PLOS ONE

RESEARCH ARTICLE

Analysis of Twitter data with the Bayesian fused graphical lasso

Mehran Aflakparast[1], **Mathisca de Gunst**[1], **Wessel van Wieringen**[1,2]*

1 Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, **2** Department of Epidemiology and Data Science, Amsterdam Public Health Research Institute, Amsterdam UMC, location VUmc, Amsterdam, The Netherlands

* w.n.van.wieringen@vu.nl

Abstract

We propose a method to simplify textual Twitter data into understandable networks of terms that can signify important events and their possible changes over time. The method allows for common characteristics of the networks across time periods and each period can comprise multiple unknown sub-networks. The networks are described by Gaussian graphical models and their parameter values are estimated through a Bayesian approach with a *fused lasso*-type prior on the precision matrices of the underlying mixtures of the sub-models. A flexible data allocation scheme is at the heart of an MCMC algorithm to recover mean and covariance parameters of the mixture components. Several implementations of the outlined estimation procedure are studied and compared based on simulated data. The procedure with the highest predictive power is used for mining tweets regarding the 2009 Iranian presidential election.

Citation: Aflakparast M, de Gunst M, van Wieringen W (2020) Analysis of Twitter data with the Bayesian fused graphical lasso. PLoS ONE 15(7): e0235596. https://doi.org/10.1371/journal.pone.0235596

Editor: Lei Shi, Yunnan University of Finance and Economics, CHINA

Received: September 6, 2019

Accepted: June 19, 2020

Published: July 27, 2020

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This research was supported by NWO-STAR grant 613.009.014 from the Netherlands Organization for Scientific Research.

Competing interests: The authors have declared that no competing interests exist.

1 Introduction

Twitter is a prominent social media tool that provides a rich resource of information. The huge volume of gathered information calls for powerful methods to translate large and complex data into small chunks of understandable signals that can be used in several areas ranging from social sciences and health research to marketing and e-commerce. As an example, studied later in detail, more than one million tweets related to the social upheaval surrounding the 2009 Iranian presidential election may be compressed into an accessible visual summary. Such summary information can entail different topics that are highlighted in a certain period of time and have evolved over time. This can be viewed as a form of network reconstruction where collections of linked words, concepts or terms represent highlighted topics at a certain time-stamp. Any changes in such topics over time, from becoming outdated, expanded or created, can be explained by evolution of the links between the words.

Our interest is in the reconstruction of networks of words/terms from *multiple* Twitter data sets corresponding to specific time periods, where different networks share topological similarities. Next to this, data from a particular time period may be heterogeneous in that they cannot be represented by a single network. This phenomenon might be explained by differences among (parts of) the networks across different time periods, or can originate from hidden sub-

networks within each time period. Naturally, to end up with interpretable networks we aim to reconstruct networks in which only a few terms or predictors play an important role. This means that we search for *sparse* networks, networks with relatively few links. Therefore we *a)* propose a framework to simultaneously reconstruct multiple sparse networks with a possibly shared structure and *b)* extend this idea to the case where there is more than one network for a given time period.

The problem of network reconstruction is operationalized here as the estimation of a number of Gaussian graphical models (GGMs) for which the nonzero elements of the precision matrices (inverse covariance matrices) correspond to edges of the network. Estimation of a single GGM, especially in a high-dimensional setting where the data dimension is larger than the sample size, often proceeds in a regularized fashion (see, for example, [1–5]). These methods typically minimize the log-likelihood of the data augmented with an ℓ_1-penalty on the elements of the precision matrix. For instance, Meinshausen and Bühlmann [1] proposed to identify the edges of a GGM by sparse estimation of the precision matrix through lasso regressions of each random variable on all other variables. Friedman et al. [4] presented the *graphical lasso*, a procedure to estimate sparsely the precision matrix directly. In [2, 3, 5] different estimation algorithms for the graphical lasso method are treated. An alternative approach is provided in [6], where it is illustrated that in cases where the true graphical model does not need to be extremely sparse in terms of containing many zero elements, ridge penalties coupled with post-hoc selection may outperform the lasso. A common challenge with these approaches is that they do not take into consideration the uncertainty of the parameter estimates and require selection of the penalty parameters that control sparsity. Wang [7] proposed a Bayesian counterpart to the graphical lasso that provides a solution for such shortcomings, and developed a fast algorithm to estimate a moderately large precision matrix. However, this method does not serve our purpose in the face of multiple networks with possibly evolving structures.

There is a number of methods that consider simultaneous estimation of multiple graphical models corresponding to more than one data set (see for instance [8–12]). Guo et al. [8] extended the graphical lasso with a certain parametrization of multiple precision matrices whose elements are expressed as a product of shared and class-specific factors. Through a hierarchical penalty on both the shared and the class-specific factors their method shrinks some elements in the inverse covariance matrices to zero. Danaher et al. [9] proposed a general framework with arbitrary type of penalty and derived fused lasso and group lasso estimators, where the fused lasso estimation encourages shared structure and/or equal values for the elements across the precision matrices, while the group lasso estimation emphasizes only a shared sparse structure. More recently, a fused ridge version of multiple graphical model estimation has been proposed [11]. As another example, Zhu et al. [10] adapted the truncated ℓ_1-penalization of [13] to stimulate elements of the precision matrices across data sets to be similar. Bayesian counterparts include [12, 14]. These methods give proper consideration to common characteristics of the data sets while simultaneously estimating them. However, they lack the flexibility to account for heterogeneity within each data set. In [15] the problem of learning the evolution of an interaction network, modeled as a GGM, from cross-sectional, high-dimensional data in the face of heterogeneity was addressed through fused ridge penalized estimation of a combination of mixtures of GGMs. Here we consider this problem from a Bayesian perspective.

In this paper we present a novel Bayesian approach to the joint estimation of multiple graphical models, that takes into account both shared topological structures between the networks, and heterogeneity within the networks. In particular, we propose a Bayesian Gaussian fused graphical lasso estimation algorithm to estimate group-wise precision matrices that may exhibit network similarities, and augment this with a mixture model to account for

heterogeneity of the data within a network. This is done in the spirit of the data integrative Bayesian inference method that we proposed in [16], by forming a new prior distribution on the elements of the precision matrices and obtaining a posterior distribution that resembles the ℓ_1-penalized likelihood plus a fused penalty. A data allocation scheme is employed to simultaneously uncover the hidden clustering components of the mixture model while estimation of the cluster-specific precision matrices is achieved through column-wise block Gibbs sampling. In the application that we consider the different networks originate from multiple time periods, however, for the method the ordering over time is irrelevant. This makes our method widely applicable.

The paper is organized as follows. In Section 2.1 we propose a Bayesian Gaussian graphical network reconstruction method for data from multiple time periods or multiple groups. This is extended in Section 2.2 to allow for heterogeneity in the sense that each time period may encompass data from more than one (unknown) sub-population. In Section 3.1, these approaches are evaluated and compared by simulation. Section 3.2 illustrates the application of the proposed method in analyzing tweets regarding the 2009 Iranian presidential election. We conclude in Section 4 with discussing future improvements.

2 Materials and methods

Throughout the paper we will use capital letters to denote random variables, random vectors or random matrices; bold type will be used for vectors and matrices. The symbol $\propto$ stands for "is proportional to". To emphasize that for the proposed method the ordering of the time periods is irrelevant, throughout this section we will use the word *group* instead of time period, and the unknown sub-populations belonging to one time period, will be called *subgroups*, *clusters*, or *components*.

2.1 Bayesian fused graphical lasso

Characteristics from a sample of n individuals comprising T groups have been observed. For $t = 1, \ldots, T$, the number of individuals in group t will be denoted by n_t, and for $i = 1, \ldots, n_1, n_1 + 1, \ldots, n_1 + n_2, n_1 + n_2 + 1, \ldots, \sum_{t=1}^{T} n_t = n$, the random vector $\mathbf{Y}_i$ represents the p-dimensional vector of characteristics of individual i. In the sequel we will write $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_n^T)$ for the complete $n \times p$ data matrix. For later use we also define $n_{<t} = \sum_{s=1}^{t-1} n_s$ and $n_{<1} = 0$. The grouping of individuals is exhaustive and exclusive in the sense that an individual appears in a single group only.

The random vectors of characteristics are assumed to be independent and to follow a group-wise Gaussian law,

$$\mathbf{Y}_i \sim \mathcal{N}_p(\boldsymbol{\mu}_t, \boldsymbol{\Omega}_t^{-1}), \;\; i = n_{<t} + 1, \ldots, n_{<t} + n_t, \;\; t = 1, \ldots, T.$$

We consider the joint estimation the groups' precision matrices $\Omega = \{\Omega_1, .., \Omega_T\}$. For the purpose of interpretability sparse estimates are sought for, while the context suggests that the structure of the precision matrices may be shared between groups. In a frequentist setting all requirements (high-dimensionality, sparsity and a possibly common structure of the precision matrices) are catered for by the fused graphical lasso estimator [9], which maximizes the following penalized joint log-likelihood

$$\sum_{t=1}^{T} \log(|\boldsymbol{\Omega}_t|) - \mathrm{tr}(\mathbf{S}_t \boldsymbol{\Omega}_t) - \sum_{t=1}^{T} \lambda_t \parallel \boldsymbol{\Omega}_t \parallel_1 - \sum_{t_1 < t_2} \lambda_{t_1, t_2} \parallel \boldsymbol{\Omega}_{t_1} - \boldsymbol{\Omega}_{t_2} \parallel_1, \tag{1}$$

with respect to $\boldsymbol{\Omega}$. In (1), $\mathbf{S}_t = \mathbf{S}_t(\mathbf{Y})$ denotes the sample covariance matrix of group t. Note that

the estimator above generalizes the one originally proposed in [9] which uses $\lambda_{t_1,t_2} = \lambda_f$ for all t_1 and t_2 (except $t_1 \neq t_2$). The last two summands of the penalized log-likelihood (1) comprise the fused graphical lasso penalty. The convexity of the penalty tackles the high-dimensionality, while its first summand (i.e., the lasso penalty) induces sparsity, and, finally, its second summand (i.e., the fused penalty) shrinks the precision matrices towards a common structure for large values of the penalty parameter λ_{t_1,t_2}.

Here we present a Bayesian interpretation of the fused graphical lasso. This requires the evaluation of the joint posterior distribution $p(\boldsymbol{\Omega}|\mathbf{Y}, \boldsymbol{\Lambda})$ of the $\boldsymbol{\Omega}_\mathbf{t}$. To this end we denote by $\{\boldsymbol{\Omega}\}_{-t}$ the set of precision matrices with $\boldsymbol{\Omega}_t$ excluded, and define for $t = 2, \ldots, T$, the prior distribution of each precision matrix given the others as

$$\begin{aligned} p(\boldsymbol{\Omega}_t \,|\, \{\boldsymbol{\Omega}\}_{-t},\, \boldsymbol{\Lambda}) \;\propto \prod_{j_1<j_2} \frac{\lambda_t}{2} \exp\left(-\lambda_t |\omega^t_{j_1 j_2}|\right) \prod_{j=1}^{p} \frac{\lambda_t}{2} \exp\left(-\frac{\lambda_t}{2}\,\omega^t_{jj}\right) \\ \times \prod_{t' \neq t} \prod_{j_1<j_2} \frac{\lambda_{t',t}}{2} \exp\left(-\lambda_{t',t}\,|\omega^{t'}_{j_1 j_2} - \omega^t_{j_1 j_2}|\right) I(\boldsymbol{\Omega}_t \succ 0). \end{aligned} \tag{2}$$

In the above ω^t_{ij} denotes the i, j-th element of $\boldsymbol{\Omega}_t$. Note that the (2) is invariant to the order of conditioning. Furthermore, the diagonal and off-diagonal elements of the precision matrices are thus assumed to follow *a priori* an exponential and a double exponential distribution, respectively, (see for example [17] and [7] for similar approaches). The differences between corresponding precision elements of any pair of groups also obey a double exponential law. The term $I(\boldsymbol{\Omega}_t \succ 0)$ limits the support of the prior to the positive definite matrices. With the fused graphical lasso prior (2), the posterior distribution $\boldsymbol{\Omega}_\mathbf{t}$ is not a well-known standard distribution, but an efficient Gibbs sampling scheme can be designed. This extends the work of [7] for the Bayesian graphical lasso. In a nutshell, the Gibbs sampler amounts to iteratively sampling one column of $\boldsymbol{\Omega}_t$ at the time which guarantees positive definiteness.

As a first step towards our Gibbs sampler we derive a tractable formulation of the conditional posterior of each precision matrix given the others. Application of the definition of conditional probability and subsequent insertion of the equality $p(\boldsymbol{\Omega}|\boldsymbol{\Lambda}) = p(\boldsymbol{\Omega}_t|\{\boldsymbol{\Omega}\}_{-t}, \boldsymbol{\Lambda})p(\{\boldsymbol{\Omega}\}_{-t}|\boldsymbol{\Lambda})$ yields

$$p(\boldsymbol{\Omega}_t|\, \mathbf{Y}_t, \{\boldsymbol{\Omega}\}_{-t}, \boldsymbol{\Lambda}) \;\;\propto\;\; p(\mathbf{Y}_t \,|\, \boldsymbol{\Omega}_t)\, p(\boldsymbol{\Omega}_t \,|\, \{\boldsymbol{\Omega}\}_{-t}, \boldsymbol{\Lambda}).$$

An analytic expression is now readily available from the normality assumption of the data together with the prior (2). Gibbs sampling, however, is still hampered by the double exponential distributions employed in the prior of the precision elements. This is circumvented by a hierarchical representation of these distributions by a scale mixture of normal distributions [18],

$$\frac{\lambda}{2} \exp(-\lambda\,|\omega|) \;\;=\;\; \int_0^\infty \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{\omega^2}{2\tau}\right) \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2\tau}{2}\right) d\tau. \tag{3}$$

Define, corresponding to each double exponential distribution in the prior, a latent scale parameter $\tau^t_{j_1 j_2}$ in accordance with the scale mixture representation (3) above. Furthermore, let $\boldsymbol{\tau}_t = \{\tau^t_{j_1 j_2}\}_{j_1<j_2}$ denote the independent latent scale parameters corresponding to group t, and let $\boldsymbol{\tau} = \{\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_T\}$.

Finally, we endow λ_t and $\lambda_{t',t}$ with independent gamma priors with shape parameter s and rate parameter r. Conditioning on the latent scale mixture parameters, and rewriting the

hierarchical model, we obtain for $i = n_{<t} + 1, \ldots, n_{<t} + n_t$, $t = 1, \ldots, T$,

$$\begin{cases} \mathbf{Y}_i \,|\, \boldsymbol{\mu}_t,\, \boldsymbol{\Omega}_t & \sim & \mathcal{N}_p(\boldsymbol{\mu}_t, \boldsymbol{\Omega}_t^{-1}) \\ p(\boldsymbol{\Omega}_t |\, \{\boldsymbol{\Omega}\}_{-t}, \boldsymbol{\Lambda},\, \tau) & = & \prod_{j=1}^{p} \frac{1}{2}\lambda_t \exp(-\frac{1}{2}\lambda_t \omega_{jj}^t) \prod_{j_1<j_2} \phi_{0,\tau^g_{j_1j_2}}(\omega^t_{j_1j_2}) \\ & & \prod_{t'\neq t} \prod_{j_1<j_2} \phi_{\omega^{t'}_{j_1j_2},\tau^{t'}_{j_1j_2}}(\omega^t_{j_1j_2})\, I(\boldsymbol{\Omega}_t \succ 0) \\ p(\tau_t |\, \boldsymbol{\Lambda}) & = & \prod_{j_1<j_2} \frac{1}{2}\lambda_t^2 \exp(-\frac{1}{2}\lambda_t^2 \tau^t_{j_1j_2}) \\ p(\tau_{t'} |\, \boldsymbol{\Lambda}) & = & \prod_{j_1<j_2} \frac{1}{2}\lambda^2_{t',t} \exp(-\frac{1}{2}\lambda^2_{t',t} \tau^{t'}_{j_1j_2}),\ t' \neq t, \\ \lambda_t & \sim & \mathcal{G}(r,s), \\ \lambda_{t',t} & \sim & \mathcal{G}(r,s) \text{ for } t' \neq t, \end{cases} \tag{4}$$

where $\phi_{a,b}$ stands for the density function of the normal distribution with mean $a = 1$ and variance $b = 1$.

The mean parameters $\boldsymbol{\mu}_t$ are assumed to have independent priors (see Section 2.2.2). Integrating out the scale mixture parameters in the hierarchical model above will yield the double exponential distributions.

To arrive at an efficient posterior sampling procedure, the precision matrix posterior needs to be broken down further. For this we let $Y^t = \{Y_i\}_{n_{<t}+1,\ldots,n_{<t}+n_t}$ denote the set of all Y_i belonging to group t, $t = 1, \ldots, T$. After putting all components of the hierarchical model (4) together, we find that the posterior of the precision matrix of the t-th group satisfies,

$$\begin{aligned} p(\boldsymbol{\Omega}_t |\, \mathbf{Y}^t, \{\boldsymbol{\Omega}\}_{-t}, \boldsymbol{\Lambda}, \boldsymbol{\tau}) \ \propto |\boldsymbol{\Omega}_t|^{n_t/2} \exp\Big[-\tfrac{1}{2}\operatorname{tr}(\mathbf{S}_t\boldsymbol{\Omega}_t)\Big] \prod_{j=1}^{p} \frac{1}{2}\lambda_t \exp(-\frac{1}{2}\lambda_t\omega^t_{jj})\, I(\boldsymbol{\Omega}_t \succ 0) \\ \times \exp\left\{-\frac{1}{2}\sum_{j_1<j_2}[(\omega^t_{j_1j_2})^2 (\mathbf{A}_t)_{j_1j_2} - 2\sum_{j_1<j_2} \omega^t_{j_1j_2} (\mathbf{B}_t)_{j_1j_2}]\right\}, \end{aligned} \tag{5}$$

in which $\mathbf{A}_t$ and $\mathbf{B}_t$ are zero-diagonal and symmetric matrices with off-diagonal entries

$$(\mathbf{A}_t)_{j_1j_2} = \frac{1}{\tau^t_{j_1j_2}} + \sum_{t'\neq t} \frac{1}{\tau^{t'}_{j_1j_2}} \quad \text{and} \quad (\mathbf{B}_t)_{j_1j_2} = \sum_{t'\neq t} \frac{\omega^{t'}_{j_1j_2}}{\tau^{t'}_{j_1j_2}}. \tag{6}$$

From this we derive the column-(and row-)wise posterior of the matrix $\boldsymbol{\Omega}_t$. Without loss of generality we illustrate this for the last column (row). Hereto denote the 2×2 block partition of a matrix $\mathbf{X}$ by

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{11} & \mathbf{x}_{12} \\ \mathbf{x}_{12}^\top & x_{22} \end{pmatrix}, \tag{7}$$

with $\mathbf{X}_{11}$ and $\mathbf{x}_{12}$ a $(p-1) \times (p-1)$ and $(p-1) \times 1$ dimensional matrix respectively, while x_{22} a scalar. Applying this notation to the matrices involved in the posterior (5), using that by the Schur decomposition it holds that

$$|\boldsymbol{\Omega}_t| = |\boldsymbol{\Omega}^t_{11}||\omega^t_{22} - \boldsymbol{\omega}^t_{12}\top(\boldsymbol{\Omega}^t_{11})^{-1}\omega^t_{12}|,$$

and using the identity

$$\mathrm{tr}(\mathbf{S}_t\mathbf{\Omega}_t) = \mathrm{tr}(\mathbf{S}_{11}^t\mathbf{\Omega}_{11}^t) + 2(s_{12}^t\top\boldsymbol{\omega}_{12}^t) + t_{22}^t\omega_{22}^t,$$

we obtain that under the assumption that $\mathbf{\Omega}_{11}^t$ is (temporarily) known, the posterior for the last column (row) of the t-th precision matrix, $p(\omega_{12}^t, \omega_{22}^t \,|\, \mathbf{Y}^t,\ \mathbf{\Omega}_{11}^t,\ \{\mathbf{\Omega}\}_{-t},\ \mathbf{\Lambda},\ \boldsymbol{\tau})$, is proportional to

$$\begin{aligned}&[\omega_{22}^t - (\boldsymbol{\omega}_{12}^t)^\top(\mathbf{\Omega}_{11}^t)^{-1}\boldsymbol{\omega}_{12}^t]^{n_t/2} \\ &\quad\times \exp\{-\tfrac{1}{2}[\boldsymbol{\omega}_{12}^t\top\mathbf{D}_{\boldsymbol{a}_{12}^t}\omega_{12}^t + 2(s_{12}^{t\top} - \boldsymbol{b}_{12}^{t\top})\omega_{12}^t + (s_{22}^t + \lambda_t)\omega_{22}^t]\},\end{aligned}$$

where $\mathbf{D}_{\boldsymbol{a}_{12}^t}$ is the diagonal matrix with $\mathbf{a}_{12}^t$ on its diagonal. When followed by the change-of-variables

$$\begin{aligned}\gamma_t &= \omega_{22}^t - (\boldsymbol{\omega}_{12}^t)^\top(\mathbf{\Omega}_{11}^t)^{-1}\omega_{12}^t, \\ \boldsymbol{\delta}_t &= \omega_{12}^t,\end{aligned} \tag{8}$$

the conditional joint distribution of δ_t and γ_t, it can easily be seen that

$$\begin{aligned}\gamma_t \,|\, \mathbf{Y}^t,\ \mathbf{\Omega}_{11}^t,\ \mathbf{\Lambda},\ \boldsymbol{\tau} &\sim \mathcal{G}\big[\tfrac{1}{2}n_t + 1,\ \tfrac{1}{2}(s_{22}^t + \lambda_t)\big], \\ \delta_t \,|\, \mathbf{Y}^t,\ \mathbf{\Omega}_{11}^t,\ \{\mathbf{\Omega}\}_{-t},\ \mathbf{\Lambda},\ \boldsymbol{\tau} &\sim \mathcal{N}_{p-1}(-\mathbf{\Sigma}_{\delta_t}[(s_{12}^t)^\top - (\boldsymbol{b}_{12}^t)^\top], \mathbf{\Sigma}_{\delta_t}),\end{aligned} \tag{9}$$

where $\mathbf{\Sigma}_{\delta_t} = [\mathbf{D}_{a_{12}^t} + (s_{22}^t + \lambda_t)(\mathbf{\Omega}_{11}^t)^{-1}]^{-1}$. Note that the positive definiteness of $\mathbf{\Omega}_t$ is guaranteed due to that of $\mathbf{\Omega}_{11}^t$ and the fact that $\gamma_t = \omega_{22}^t - \boldsymbol{\omega}_{12}^{t\prime}(\mathbf{\Omega}_{11}^t)^{-1}\boldsymbol{\omega}_{12}^t > 0$ (cf. [2]).

Next we turn to the scale mixture parameters $\boldsymbol{\tau}_t$ corresponding to the priors on the elements of the t-th group precision matrix. Gathering terms involving the τ_{j_1,j_2}^ts we find that, conditionally on $\mathbf{\Omega}_t$ and $\mathit{\Lambda}$, they follow an inverse Gaussian distribution for all $j_1 < j_2$:

$$\begin{aligned}1/\tau_{j_1j_2}^t \,|\, \mathbf{\Omega}_t, \mathbf{\Lambda} &\sim \textit{inv-Gauss}\{\lambda_t^2(\omega_{j_1j_2}^t)^{-2}]^{1/2}, \lambda_t^2\}, \\ 1/\tau_{j_1j_2}^{t'} \,|\, \mathbf{\Omega}_t,\ \mathbf{\Omega}_{t'},\ \mathbf{\Lambda} &\sim \textit{inv-Gauss}\{[\lambda_{t',t}^2(\omega_{j_1j_2}^t - \omega_{j_1j_2}^{t'})^{-2}]^{1/2}, \lambda_{t',t}^2\},\ t' \neq t,\end{aligned} \tag{10}$$

in which the inverse Gaussian distribution parametrization of [19] is used.

In a similar fashion the posterior conditional distributions of the λ_t and $\lambda_{t',t}$ can be derived. Based on their gamma priors given in the hierarchical model (4), their full conditional distributions are gamma distributions as well:

$$\begin{aligned}\lambda_t \,|\, \mathbf{\Omega}_t &\sim \mathcal{G}(r + \tfrac{1}{2}p(p+1),\ s + \sum_{j_1\leq j_2}|\omega_{j_1j_2}^t|), \\ \lambda_{t',t} \,|\, \mathbf{\Omega}_{t'}, \mathbf{\Omega}_t &\sim \mathcal{G}(r + \tfrac{1}{2}p(p+1),\ s + \sum_{j_1\leq j_2}|\omega_{j_1j_2}^{t'} - \omega_{j_1j_2}^t|),\ t' \neq t.\end{aligned} \tag{11}$$

2.1.1 Sampling from the posteriors. The posterior densities derived above facilitate sampling from the joint posterior of the precision matrices. This is achieved as described in Box 1.

The algorithm then re-iterates. After a burn-in period, when the samples are seen to be representative for the desired posterior, point estimates for the parameters are obtained from these samples through appropriate summary statistics.

Notice that the above presented Bayesian fused graphical lasso estimation procedure shrinks the elements of the precision matrices towards zero but does not actually set them to

Box 1: Bayesian fused graphical lasso algorithm

- Initialise $\mathbf{\Lambda}$, $\boldsymbol{\tau}$ and $\mathbf{\Omega}$ from their priors (4).
- For $t = 1, .., T$ do:
  - Calculate $\mathbf{A}_t$ and $\mathbf{B}_t$ from (6) using the current values of $\boldsymbol{\tau}$ and $\mathbf{\Omega}$.
  - For column (row) $i = 1, .., p$,
    1. Block partition $\mathbf{A}_t$, $\mathbf{B}_t$, $\mathbf{\Omega}_t$ and $\mathbf{S}_t$.
    2. Sample δ_t and γ_t from their posteriors (9).
    3. Update the corresponding column and row of $\mathbf{\Omega}_t$ using the change of variables (8).
- Sample $\boldsymbol{\tau}$ from the posteriors (10).
- Sample the tuning parameters $\mathbf{\Lambda}$ from the posteriors (11).

zero. Sparsity is achieved by post-hoc estimation via selection based on (quantile-based) Bayesian credible intervals.

2.2 Mixture models for multi-group data

The Bayesian fused graphical lasso algorithm presented above can be used to jointly recover graphical networks for multiple data sets. In this subsection we extend the method to the case where the data not only come from multiple groups, but also within each group there may exist multiple sub-populations.

The data stored in the $n \times p$ matrix $\mathbf{Y}$, stem from T independent *known* groups as before. Additionally, it is assumed that within each group the sample originates from a heterogeneous population. The population of group t, $t = 1, \ldots, T$, comprises K_t independent *unknown* subgroups. Let, Z_i denote the latent random variable that indicates the i-th individual's subgroup membership, $i = n_{<t} + 1, \ldots, n_{<t} + n_t$, $t = 1, \ldots, T$. In other words, for individual i belonging to group t we would have $Z_i = k_t$ if this individual would be a member of subgroup k_t. With the subgroup information unavailable, the random variable $\mathbf{Y}_i$ is assumed to follow the mixture model

$$\mathbf{Y}_i \sim \sum_{k_t=1}^{K_t} \pi_{t,k_t} \mathcal{N}_p(\boldsymbol{\mu}_{t,k_t}, \mathbf{\Omega}_{t,k_t}^{-1}), \qquad i = n_{<t} + 1, \ldots, n_{<t} + n_t, t = 1, \ldots, T \tag{12}$$

with $\pi_{t,k_t} = p(Z_i = k_t)$ being the probability that individual i belonging to group t is a member of subgroup k_t. Hence, these mixing proportions π_{t,k_t} sum to one group-wise: $\sum_{k_t=1}^{K_t} \pi_{t,k_t} = 1$. Moreover, given the component memberships Z_i, the data from each mixture component, corresponding to the subgroup×group-combinations, follow a multivariate Gaussian distribution:

$$\mathbf{Y}_i \,|\, Z_i = k_t \sim \mathcal{N}_p(\boldsymbol{\mu}_{t,k_t}, \mathbf{\Omega}_{t,k_t}^{-1}).$$

Since data within and across groups are independent, the likelihood $\mathcal{L}$ for this situation is given by

$$\mathcal{L}(\mathbf{Y}|\mathbf{Z},\Xi) \quad = \quad \prod_{t=1}^{T}\prod_{i=n_{<t}+1}^{n_{<t}+n_t}\sum_{k_t=1}^{K_t}\pi_{t,k_t}\phi_{\boldsymbol{\mu}_{t,k_t},\boldsymbol{\Omega}_{t,k_t}^{-1}}(\mathbf{Y}_i). \tag{13}$$

Here $\mathbf{Z} = (Z_1, \ldots, Z_n)$, $\Xi = \{\pi_{t,k_t}, \boldsymbol{\mu}_{t,k_t}, \boldsymbol{\Omega}_{t,k_t}^{-1} : i = n_{<t} + 1, \ldots, n_{<t} + n_t, t = 1, \ldots, T\}$, the set of all model parameters, and $\phi_{a,B}$ denotes the density function of the multivariate normal distribution with mean vector **a** and covariance matrix **B**.

Estimation of the mixture model (13) is carried out by first clustering the data points within each group. This is achieved by adopting the Bayesian clustering scheme that assigns informative priors on the component memberships as proposed in [16], and briefly described in Section 2.2.1. Next, the mixture parameters are estimated component-wise. The estimation procedure for the component-wise estimation of the mean parameters is described in Section 2.2.2. The precision matrices are estimated in either of the following two ways:

(a). Bayesian stage-wise (BS), this is separately in a group-wise manner by Gibbs sampling for a single Bayesian graphical lasso as in [7], if there is no reason to assume that the data across groups have a common structure;

(b). jointly with the Bayesian fused (BF) estimation procedure described in Section 2.1 above, if the data across groups are likely to have a common structure.

The corresponding mixture models are called Bayesian stage-wise mixture model (BSM) and Bayesian fused mixture model (BFM), respectively.

2.2.1 Estimation of component memberships. Data clustering is an important step in estimation of mixture models. To improve the estimation procedure, one may consider making use of available additional information such as cluster information or similarity measurements from other data sources. The proposed data allocation strategy, Data Integrative Chinese Restaurant Process (DI-CRP), is a generalization of the Chinese restaurant process (CRP) with additional flexibility that facilitates incorporating external sample-level information in mixture modeling with an unknown number of components [16]. Like CRP, DI-CRP is a data allocation strategy that explores the conditional distribution of the component membership of one data point given that of the rest of data points. This allows the number of mixture components to be determined adaptively. Let $\mathcal{S}^t = (s^t_{ii'})_{i\le i'=1}^{n_t}$ represent the additional information on similarity of data points in group t.

As the data allocation is independently carried out for every group, we will now drop the group index t in the notation and denote the index of the k_t-th mixture component simply by k. We assume the following conditional probabilities for the component membership variables:

$$P(Z_i = k|\boldsymbol{Z}_{-i},\alpha,\mathcal{S}) \propto \begin{cases} n^*_{-i,k}h_i(k) & \text{if } k \text{ is an existing component} \\ \\ \alpha & \text{if } k \text{ is a new component} \end{cases} \tag{14}$$

where $h_i(k)$ is a function that indicates the overall similarity of the data point i with all other data points in component k. This function can appear in different forms. Here we use the

simple form

$$h_i(k) = 1 + \sum_{i \neq i'} s_{ii'} I_{\{Z_{i'}=k\}},$$

and

$$n^*_{-i,k} = \sum_{i \neq i'} I_{\{s_{ii'} \geq T_i\}} I_{\{Z_{i'}=k\}}$$

with $I_{\{sii'} \geq T_i\}$ as the factor that indicates when a data point is considered to be similar to the rest of data points in a certain component. For example, in our application T_i is assumed to be the third quantile of the similarity values between data point i and the rest of the group-specific data points. The reason to introduce $n^*_{-i,k}$ is to diminish the influence of a minority of data points in a cluster that have possibly very large similarity values with a new data point. In other words, this criterion is to direct the clustering in such a way that a new data point becomes more likely to end up being clustered in a component of which the majority of the data points shares high similarities with the new data point.

Multiplying the likelihood (12) and the prior (14) we see that the posterior of the latent variables satisfies

$$P(Z_i = k | \mathbf{Y}_i, Z_{-i}, \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k, \alpha, \mathcal{S}) \propto \begin{cases} n^*_{-i,k} h(c_i, k)\, p(\mathbf{Y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k), & \text{if } k \text{ is an already existing component,} \\ \alpha \int p(\mathbf{Y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k) p(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k) d\boldsymbol{\mu}_k d\boldsymbol{\Omega}_k, & \text{if } k \text{ is a new component.} \end{cases} \quad (15)$$

As the integral in (15) is not analytically tractable, it can be approximated by Monte Carlo samples as described in [16].

The number of mixture components K is largely controlled by the choice of α, in that larger values lead to more components. Following [20], we assume a gamma prior with mean a and rate parameter b for the concentration parameter α. The full conditional distribution can be derived given the number of components K—which is implied by the fact that $\mathbf{Z}$ is given—following the hierarchy

$$\begin{aligned} \alpha | \zeta, K &\sim \rho_\zeta\, \mathcal{G}(a + K, b - log(\zeta)) + (1 - \rho_\zeta)\, \mathcal{G}(\alpha + K - 1, b - log(\zeta)), \\ \zeta | \alpha &\sim \mathcal{B}(\alpha + 1, n), \end{aligned} \quad (16)$$

where $\frac{\rho_\zeta}{1 - \rho_\zeta} = \frac{a + K - 1}{n(b - log(\zeta))}$.

The data allocation probabilities above form the basis for building a clustering algorithm that functions through sampling from posterior probabilities of component memberships.

2.2.2 Estimation of component means. Once the data in all groups are clustered, i.e. $\mathbf{Z}$ is known, the component-specific parameters are to be updated. In contrast to the previous sections, here the component-specific mean parameters are assumed unknown with a conditional prior distribution (again dropping group indices)

$$\boldsymbol{\mu}_k | \boldsymbol{\Omega}_k \sim \mathcal{N}_p(\boldsymbol{\mu}_0, \kappa_0^{-1} \boldsymbol{\Omega}_k), \quad k = 1, \ldots, K, \quad (17)$$

with $\boldsymbol{\mu}_k$ and Ω_k the k-th component mixture mean and precision parameters. In this prior distribution $\boldsymbol{\mu}_0$ and κ_0 are hyperparameters of vector and scalar type, respectively, and they are the same for all subgroups over all groups. The prior distribution (17) is conjugate and,

combined with the data likelihood, yields the posterior

$$\boldsymbol{\mu}_k \,|\, \mathbf{Y}, \boldsymbol{\Omega}_k \sim \mathcal{N}_p(\mathbf{m}_k, (n_k + \kappa_0)^{-1}\boldsymbol{\Omega}_k), \quad k = 1, \ldots, K, \tag{18}$$

with n_k the number of data points assigned to mixture component k and

$$\mathbf{m}_k = \frac{n_k}{n_k + \kappa_0}\bar{\mathbf{Y}}_k + \frac{\kappa_0}{n_k + \kappa_0}\boldsymbol{\mu}_0, \tag{19}$$

where $\bar{\mathbf{Y}}_k$ is the p-dimensional mean of the data vectors assigned to the k-th group. We note that we do not enforce sparsity on the component mean parameters. With a different prior setting, this could also be established.

2.2.3 Algorithm. The algorithm for computing component memberships is given in Box 2. Note that for simplicity the component indices are denoted by k or k', instead of t, k_t or t, k'_t.

To initialize clustering, we fix the maximum number of mixture components to K_{max} ($\leq n$). The first component is created by sampling parameter values from their prior distributions. Next, a data point having the largest normal density value among all data points is assigned to this component. The second component is created in the same way as the first one, but, the second data point can be assigned to either the first component or to the second one based on the maximum value of the generated density values. This continues until all data points are assigned to a finite number K ($\leq K_{max}$) of components.

One sweep of the algorithm updates the number of components, the component memberships and the component-wise parameters. At each iteration of the Metropolis algorithm the updated component memberships are the basis of the clustering, followed by inference of the component-specific mean and precision parameters. We use the posterior means to estimate the component specific parameters. As the estimate for the number of components K, we use the final number of components after the algorithm has converged.

Notice that in general the MCMC algorithm starts with multiple (known) groups and explores the structure of data within each group in order to further cluster them into smaller sub-groups (i.e. steps 1–2). Also notice the difference between step 1 and step 2 of the algorithm. While step 1 of the algorithm controls the birth of a new component or death of an existing component, step 2 attempts to update the clustering by exchanging data points.

3 Results

3.1 Simulation

The performance of the proposed methods was assessed via a simulation analysis with three objectives: i) evaluation of the performances of the two approaches BS and BF in recovering graphical networks corresponding to multiple data sets, ii) comparison of the performance of BSM and BFM in the proposed mixture context, and iii) assessment of the accuracy of the cluster assignment scheme DICRP and its comparison to that of the original CRP in the present context. Two simulation studies were conducted as described below.

The hyper-parameter values were assigned mainly based on previous studies and partly based on independent simulations. Firstly, the rate parameter s controlling the tuning parameters has to be sufficiently larger than zero to avoid computational issues, therefore it was set to unity (see [17] for a substantiation of this choice). We took the shape parameter $r = 0.001$ based on a simulation study. Fig 1 illustrates the impact of the shape parameter on the empirical posterior density of zero and non-zero elements of the precision matrix. Secondly, hyperparameters of the concentration parameter α were set to $a = b = 1$ as recommended in [21] and [22]. Lastly, the hyperparameters corresponding to the mixture means $\boldsymbol{\mu}_0$, κ_0 were set to a

Box 2: MCMC algorithm for Bayesian groupwise mixture (BSM) and Bayesian fused mixture (BFM) methods

1. Update number of components:

for $i = n_{<t-1} + 1, \ldots, n_{<t-1} + n_t$, given a current clustering $Z_i = k$, if i is not a singleton data point, create a new component k' by sampling from the prior distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$, and update $Z_i = k'$ with probability

$$min\left\{1, \frac{\alpha}{(n_t - 1)} \times \frac{\phi_{\boldsymbol{\mu}_{k'}, \boldsymbol{\Omega}_{k'}^{-1}}(\mathbf{Y}_i)}{\phi_{\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k^{-1}}(\mathbf{Y}_i)}\right\},$$

and if i is the only data point in component k (singleton), propose k' among already existing components with a probability proportional to $n^{\star}_{-i,k'}$, and update $z_i = k'$ with probability

$$min\left\{1, \frac{n_t - 1}{\alpha} \times \frac{h_i(k')\phi_{\boldsymbol{\mu}_{k'}, \boldsymbol{\Omega}_{k'}^{-1}}(\mathbf{Y}_i)}{h_i(k)\phi_{\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k^{-1}}(\mathbf{Y}_i)}\right\}.$$

2. Update component memberships:

for $i = n_{<t-1} + 1, \ldots, n_{<t-1} + n_t$, if data point i belongs to a component with more than one occupant, update its component membership with probability equal to

$$\frac{n^{\star}_{-i,k} h_i(k)\, \phi_{\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k^{-1}}(\mathbf{Y}_i)}{\sum_{k=1}^{K} n^{\star}_{-i,k} h_i(k)\, \phi_{\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k^{-1}}(\mathbf{Y}_i),},$$

otherwise do nothing.

3. Update mixture parameters:

3.1. update mixture means from posterior (18) and (19).

3.2. based on the application choose either BMS or BFM and

BSM: update precision matrices corresponding to the current group by Gibbs sampling for a single Bayesian graphical lasso as in [7].

BFM: repeat steps 1–2 and 3.1 for all groups, then jointly update all precision matrices from all sub-groups by the Gibbs sampling procedure in Box 1 of Section 2.1.1.

4. Iteration:

repeat steps 1–4 until convergence.

The software package that implements the algorithm and illustrative examples are publicly available from https://github.com/mehranaflak/IMLR_TextGGN

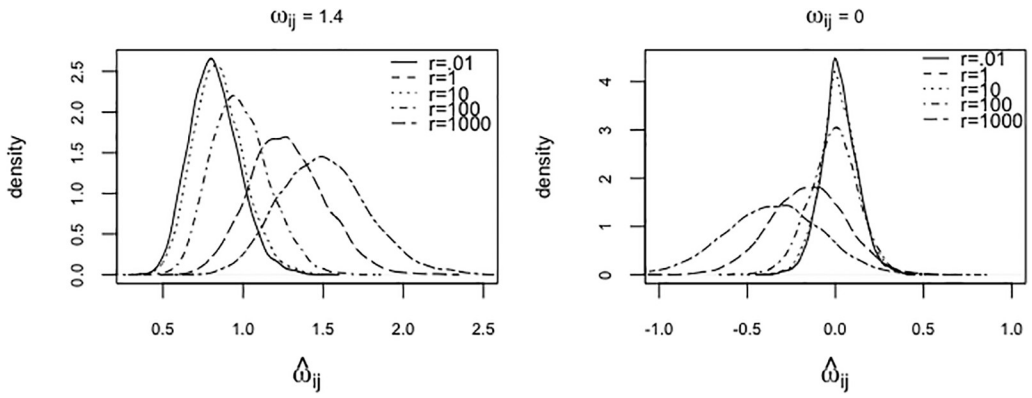

Fig 1. Element-wise empirical posterior distribution ofprecision matrices. Empirical posterior density by varying tuning parameter for zero (left) and non-zero (right) element of Ω for $p = 10$.

https://doi.org/10.1371/journal.pone.0235596.g001

zero vector of dimension p and a scalar equal to unity, respectively. The reason for assigning a less informative prior to the mixture means is that our main objective of this study which is to focus on the estimation of the precision matrices.

The MCMC algorithm used 1000 Gibbs iterations and a burn-in period of 100 iterations. These numbers of iteration were motivated from preliminary simulation studies (not shown) in which we noticed that the MCMC algorithm for a similar problem appeared to have reached convergence for significantly fewer (≈50) iterations. In our applications, after a burn-in period of 100 iterations the samples were seen to be representative for the desired posterior, see Fig 2 which shows one example of posterior sample traces from zero and non-zero elements of the ground-truth inverse covariance parameters as an illustration that there was a good mixing around the true parameter values with no particular pattern.

Point estimates of the parameters were obtained by posterior mean calculation over the iterations. Sparsification was carried out through construction of the 95% two-sided Bayesian credible intervals. Then, the presence of an edge was inferred when the corresponding credible interval did not contain zero. The performance of the estimation of the model parameters was measured by Frobenius loss, for both the mean vector and the iprecision matrices. The ability to reconstruct the underling conditional independence graph was evaluated by means of the F-score (i.e. the harmonic mean of precision and sensitivity).

Several settings were fixed for both simulation studies. In the first simulation study the BS and BF methods were compared with respect to their ability to estimate various (six) precision matrices. These matrices either had conditional independence graphs with similar structures as depicted in Fig 3, or were randomly generated positive definite matrices without any imposed similarity constraints. Both cases defined six p-variate Gaussian graphical models. All models had zero mean vectors. For each case 100 data sets of size n were generated from each of the six p-variate Gaussian models. This amounted to 600 simulated data vectors in each case. This was repeated for several combinations of sample sizes and dimensions, $n \in \{50, 100, 500\}$ and $p \in \{5, 10, 20, 30\}$. The simulated data were analyzed by both BF and BS methods in order to derive estimates of the precision matrices.

The performance of BS and BF was measured based on average squared estimation errors over the 100 simulated data vectors and over all 6 precision matrices. Figs 4 and 5 present boxplots of the estimation errors for the different combinations of sample sizes and dimensions. From Fig 4 it is evident that the BF procedure yields smaller Frobenius error than the BS procedure for precision matrices with similar conditional independence graphs. For the randomly

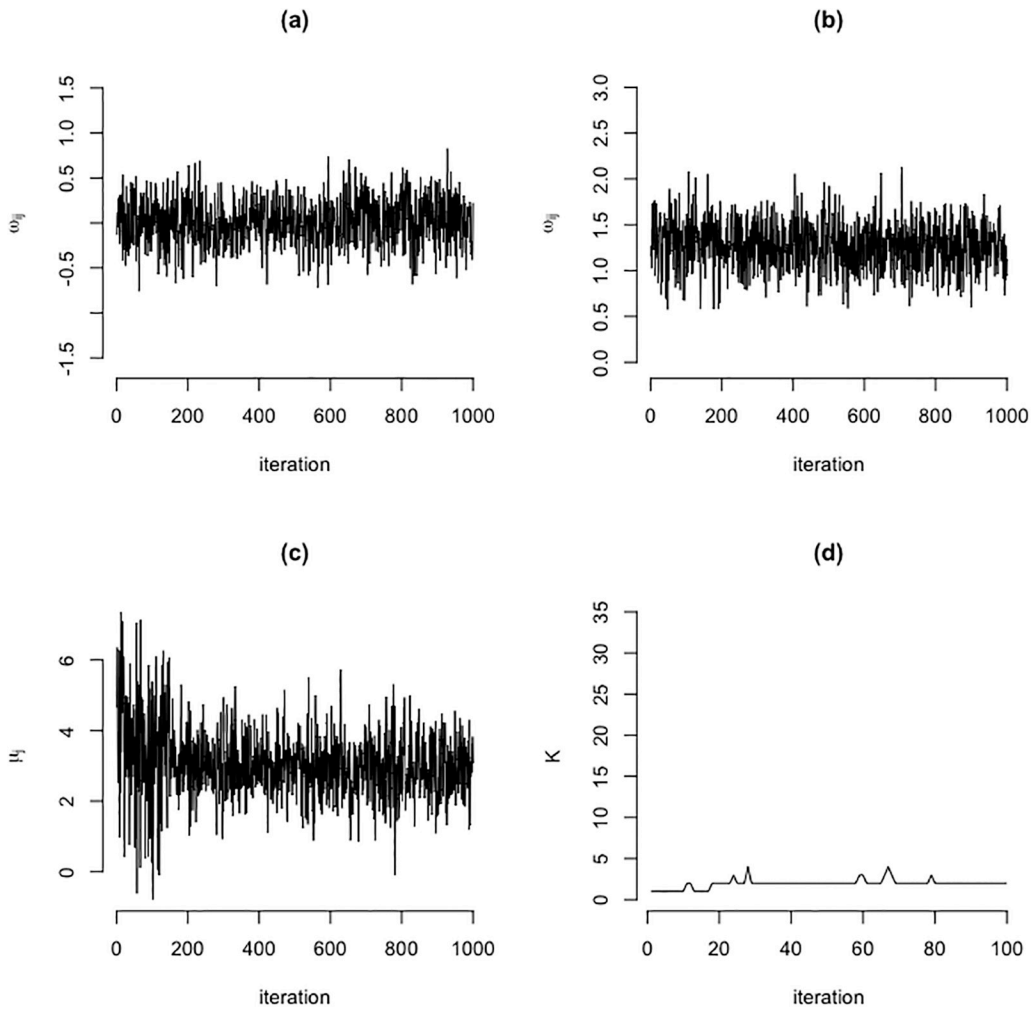

Fig 2. Mixing of MCMC samples. Trace of Gibbs samples of zero(a) and non-zero (b) element of precision matrix, mixture mean (c), and the number of components (d).

https://doi.org/10.1371/journal.pone.0235596.g002

generated networks without similarity structure, the two methods produce similar results, see Fig 5.

In the second simulation study the performance of the BSM and BFM procedures for the estimation of a mixture of Gaussian graphical models, as given by (12), were assessed. The hyper-parameters regarding the precision matrices were selected as above, and we set $\boldsymbol{\mu}_0 = 0$ and $\kappa_0 = 1$ for the sake of simplicity. The simulation study considered three consecutive time periods. For each period data were generated from (12) such that the data exhibited more heterogeneity in subsequent periods: the data in the first period were homogeneous, whereas data of the second and third period stemmed from two and three sub-populations, respectively. The conditional independence graphs associated with the precision matrices were topologically related to mimic a simple evolution as in Fig 3. The data were generated according to the parameter settings in the Supporting Information with varying sample sizes $n \in \{100, 200, 300, 400, 500, 1000\}$ and dimensions $p \in \{10, 20, 30, 40\}$ with 50 independent data sets for each (n, p) combination. We employed BSM and BFM methods on all data sets to fit mixture models.

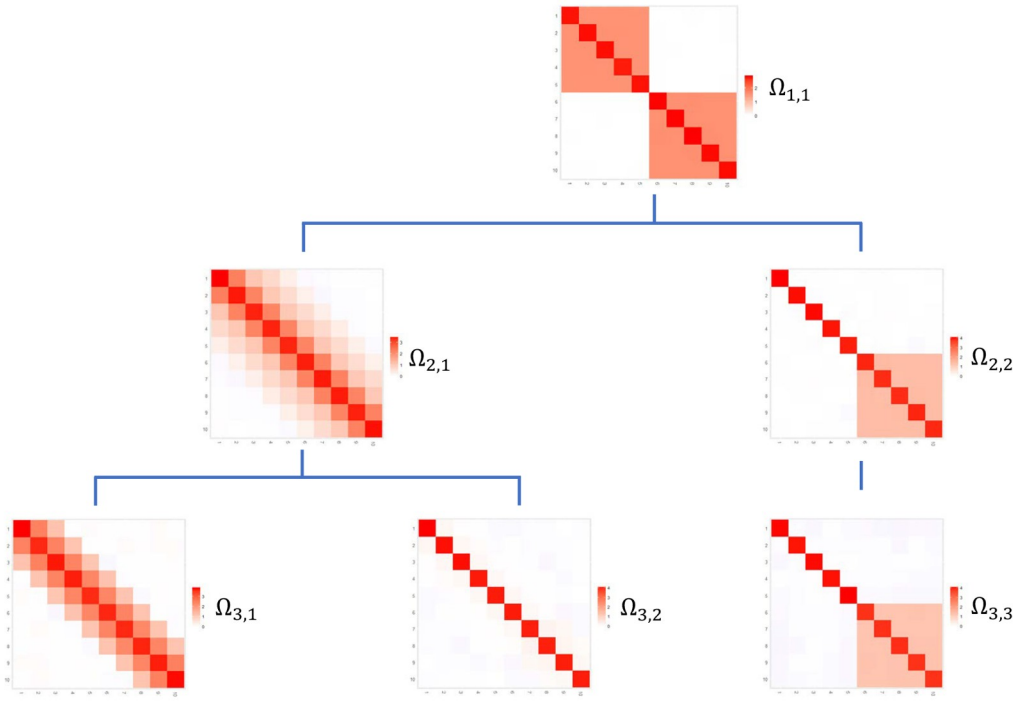

Fig 3. Structure of precision matrices. Graphical network structure for a dataset with 3 stages that was used for data generation.

https://doi.org/10.1371/journal.pone.0235596.g003

The performance of the BSM and BFM methods for the estimation of the mixture means and precision matrices, as well as the quality of the edge presence/absence classification as obtained from this simulation study is summarized in Figs 6 and 7. Each point in these plots represents an average taken over 50 independent results. From these Figs we conclude that a larger sample size generally tends to increase the accuracy of the estimation and to alleviate estimation errors. Larger data dimensions, on the other hand, have a reverse influence on accuracy of the results. With respect to precision matrix estimation Fig 7 suggests a slightly better performance for the BFM method over the BSM one, especially for higher dimensions. But for mixture mean estimation or graph recovery there is no strong evidence of superiority for any of the two methods, see Figs 6 and 8.

The influence of the proposed component membership priors, as specified in (15), on the cluster assignment was assessed. The clustering approaches were compared for BSM only, as the clustering in both the BSM and BFM procedures is carried out for each period separately. The data comprising six subgroups were drawn from the mixture model (12) with the settings as specified in the Supporting Information, assuming various sample sizes and dimensions: $n \in \{50, 100, 300\}$ and $p \in \{5, 10, 16, 20\}$. For each (n, p) combination 50 independent data sets were generated. For the assessment two different scenarios were considered: *a)* no additional information on the samples is available: BSM-CRP, and *b)* external evidence on the samples' similarity is available: BSM-DICRP. In scenario *a)* the priors (15) were equivalent to those of CRP as $\mathcal{S} = 0$. In scenario *b)* priors (15) were used with a non-zero similarity matrix. For the data in each stage, the similarity matrix $\mathcal{S}$ was generated based on the true clustering of the data points, where $s_{ij} = 1$ if data point i lies in the same cluster as data point j, and $s_{ij} = 0$ otherwise. These methods were applied to all of the simulated data sets to measure the performance of the clustering scheme when additional clustering information is available.

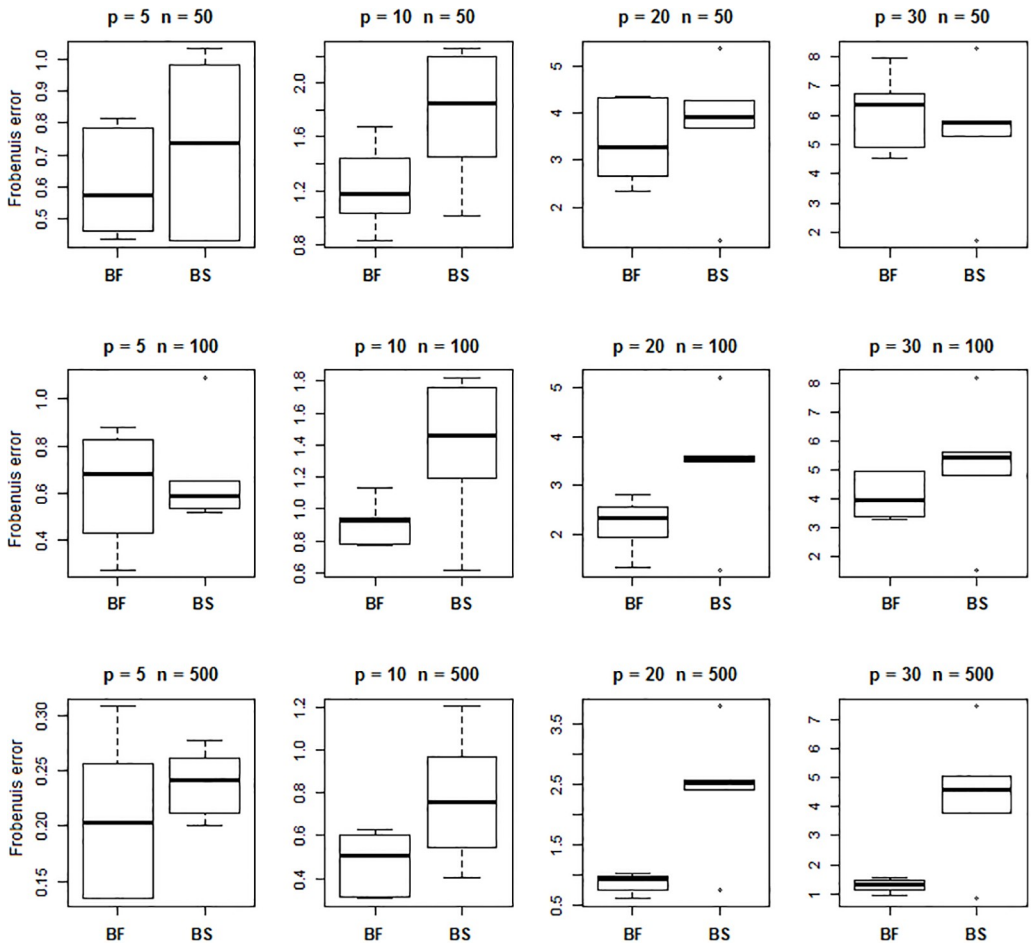

Fig 4. Prediction loss of precision matrices. Frobenuis errors for estimation of precision matrices with a defined *shared* structure by BF and BS methods.

https://doi.org/10.1371/journal.pone.0235596.g004

Next, the clustering performance is measured by the proportion of correctly identified ground-truth mixture clusters. The results are summarized in Fig 9 where CRP or DICRP are compared. For example, the top row of this figure shows that in 100% of the simulations the estimated number of components is equal to the ground-truth number of components (= 1). Generally, these results show that, when additional clustering information is available, BSM combined with DICRP yields a better (clustering) performance than BSM combined with the CRP prior on the component memberships.

Finally, we measured the computational time of the algorithm in order to illustrate the scalability of our approach. The results are shown only for the BFM approach using the abovementioned simulation settings for varying data dimension and sample size. In summary, as depicted in Fig 10, despite using an efficient block sampling technique the scalabilty becomes an issue for higher dimensions, while this is not the case for increasing number of samples.

3.2 Analysis of Twitter data

In this section, we illustrate an example on how to summarize Twitter data into networks of terms or 'words', using the BSM and BSF methods. The inferred conditional independence

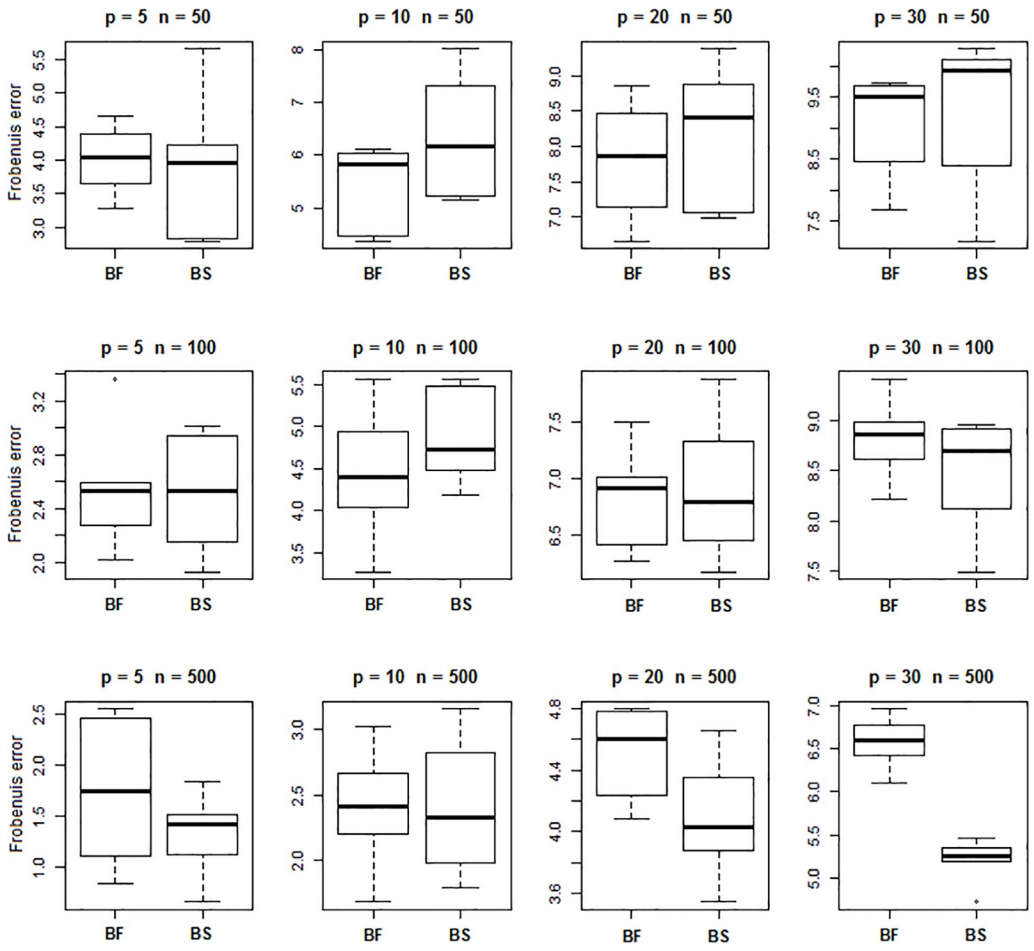

Fig 5. Prediction loss of precision parameter estimation. Frobenuis errors for estimation of precision matrices with *random* structure by BF and BS methods.

https://doi.org/10.1371/journal.pone.0235596.g005

graphs are then studied to signify topics and their evolution through time. To this end, we analyzed tweets regarding the Iranian 2009 presidential election.

3.2.1 Context. The two main candidates of the Iranian presidential election of June 2009 were Mir-Hossein Mousavi of the reformist "Green Movement" and president Mahmoud Ahmadi-Nejad, running for a new term. The latter won the election, but the result was disputed by alleged voting irregularities and fraud. These allegations gave rise to mass protests by supportors of Mir-Hossein Mousavi's "Green Movement" against the president-elect.

A time line of events crucial to the 2009 Iranian presidential election is summarized below for three consecutive periods. Words in italic are most frequently used in the tweets of the three time periods.

Period I

2009-05-12 Election day.

2009-05-15 First mass rally to *protest* against the election results with demonstrators chanting words like '*Allaho Akbar*'.

2009-05-19 Speech by *Khamenei*, the supreme leader, on *Friday*'s (weekend day in Iran) prayer.

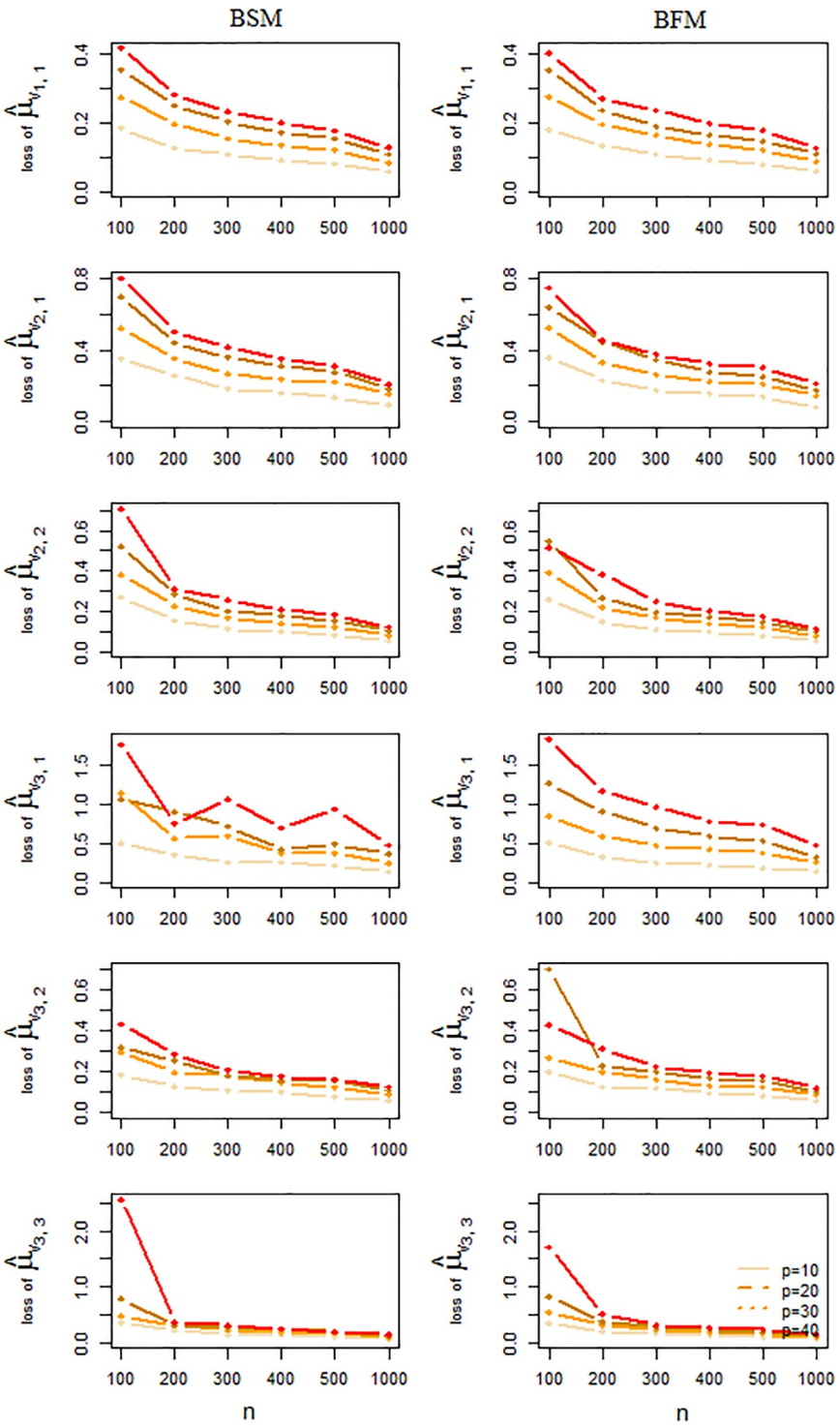

Fig 6. Prediction loss of component-wise mean parameterestimation. Frobenuis errors of mixture mean estimation averaged over 50 independent simulated datasets by BSM and BFM methods.

https://doi.org/10.1371/journal.pone.0235596.g006

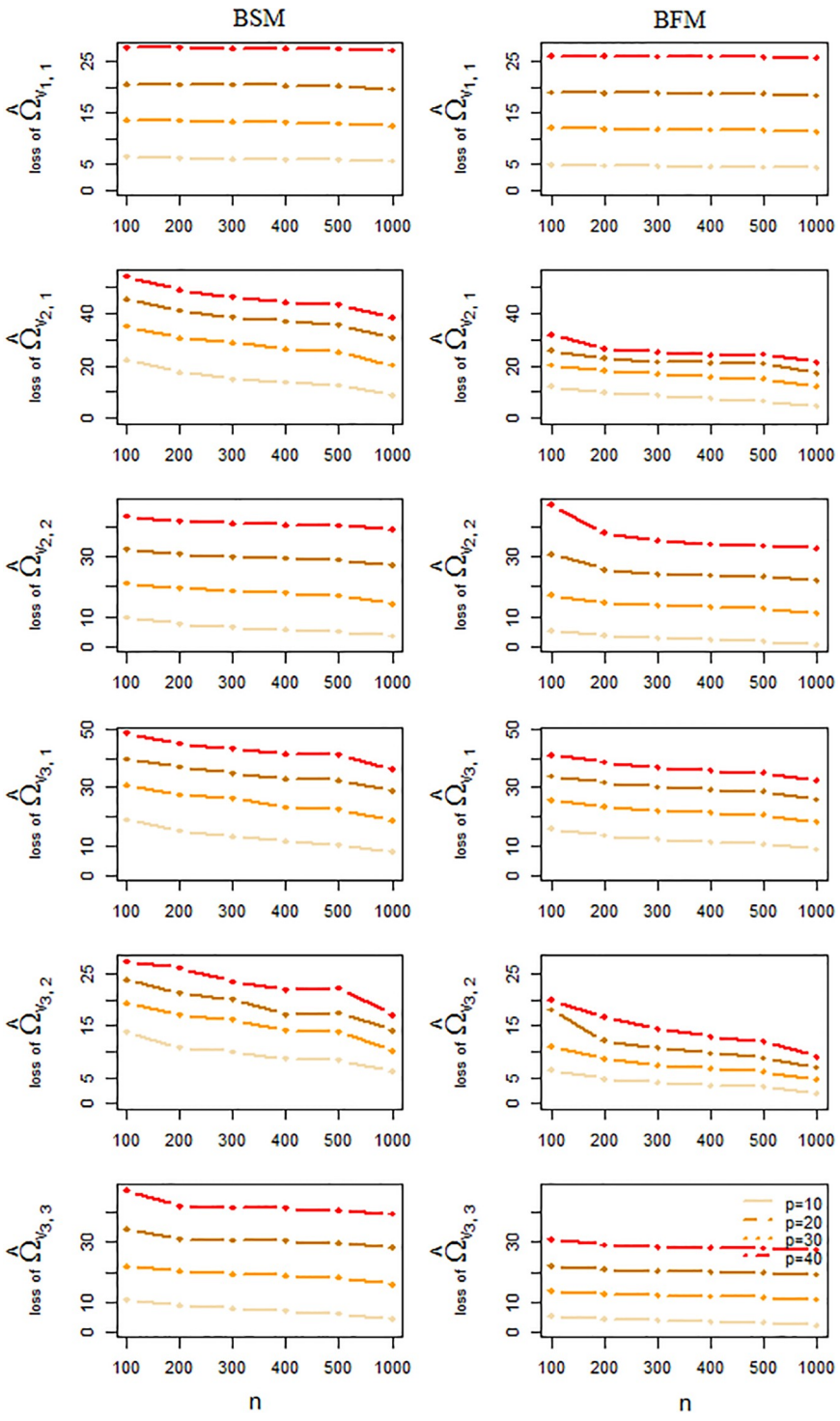

Fig 7. Prediction loss of component-wise precision parameter estimation. Frobenuis errors of precision parameter estimation averaged over 50 independent simulated datasets by BSM and BFM methods.

https://doi.org/10.1371/journal.pone.0235596.g007

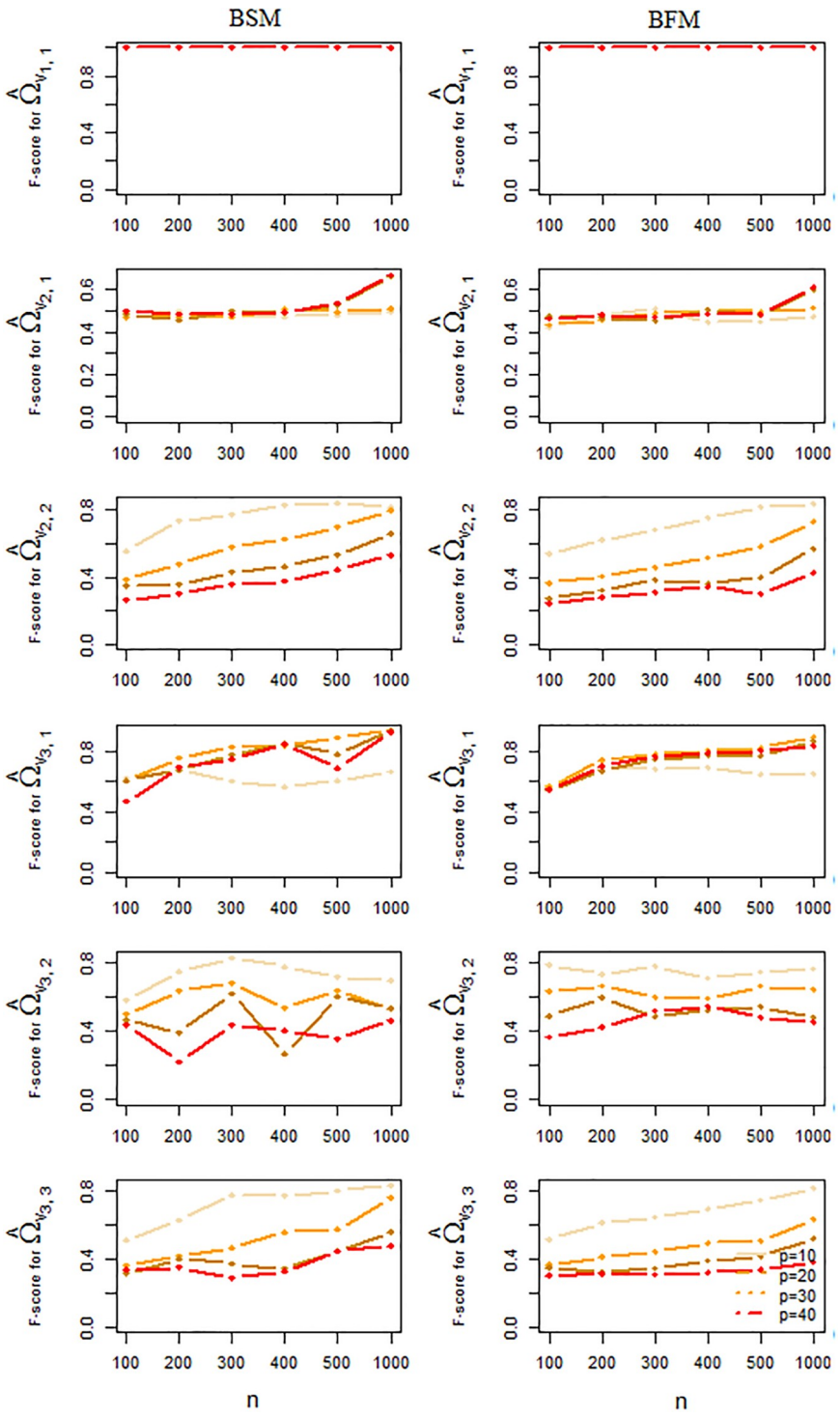

Fig 8. Classification accuracy. F-scores corresponding to estimation of sparsified precision matrices averaged over 50 independent simulated datasets by BSM and BFM.

https://doi.org/10.1371/journal.pone.0235596.g008

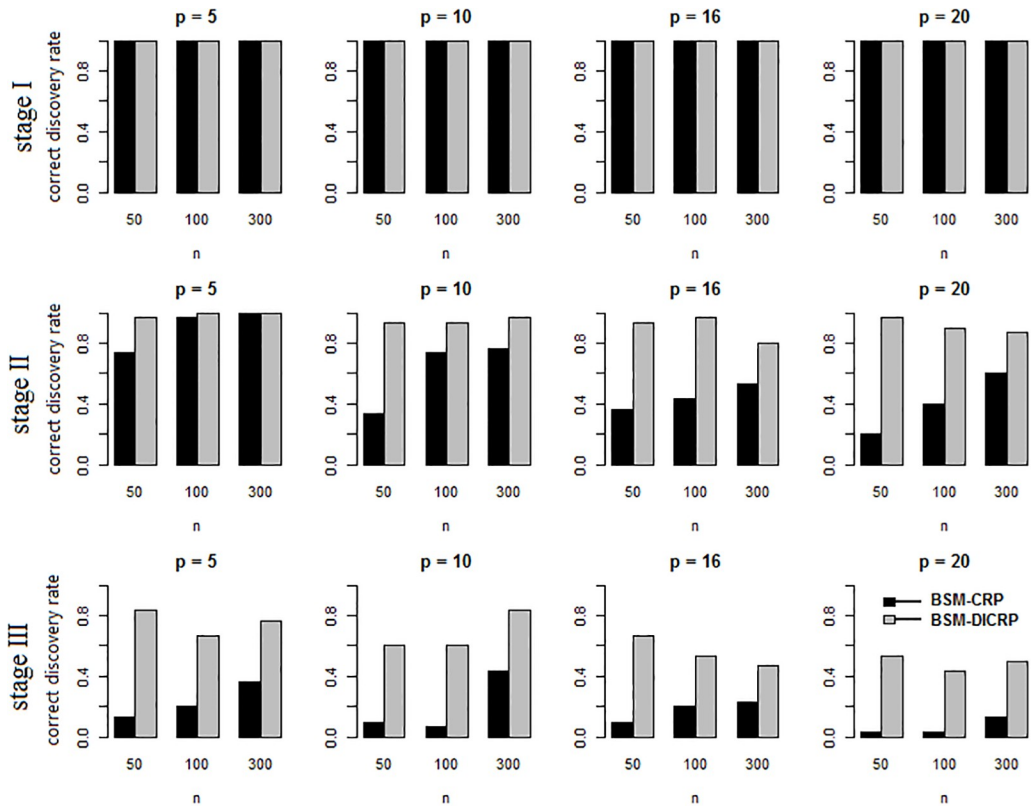

Fig 9. Mixture clustering accuracy. Estimation of number of components in a 3-stage simulated dataset with number of components $K_1 = 1$, $K_2 = 2$ and $K_3 = 3$.

https://doi.org/10.1371/journal.pone.0235596.g009

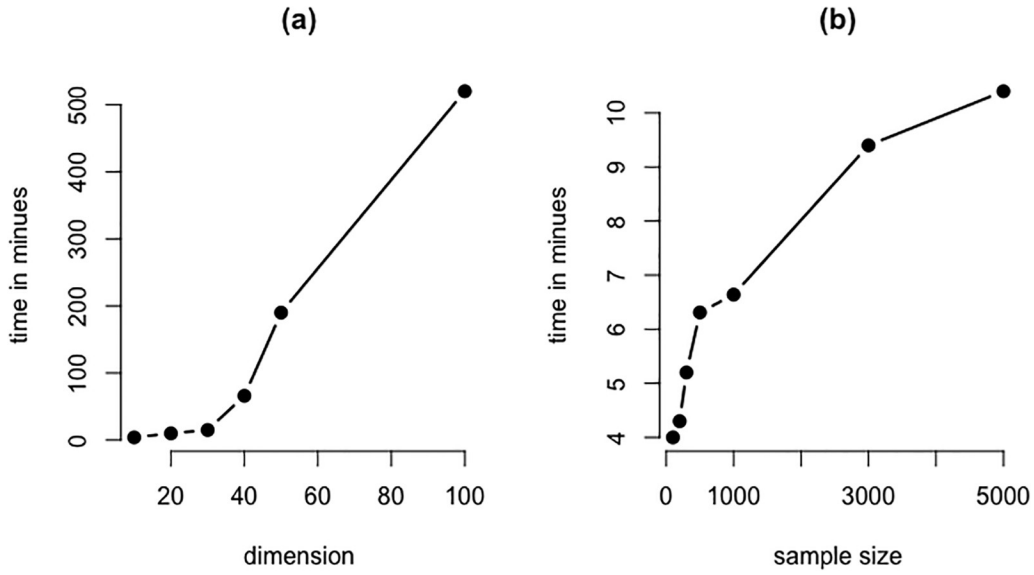

Fig 10. Computational time. Computational time to estimate mixture graphical networks as a function of (a) the data dimension with a fixed sample size of 100 and (b) the sample size with a fixed dimension of 10.

https://doi.org/10.1371/journal.pone.0235596.g010

2009-05-20 Neda (a *Mousavi*-supporter was *shot* and many more supporters were imprisoned at the *Evin prison*. *human rights* groups and the president of the United States issued statements urging the end to violence against protesters.

2009-06-09 Demonstrations on the tenth anniversary of the 18-Tir uprising. This refers to the Iranian Student Protests of July 1999 (also known as 18th of Tir (7–13 July) and Kuye Daneshgah Disaster in Iran) were the most widespread and violent public protests to occur in Iran since the early years of the Iranian Revolution. During the 18-Tir uprising students asked for greater freedom of speech, among others.

2009-06-17 Friday's prayer sermon by *Akbar Rafsanjani* attended by $\geq$ two million people, including the leaders of the *"Green Movement"*. Among the people that were *arrested* and taken to the *Evin prison* was *Shadi Sadr*, an Iranian women rights activist who featured prominently in the news.

2009-08-01 Start of the trial against the people arrested in the protests. These trials were condemned by the protesters associated with the *"Green Movement"*.

2009-09-18 On Friday Quds day, an annual event held on the last Friday of the Ramadan that was initiated by the Islamic Republic of Iran in 1979, the second wave of major protests in Tehran and other cities took place.

Period II

2009-11-04 Students' day, which refers to the day that the USA embassy was conquered in Iran (November 04, 1379), saw another large demonstration with people chanting among others slogans like "*Allaho Akbar*", "A green Iran doesn't need *nuclear weapons*", and "*death* to dictator", among others. On the same day the reformist candidate *Mousavi* was grounded and could no longer leave house.

2009-12-07 Iranian scholar day, which is the anniversary of the murder of three students of University of Tehran on December 7, 1953 by Iranian police., show thousands of students protesting against the government and demanding a regime change. In a speach *Akbar Rafsanjani* criticized the strict measures taken by the Iranian government.

2009-12-19 Ayattollah Ali *Montazeri*, an influential figure of the *"Green Movement"*, died.

2009-12-21 A rally took place on the occasion of the funeral ceremony for Ayatollah *Montazeri*. Simultaneously, reports were published claiming *torture* and rape of prisoners associated with the *"Green Movement"*. The clashes between supporters of the '*"Green Movement"* and the police continued for a few days in the cities of Isfahan and Qom.

2009-12-27 At the day of "Ashura" another big demonstration took place during which many people were shot and killed, including Seyed Ali *Mousavi*, the nephew of Mir-Hossen Mousavi. Ashura is the tenth day of Muharram in the Islamic calendar, and commemorates the death of Husayn ibn Ali, the grandson of Muhammad. For a majority of Shia's Muslim Ashura has become a ceremonial mourning day.

2009-12-28 Western countries condemned the violations of the protesters' *human rights* by the Iranian government.

2010-02-17 Signing of a petition by supporters of the *"Green Movement"* against a law that would limit *women's rights*.

Period III

2010-05-10 For fear of violence by the leaders of the *"Green Movement"*, the planned demonstrations on the first anniversary of the disputed election was cancelled. Police violence against women who allegedly were improperly clothed in public.

2010-05-11 Citizens were warned of the probable consequences of participation in the anniversary of the election.

2010-05-12 Minor demonstrations took place, but ended prematurely due to governmental interference.

3.2.2 Data. The data consist of tweets pertaining to this presidential election and the follow-up events that were previously collected and studied in [23]. This amounts to 1,532,289 tweets in total that were retrieved from one month before the elections until roughly 15 months after, to be precise: from 5 May 2009 until 8 August 2010. Duplicated and empty tweets, e.g. tweets containing only website links, were removed, resulting in 1,004,428 remaining tweets. We divided these 15 months into 3 periods of roughly 5 months each. Fig 11 presents the frequency of the tweets per day.

Next, the tweets in each period were divided into $n_t = 500$, $t = 1, 2, 3$ 'bags of words' or so called *documents*. These were subjected to standard text normalization techniques such as stemming and space/stopword/punctuation removal. We first identified the most frequent terms shared between the 3 periods and selected $p = 30$ co-occurring terms. Then, for each period a *term-document matrix* with rows and columns representing documents and (shared) terms, respectively, was created. Finally, the term-document matrices were mapped into continuous valued data matrix $\mathbf{Y}$ of the three periods $t = 1, 2, 3$ with, for $i = n_{<t} + 1, \ldots, n_{<t} + n_t$,

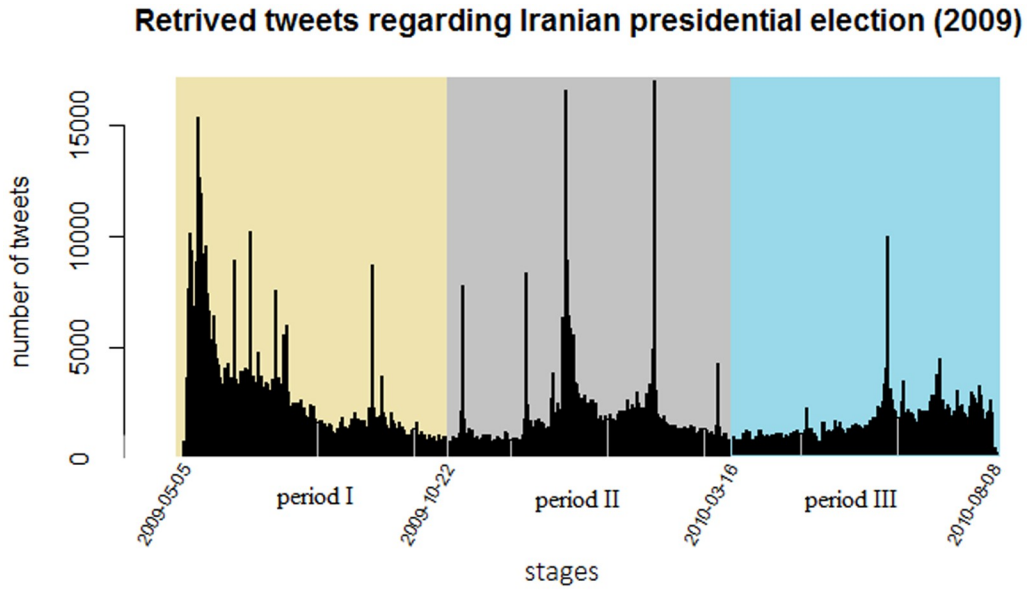

Fig 11. Tweets Frequency. Over time frequency of tweets regarding Iranian presidential election (2009). The whole period is divided into three equal length periods. High peaks majorly refer to important events of the time such as mass rallies and protests.

https://doi.org/10.1371/journal.pone.0235596.g011

and $t = 1, 2, 3$,

$$\mathbf{Y}_i = \{tf_i(w_1)\log[n_t/df(w_1)], \ldots, tf_i(w_p)\log[n_t/df(w_p)]\} \quad \text{for}$$

where $tf_i(w_j)$ represents the *term frequency* of w_j, this is the frequency with which the word w_j occurs in the i-th document, and $df(w_j)$ the *document frequency* of w_j, this is number of documents in which w_j appears.

We constructed the similarity matrix $\mathcal{S}^t = (s^t_{ii'})^{n_t}_{i\leq i'=1}$ based on the number of days between the posting dates of tweets. The impact of this measure on the clustering in mixture models is studied in an earlier paper [24] where several similarity measures are compared.

3.2.3 Results. The primary purpose of this analysis was to reconstruct networks that signify important topics in each period through linking terms or words. Different versions of the Bayesian fused graphical lasso estimation that were presented in the preceding sections, as well as the Bayesian lasso, were applied to the data. For clarity these are briefly recapped:

`non-mixture`: Data of each period are assumed to follow a multivariate normal distribution and their precision matrices are estimated by the Bayesian graphical lasso [7].

`BSM-CRP`: To account for heterogeneity, data from each period are assumed to follow the mixture model (12). The Bayesian stage-wise mixture (BSM) algorithm is used to estimate the model parameters. This algorithm uses a data allocation scheme that is equivalent to the CRP.

`BSM-DICRP`: Identical to the BSM-CRP approach but with DIRCP used for data allocation in which the number of days between documents serves as external information (see Section 2.2.1). The latter accommodates the possibility of documents contiguous in time to entail more similar information than documents well-separated in time.

`BFM-CRP`: As before a mixture model per period is assumed to be a good description of the data. In the estimation of these mixture models the precision matrices of the mixture components may now inherit or share structure within or between time periods. This is achieved by the Bayesian fused graphical (BFM) mixture approach (see Section 2.2.3). The CRP is used for data allocation in the mixture estimation.

`BFM-DICRP`: As `BFM-CRP`: but with the DICRP used for data allocation in which the number of days between days serves as external information.

These methods are compared for the Twitter data and compared with respect to their predictive power established through a 5-fold cross-validation method. The hyperparameter values are equal to those described in the simulation section. The results of this comparison study are summarized in Fig 12.

A first observation inferable from Fig 12 is the substantial difference between non-mixture and mixture estimation methods. The heterogeneity assumption thus seems to be a crucial one. Secondly, among the mixture estimation approaches those equipped with a DICRP data allocation scheme reveal a slight estimation improvement compared to the approach with the CRP scheme. This suggests that information of the number of days between tweets aids in the mixture component assignment of individual terms. Finally, the BFM approaches show a slight performance improvement, in terms of prediction accuracy, over their BSM counterparts. Final results of the Twitter data are thus based on the `BFM-DICRP` analysis, which are contrasted to those originating from the `non-mixture` approach. The corresponding networks are displayed in Fig 13. These are based on the partial correlation matrices obtained by

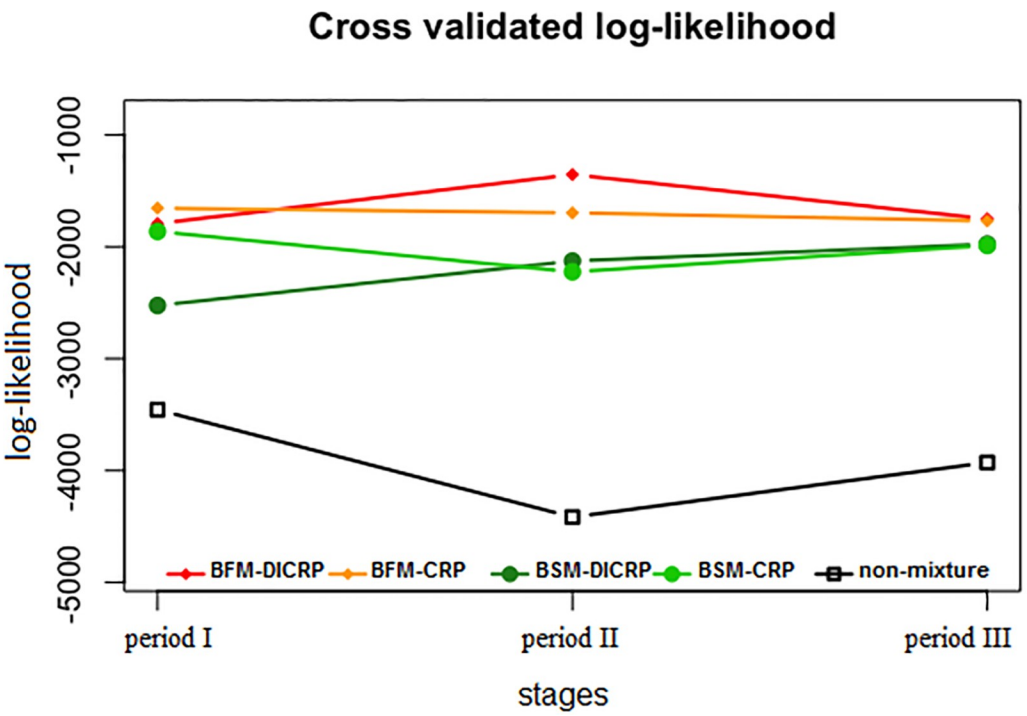

Fig 12. Performance of estimation approaches. Predictive log-likelihoods obtained by 5-fold cross-validation corresponding to three stage twitter data analysis with a `non-mixture` estimation, mixture estimations without additional similarity data (`BSM-CRP` and `BFM-CRP`), and mixture estimations taking into account external information on similarity of consequent tweets (`BSM-DICRP` and `BFM-DICRP`).

https://doi.org/10.1371/journal.pone.0235596.g012

standardization from the estimated precision matrices. The networks comprise 30 terms (nodes) linked by edges whose 95% Bayesian credible interval does not contain zero.

The resulting networks of Fig 13 reveal groups of terms or words that are more frequently linked within than between groups:

1. (*Colored green*.) The key words of speeches, open statements of political leaders, or demands of protesters such as 'human', 'right', 'change', 'support', 'green' (for "Green Movement"), 'democracy', 'tweet', 'women rights', 'freedom'.
2. (*Colored purple*.) The key words that can either represent strict measures taken by the government to suppress the protests such as 'torture', 'prison', 'Evin' (name of a prison in Tehran), 'protest', 'shot', 'forces', 'arrest', 'trial', or chanting words such as 'Allah', 'Akbar', 'death'.
3. (*Colored blue*.) The names of significant political leaders or influential groups/individuals such as 'Ahmadi-Nejad', 'Khamenei', 'Montazeri', 'Mousavi', 'Neda', 'Rafsanjani', 'Shadi', 'student', and 'women'.

This grouping is best recognized by mixture estimation, particularly for the first and second period. This may be due to the fact that most demonstrations and the subsequent reaction by the police took place in the first two time periods.

The reconstructed networks of words across tweets reflect the evolution of the Iranian political situation at the time. Several pairs of words, such as 'human–right', 'Evin–prison", 'Allah–Akbar', 'support–green (movement)' are connected in the networks of all three periods. Both words within each pair stem from the same semantic category, irrespective of the method

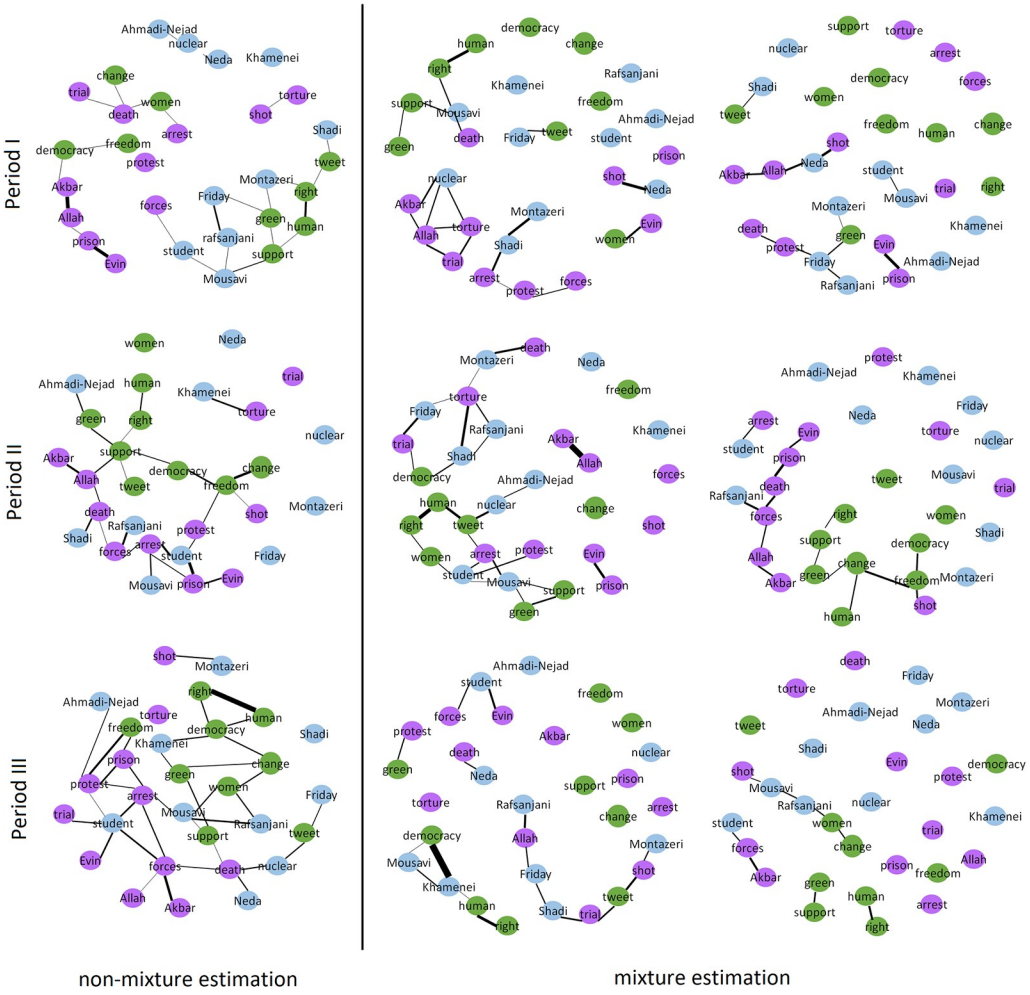

Fig 13. Reconstructed networks of terms. Reconstructed term networks from tweets regarding Iranian presidential elections 2009, using a `BSM-DICRP` and `non-mixture` estimation approaches.

https://doi.org/10.1371/journal.pone.0235596.g013

used. Such a semantic grouping is, however, more prominent in the results of period I and II from the mixture model. This is most likely due to the fact that these two periods saw most demonstrations, with possible reaction of the state. More specifically, period-specific meaningful, i.e., coinciding with the events listed in the above presented time line, links can be identified from the reconstructed networks. For example, the 'Neda'-'shot' and 'Shadi'-'arrest' link appears only in the first time period. Further, the 'Montazeri–death' link in the period II networks reflect the death of Ayatollah Montazeri during this period. Finally, the word 'student' takes a more central place in the period II networks due to two student events, related arrests and claimed torture of students that took place then.

The mixture modeling identifies two different groups of tweets for each period (see Fig 13). These groups may be interpreted by means of the cohesion among words and their semantics. For example, in period I the conditional independence graph of the first mixture component reveals word clusters that broadly combine topics of "public tumult" and "political demands", while that of the other mixture component appears to connect words only if they relate to actual events. As such the two groups may loosely represent tweets originating from twitter accounts aliased with the protest and media, respectively. For period II a different contrast

between the cohesion of words within tweets becomes apparent. The first mixture component is hard to interpret and may represent a mixed bag of tweets representing the general turmoil of the period. The second mixture component, however, nicely shows two distinct clusters each representing a different semantic category: tweets with a single uniform message filled either with the protester's demands or with an account of the negative events taking place.

4 Discussion

This paper transfers and extends the Bayesian graphical lasso for network reconstruction to the field of (chronological) textual social media data analysis. Twitter data from several time periods related to the 2009 Iranian presidential elections are used to show the potential of the approach. The data are studied from a graphical network estimation perspective and identifies the relation (and their variation over time) among topics. Statistically, the problem amounts to simultaneous estimation of precision matrices which is solved by the Bayesian graphical lasso, and which is extended here to *i)* account for heterogeneity in the data, *ii*) incorporate external information in the unravelling of this heterogeneity, and *iii)* borrow network similarities among identified groups. Extraction of summary information from one and a half million tweets related to the aforementioned election shows promise. Moreover, the flexibility of this Bayesian framework enables several approaches to address different assumptions on the data structure.

A possible useful inroad for future research might be to address high-dimensionality. The presented Twitter data analysis was limited to $p = 30$ terms. The proposed method is applicable to larger p, but an increase of the dimension may prohibit the interpretation of the reconstructed network. For comprehensive interpretation of large networks it might be crucial to develop a complementary semantic analysis, possibly based on a community finding method.

Supporting information

S1 Data.
(ZIP)

S1 Table. Parameter values used in the second simulation study.
(PDF)

Author Contributions

Conceptualization: Mehran Aflakparast, Mathisca de Gunst, Wessel van Wieringen.

Data curation: Mehran Aflakparast.

Formal analysis: Mehran Aflakparast.

Funding acquisition: Mathisca de Gunst.

Investigation: Mehran Aflakparast.

Methodology: Mehran Aflakparast, Mathisca de Gunst, Wessel van Wieringen.

Software: Mehran Aflakparast.

Supervision: Mathisca de Gunst, Wessel van Wieringen.

Validation: Mehran Aflakparast.

Visualization: Mehran Aflakparast.

Writing – original draft: Mehran Aflakparast.

Writing – review & editing: Mathisca de Gunst, Wessel van Wieringen.

References

1. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. The annals of statistics. 2006; p. 1436–1462. https://doi.org/10.1214/009053606000000281
2. Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika. 2007; 94(1):19–35. https://doi.org/10.1093/biomet/asm018
3. Banerjee O, Ghaoui LE, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. Journal of Machine Learning Research. 2008; 9(Mar):485–516.
4. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9(3):432–441. https://doi.org/10.1093/biostatistics/kxm045 PMID: 18079126
5. Rothman AJ, Bickel PJ, Levina E, Zhu J. Sparse permutation invariant covariance estimation. Electronic Journal of Statistics. 2008; 2:494–515. https://doi.org/10.1214/08-EJS176
6. van Wieringen WN, Peeters CF. Ridge estimation of inverse covariance matrices from high-dimensional data. Computational Statistics & Data Analysis. 2016. https://doi.org/10.1016/j.csda.2016.05.012
7. Hao W. Bayesian graphical lasso models and efficient posterior computation. Bayesian Analysis. 2012; 7(4):867–886. https://doi.org/10.1214/12-BA729
8. Guo J, Levina E, Michailidis G, Zhu J. Joint estimation of multiple graphical models. Biometrika. 2011; 98:1–15. https://doi.org/10.1093/biomet/asq060 PMID: 23049124
9. Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2014; 76 (2):373–397. https://doi.org/10.1111/rssb.12033
10. Zhu Y, Shen X, Pan W. Structural pursuit over multiple undirected graphs. Journal of the American Statistical Association. 2014; 109(508):1683–1696. https://doi.org/10.1080/01621459.2014.921182 PMID: 25642006
11. Bilgrau AE, Peeters CF, Eriksen PS, Bøgsted M, van Wieringen WN. Targeted Fused Ridge Estimation of Inverse Covariance Matrices from Multiple High-Dimensional Data Classes. Journal of Machine Learning Research. 2020; 21(26):1–52.
12. Peterson C, Stingo FC, Vannucci M. Bayesian inference of multiple Gaussian graphical models. Journal of the American Statistical Association. 2015; 110(509):159–174. https://doi.org/10.1080/01621459.2014.896806 PMID: 26078481
13. Shen X, Pan W, Zhu Y. Likelihood-based selection and sharp parameter estimation. Journal of the American Statistical Association. 2012; 107(497):223–232. https://doi.org/10.1080/01621459.2011.645783 PMID: 22736876
14. Yajima M, Telesca D, Ji Y, Muller P. Differential patterns of interaction and Gaussian graphical models. Preprint. 2012.
15. Aflakparast M, de Gunst MC, van Wieringen WN. Reconstruction of molecular network evolution from cross-sectional omics data. Biometrical Journal. 2018; 60(3):547–563. https://doi.org/10.1002/bimj.201700102 PMID: 29320604
16. Aflakparast M, de Gunst M. Data integrative Bayesian inference for mixtures of regression models. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2019; 68(4):941–962. https://doi.org/10.1111/rssc.12346
17. Park T, Casella G. The Bayesian lasso. Journal of the American Statistical Association. 2008; 103 (482):681–686. https://doi.org/10.1198/016214508000000337
18. Andrews DF, Mallows CL. Scale mixtures of normal distributions. Journal of the Royal Statistical Society Series B (Methodological). 1974; p. 99–102. https://doi.org/10.1111/j.2517-6161.1974.tb00989.x
19. Chhikara R. The Inverse Gaussian Distribution: Theory: Methodology, and Applications. vol. 95. CRC Press; 1988.
20. Escobar M, West M. Bayesian prediction and density estimation. J Amer Statist Assoc. 1995; 90:577–588. https://doi.org/10.1080/01621459.1995.10476550
21. Rasmussen CE. The infinite Gaussian mixture model. In: Advances in neural information processing systems; 2000. p. 554–560.
22. Rasmussen CE, Ghahramani Z. Infinite mixtures of Gaussian process experts. Advances in Neural Information Processing Systems. 2002; 2:881–888.

23. Tabatabaei SA, Asadpour M. Study of influential trends, communities, and websites on the post-election events of Iranian presidential election in Twitter. In: Social Network Analysis-Community Detection and Evolution. Springer; 2014. p. 71–87.

24. Aflakparast M, Geeven G, de Gunst MC. Bayesian mixture regression analysis for regulation of Pluripotency in ES cells. BMC bioinformatics. 2020; 21(1):1–13.