{\frac{|\mathsf{nbd}_G(i)|}{2}}. \tag{5.15}$$

The birth and death rates in (5.5) based on the marginal pseudo-likelihood for an edge $e = (i, j) \in V \times V$ are calculated from

$$\widehat{R}_e(G) = \min\left\{ \frac{\mathsf{P}(\mathbf{x}_i \mid \mathbf{x}_{\mathsf{nbd}_{G^*}(i)})}{\mathsf{P}(\mathbf{x}_i \mid \mathbf{x}_{\mathsf{nbd}_G(i)})} \frac{\mathsf{P}(\mathbf{x}_j \mid \mathbf{x}_{\mathsf{nbd}_{G^*}(j)})}{\mathsf{P}(\mathbf{x}_j \mid \mathbf{x}_{\mathsf{nbd}_G(j)})} \left(\frac{\beta}{1-\beta}\right)^{\delta}, 1 \right\},$$

where for the birth of edge e we take $G^* = (V, E \cup \{e\})$, $\delta = 1$, and for the death of edge e we take $G^* = (V, E \setminus \{e\})$, $\delta = -1$.

5.4. *Dirichlet prior specification.* Consider the observed categorical variable X_i and the variables $\{X_j : j \in \mathsf{nbd}_G(i)\}$ that are its neighbors in G. The marginal table associated with $\mathbf{X}_{\{i\}\cup\mathsf{nbd}_G(i)}$ has counts $\{n_{i,kl} : k \in \mathcal{X}_i, l \in \mathcal{X}_{\mathsf{nbd}_G(i)}\}$. The counts in the slice of this marginal table defined by the combination of levels $l \in \mathcal{X}_{\mathsf{nbd}_G(i)}$ are $\{n_{i,kl} : k \in \mathcal{X}_i\}$; the sum of these counts is $n_{i,\cdot l}$. The parameters $\{\alpha_{i,kl} : k \in \mathcal{X}_i\}$ of the Dirichlet prior (5.11) can be interpreted as the values of a fictive vector of counts of the same dimension as the vector of observed marginal counts $\{n_{i,kl} : k \in \mathcal{X}_i\}$. By setting

$$\alpha_{i,kl} = \frac{1}{2} \qquad \text{for all } i = 1, \ldots, p; k \in \mathcal{X}_i; l \in \mathcal{X}_{\mathsf{nbd}_G(i)}, \tag{5.16}$$

we choose the Jeffrey's prior for the Multinomial distribution associated with the marginal counts $\{n_{i,kl} : k \in \mathcal{X}_i\}$ [Pensar et al. (2017)]. We note that the sum of the fictive vector of counts $\{\alpha_{i,kl} : k \in \mathcal{X}_i\}$ is $\alpha_{i,\cdot l} = \frac{r_i}{2}$. Alternatively, we can construct a prior specification by starting with a fictive p-dimensional table with cells indexed by $\mathcal{X}_{\{1,2,\ldots,p\}}$ that has a grand total equal with $\alpha > 0$ and equal cell values

$$\alpha(r_1 r_2 \cdot \ldots \cdot r_p)^{-1}. \tag{5.17}$$

The corresponding fictive vector of counts from (5.16) is:

$$\alpha_{i,kl} = \alpha\Bigg[\prod_{j\in\{i\}\cup\mathsf{nbd}_G(i)} r_j\Bigg]^{-1} \qquad \text{for all } i = 1, \ldots, p; k \in \mathcal{X}_i; l \in \mathcal{X}_{\mathsf{nbd}_G(i)}. \tag{5.18}$$

In this case, the sum of the fictive vector of counts $\{\alpha_{i,kl} : k \in \mathcal{X}_i\}$ is

$$\alpha_{i,\cdot l} = \alpha\Bigg[\prod_{j\in\mathsf{nbd}_G(i)} r_j\Bigg]^{-1}.$$

Fictive tables as in (5.17) define Diaconis–Ylvisaker conjugate prior distributions for parameters of hierarchical loglinear models [Massam, Liu and Dobra (2009)]. The effect of the choice of the values of α in (5.17) on the loglinear models selected based on Bayes factors determined by Diaconis–Ylvisaker conjugate priors has been studied empirically in Massam, Liu and Dobra (2009), and theoretically from a geometrical perspective in Letac and Massam (2012). These papers found that, for larger values of α, more interaction terms appear in the hierarchical loglinear models with the largest posterior probabilities. When α becomes smaller with values close to 0, the hierarchical loglinear models selected contain fewer interaction terms that involve a smaller number of variables. Since the Diaconis–Ylvisaker conjugate priors are more general versions of the Dirichlet priors in (5.11), it follows that α in (5.17) and (5.18) acts as a regularization parameter: smaller values of α that are close to 0 will lead to sparser graphs G with high posterior probabilities, while larger values of α will lead to denser graphs G. Choosing $\alpha = 1$ in (5.18) means augmenting the counts in the observed contingency table with another contingency table that has equal counts, and a grand total of 1.

From (5.11) it can be seen that, if $\alpha_{i,kl}$ are small (but strictly positive) compared to $n_{i,kl}$, then the sensitivity of the choice of values of the Dirichlet priors should be minimal. In the simulation studies and the analysis of the geolocated Twitter data from Section 7 we use the Jeffrey's prior (5.16). This choice is reasonable because the sample sizes involved in each example yield counts that are much larger than the corresponding Dirichlet parameters. However, for applications in which the sample size is considerably smaller, the sensitivity of the graphs selected with respect to different choices of Dirichlet parameters $\{\alpha_{i,kl} : k \in \mathcal{X}_i\}$ needs to be investigated.

5.5. *Speeding up the BDMCMC algorithm.* The key bottleneck of the BDMCMC algorithm is the computation at every iteration of the birth and death rates (5.5) for all the $p(p-1)/2$ possible edges. Fortunately, the rates associated with one edge can be calculated independently of the rates associated with the other edges, and can be performed in parallel which represents a first key computational improvement. We implemented parallel computations of the birth and death rates in the current version of the `R` package `BDgraph` [Mohammadi, Wit and Dobra (2018)] using `OpenMP` [OpenMP Architecture Review Board (2008)]. Most code in this package is written in `C++` and interfaced in `R`.

A second key computational improvement is possible when the marginal likelihood is replaced with the marginal pseudo-likelihood as detailed in Section 5.3. Since at each step of the BDMCMC algorithm one edge $e = (i, j)$ is selected for addition or removal, only the marginal likelihood (5.13) of the full conditionals of the two vertices i and j will change. Thus, we need to recalculate the $(p-1)+(p-1)-1 = 2p-3$ rates that correspond with these two vertices. The remaining rates will stay the same. As such, at each iteration we update $2p-3$ rates instead of $p(p-1)/2$ rates. This represents a huge computational saving especially for graphs with many vertices. For example, for the Twitter mobility data we analyze in Section 7, we look at graphs with $p = 214$ vertices. Instead of computing 22,791 rates at each step of the BDMCMC algorithm, we only need to determine 422 rates which means that a single edge update can be done approximately 54 times faster.

A third key computational improvement comes from allowing multiple edge updates at each iteration. The vast majority of the MCMC and stochastic search algorithms that have been developed in the Bayesian graphical models literature are based on adding or removing one edge at each iteration [Jones et al. (2005), Lenkoski and Dobra (2011), Scott and Carvalho (2008), Wang and Li (2012), Mohammadi et al. (2017), Mohammadi and Wit (2015), Mohammadi, Massam and Letac (2017), Cheng and Lenkoski (2012)]. These single edge updates are in part responsible for making these structural learning algorithms quite slow for datasets that comprise a larger number of variables p. Multiple birth–death sampling approaches have been used to address image processing problems that aim to

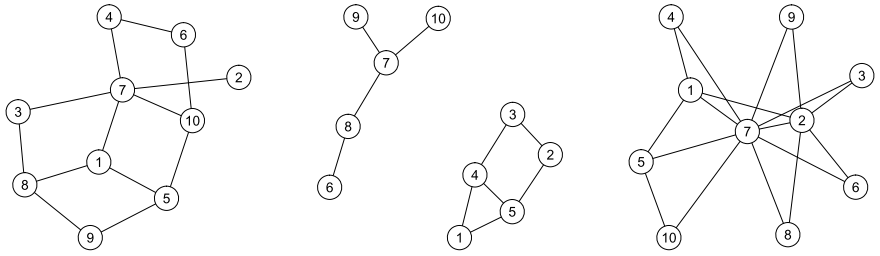
detect a configuration of objects from a digital image, and have been found to outperform the convergence speed of competing reversible jump MCMC algorithms [Descombes, Minlos and Zhizhina (2009), Gamal-Eldin, Descombes and Zerubia (2010), Gamal-Eldin et al. (2011)].

By following this idea, it is possible to transform Algorithm 1 into a multiple birth–death MCMC algorithm based on a multiple birth–death process. At each iteration, after computing and ranking the birth and deaths rates (5.5), we select not one but a fixed number $N_0 \geq 2$ of edges to be added or removed from the graph. By doing so, N_0 edges are updated at no computational cost compared to a single edge update. Through multiple edge updates which we have also implemented in the `R` package `BDgraph` [Mohammadi, Wit and Dobra (2018)], the BDMCMC algorithm can quickly move to regions with high posterior probability in the graph space $\mathcal{G}_p$. The ability to move towards high posterior probability graphs in a smaller number of iterations is especially important in applications in which the ratio between the number of samples available and the number of variables is small. Choosing N_0 can be done with respect to the amount of time required to complete one iteration of the algorithm. For example, setting $N_0 = 100$ means that 100 edges can be updated in approximately the same time Algorithm 1 would require to update a single edge. This means that one can employ $N_0 = 100$ times fewer iterations to move towards graphs with comparable complexity. However, performing multiple edge updates at each iteration of the BDMCMC algorithm does not have any theoretical guarantees related to sampling from the correct target posterior distribution (5.1). For this reason, multiple edge updates should be performed only for a reduced number of iterations to identify several graphs that have higher posterior probabilities compared to the empty graph, the full graph or a random graph sampled from $\mathcal{G}_p$. These graphs can be subsequently used as starting points for Algorithm 1 with single edge updates. For the simulation study we present next, and for the analysis of the geolocated Twitter data from Section 7 we did not need to employ multiple edge updates to identify better starting points for Algorithm 1: our sampler with single edge updates was sufficiently fast to not require us starting it from higher posterior probability graphs.

6. Simulation study. We investigate the performance of the BDMCMC algorithm in recovering the graph structure from categorical data by comparing it to the hill-climbing (HC) algorithm proposed by Pensar et al. (2017). While the BDMCMC algorithm samples from the marginal posterior distribution (5.15), the HC algorithm solves the optimization problem $\max\{\mathsf{P}_{\text{pl}}(G \mid \mathbf{x}) : G \in \mathcal{G}_p\}$ using a method that involves two phases.

We consider three types of graphs (see Figure 3):

1. *Random*: A graph in which edges were randomly generated from the prior (5.14) with $\beta = 0.4$.



(a) Random (b) Cluster (c) Scale-free

FIG. 3. *Example graphs with* $p = 10$ *vertices used in the simulation study from Section 6.*

2. *Cluster*: A graph with two clusters (connected components) each with $p = 5$ vertices. The edges in both clusters were randomly generated from the prior (5.14) with $\beta = 0.6$.

3. *Scale-free*: A graph sampled from a power-law degree distribution with the Barabási–Albert algorithm [Albert and Barabási (2002)].

We also consider graphs with $p = 20$ vertices that have two connected components with 10 vertices and the same edge structure of type "Random", "Cluster", or "Scale-free". We simulated binary contingency tables with $p \in \{10, 20\}$ variables that comprise $n \in \{200, 500, 1000\}$ samples from random graphs of these three types. We repeated the simulation experiment that involves the generation of 18 contingency tables 50 times. We performed all computations with the `R` package `BDgraph` [Mohammadi, Wit and Dobra (2017, 2018)]. For each contingency table we generated, we ran the BDMCMC and the HC algorithms using the prior (5.14) with $\beta = 0.5$ starting from the empty graph. The BDMCMC algorithm was run for 100,000 iterations. The first 60,000 iterations were discarded as burn-in.

We note that, for $p = 10$, the expected number of edges for random graphs of type "Random" and "Cluster" is $0.4 \cdot \binom{10}{2} = 18$ and $2 \cdot 0.6 \cdot \binom{5}{2} = 12$. Under our assumed prior (5.14) with $\beta = 0.5$ on $\mathcal{G}_{10}$, the expected number of edges of a random graph is $0.5 \cdot \binom{10}{2} = 22.5$. For $p = 20$, the expected number of edges of random graphs of type "Random" and "Cluster" is 36 and 24 since these graphs are generated by putting together two random graphs with $p = 10$ vertices. Under the prior (5.14) with $\beta = 0.5$ on $\mathcal{G}_{20}$, the expected number of edges of a random graph is $0.5 \cdot \binom{20}{2} = 95$. As such, our choice of priors on $\mathcal{G}_{10}$ and $\mathcal{G}_{20}$ put the BDMCMC and the HC structural learning algorithms at a disadvantage since the true graphs are on average sparser than the graphs under the prior.

We estimated the structure of the true graph based on model averaging [Madigan et al. (1996)] of the graphs sampled by the BDMCMC algorithm. We calculate the posterior inclusion probabilities of edges (5.6), and determine the median graph whose edges have posterior inclusion probabilities greater than 0.5. The structure

of the true graph was estimated with the HC algorithm based on the "and" and the "or" criteria in the first phase of the algorithm [Pensar et al. (2017)].

We evaluate the performance of the two algorithms in recovering the structure of the true graphs using the F_1-score measure [Baldi et al. (2000)],

$$F_1\text{-score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \tag{6.1}$$

and the Structure Hamming distance (SHD) [Tsamardinos, Brown and Aliferis (2006)],

$$\text{SHD} = \text{FP} + \text{FN}, \tag{6.2}$$

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively. The values of the F_1-score range between 0 and 1, and the values of the SHD are positive. A better performance in recovering the true graph is associated with larger values of the F_1-score, and with smaller values of the SHD.

The results are summarized in Table 1 in the Supplementary Material [Dobra and Mohammadi (2018)]. For most simulation experiments, the BDMCMC algorithm has an advantage over the HC algorithm especially for the F_1-score. ROC curves showing the performance of the BDMCMC algorithm are presented in Figure 1 in the Supplementary Material [Dobra and Mohammadi (2018)].

7. Analysis of the geolocated Twitter data. We come back to the $p = 214$ dimensional binary contingency table constructed from geotweets that was described in Section 4. We use the BDMCMC algorithm to sample graphs from the marginal posterior distribution (5.15) on $\mathcal{G}_{214}$. We employ the prior (5.14) with $\beta = 1/\binom{214}{2} = 4.388 \times 10^{-5}$. Under this prior, the expected number of edges is 1, thus sparser graphs receive larger prior probabilities compared to denser graphs. We performed all computations on a cluster with 7 compute nodes, each with 48 Intel Xeon 2.6 GHz cores with a Linux operating system.

First, we want to gain some understanding of the ability of the BDMCMC algorithm to move towards graphs with large posterior probabilities in $\mathcal{G}_{214}$. We sample 20 graphs $\{G_i\}_{i=1}^{20}$ having increasing number of edges: G_i has a number of edges randomly sampled from $(200(i-1), 200i)$. The resulting set of graphs ranges from most sparse (G_1) to most dense (G_{20}). Starting from each graph G_i, $i = 1, \ldots, 20$, we ran the BDMCMC algorithm for 10,000 iterations. Figures 2 and 3 in the Supplementary Material [Dobra and Mohammadi (2018)] show the sum of the estimated posterior edge inclusion probabilities and the number of edges included in the sampled graphs against iteration number. After 7000 iterations in each of the 20 runs, the BDMCMC algorithm seems to have reached the same neighborhood of graphs. Thus, although the number of graphs in $\mathcal{G}_{214}$ is extremely large ($\approx 10^{6861}$), the BDMCMC algorithm seems to be very efficient in identifying graphs with high posterior probability.

Next, we ran the BDMCMC algorithm for 400,000 iterations using parallel calculations of the birth and death rates from a starting graph sampled from the prior (5.14) on $\mathcal{G}_{214}$. The first 200,000 iterations were discarded as burn-in. Figures 4 and 5 in the Supplementary Material [Dobra and Mohammadi (2018)] show the BDMCMC algorithm seems to have reached convergence in less than 10,000 iterations.

We estimate the posterior inclusion probabilities (5.6) of the $\binom{214}{2} = 22{,}791$ edges. Figure 4 is a heatmap of the matrix of the estimated posterior edge inclusion probabilities. Most of the estimated posterior edge inclusion probabilities are zero: 21,138 (92.78%). A number of 12, 5, and 7 edges have estimated posterior inclusion probabilities in $(0, 0.5)$, $[0.5, 0.9)$ and $[0.9, 0.1)$, respectively. The remaining 1522 (6.65%) have estimated posterior inclusion probabilities equal to 1. We use the median graph which includes the 1534 edges with estimated posterior inclusion probabilities greater than 0.5 as our estimate of the conditional independence graph. Henceforth we refer to this graph as the South Africa (SA) Twitter graph.

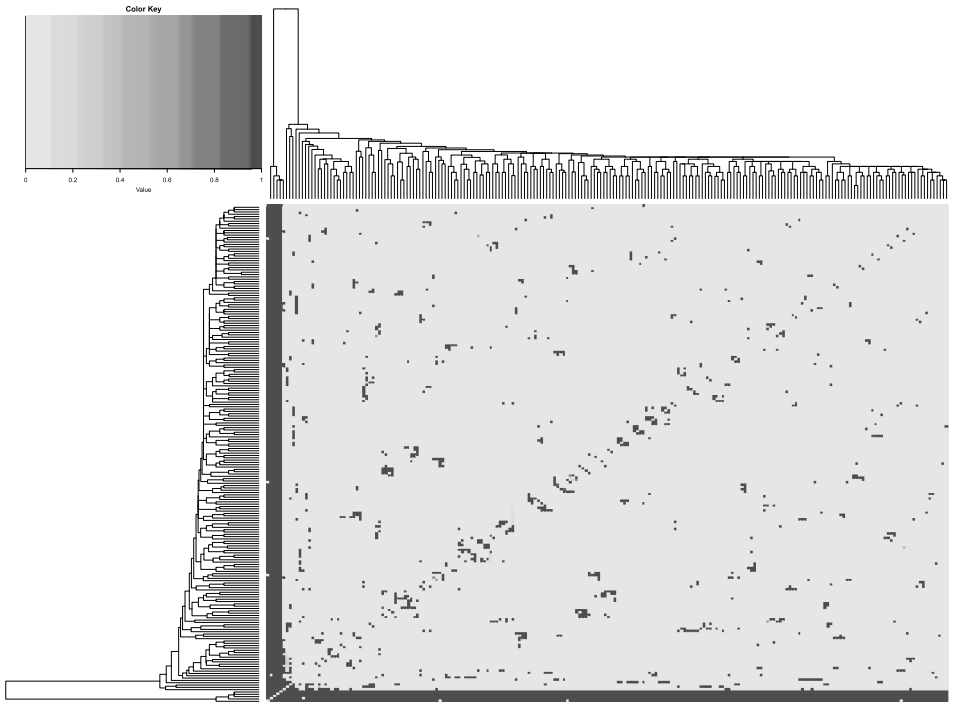


FIG. 4. *Heatmap of the* 214×214 *matrix of posterior inclusion probabilities of edges for the Twitter data. Darker shades of grey mark edges with posterior inclusion probability closer to* 1. *The bottom five rows and the leftmost five columns correspond with the five hub municipalities JHB, EKU, TSH, ETH, and CPT.*

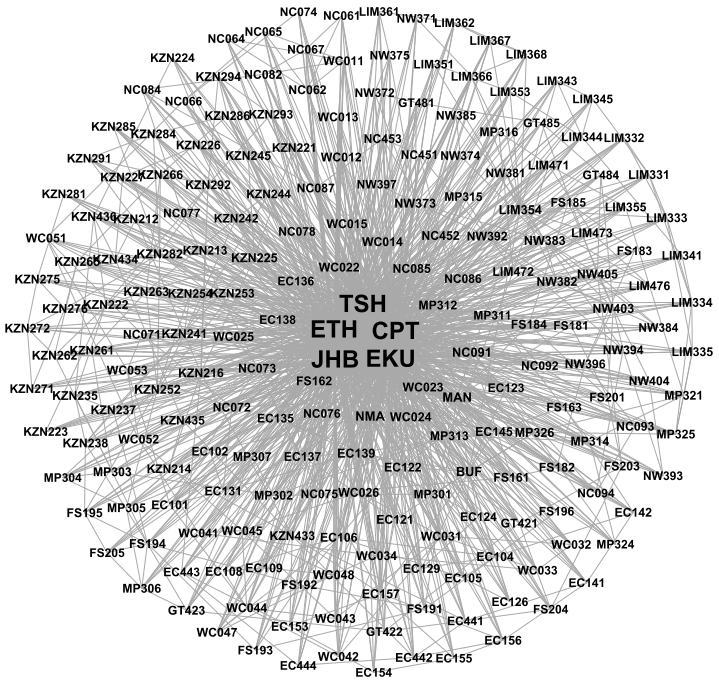


FIG. 5. *Rendering of the SA Twitter graph. The five hub municipalities* (*JHB*, *EKU*, *TSH*, *ETH*, *and CPT*) *are shown in the center of the graph. The size of the fonts is proportional with the degree of each vertex.*

A rendering of the SA Twitter graph is presented in Figure 5. The 213 municipalities of South Africa are denoted by their identifiers—see Table 5 in the Supplementary Material [Dobra and Mohammadi (2018)]. This table provides the identifier, complete name, province to which it belongs, area, population size, and density for each municipality. The 214th vertex of this graph is associated with the Local (yes/no) variable. We explore the SA Twitter graph using four centrality measures [Imai (2017)] that capture the extent to which a vertex is connected to other vertices, and occupies a central position in the structure of the network: (i) *degree* counts the number of edges that originate from a given vertex; (ii) *closeness* measures how close is a vertex from each one of the other vertices; (iii) *betweenness* finds vertices that connect other vertices (i.e., belong to the shortest paths connecting pairs of vertices); and (iv) *page rank* defines more central vertices based on a voting process which allocates votes to a vertex based on other connected vertices, and it is determined through an iterative algorithm. Barplots of the largest 10 values of each of the four centrality measures are presented in Figures 6, 7, 8, and 9 in the Supplementary Material [Dobra and Mohammadi (2018)].

For each of the four centrality measures, their top five largest values correspond with the following municipalities: Johannesburg (JHB, Gauteng), Ekurhu-

leni (EKU, Gauteng), Tshwane (TSH, Gauteng), eThekwini (ETH, KwaZulu-Natal), and Cape Town (CPT, Western Cape). The next five largest values also correspond with the same five municipalities for all four measures: Mangaung (MAN, Free State), Nelson Mandela Bay (NMA, Eastern Cape), Polokwane (LIM354, Limpopo), Buffalo City (BUF, Eastern Cape), and Sol Plaatjie (NC091, Northern Cape). We remark that the values of centrality measures for JHB, EKU, TSH, ETH, and CPT are significantly larger than the values for MAN, NMA, LIM354, BUF, and NC091. For example, the degrees for the first group are 213, 210, 213, 212, and 213, while the degrees for the second group are 25, 24, 21, 20, and 19.

As such, JHB, EKU, TSH, ETH, and CPT are the five key hubs of the SA Twitter graph. Their geographical location is mapped in Figure 6, while Table 1 gives summary information about them. From Figure 4 we see that only few of the posterior edge inclusion probabilities are strictly positive among edges that do not involve the five hubs. Three hubs (JHB, TSH, EKU) are located in the Johannesburg/Soweto/Pretoria area which represents the region of South Africa in which more than 11 million people reside (2015 South African National Census) either permanently, or temporarily to find employment in factories or gold mines. The other two hubs are located around the cities of Cape Town and Durban which, together with Johannesburg, Soweto, and Pretoria, are among the largest South African cities. A great number of local and international travelers visit these five hubs for shorter or longer periods of times. Based on the predictive interpretation

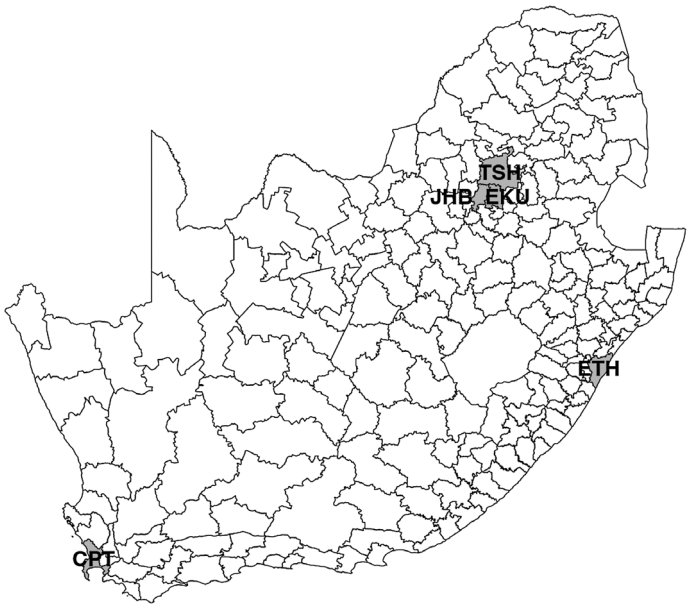


FIG. 6. *Map of South Africa showing the five hub municipalities* (*JHB*, *EKU*, *TSH*, *ETH*, *and CPT*) *of the SA Twitter graph*.

TABLE 1
Summary geographic and demographic information about the five hub municipalities in the SA Twitter graph. Population data extracted from the 2016 *Community Survey, Statistics South Africa. Retrieved from https://interactive2.statssa.gov.za/webapi*

Id.	Municipality name	Province	Area (km^2)	Population	Density
TSH	City of Tshwane	Gauteng	6298	3,275,152	520
ETH	eThekwini	KwaZulu-Natal	2556	3,702,231	1448.50
CPT	City of Cape Town	Western Cape	2446	4,005,016	1637.60
JHB	City of Johannesburg	Gauteng	1645	4,949,347	3008.80
EKU	Ekurhuleni	Gauteng	1975	3,379,104	1710.60

of the SA Twitter graph, the presence or absence of a Twitter user from one of the five hubs is predictive of the presence or absence of this user from almost all the other municipalities. Furthermore, the presence or absence of an user from almost all the municipalities that are not hubs is predictive of their presence or absence from each of the hubs.

Figure 7 shows the relationships between the number of people living in each municipality, the number of Twitter users that posted geolocated messages in each municipality, and the degree of the vertices associated with each municipality in the SA Twitter graph. The five hub municipalities stand out in their own cluster clearly separated from the rest: they represent the municipalities with the largest population, the largest number of geolocated Twitter users, and the largest degree.

The vertex associated with the variable Local is not central in the structure of the SA Twitter graph. Its degree is 12, and the other centrality measures are also significantly smaller compared to those of the five hubs. The 12 municipalities that are connected with an edge with the Local vertex are mapped in Figure 8. In addi-

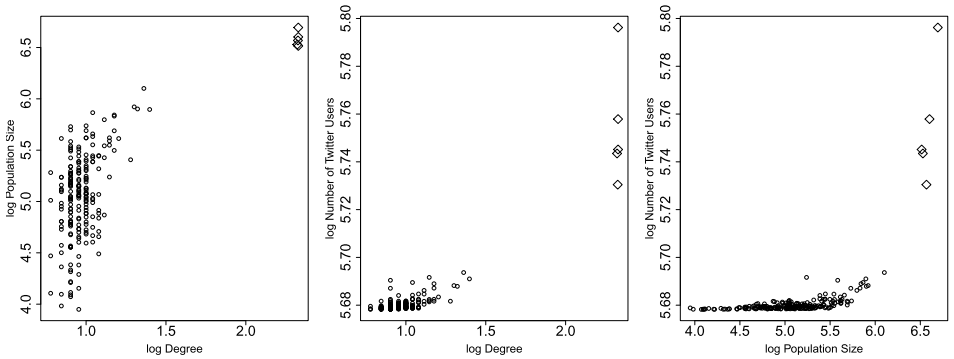


FIG. 7. *Pairwise relationships between population size, number of Twitter users, and degree in the SA Twitter graph of each municipality. The five hub municipalities (JHB, EKU, TSH, ETH, and CPT) are marked with diamonds.*

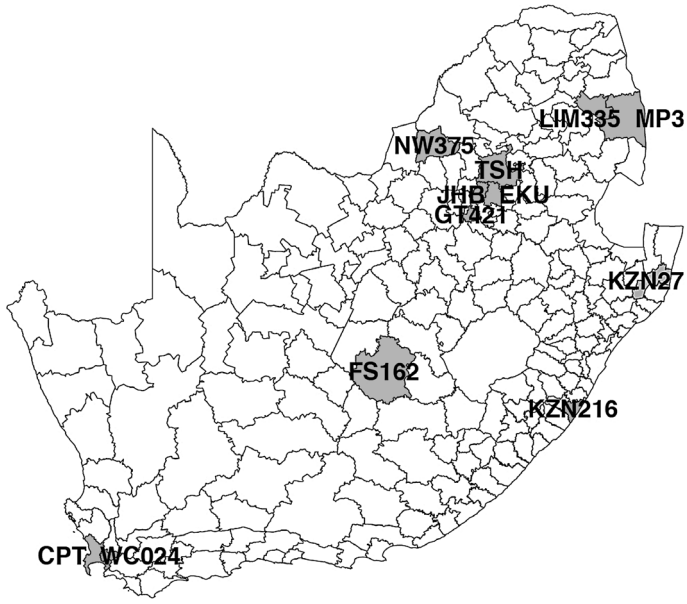


FIG. 8. *Map of South Africa showing the* 12 *municipalities that are linked by an edge with the vertex associated with the Local variable in the SA Twitter graph.*

tion to four of the hubs (JHB, EKU, TSH, CPT), the presence or absence patterns of a Twitter user from the following municipalities are predictive of whether this user is local to South Africa: Kopanong (FS162, Free State), Ray Nkonyeni and Big Five Hlabisa (KZN216 and KZN276, KwaZulu-Natal), Maruleng (LIM335, Limpopo), Bushbuckridge (MP325, Mpumalanga), Moses Kotane (NW375, North West), Emfuleni (GT421, Gauteng), and Stellenbosch (WC024, Western Cape). It is quite interesting to examine the spatial distribution of these 12 municipalities: CPT and WC024 are adjacent municipalities around Cape Town; TSH, JHB, and EKU define a spatially contiguous region in the Johannesburg/Soweto/Pretoria area; while LIM335 and MP325 are adjacent municipalities at the border between South Africa and Mozambique. In the KwaZulu-Natal province, the municipalities KZN216 and KZN276 that are located to the south and to the north of the city of Durban are among the neighbors of Local, but the ETH municipality in which Durban is located is not (quite surprisingly) among the neighbors of Local. The FS162 municipality is located south of Bloemfontein—a major city in South Africa known for its mining industry. The NW375 municipality is located north west of the Johannesburg/Pretoria area, and it comprises Sun City and a major national park—both key touristic destinations.

We determine the effect of the presence and absence patterns of Twitter users from these 12 municipalities on the odds of being local to South Africa by fitting a logistic regression model for the Local variable with 12 explanatory variables associated with these municipalities. The estimated adjusted odds ratios are given

in Tables 2 and 3 in the Supplementary Material [Dobra and Mohammadi (2018)]. A number of 10 municipalities have adjusted odds ratios significantly smaller than 1 at significance level $\alpha = 0.05$. Given the same presence and absence pattern in the remaining 11 municipalities, a Twitter user that posted geotweets from one of these municipalities has smaller odds of being local to South Africa compared to another Twitter user that did not post geotweets from that municipality. However, the TSH and GT421 municipalities located to the north and to the south of the Johannesburg/Soweto/Pretoria area have estimated adjusted odds ratios significantly greater than 1 at significance level $\alpha = 0.05$. Given the same presence and absence pattern in the remaining 11 municipalities, the odds of being local to South Africa of a Twitter user that was present in GT421 (TSH) are 5.414 (2.347) times larger than the odds of being local to South Africa of another Twitter user that was absent from GT421 (TSH). It is known that a considerable number of Mozambicans come to work in the mines in the Johannesburg/Pretoria area for extended periods of time [Baltazar et al. (2015)]. Their residences might be located in the GT421 and TSH municipalities where they could exceed the number of South African Twitter users.

It is also interesting to examine the interaction structure of the graphical loglinear model induced by the SA Twitter graph. The interaction terms present in this model are complete subgraphs or cliques of the SA Twitter graph. The generators of this model [Edwards and Havránek (1985)] are the maximal interaction terms, that is, terms that are not a subset of other interaction terms that are also present in the model. The generators of a graphical loglinear model are the maximal cliques of the graph that defines that model. The SA Twitter graph has 251 maximal cliques—see Table 4 in the Supplementary Material [Dobra and Mohammadi (2018)]. There are 11 generators of size 6, 94 generators of size 7, 126 generators of size 8, and 20 generators of size 20. All five hub municipalities appear in the 20 largest generators—see Table 2. Three of them, namely JHB, CPT, and TSH appear in all the 251 generators. ETH and EKU appear in 244 and 242 generators, respectively. The municipalities NMA, MAN, NC091, and BUF appear in 14, 14, 10, and 9 generators, respectively. The rest of the municipalities appear in 8 generators or less.

8. Conclusions. This paper makes several contributions. First, it generalizes the birth–death Markov chain Monte Carlo (BDMCMC) algorithm introduced by Mohammadi and Wit (2015) in the context of Gaussian graphical models to general undirected graphical models. Second, based on marginal pseudo-likelihood for categorical data of Pensar et al. (2017), we show how to efficiently calculate the birth and death rates for the BDMCMC algorithm for arbitrary undirected graphs. Third, we use our methodology to analyze a 214-dimensional contingency table that captures the mobility patterns of Twitter users in South Africa. This is a dataset we collected at the University of Washington which has never been analyzed before. We learned that five municipalities are hubs for the mobility patterns of Twitter

TABLE 2
The largest cliques of the SA Twitter graph. The hub municipalities (JHB, EKU, TSH, ETH, and CPT) appear in bold

Id.	Size	Clique
1	9	**CPT EKU ETH JHB TSH** FS181 FS184 FS201 MAN
2	9	**CPT EKU ETH JHB TSH** FS194 KZN235 KZN237 KZN238
3	9	**CPT EKU ETH JHB TSH** BUF EC121 EC122 EC157
4	9	**CPT EKU ETH JHB TSH** BUF EC124 EC129 EC139
5	9	**CPT EKU ETH JHB TSH** BUF EC104 EC105 EC126
6	9	**CPT EKU ETH JHB TSH** FS161 FS182 MAN NC091
7	9	**CPT EKU ETH JHB TSH** KZN282 KZN284 KZN291 KZN292
8	9	**CPT EKU ETH JHB TSH** LIM331 LIM332 LIM333 LIM354
9	9	**CPT EKU ETH JHB TSH** LIM354 LIM355 LIM473 LIM476
10	9	**CPT EKU ETH JHB TSH** LIM354 LIM366 LIM367 LIM368
11	9	**CPT EKU ETH JHB TSH** NC071 NC072 NC073 WC053
12	9	**CPT EKU ETH JHB TSH** BUF EC104 EC105 NMA
13	9	**CPT EKU ETH JHB TSH** EC104 EC105 EC106 NMA
14	9	**CPT EKU ETH JHB TSH** BUF LIM354 MAN NMA
15	9	**CPT EKU ETH JHB TSH** BUF MAN MP326 NMA
16	9	**CPT EKU ETH JHB TSH** NW373 NW374 NW383 NW385
17	9	**CPT EKU ETH JHB TSH** NW381 NW382 NW383 NW392
18	9	**CPT EKU ETH JHB TSH** WC012 WC013 WC014 WC015
19	9	**CPT EKU ETH JHB TSH** WC014 WC015 WC023 WC024
20	9	**CPT EKU ETH JHB TSH** WC043 WC044 WC045 WC048

users in South Africa, and that the presence or absence of Twitter users from 12 municipalities are predictive of users being locals or visitors of South Africa.

The hill-climbing (HC) algorithm [Pensar et al. (2017)] determines graphs with high posterior probability using a greedy hill-climbing optimization algorithm. For this reason, the HC algorithm will inevitably end up in a local maximum. Which local maximum the HC algorithm will find depends on the choice of starting graph. The results of the simulation study from Section 6 were obtained by starting the HC algorithm from empty graphs. Since the true graphs were sparse, the HC algorithm recorded a good performance that was comparable with the performance of the BDMCMC algorithm. However, if we would have started the HC algorithm from random graphs that contained a larger number of edges, the HC algorithm might have been at a disadvantage. As we illustrated in Section 7, starting the BDMCMC algorithm from sparser or denser graphs led to the identification of the same neighborhood of graphs with high posterior probabilities. The BDMCMC algorithm has a key advantage over the HC algorithm in terms of its ability to visit graphs with lower posterior probability in order to escape local optima, and move towards other graphs with larger posterior probabilities.

Our applied results give an understanding of the movements of 476,601 individuals that used geolocated tweets in South Africa between 2011 and 2016. It is true

that the movements of this specific group of people might not be representative of major flows of movement of South Africans, or of the visitors of this country. And, due to the selected locations Twitter users choose to post their tweets from, it is possible that even the travel trajectories of these individuals could be only partially captured. However, to the best of our knowledge, there is no other study on human mobility that involves a larger number of individuals in South Africa, and comprises a larger number of recorded locations (>46 millions). While our findings must be interpreted with care from a sociodemographic perspective, the methodology we introduce in this article can be successfully applied to modeling patterns of repeated across regions movement that span entire countries, and comprise a large number of individuals.

Our modeling approach is based on multi-way contingency tables that cross-classify presence and absence patterns from regions of interest, together with other relevant categorical factors. Our framework goes beyond methods that focus exclusively on modeling flows of migration between origin and destination areas. However, our methodology has several limitations. A significant loss of information occurred when the latitude and longitude coordinates of the geotweets were mapped into municipalities. Furthermore, we took into consideration only the presence or absence of an individual in each municipality: a single tweet in a municipality marked an individual as present in that area. The actual frequency of tweets of an individual in a municipality has not been accounted for, as well as the temporal sequence in which the geotweets have been posted. This means that the order in which an individual visited municipalities was not modeled in our analysis. The spatial distribution of municipalities has also been overlooked. It is plausible that an individual will most often visit municipalities that are spatially close to each other as opposed to more distant municipalities. We are currently working on extending our modeling framework to address some of these limitations.

In a companion short paper [Mohammadi and Dobra (2017)], we provide code and explanations on the use of the `R` package `BDgraph` [Mohammadi, Wit and Dobra (2018)] to analyze contingency tables. In particular, we analyze a well known six-way contingency table called the Czech autoworkers data [Edwards and Havránek (1985)], and compare the interaction graph identified with the BDMCMC algorithm with the graphs determined using the `R` package `gRim` [Højsgaard, Edwards and Lauritzen (2012)], the MCMC algorithm of Massam, Liu and Dobra (2009), and the loglinear model determination method of Edwards and Havránek (1985). We also give example code for analyzing the geolocated Twitter data contingency table from Sections 4 and 7.

We empirically demonstrated that our version of the BDMCMC algorithm can efficiently determine conditional independence graphs with 214 categorical variables. To the best of our knowledge, this is the highest-dimensional contingency table analyzed so far with loglinear models. These developments would not have been possible without Steve Fienberg's visionary life long work which led to the birth of a research community that spans several disciplines (social sciences, health

and medical sciences, computer science, and statistics) and will continue to generate fundamental scientific knowledge for many generations to come.

Acknowledgments. The authors thank Johan Pensar for providing some of the code used in the simulation study and Sven Baars for his suggestions related to parallel coding in C++. We also would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

SUPPLEMENTARY MATERIAL

Additional proofs, maps, figures and tables (DOI: 10.1214/18-AOAS1164SUPP; .pdf). In this online supplementary material, we provide the proof for Theorem 5.1, together with additional maps, figures, and tables referenced in this article.

REFERENCES

AGRESTI, A. (1990). *Categorical Data Analysis*. Wiley, New York. MR1044993

ALBERT, R. and BARABÁSI, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Modern Phys*. **74** 47–97. MR1895096

BALDI, P., BRUNAK, S., CHAUVIN, Y., ANDERSEN, C. A. and NIELSEN, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics* **16** 412–424.

BALTAZAR, C. S., HORTH, R., INGUANE, C., SATHANE, I., CÉSAR, F., RICARDO, H., BOTÃO, C., AUGUSTO, Â., COOLEY, L., CUMMINGS, B., RAYMOND, H. F. and YOUNG, P. W. (2015). HIV prevalence and risk behaviors among Mozambicans working in South African mines. *AIDS Behav.* **19** 59–67.

BECKER, R., CÁCERES, R., HANSON, K., ISAACMAN, S., LOH, J. M., MARTONOSI, M., ROWLAND, J., URBANEK, S., VARSHAVSKY, A. and VOLINSKY, C. (2013). Human mobility characterization from cellular network data. *Commun. ACM* **56** 74–82.

BESAG, J. (1975). Statistical analysis of non-lattice data. *J. R. Stat. Soc., Ser. D Stat.* **24** 179–195.

BESAG, J. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika* **64** 616–618. MR0494640

BHATTACHARYA, A. and DUNSON, D. B. (2012). Simplex factor models for multivariate unordered categorical data. *J. Amer. Statist. Assoc.* **107** 362–377. MR2949366

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA. With the collaboration of Richard J. Light and Frederick Mosteller. MR0381130

BROCKMANN, D., HUFNAGEL, L. and GEISEL, T. (2006). The scaling laws of human travel. *Nature* **439** 462–465.

CALABRESE, F., DIAO, M., LORENZO, G. D., FERREIRA JR., J. and RATTI, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transp. Res., Part C, Emerg. Technol.* **26** 301–313.

CANALE, A. and DUNSON, D. B. (2011). Bayesian kernel mixtures for counts. *J. Amer. Statist. Assoc.* **106** 1528–1539. MR2896854

CAPPÉ, O., ROBERT, C. P. and RYDÉN, T. (2003). Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 679–700. MR1998628

CHENG, Y. and LENKOSKI, A. (2012). Hierarchical Gaussian graphical models: Beyond reversible jump. *Electron. J. Stat.* **6** 2309–2331. MR3020264

CLYDE, M. and GEORGE, E. I. (2004). Model uncertainty. *Statist. Sci.* **19** 81–94. MR2082148

DELLAPORTAS, P. and FORSTER, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86** 615–633. MR1723782

DELLAPORTAS, P. and TARANTOLA, C. (2005). Model determination for categorical data with factor level merging. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 269–283. MR2137325

DESCOMBES, X., MINLOS, R. and ZHIZHINA, E. (2009). Object extraction using a stochastic birth-and-death dynamics in continuum. *J. Math. Imaging Vision* **33** 347–359. MR2480967

DOBRA, A. and LENKOSKI, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Stat.* **5** 969–993. MR2840183

DOBRA, A., LENKOSKI, A. and RODRIGUEZ, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *J. Amer. Statist. Assoc.* **106** 1418–1433. MR2896846

DOBRA, A. and MASSAM, H. (2010). The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. *Stat. Methodol.* **7** 240–253. MR2643600

DOBRA, A. and MOHAMMADI, R. (2018). Supplement to "Loglinear model selection and human mobility." DOI:10.1214/18-AOAS1164SUPP.

DOBRA, A., WILLIAMS, N. E. and EAGLE, N. (2015). Spatiotemporal detection of unusual human population behavior using mobile phone data. *PLoS ONE* **10** 1–20.

DOBRA, A., BÄRNIGHAUSEN, T., VANDORMAEL, A. and TANSER, F. (2017). Space-time migration patterns and risk of HIV acquisition in rural South Africa. *AIDS* **31** 37–145.

DONATO, K. M. (1993). Current trends and patterns of female migration: Evidence from Mexico. *Int. Migr. Rev.* **27** 748–771.

DRTON, M. and MAATHUIS, M. H. (2017). Structure learning in graphical modeling. *Annu. Rev. Statist. Appl.* **4** 365–393.

DUNSON, D. B. and XING, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Amer. Statist. Assoc.* **104** 1042–1051. MR2562004

DURAND, J., KANDEL, W., PARRADO, E. A. and MASSEY, D. S. (1996). International migration and development in Mexican communities. *Demography* **33** 249–264.

EDWARDS, D. and HAVRÁNEK, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika* **72** 339–351. MR0801773

FIENBERG, S. E. (1970). The analysis of multidimensional contingency tables. *Ecology* **51** 419–433.

FIENBERG, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*, 2nd ed. MIT Press, Cambridge, MA. MR0623082

FIENBERG, S. E. and RINALDO, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *J. Statist. Plann. Inference* **137** 3430–3445. MR2363267

FIENBERG, S. E. and RINALDO, A. (2012). Maximum likelihood estimation in log-linear models. *Ann. Statist.* **40** 996–1023. MR2985941

GAMAL-ELDIN, A., DESCOMBES, X. and ZERUBIA, J. (2010). Multiple birth and cut algorithm for point process optimization. In 2010 *Sixth International Conference on Signal-Image Technology and Internet-Based Systems* (*SITIS*) 35–42. IEEE, Los Alamitos, CA.

GAMAL-ELDIN, A., DESCOMBES, X., CHARPIAT, G. and ZERUBIA, J. (2011). A fast multiple birth and cut algorithm using belief propagation. In 2011 18*th IEEE International Conference on Image Processing* 2813–2816. IEEE, Los Alamitos, CA.

GONZALEZ, M. C., HIDALGO, C. A. and BARABASI, A.-L. (2008). Understanding individual human mobility patterns. *Nature* **453** 779–782.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

GUERZHOY, M. and HERTZMANN, A. (2014). Learning latent factor models of travel data for travel prediction and analysis. In *Advances in Artificial Intelligence*. *Lecture Notes in Computer Science* **8436** 131–142. Springer, Cham. MR3218638

HARRIS, J. R. and TODARO, M. P. (1970). Migration, unemployment and development: A two-sector analysis. *Am. Econ. Rev.* **60** 126–142.

HOFF, P. D. (2008). Multiplicative latent factor models for description and prediction of social networks. *Comput. Math. Organ. Theory* **15** Art. ID 261.

HÖFLING, H. and TIBSHIRANI, R. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.* **10** 883–906. MR2505138

HØJSGAARD, S., EDWARDS, D. and LAURITZEN, S. (2012). *Graphical Models with R*. Springer, New York. MR2905395

IMAI, K. (2017). *Quantitative Social Science*: *An Introduction*. Princeton Univ. Press, Princeton, NJ.

JONES, B., CARVALHO, C., DOBRA, A., HANS, C., CARTER, C. and WEST, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci.* **20** 388–400. MR2210226

JURDAK, R., ZHAO, K., LIU, J., ABOUJAOUDE, M., CAMERON, M. and NEWTH, D. (2015). Understanding human mobility from Twitter. *PLoS ONE* **10** 1–16.

KUNIHAMA, T. and DUNSON, D. B. (2013). Bayesian modeling of temporal dependence in large sparse contingency tables. *J. Amer. Statist. Assoc.* **108** 1324–1338. MR3174711

LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. The Clarendon Press, Oxford Univ. Press, New York. MR1419991

LEETARU, K., WANG, S., CAO, G., PADMANABHAN, A. and SHOOK, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday* **18**. Available at http://firstmonday.org/ojs/index.php/fm/article/view/4366/3654.

LENKOSKI, A. and DOBRA, A. (2011). Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *J. Comput. Graph. Statist.* **20** 140–157. Supplementary material available online. MR2816542

LETAC, G. and MASSAM, H. (2012). Bayes factors and the geometry of discrete hierarchical log-linear models. *Ann. Statist.* **40** 861–890. MR2985936

MADIGAN, D. and RAFTERY, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89** 1535–1546.

MADIGAN, D. and YORK, J. (1995). Bayesian graphical models for discrete data. *Int. Stat. Rev.* **63** 215–232.

MADIGAN, D. and YORK, J. C. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika* **84** 19–31. MR1450189

MADIGAN, D., RAFTERY, A. E., VOLINSKY, C. and HOETING, J. (1996). Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models* 77–83.

MASSAM, H., LIU, J. and DOBRA, A. (2009). A conjugate prior for discrete hierarchical log-linear models. *Ann. Statist.* **37** 3431–3467. MR2549565

MASSEY, D. S. (1990). Social structure, household strategies, and the cumulative causation of migration. *Popul. Index* **56** 3–26.

MASSEY, D. S. and ESPINOSA, K. E. (1997). What's driving Mexico–U.S. migration? A theoretical, empirical, and policy analysis. *Am. J. Sociol.* **102** 939–999.

MASSEY, D. S., ARANGO, J., HUGO, G., KOUAOUCI, A., PELLEGRINO, A. and TAYLOR, J. E. (1993). Theories of international migration: A review and appraisal. *Popul. Dev. Rev.* **19** 431–466.

MASSEY, D. S., WILLIAMS, N., AXINN, W. G. and GHIMIRE, D. (2010). Community services and out-migration. *Int. Migr.* **48** 1–41.

MOHAMMADI, A. and DOBRA, A. (2017). The R package BDgraph for Bayesian structure learning in graphical models. *ISBA Bull.* **4** 11–16.

MOHAMMADI, A., MASSAM, H. and LETAC, G. (2017). The ratio of normalizing constants for Bayesian graphical Gaussian model selection. Preprint. Available at arXiv:1706.04416.

MOHAMMADI, A. and WIT, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Anal*. **10** 109–138. MR3420899

MOHAMMADI, R. and WIT, E. C. (2017). BDgraph: An R package for Bayesian structure learning in graphical models. Preprint. Available at arXiv:1501.05108v4.

MOHAMMADI, R. and WIT, E. C. and DOBRA, A. (2018). BDgraph: Bayesian structure learning in graphical models using birth–death MCMC. R package version 2.49.

MOHAMMADI, A., ABEGAZ, F., VAN DEN HEUVEL, E. and WIT, E. C. (2017). Bayesian modelling of Dupuytren disease by using Gaussian copula graphical models. *J. R. Stat. Soc. Ser. C. Appl. Stat*. **66** 629–645. MR3632345

NARDI, Y. and RINALDO, A. (2012). The log-linear group-lasso estimator and its asymptotic properties. *Bernoulli* **18** 945–974. MR2948908

NEUBAUER, G., HUBER, H., VOGL, A., JAGER, B., PREINERSTORFER, A., SCHIRNHOFER, S., SCHIMAK, G. and HAVLIK, D. (2015). On the volume of geo-referenced tweets and their relationship to events relevant for migration tracking. In *Environmental Software Systems. Infrastructures, Services and Applications*: 11*th IFIP WG* 5.11 *International Symposium, ISESS* 2015, *Melbourne, VIC, Australia, March* 25–27, 2015. *Proceedings* (R. Denzer, R. M. Argent, G. Schimak and J. Hřebíček, eds.) 520–530. Springer, Cham.

OPENMP ARCHITECTURE REVIEW BOARD (2008). OpenMP application program interface version 3.0.

PENSAR, J., NYMAN, H., NIIRANEN, J. and CORANDER, J. (2017). Marginal pseudo-likelihood learning of discrete Markov network structures. *Bayesian Anal*. **12** 1195–1215. MR3724983

PRESTON, C. (1975). Spatial birth-and-death processes. *Bull. Inst. Int. Stat*. **46** 371–391, 405–408 (1975). With discussion. MR0474532

RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist*. **38** 1287–1319. MR2662343

RAYMER, J., ABEL, G. and SMITH, P. W. F. (2007). Combining census and registration data to estimate detailed elderly migration flows in England and Wales. *J. Roy. Statist. Soc. Ser. A* **170** 891–908. MR2408983

RAYMER, J., WIŚNIOWSKI, A., FORSTER, J. J., SMITH, P. W. F. and BIJAK, J. (2013). Integrated modeling of European migration. *J. Amer. Statist. Assoc*. **108** 801–819. MR3174664

SCOTT, J. G. and CARVALHO, C. M. (2008). Feature-inclusion stochastic search for Gaussian graphical models. *J. Comput. Graph. Statist*. **17** 790–808. MR2649067

SMAPP (2017). smappR package: Tools for analysis of Twitter data, Social Media and Participation, New York University. Available at https://github.com/SMAPPNYU/smappR.

SMITH, P. W. F., RAYMER, J. and GIULIETTI, C. (2010). Combining available migration data in England to study economic activity flows over time. *J. Roy. Statist. Soc. Ser. A* **173** 733–753. MR2759963

STARK, O. and BLOOM, D. E. (1985). The new economics of labor migration. *Am. Econ. Rev*. **75** 173–178.

STARK, O. and TAYLOR, J. E. (1985). Migration incentives, migration types: The role of relative deprivation. *Econ. J*. **101** 1163–1178.

STOPHER, P. R. and GREAVES, S. P. (2007). Household travel surveys: Where are we going? *Transp. Res., Part A Policy Pract*. **41** 367–381.

TARANTOLA, C. (2004). MCMC model determination for discrete graphical models. *Stat. Model*. **4** 39–61. MR2037813

TATEM, A. J. (2014). Mapping population and pathogen movements. *Int. Health* **6** 5–11.

TAYLOR, J. E. (1987). Undocumented Mexico–U.S. migration and the returns to households in rural Mexico. *Am. J. Agric. Econ*. **69** 616–638.

TODARO, M. P. (1969). A model of labor migration and urban unemployment in less developed countries. *Am. Econ. Rev*. **59** 138–148.

TODARO, M. P. and MARUSZKO, L. (1987). Illegal immigration and U.S. immigration reform: A conceptual framework. *Popul. Dev. Rev.* **13** 101–114.

TSAMARDINOS, I., BROWN, L. E. and ALIFERIS, C. F. (2006). The max–min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.* **65** 31–78.

TWITTER, INC. (2017). Twitter REST APIs. Available at https://dev.twitter.com/rest/public.

VANWEY, L. K. (2005). Land ownership as a determinant of international and internal migration in Mexico and internal migration in Thailand. *Int. Migr. Rev.* **39** 141–172.

WAINWRIGHT, M. and JORDAN, M. (2008). Graphical models, exponential families and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.

WANG, H. and LI, S. Z. (2012). Efficient Gaussian graphical model determination under G-Wishart prior distributions. *Electron. J. Stat.* **6** 168–198. MR2879676

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester. MR1112133

WILLIAMS, N. (2009). Education, gender, and migration in the context of social change. *Soc. Sci. Res.* **38** 883–896.

WILLIAMS, N. E., THOMAS, T. A., DUNBAR, M., EAGLE, N. and DOBRA, A. (2015). Measures of human mobility using mobile phone records enhanced with GIS data. *PLoS ONE* **10** 1–16.

WOLF, J., OLIVEIRA, M. and THOMPSON, M. (2003). Impact of underreporting on mileage and travel time estimates: Results from global positioning system-enhanced household travel survey. *Transp. Res. Rec.* **1854** 189–198.

DEPARTMENT OF STATISTICS,
DEPARTMENT OF BIOBEHAVIORAL NURSING AND
HEALTH INFORMATICS
AND CENTER FOR STATISTICS AND
THE SOCIAL SCIENCES
UNIVERSITY OF WASHINGTON
BOX 354322
SEATTLE, WASHINGTON 98195
USA
E-MAIL: adobra@uw.edu

DEPARTMENT OF OPERATION MANAGEMENT
FACULTY OF ECONOMICS AND BUSINESS
UNIVERSITY OF AMSTERDAM
AMSTERDAM
THE NETHERLANDS
E-MAIL: a.mohammadi@uva.nl