Learning social networks from text data using covariate information

Xiaoyi Yang¹ · Nynke M. D. Niezink¹ · Rebecca Nugent¹

Accepted: 30 July 2021 / Published online: 18 September 2021
© The Author(s) 2021

Abstract

Accurately describing the lives of historical figures can be challenging, but unraveling their social structures perhaps is even more so. Historical social network analysis methods can help in this regard and may even illuminate individuals who have been overlooked by historians, but turn out to be influential social connection points. Text data, such as biographies, are a useful source of information for learning historical social networks but the identification of links based on text data can be challenging. The Local Poisson Graphical Lasso model models social networks by conditional independence structures, and leverages the number of name co-mentions in the text to infer relationships. However, this method does not take into account the abundance of covariate information that is often available in text data. Conditional independence structure like Poisson Graphical Model, which makes use name mention counts in the text can be useful tools to avoid false positive links due to the co-mentions but given historical tendency of frequently used or common names, without additional distinguishing information, we may introduce incorrect connections. In this work, we therefore extend the Local Poisson Graphical Lasso model with a (multiple) penalty structure that incorporates covariates, opening up the opportunity for similar individuals to have a higher probability of being connected. We propose both greedy and Bayesian approaches to estimate the penalty parameters. We present results on data simulated with characteristics of historical networks and show that this type of penalty structure can improve network recovery as measured by precision and recall. We also illustrate the approach on biographical data of individuals who lived in early modern Britain between 1500 and 1575. We will show how these covariates affect the statistical model's performance using simulations, discuss how it helps to better identify links for the people with common names and those who are traditionally underrepresented in the biography text data.

Keywords Local Poisson Graphical Lasso model · Social networks · Text data · L1 penalty factor · Bayesian penalty estimation

1 Introduction

The structure of social relationships among individuals can be represented by networks. Studying currently existing social networks is relatively simple, since information about them can often be obtained from multiple sources. For example, to learn individuals' personal social network, we can survey them or use information from social media platforms (Robins 2015). In an academic setting, co-authorship information from curriculum vitae, journals, and online archives can provide information about collaborations (Newman 2004). However, when the goal is to learn the structure of a historical social network, the potential sources of information are limited. In this case, text data is often the most easily accessible source of information. Networks inferred based on (historical) texts may help history scholars to identify individuals who are nowadays less well-known, but in their time occupied interesting social positions or positions of influence. In this paper, we present a new methodology to unravel (historical) social networks, based on text data.

Previous work on inferring social networks from text data spans various humanities and social science applications. For example, Marsden (1990) and Üsdiken and Pasadeos (1995) use structured surveys and citations respectively to estimate collaboration networks. Almquist and Bagozzi (2019) uncovered the underlying network structure of radical activist groups based on British radical environmentalist texts that appeared from 1992 to 2003. Their work primarily concentrated on the application of topic models to analyze these text, and they inferred a network by counting how often the names of activists co-occurred in the text. The China Biographical Database Project (Harvard University et al. 2021) is also a great example of how networks can be extracted from historical documents. For this labor-intensive project, first all the possible expressions of human relations (e.g. "A is friends with B") were manually listed, and then pattern matching was used on the text to detect social relations. Bonato et al. (2016) extracted and analyzed the social network from three best-selling novels, defining a link between two characters if their names co-appear within 15 words.

1.1 Six Degrees of Francis Bacon

Six Degrees of Francis Bacon (SDFB) is a recent historical network project, focusing on estimating the social network in early modern Britain during 1500–1700 (Warren et al. 2016; SDFB 2021). As their text source, SDFB uses biographies from the *Oxford Dictionary of National Biography* Matthew et al. (2014) and infers possible relationships between people from the number of times the name of one person occurs in a section of the other person's biography, under the assumption that if two people knew each other as more than just acquaintances and/or were colleagues, they are more likely to show up in each other's biographies. Warren et al. (2016) uses a Local Poisson Graphical Lasso model (Allen and Liu 2012) to estimate the social network using these count data, arguing that a conditional independence structure should be considered when constructing these

types of historical social networks to help distinguish whether two people likely knew each other or just happen to be co-mentioned in a document, potentially because they had a common acquaintance. In comparison to the use of co-occurrence without additional constraints, Warren et al. (2016) claims that conditional independence structures tend to avoid the false positive detection of links caused by confounding factors such as mutual acquaintances. One potential reason for this is that in the ODNB biographies, the use appositive clauses to explain people's relations is common. For example, in Francis Bacon's biography, we have

“Bacon, Francis, Viscount St Alban (1561-1626), lord chancellor, politician, and philosopher, was born on 22 January 1561 at York House in the Strand, London, the second of the two sons of Sir Nicholas Bacon (1510-1579), lord keeper, and his second wife, Anne (c.1528-1610) [see Bacon, Anne], daughter of Sir Anthony Cooke, tutor to Edward VI, and his wife, Anne, née Fitzwilliam.”

In this paragraph, Edward VI is mentioned to explain who Sir Anthony Cooke is, but he has no directly stated connection with Francis Bacon. If we had only used the co-occurrence of names as a proxy for a social tie, we could have misinterpreted this relation. Conditional independence can help us to identify whether Francis Bacon and Edward VI knew each other, given all other people's mentions, such as those of Anthony Cooke.

1.2 Including covariate information

There is no doubt that the SDFB project has contributed a rich resource to support humanities research on early modern Britain. However, despite this auspicious start, there is room for improvement. For example, in their validation of precision and recall among 12 non-random people, Warren et al. (2016) find that the SDFB approach tends to have high precision but relatively low recall. One possible reason for this behavior is that the model only makes use of the co-mention counts but ignores other information available in the text such as individual characteristics (e.g., occupation, social group). According to homophily theory, similar individuals are more likely to connect to each other than the dissimilar ones (McPherson et al. 2001). For example, several studies have shown that people who share similar age, education level (Kossinets and Watts 2009), occupation (Calvo-Armengol and Jackson 2004), gender and economic status (McPherson and Smith-Lovin 1982) are more likely to be connected. Given these results, we might expect that pairs of people linked in the estimated SDFB network¹ also may share common characteristics. Yet, the individual covariate information available in biography data was not taken into account by Warren et al. (2016). In a subsequent study, Mohamed (2020) used a logistic model to predict whether two people know each

¹ Note that ‘the’ SDFB network is actually an aggregate of 100 networks estimated based on 100 imputed document-by-person data sets—imputation was needed due to the existence of partial and duplicated names. Two individuals are considered related if their link is identified in more than half of the estimated networks. The final reconstruction additionally includes a few relationships manually added by expert historians.

other using features of the estimated SDFB network (e.g., common links) and pairwise covariates (e.g., same gender or social group), and finds that at least one third of the false positive links (i.e., the links that the model predicts with a high probability but do not exist in the estimated SDFB network) have supporting historical sources. This indicates that using covariate information within the model may help us to improve estimation of historical networks.

The idea of incorporating additional information into a Lasso regression model is not new. Yuan and Lin (2006) proposed group Lasso to add penalties to groups rather than individuals. Li et al. (2015) extended this method to a multivariate sparse group Lasso to incorporate arbitrary and group structures in the data. Their model provided a unique penalty for each node but also included a penalty for each group where the groups could overlap and even be nested. Zou (2006) proposed the adaptive LASSO which uses initial coefficient estimates without regularization to inform starting penalty weights. However, these papers did not include approaches for incorporating additional information into the penalties beyond group structure.

On the other hand, Boulesteix et al. (2017) proposed IPF-Lasso which assigns different penalty factors to all independent variables in their model that are a function of external information, and used cross validation to select penalty parameters based on model performance. In a similar vein, Zeng et al. (2021) outlined the Bayesian interpretation of penalized regression with covariate-dependent penalty parameters, re-formulating Lasso regression as a Bayesian model. However, these approaches have not been implemented in the Poisson case, which involves different estimation challenges.

In this paper, we extend the methods proposed by Boulesteix et al. (2017) and Zeng et al. (2021) to the Local Poisson Graphical Lasso model, and apply our extension in the context of the SDFB project. We will show (1) how to implement individual-level (node-level) covariates into the network model's penalty factors, (2) two methods for estimating penalty factors for potentially a large number of covariates, and (3) how the inclusion of additional information into penalty estimation can significantly improve precision and recall.

2 Model

In this paper, we aim to reconstruct a social network from text data. We represent this network by an undirected graph $G = (V, E)$, defined by a set of nodes V (in this case, individuals) of size $|V| = p$ and undirected edges E (social relations). We aim to learn this network based on text data consisting of n documents of similar length. For each of the documents, we count how often it mentions each of the p individuals. This type of information can be obtained by manual coding or using natural language processing techniques. If two people are mentioned in the same document, this could be an indication that they knew (or know) each other. However, this implication is not definite: they could, for example, have had a common acquaintance, and be co-mentioned in a document as a result of that. For this reason, conditional independence structures are a natural tool for estimating social networks from text data. In particular, if two people's counts of name

mentions in a document are positively correlated, conditional on all other people's mentions, then this indicates that they may have known each other.

There are several methods to infer conditional independence structures. Among these, the Gaussian graphical model is most popular—it is well defined and has a computationally efficient estimation method (Lauritzen 1996; Friedman et al. 2008). However, it is designed for continuous-valued, Gaussian distributed data instead of count data. The Poisson Log-normal model was developed for count data (Aitchison and Ho 1989), but is hard to estimate on large data sets. We found that its computation time is hundred times that of the Gaussian graphical model or the Local Poisson Graphical model (Allen and Liu 2012). Therefore, we here use the Local Poisson Graphical model, as was done by Warren et al. (2016).

The Poisson graphical model is a model for count data constructed such that the conditional distribution of each node (e.g., name count of a person), conditional on all the other nodes, is univariate Poisson (Yang et al. 2012). Unfortunately, for its density to be normalizable, all the conditional dependencies between variables in the model need to be non-positive. The Local Poisson Graphical Lasso model does not have this restriction—it is a variant of the Poisson graphical model that enforces sparsity and is estimated locally (Allen and Liu 2012). The model is called “local” because it uses a neighborhood selection scheme, as proposed by Meinshausen et al. (2006), estimating the conditional independence restrictions separately for each node in the graph.

Let $Y \in \mathbb{N}_0^{n \times p}$ be the document by person matrix, where Y_{ij} indicates how many times person j is mentioned in document i . Each row indicates how often each name is mentioned in a document, and each column indicates the name mentions for one person across all documents. We denote an observed document by person matrix by y , while Y denotes the random matrix. For document i , let $Y_{i,\neq j}$ denote the vector of name counts for all individuals other than j . The Poisson Graphical model can be expressed as

$$Y_{ij} \mid Y_{i,\neq j} = y_{i,\neq j}, \theta, \Theta \sim \text{Poisson} \left(e^{\theta_j + \sum_{k \neq j} y_{ik} \Theta_{kj}} \right), \quad (1)$$

where $\theta \in \mathbb{R}^p$ and $\Theta \in \mathbb{R}^{p \times p}$, with $\Theta_{ii} = 0$ for all i . The node parameter θ_j serves as the intercept in this model and we use the edge parameters Θ_{jk} to infer the relation between individuals j and k . If $\Theta_{jk} > 0$, then if individual k is mentioned often in a document, j is likely to be mentioned often as well, which is suggestive of a social tie between j and k . The opposite is true when Θ_{jk} is negative. We thus consider $\Theta_{jk} > 0$ as an indication of a social tie between individuals j and k .

To enforce sparsity, we add a Lasso penalty term to model (1). The value of $\Theta_j = (\Theta_{1j}, \dots, \Theta_{pj})$ that maximizes the penalized log-likelihood is

$$\hat{\Theta}_j = \arg \max_{\Theta_j} \sum_{i=1}^n \left[y_{ij}(y_{i,\neq j} \Theta_{\neq j,j}) - e^{y_{i,\neq j} \Theta_{\neq j,j}} \right] - \sum_{k \neq j} \rho |\Theta_{kj}|, \quad (2)$$

where the tuning parameter ρ is used to control the sparsity of the network. While Warren et al. (2016) penalized all edges equally, In the next section, we will

incorporate covariate information in ρ to differentially penalize edges depending on node and edge covariates.

2.1 Penalty factors

The Poisson Graphical Lasso model leverages the name count data in text to learn social network information. However, generally, the text in which names are embedded is rich with other information that could be indicative of social ties. For example, it is useful to consider available demographic information when reconstructing social networks from text data. Homophily theory indicates that people with common characteristics are generally more likely to be connected than those who are not alike (McPherson et al. 2001). Therefore, it might be relevant to know, e.g., whether individuals were part of the same family or social group/club, worked for the same company, and whether they lived geographically close to one another.

Here we extend the Local Poisson Graphical Lasso model with a multiplicative factor for the penalty term that depends on individual covariate information inferred from the text. For person j , we define the covariate matrix $Z^j \in \{0, 1\}^{p \times m}$, with m covariates, by

$$Z_{kh}^j = \begin{cases} 1 & \text{if persons } j \text{ and } k \text{ have an equal value for covariate } h, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

We here consider binary-valued matrices Z^j , but the approach proposed in this paper is also applicable to real-valued covariates. Examples of these include last name similarity and last name or social group commonality scores. In this case, vector Z^j would no longer be binary, but would also contain continuous-valued similarity scores. For example, if we want to account for misspelling when we are comparing the last names, instead of considering whether person j and person k have exactly the same last name, we can use their last name similarity, e.g., the Jaro–Winkler similarity the last names of person j and person k (Winkler 1990).

For each covariate we include a different penalty factor. Thus, for each person j , the estimators are given by

$$\begin{aligned} \hat{\theta}_j = \arg \max_{\theta_j} & \sum_{i=1}^n [y_{ij}(y_{i,\neq j} \theta_{\neq j,j}) - e^{y_{i,\neq j} \theta_{\neq j,j}}] \\ & - \sum_{k \neq j} \rho_{kj} |\theta_{kj}| \quad \text{with } \log(\rho_{\neq j,j}) = Z^{j*} \alpha, \end{aligned} \quad (4)$$

where Z^{j*} is the matrix Z^j with the j th row excluded and prefixed by an all-one column vector, and $\alpha \in \mathbb{R}^{m+1}$ denotes the penalty factor. The first element of α is α_0 , an intercept controlling the overall shrinkage. If two individuals k and j share a common value on a covariate h , the penalty for parameter θ_{jk} , indicating the link between them, is e^{α_h} times the overall penalty. Therefore, if having covariate h in common makes two people more likely to be connected, then α_h will be negative. Otherwise, it will be positive.

To illustrate this setup, suppose we have two covariates—last name and occupation—and consider the model for the name mentions Y_{ij} of Francis Bacon (person j) in document i . The $p \times 2$ covariate matrix Z^j indicates for the p individuals in the data whether they share their last name and occupation with Francis Bacon. An example of this matrix is shown in Table 1. Matrix Z^{j*} equals matrix Z^j , but with the row of Francis Bacon taken out and prefixed by an all-one column vector. The penalty factor in this case is given by $\alpha = (\alpha_0, \alpha_{\text{LN}}, \alpha_{\text{OC}})$, where α_0 is the penalty intercept and α_{LN} and α_{OC} are the penalty factors corresponding to sharing a last name and sharing an occupation, respectively. Their effects on the penalty for parameter θ_{jk} are given in Table 2.

Birth and death date are covariates that deserve special treatment in this framework, since if two individuals were not alive at the same time, they could not have had a social connection. To address this, Warren et al. (2016) removed the links between people who were not alive at the same time post-network estimation. Given our penalty factor structure, we can instead include birth and death year information directly into the model. We set the penalty factor for the lifespan overlap covariate to be infinity and so do not link people with non-overlapping lifespans. Including infinite penalties into the model serves the same purpose as the post-modeling removal of ‘impossible’ links, but will largely decrease the computational complexity as it decreases the number of parameters θ_{ij} that need to be estimated.

3 Estimation

For each person j , we fit a Poisson regression model including an L1 penalty to enforce sparsity. We estimate model parameters via penalized maximum likelihood using cyclical coordinate descent, as implemented in the R package `glmnet` (Friedman et al. 2010). This method consecutively optimizes the objective function, given as part of expression (4), over each parameter while keeping the others fixed, and cycles until convergence.

Table 1 Example excerpt of covariate matrix Z^j , when j refers to Francis Bacon

	Same last name as Francis Bacon	Same occupation as Francis Bacon
Nicolas Bacon	1	1
Anne Bacon	1	0
Francis Bacon	1	1
Walter Raleigh	0	1
Queen Elizabeth I	0	0
⋮	⋮	⋮

Table 2 Lasso penalty parameters ρ_{kj} as in (4) for parameters θ_{jk} in a model with penalty factors depending on last name and occupation

Person j and k .	Penalty ρ_{jk}
Share no common covariate	e^{α_0}
Only have the same last name	$e^{\alpha_0 + \alpha_{ln}}$
Only have the same occupation	$e^{\alpha_0 + \alpha_{oc}}$
Share both last name and occupation	$e^{\alpha_0 + \alpha_{ln} + \alpha_{oc}}$

After estimating the edge parameters θ_{jk} , we only interpret positive estimates as an indication of the existence of a link, as proposed by Warren et al. (2016). A negative θ_{jk} would imply that if a document mentions person j more, it would mention person k less: this is not indicative of a relationship between persons j and k . Also, note that both θ_{jk} and θ_{kj} reflect the relation between persons j and k . Here, we adopt the “OR” rule for determining links. That is, after estimating the edge parameter vectors for persons j and k , we say that there is a social tie between j and k when at least one of $\hat{\theta}_{jk}$ and $\hat{\theta}_{kj}$ is positive. The “AND” rule would require both $\hat{\theta}_{jk}$ and $\hat{\theta}_{kj}$ to be positive to claim a social tie, likely resulting in higher specificity, but lower recall. Choosing the “OR” rule instead of the “AND” rule also helps to resolve situations where there is a social tie between two individuals (j and k), but a third person (l) impacts the estimation of this tie. For example, suppose we are modeling the name mentions of person j —estimating θ_j —and there exists a third person l whose mentions are highly correlated with those of person k . If the Lasso algorithm would select the edge to person l over that to person k , the link between individuals j and k would not be identified in case the “AND” rule was used. In this case, however, the “OR” rule could still capture the link between persons j and k through the estimation of θ_k .

Estimating the value of penalty vector α is essential for determining the edge parameters. In the following two sections, we discuss two approaches to estimate the penalty vector: using greedy search (Sect. 3.1) and using the reformulation of Lasso regression in the Bayesian framework (Sect. 3.2).

3.1 Greedy search

One way to estimate the penalty factor α is by defining a grid of penalty parameter values and evaluating the corresponding models, selecting the values that minimize the prediction error, as measured by the mean square error (Boulesteix et al. 2017). The focus of this work is improving the overall model performance, specifically in identifying the set of true links. In this context, we think prediction error is a suitable goodness-of-fit measure. Even though we do not include a formal comparison, we believe other suitable measures of goodness-of-fit (AIC, BIC etc.) should also perform well if the focus lies more on model complexity.

The approach discussed by Boulesteix et al. (2017) evaluates the model for all combinations of penalty parameter values on the search grid, and therefore is generally computationally feasible only when the number of covariates is small (say, no more than four). However, greedily searching the parameter space allows

for inclusion of more covariates. Our proposed greedy algorithm for α can be described as follows (see “Appendix 1” for the pseudo-code). Starting with all $\alpha_h = 0$ for $h \neq 0$ (i.e. no penalty adjustment), the algorithm first iterates over all covariates in random order. For each covariate h and a gridded range of pre-specified α_h values, we use cross-validation to choose the baseline $\hat{\alpha}_0$ (holding all other α_h penalty parameters fixed) and calculate the corresponding MSE. We then choose the $\hat{\alpha}_h$ that corresponds to the lowest MSE. The algorithm repeatedly randomly iterates through all covariates in this way, looking to update the $\hat{\alpha}_h$ values in such a way to decrease the MSE, and stopping when no further α_h tuning leads to a decrease in MSE.

To use this algorithm, we need to specify the search range for α_h and the step size d , i.e., the distance between the searching values. We recommend starting with a search range for α_h such as $[-1.2, 0]$ (values outside that range have diminishing impact on the multiplicative factor value; the difference between $e^{-1.2}$ and $e^{-1.3}$ is much smaller than that between $e^{-0.1}$ and $e^{-0.2}$) and a relatively large step size d (e.g., 0.1). The search range can be enlarged if the margins are hit during the initial estimation. Decreasing the step size d can of course lead to a more fine-grained solution but will also lead to a higher computation time. We could also choose the search range $[a_n, b_n]$ using prior information, such as which covariates are expected to be influential and approximately how they might affect the chance of two individuals to be connected. For example, if we know a covariate h is likely associated with an increased chance of a link, we could initially limit the search range of α_h to the negative numbers, and consequently reduce the computational time.

3.2 Bayesian estimation

Lasso estimates can equivalently be derived as the Bayesian posterior modes under independent Laplace priors for the parameters to which shrinkage is applied (Tibshirani 1996). Therefore, we can use the Bayesian framework to estimate the penalty parameters α . To this end, we complement model (1) with a Laplace prior for the edge parameters: for $k \neq j$,

$$\theta_{kj} \sim \text{Laplace}(0, b_{kj}), \quad b_{kj} \propto \rho_{kj}, \quad \log(\rho_{\neq jj}) = Z^{j*} \alpha. \quad (5)$$

Notice that α only influences the penalties on the edges and not the node parameters θ_j . We will specify the exact form of b_{kj} later in this section. We here extend work by Zeng et al. (2021) on incorporating covariate-dependent penalty factors in the Lasso term in linear regression and linear discriminant analysis (LDA) models to the Local Poisson Graphical Lasso model.

We use an empirical Bayesian approach to estimate the penalty parameters α . First, for each person j , we approximate the marginal log-likelihood of α , denoted by $l_j(\alpha)$, marginalizing over the coefficients $\theta_{\neq jj}$. The estimate of α is given by

$$\hat{\alpha} = \arg \max_{\alpha} \sum_{j=1}^p l_j(\alpha). \quad (6)$$

Note that we maximize the sum of the marginal distributions, because we need a global penalty factor over all people instead of for one specific person j . Since the $l_j(\alpha)$ are not convex, we use a Majorization Minimization procedure to estimate $\hat{\alpha}$ (Zeng et al. 2021). We then use the estimate $\hat{\alpha}$ as input for the penalized maximum likelihood estimation, as summarized at the start of Sect. 3.

Since the Poisson regression likelihood and the Laplace prior are not conjugate pairs, there is no closed form expression for the marginal likelihood of α . We here present a general outline of how we approximated $l_j(\alpha)$, approximating both the Poisson regression likelihood and the Laplace prior—see “Appendix 1” for the full derivation. First, we apply the log-gamma transformation to approximate the Poisson regression likelihood by a multivariate Gaussian distribution (Chan and Vasconcelos 2009). In order to avoid $\log(0)$ in our derivation, we add 1 to all the observed outcomes y_{ij} , that is, define $y_{ij}^* = y_{ij} + 1$. Second, we assume θ_{kj} follows the Laplace prior $\theta_{kj} \sim \text{Laplace}(0, \frac{\rho_{kj}}{2\sigma_j^2})$, where $\hat{\sigma}_j^2 = \sum_{i=1}^n \frac{1}{y_{ij}^*}$ is the estimated variance in the Gaussian distribution approximating the Poisson likelihood. We approximate this prior by a normal distribution with the same variance (Zeng et al. 2021), yielding

$$\theta_{\neq j,j} \sim \mathcal{N}(0, V^j) \quad (7)$$

where $V^j \in \mathbb{R}^{(p-1) \times (p-1)}$ is a diagonal matrix with $V_{kk}^j = 2\sigma_j^2 \exp^{-2Z_k^{j*}\alpha}$, in which Z_k^{j*} is the k th row of the covariate matrix Z^{j*} .

Combining the two, we can approximate the log-likelihood of α for person j and find

$$-l_j(\alpha) \propto \log |C_\alpha| + \log(y_j^*)^\top C_\alpha^{-1} \log(y_j^*) \quad (8)$$

where $C_\alpha = \sigma_j I^2 + y_{\neq j} V^j y_{\neq j}^\top$, with $y_{\neq j}$ denoting data matrix y excluding the j th column, and where $\log(\cdot)$ is applied element-wise to $y_j^* = (y_{1j}^*, \dots, y_{nj}^*)^\top$. Integrating this in expression (6), we can estimate the penalty factors.

4 Simulation study

In order to compare the methods quantitatively, we created a small community (network) as the ground truth for the simulation. The community is similar by design to the SDFB social network, and we assume to have similar available demographic information. When creating the network, we considered three covariates: last name, group membership and birth/death year overlap. Note that we assume that if the lifespan of two people does not overlap (i.e., it is impossible that they physically met each other), then they should not be linked, regardless of all other factors. Below is a description of the network simulation design:

1. We generate 50 families in the community with 30 different last names (i.e., people with the same last name can be from different families).
2. For each family, we randomly generate 5–12 people, each with a birth and death year between 1500 and 1600 and a life length varying from 5 to 70 years.
3. Within a family, among those people whose lifespan overlaps, 50% know each other.
4. There are three social groups, A, B and C. Each person is randomly assigned to one of the groups with probability 0.5, 0.25 and 0.25, respectively.
5. Among those people whose lifespan overlaps, we additionally create 100, 100 and 50 links within group A, B and C, respectively.
6. Finally, we add 300 random links to the community.

This design yields 464 people and 1164 links. Figure 1 illustrates a subset of the community with ten families, 100 people, and 158 links. Since our network design yields higher link density within a family/group than average in the network, we anticipate that all α will be negative, corresponding to a smaller penalty if two people share the same last name or are in the same group (we will estimate separate penalty factors for each group).

Figure 1 shows that there are indeed many family ties and we thus expect the absolute value of the penalty factor for last name to be large. The social group covariates might also be useful to predict the links but should have a smaller effect than last name. Since group B is the densest one, we expect that its penalty factor may be larger compared to the other two groups. Recall that if two people's lifespan does not overlap, we set the penalty on their link to infinity.

Using a simulation framework adopted from Allen and Liu (2012), we generate 10 different document-by-person matrices, each with 2000 documents (i.e. of similar size as SDFB). For each matrix, we run the greedy algorithm and the Bayesian approach to estimate the penalty factor α . Since the matrices are all

Fig. 1 Sub-community composed of 10 families (100 people, 158 links). Last names are represented by colors and social groups by shapes. The network shows clear family structure with some social group structure and additional random links

generated from the same network, we expect the $\hat{\alpha}$ to be similar and reflective of the designed network structure (e.g., relatively larger values for last name and group B). We then compare the network estimated with no penalty adjustment to the networks estimated with penalty adjustment (for both penalty factor estimations methods) by calculating their best precision and recall.

We define the “best precision and recall” to be the highest sum of precision and recall generated from the estimated network, where precision and recall are defined as

$$\text{precision} = \frac{TP}{TP + FP} \quad \text{recall} = \frac{TP}{TP + FN}, \quad (9)$$

where TP denotes the number of true positive links, FP the number of false positive links, and FN the number of false negative links. Our expectation is that the penalty adjustment allowing for incorporating covariate information into the network estimation will be associated with an improved average of precision and recall. We also examine the predicted 10-family sub-communities in an attempt to characterize how the penalty factors impact the false positives and false negatives.

All the simulations were run on a personal laptop with an Intel(R) Core(TM) i7-10510U CPU. Both estimation approaches require approximately 4-5 hours per simulation. Even though the estimation of α is global, the estimation of Local Poisson Graphical Lasso model could be parallelized to reduce computation time. For the Greedy approach, if we have prior information about how the covariate will affect the linking probability, we can also limit the search space to reduce computation time.

Figure 2 shows the estimates of $|\hat{\alpha}_h|$ for the greedy and Bayesian approach. All $\hat{\alpha}_h$ are negative, indicating if two people have the same last name or social group membership, they are more likely to be linked. (Note that we here plot the magnitudes of the penalty factors.) The larger the absolute value of $\hat{\alpha}_h$, the stronger the effect of the covariate is on the penalty. Here we see that the Bayesian approach gives more similar $\hat{\alpha}$ values across the ten runs, correctly identifying last name as the most important covariate and group B as having a slightly stronger effect than the other two groups. The greedy method gives $\hat{\alpha}$ values that are more varied and do not reflect the network design. For example, although last name has a non-zero effect, it is not substantially larger than the other $\hat{\alpha}_h$ for the social groups. One potential reason for the consistency differences is that the Bayesian approach is trying to optimize the log likelihood of α while the greedy algorithm tries to optimize the model performance based on the MSE which may find multiple combinations of $|\hat{\alpha}_h|$ that lead to similar results. For example, if two people share the same last name and the same social group, a smaller penalty on either last name or social group or both can help to recover the link, depending on which covariate is added to the model first.

For each generated document by person matrix, we also use the $\hat{\alpha}_h$ for both the greedy and Bayesian approaches to estimate the network and calculate the corresponding precision and recall, comparing these values to those for the model without penalty adjustment. Figure 3 shows the resulting distributions for precision, recall, and the average of the two. We see that the model with penalty adjustment

Fig. 2 The values of $|\hat{\alpha}_h|$ estimated by the Bayesian method (top figure) and the greedy method (bottom) for ten runs. Both approaches generally pick the last name as the most important covariate. The Bayesian approach yields more consistent values, while the estimates obtained by the greedy approach have a larger variance

has improved precision, regardless of estimation approach, while the recall for all three options remains similar. The slight improvements in the average of the two follow.

Now examining the predicted network structure, we see that all three model/estimation approaches overestimated the true number of links in the original simulated network (1164). On average across the ten document by person matrices, the model without penalty adjustment detects 1662.1 links. With penalty adjustment, the greedy estimation approach estimates 1434.7 links on average, and the Bayesian approach 1358.1 links, both an improvement over the original model.

We then take a closer look at the estimated network structure for our ten family, 100 people sub-community (Fig. 1) for two of the simulated document by person matrices. For Run 10 (top row of Fig. 2), the greedy and Bayesian estimation approaches give similar $|\hat{\alpha}_h|$ values for last name, but the greedy approach gives slightly larger $|\hat{\alpha}_h|$ values for the social group covariates. Therefore, we expect more links between people with the same group membership when using the estimates from the greedy approach compared to those of the Bayesian approach.

The relevant estimated networks for Run 10 can be seen in Fig. 4 and corresponding number of links are given in Table 3. We can see that for the network estimated by the model without a penalty adjustment, the false positive links exist across the whole network but for the network estimated with penalty adjustment, both the numbers of false positive and false negative links decrease. The predicted networks with $\hat{\alpha}$ from the greedy and Bayesian approach are similar, but there are slightly more false positive links across groups A and B for this sub-community

Fig. 3 The distribution of precision and recall for the model without penalty factor, with penalty factor estimated using the greedy approach, and with the penalty factor estimated using the Bayesian approach. With penalty adjustment, both estimation approaches show an improvement in precision without a substantial change in recall. There is no significant difference on the average of precision and recall between the two estimation approaches

with $\hat{\alpha}$ from the Bayesian approach and more false positive links within group C for the greedy approach. Note that this is in line with the observation that the absolute values of group-related penalty factors, for Run 10, are much larger for the greedy approach for group B and C. Thus the greedy approach here tends to pick up more within-group links while the Bayesian may picks more between-groups links.

We also examine Run 1 where the $|\hat{\alpha}|$ are quite different between the two estimation approaches (see Fig. 2). The greedy method tends to give a much smaller penalty change for last name but a larger penalty change on social groups A and C, although we do note that the $|\hat{\alpha}_h|$ for group B is incorrectly estimated to be smaller than those for groups A, C. The corresponding predicted networks are depicted in Fig. 5 and the corresponding number of links are in Table 4. The networks corresponding to the penalties estimated by the greedy and Bayesian approach are more dissimilar for for Run 1 than for Run 10, like the penalties themselves. Compared to Bayesian method, the absolute value of group A penalties factor are larger for the greedy approach, leading to the detection of more links between people within group A in this subset.

In summary, our simulation study gives some evidence that including covariate information through penalty adjustment can improve the performance of Local Poisson Graphical Lasso model in the context of estimating social networks from co-mention/count data derived from text. With respect to differences in the two estimation approaches, we see that the Bayesian approach tends to give more consistent results. However, we note that, given its global estimation and computational tasks (e.g. matrix inverse calculations), it will not be faster than the greedy algorithm.

Fig. 4 Predicted ten-family community network for Run 10 for the model without penalty factor, with the penalty factor estimated using the greedy approach and with the penalty factor estimated using the Bayesian approach. Grey line: true positive; blue line: false positive; red line: false negative. Node color: last name; node shape: social groups. In Run 10, the $\hat{z}_h$ are similar for last name. The greedy approach gives slightly higher values for group covariates. We see fewer false positives using the greedy approach

Table 3 The number of true positive, false positive, and false negative links for Run 10

	True positive	False positive	False negative
No penalty factor	129	20	29
Greedy	137	12	21
Bayesian	139	12	19

Fig. 5 Predicted ten-family community network for Run 1 for the model without penalty factor, with the penalty factor estimated using the greedy approach and with the penalty factor estimated using the Bayesian approach. Grey line: true positive; blue line: false positive; red line: false negative. Node color: last name; node shape: group membership. In Run 1, the greedy approach tends to give a smaller penalty on last name but a larger penalty on social groups which gives us more false positives and a few fewer false negatives within social groups

Table 4 The number of true positive, false positive, and false negative links for Run 1

	True positive	False positive	False negative
No penalty factor	121	17	33
Greedy	139	21	19
Bayesian	138	14	21

5 Six Degrees of Francis Bacon: 1500–1575

We illustrate the model proposed in this paper by an application to a part of the data used in the SDFB project (Warren et al. 2016), focusing on the period between 1500 and 1575. We compare the results of the models with and without covariate-dependent penalty factors. We consider the interpretability of the penalty factors, how they affect which network links are estimated, and approximate the precision of the models with and without penalty factors using Wikipedia as a reference.

We first extract all documents from the SDFB database that contain references to individuals who were born and passed away between 1500 and 1575. This results in 2003 documents on 420 people. Over 83% of them (394) are male, about 8% (34) are female, and for the rest the gender is unknown. Women who appear in these data are usually associated with men in the data through family or marriage.

Apart from last name and birth and death year, we here consider three other covariates, related to individuals' occupation. We distinguish three groups: the Writer group (the occupation variable in the data contains the words “poet”, “writer” or “author”), the Church group (occupation contains “church”, “religious”, “bishop” or “catholic”), and the Royal group (occupation contains “royal”, “king”, “queen” or “regent”).

Table 5 includes some descriptives of the data. Since we have limited the data to people who were alive in a period of 75 years, the lifespan of most pairs of people overlapped. Compared to the simulated data, the proportion of pairs with shared last name is much smaller. This indicates a more diverse last name distribution (the most common last name “Stewart” is the last name of royalty during this period and appeared for only 9 individuals, while other last names appeared for no more than 5 individuals), but also suggests that as long as two people shared the same last name, the chances of them belonging to the same family and knowing each other are high. Among all occupations that were listed in the data, the writer- and the church-related occupations are most popular. Individuals with a royal-related occupation tend to be closer connected than other people, which is why we consider this group, even though not that many people are part of it. People can have multiple group memberships across the three groups. Five individuals are part of more than one

Table 5 Descriptive statistics of the SDFB data on people from the period 1500–1575

Group	Number of people	
Writer	40	
Church	49	
Royal	19	
In multiple groups	5	
	Number of pairs	
Same last name	117	(0.13%)
Lifespan overlaps	81625	(92.8%)

group, like Roger Ascham, who was an author and a royalty tutor, and John Seton, who was a Roman Catholic priest as well as a writer on logic.

We estimated the penalty parameters α using the Bayesian approach outlined in Sect. 3.2. We find that

$$\begin{aligned}\hat{\alpha}_{\text{lastname}} &= -1.853 \\ \hat{\alpha}_{\text{writer}} &= 0.369 \\ \hat{\alpha}_{\text{church}} &= -1.262 \\ \hat{\alpha}_{\text{royal}} &= -0.801.\end{aligned}\tag{10}$$

From the size of the penalty factors, the last name is the most important covariate, indicating that if two people share the same last name, this is a strong indication that they may know each other. It is interesting that not for all groups the penalty factor is negative: if two people are both a writer, they are less likely to be connected. It is possible that being a writer is an occupation for which little collaboration is required, so that the writers did not socialize much with their peers. On the other hand, if two people are both related to the church or the royal family, this increases their chance of being linked.

Next, we compare the networks generated by the Local Poisson Graphical Lasso model with and without penalty adjustment. The overall penalty level for both models is the one minimizing the MSE. For the model without penalty adjustment, the estimated network consists of 156 links and for the model with penalty adjustment, the estimated network consists of 135 links. Although they partially overlap, the two networks also contain many different links. There are 40 links that are only picked up by the model with penalty adjustment and 61 links that are picked up only by the model without penalty adjustment.

How do the penalty factor values α relate to the difference between the two estimated network? To answer this question, we consider the percentage of links estimated by the two that had covariates in common (see Table 6).

As expected based on the negative penalty factor estimate for the last name and the royal group, the model with penalty adjustment picks up more links between people with the same last name or both related to the royal family. To be more specific, the model with penalty adjustment detects four additional links without losing the seven links that were estimated by the model without penalty adjustment

Table 6 Numbers and percentages of links estimated by the models with and without penalty adjustment, for whom the corresponding people had a common covariate

Covariate	With penalty adjustment model	Without penalty adjustment model
Last name	19 (8.9%)	16 (5.8%)
Writer group	4 (3.0%)	4 (2.6%)
Church group	2 (3.0%)	6 (5.2%)
Royal group	8 (5.9%)	4 (2.6%)

model. However, the proportion of links between individuals from the writer or the church group does not differ much between the two models. Both models select one link between two people in the writer group. The model with penalty adjustment even picks one link less within the church group, even though the negative penalty factor α_{church} indicates that links between people within the church group are penalized less. Note that the difference in within-group estimated links only contributes a small portion of the difference among the estimated networks. This suggests that changing the penalty on the links between people within the same groups also affects the links that are not within those groups.

Finally, we approximate the precision of the estimated networks by looking for evidence for links on Wikipedia. For a link involving two people, as long one of the persons' Wikipedia document contains the other one's name, we consider this as evidence that a link exists. Of the 135 links that are picked up by model with penalty adjustment, we find evidence for 62 (45.9%). Of the 156 links that are picked up by the model without penalty adjustment, we find evidence for 67 (42.6%).

There are 95 links that are detected by both models. For the 61 links that are only detected by the model without penalty adjustment, we notice that they repeatedly involve the same individuals. George Wishart, who is listed as "evangelical preacher and martyr" and Thomas Wynter, who is listed as "clergyman" should both belong to the church group. However, when we first defined the group, we did not pick up the word like "preacher" and "clergyman" to be included in the in the Church group, which causes the model with penalty adjustment to not pick up the links for them and also may lead to the lower linking rate in the church group, reported in Table 6. This finding indicates that it is important to systematically and meticulously define the groups, as many different words may have a similar meaning. On the other hand, for the 40 links that are only detected by the model with penalty adjustment, there are also some individuals who appear repeatedly, like Katherine Seymour, who belongs to the royal family. This is likely related to the decreased penalty for ties between royal family members. We also have Margaret Roper and Nicholas Udall in the group, who are authors closely related to the royal family.

There is no doubt that an in-depth analysis of these results would require the help from experts on British history, but from these preliminary analyses, it seems that the model with penalty adjustment yields a more precise and conservative estimate of the relationships.

6 Discussion

In this paper, we have shown promising results to support adding covariate information when estimating social networks from text data using a Local Poisson Graphical Lasso model. This covariate information is incorporated through the L1 penalty: we penalize the parameters representing the edges between two individuals depending on the extent to which they have covariates in common. To estimate the penalty factors, we have discussed two approaches: a greedy algorithm and a

Bayesian framework. Both simulation and real data example are implemented to show the validation of the approach.

There are several direction in which this work could be further developed in the future. In the Bayesian approach, we currently approximate both the Poisson regression likelihood and the Laplace prior to find the marginal likelihood for the penalty factor. We have not analytically evaluated the effect of this double approximation. Even though the simulation study yields that the penalty factor estimated by the Bayesian approach gives results comparable to those for the greedy approach, we are considering other approximation methods, for example, one Laplace approximation, instead of doing the approximation on Poisson likelihood.

Here, we have applied the model to a subset of the SDFB data. The complete SDFB data contain over 19,000 documents with references to over 13,000 people. Both approaches we have proposed to estimate the penalty factor, and especially the Bayesian approach, will be slow when dealing with large data. Considering other optimization approaches to improve the computational efficiency is an interesting avenue for future research.

In this paper, we only interpret the positive edge coefficients in the Local Poisson Graphical Lasso model as an indication for the existence of a social tie between two individuals. Note that when we ran the same models with a non-negativity constraint on the parameters (instead of disregarding negative ties as a post-processing step), we obtained roughly the same results in terms of precision and recall. It would therefore be interesting to explore alternative priors (e.g., a Gamma prior) for the edge parameters.

Finally, both in our simulation study and in our real data analysis, we only focus on binary covariates. However, there are several continuous covariates that are worth considering when estimating social networks from text data. For example, historical text often contains typos and sometimes spelling variation. Therefore, when identifying whether two people are from the same family, instead of directly comparing their last names, we could consider the similarity of their last names. By measuring name similarity using the Jaro–Winkler distance (Winkler 1990) and including this in the L1 penalty, a missing letter or a word with the same pronunciation but a different spelling (e.g., “Askham” and “Ascham”) would not be over-penalized. Also, in the real data example, we currently only consider how being in the same occupation group should affect the L1 penalty. However, this analysis does not need to be limited to the within- versus between-occupation comparison. We could also include a continuous covariate representing the similarity between occupations (or introduce penalty parameters corresponding to the links between occupation groups). The proposed Local Poisson Graphical Lasso model with covariate-dependent penalty parameters thus provides a rich framework for learning social networks from text data.

Appendix 1: the algorithm of greedy method

Input : document-by-person matrix Y , covariate matrices Z^{j*} for $j = 1, \dots, p$,
function $MSE(\alpha)$ evaluating the MSE for the model with penalty α ,
search range $[a_h, b_h]$ for $h = 1, \dots, m$, grid size d

Output: $\hat{\alpha} = [\hat{\alpha}_1, \dots, \hat{\alpha}_m]$

Initialization: $\hat{\alpha} \leftarrow [0, 0, \dots, 0]$, repeat $\leftarrow \text{true}$,
 $MSE_{old} \leftarrow MSE(\hat{\alpha})$, $MSE_{new} \leftarrow \infty$

```

while repeat do
  repeat  $\leftarrow \text{false}$ 
  order  $\leftarrow$  a random permutation of 1 to  $m$ 
  for (  $s \leftarrow 1$  to  $m$  by 1 ) do
     $h \leftarrow \text{order}[s]$ 
    for (  $\hat{\alpha}_h^* \leftarrow a_h$  to  $b_h$  by  $d$  ) do
       $\hat{\alpha}^* \leftarrow [\hat{\alpha}_1, \dots, \hat{\alpha}_h^*, \dots, \hat{\alpha}_m]$ 
       $MSE_{new} \leftarrow MSE(\hat{\alpha}^*)$ 
      if  $MSE_{new} < MSE_{old}$  then
         $MSE_{old} \leftarrow MSE_{new}$ 
         $\hat{\alpha} \leftarrow \hat{\alpha}^*$ 
        repeat  $\leftarrow \text{true}$ 
      end
    end
  end
end
end

```

Appendix 2: approximating the marginal likelihood of the penalty parameters

As mentioned in Eq. (1), we model the number of times person j appears in document i using Poisson regression,

$$Y_{ij} | Y_{i,\neq j} = y_{i,\neq j}, \theta, \Theta \sim \text{Poisson}(e^{\lambda(y_{i,\neq j})}), \quad (11)$$

where

$$\lambda(y_{i,\neq j}) = \theta_j + \sum_{k \neq j} y_{ik} \Theta_{kj}, \quad (12)$$

with a covariate-depended Lasso penalty on the Θ_{kj} or, equivalently, a Laplace prior (see expression 5). To estimate the values of penalty parameters α in a Bayesian framework, we here approximate their marginal likelihood.

We first approximate the Poisson likelihood by a normal distribution, using the log-gamma approximation (Bartlett and Kendall 1946; Prentice 1974; Chan and Vasconcelos 2009). Recall a Gamma random variable $\mu \sim \text{Gamma}(a, b)$ with distribution

$$p(\mu | a, b) = \frac{1}{\Gamma(a)b^a} \mu^{a-1} \exp^{-\frac{\mu}{b}}, \quad (13)$$

then the transformed random variable $\log(\mu)$ has a log-gamma distribution and for large a , the log-gamma distribution is approximately to $\mathcal{N}(\log(a) + \log(b), a^{-1})$. Let $b = 1$ and $a \in \mathcal{Z}^+$, and let $\eta = \log(\mu)$ then we have

$$\begin{aligned}
 p(\eta \mid a, 1) &= p(\mu = \exp^\eta \mid a, 1) \times \frac{\partial}{\partial \eta} \exp^\eta \\
 &= \frac{1}{(a-1)!} \exp^{\eta a} \exp^{-\exp^\eta} \\
 &\approx G(\eta \mid \log(a), a^{-1}),
 \end{aligned} \tag{14}$$

where $G(x \mid \mu, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp(\frac{1}{2} \|x - \mu\|_\Sigma^2)$ is the equation of a multivariate Gaussian distribution, $\|x\|_\Sigma^2 = x^T \Sigma^{-1} x$.

Since our data y are often sparse, to avoid $\log(0)$ in the remainder of this derivation, we add 1 to all response values y_{ij} and define $y_{ij}^* = y_{ij} + 1$. Using the approximation derived in Eq. (14), we find that

$$\begin{aligned}
 &\frac{1}{(y_{ij}^* - 1)!} e^{\lambda(y_{i,\neq j}) y_{ij}^*} e^{-e^{\lambda(y_{i,\neq j})}} \\
 &\approx G\left(\lambda(y_{i,\neq j}) \mid \log(y_{ij}^*), \frac{1}{y_{ij}^*}\right).
 \end{aligned} \tag{15}$$

Then the Poisson likelihood can be written as

$$\begin{aligned}
 \prod_{i=1}^n p(Y_{ij}^* \mid Y_{i,\neq j} = y_{i,\neq j}, \Theta_j) &= \prod_{i=1}^n \frac{1}{y_{ij}^*!} e^{\lambda(y_{i,\neq j}) y_{ij}^*} e^{-e^{\lambda(y_{i,\neq j})}} \\
 &\approx \prod_{i=1}^n G\left(\lambda(y_{i,\neq j}) \mid \log(y_{ij}^*), \frac{1}{y_{ij}^*}\right)
 \end{aligned} \tag{16}$$

Since the variance $1/y_{ij}^*$ is likely to be similar across the documents i , we set $\hat{\sigma}_j^2 = \frac{1}{n} \sum_{i=1}^n \frac{1}{y_{ij}^*}$.

At this point, we specify the Laplace prior defined in Eq. (5) by

$$\Theta_{kj} \sim \text{Laplace}\left(0, \frac{\rho_{kj}}{2\sigma_j^2}\right). \tag{17}$$

We approximate this Laplace prior by a normal distribution with the same variance (Zeng et al. 2021). That is, we can approximate $\beta_j \sim \mathcal{N}(0, \frac{2}{\tau_j^2})$ by $\beta_j \sim \text{Laplace}(0, \tau_j)$. Therefore, we can approximate the distribution of edge parameters $\Theta_{\neq j,j}$ by

$$\Theta_{\neq j,j} \sim \mathcal{N}(0, V^j) \tag{18}$$

where $V^j \in \mathbb{R}^{(p-1) \times (p-1)}$ is a diagonal matrix with $V_{kk}^j = \frac{2\sigma_j^2}{e^{2Z_k^{j*} \alpha}}$ in which Z_k^{j*} is the k th row of the covariate matrix Z^{j*} . Now we can write out the marginal likelihood of α :

$$\begin{aligned}
L_j(\alpha) &= \int_{\mathbb{R}^p} \prod_{i=1}^n p(Y_{ij}^* \mid Y_{i,\neq j} = y_{i,\neq j}, \Theta_j) \prod_{k \neq j} p(\Theta_{kj} \mid \alpha) d\Theta_j \\
&= \int_{\mathbb{R}^p} \prod_{i=1}^n \frac{1}{y_{ij}^*!} e^{\lambda(y_{i,\neq j}) y_{ij}^*} e^{-e^{\lambda(y_{i,\neq j})}} \prod_{k \neq j} \frac{e^{(Z_k^* \alpha)} e^{-\frac{e^{(Z_k^* \alpha)}}{2\sigma_k^2}}}{4\sigma_k^2} |\Theta_{kj}| d\Theta_j \quad (19) \\
&\approx \int_{\mathbb{R}^p} \frac{|\sigma_j I_n|^{-1/2}}{(2\pi)^{\frac{N}{2}}} e^{-\frac{1}{2} \|\lambda(y_{\neq j}) - \log(y_j^*)\|_{\sigma_j I_n}^2} \frac{|V^j|^{-1/2}}{(2\pi)^{\frac{p}{2}}} e^{-\frac{1}{2} \|\Theta_j\|_{V^j}} d\Theta_j.
\end{aligned}$$

where $y_j^* = (y_{1j}^*, \dots, y_{nj}^*)^\top$, $y_{\neq j}$ denotes y excluding the j th column, and within the norm, $\lambda(\cdot)$ operates on the columns of $y_{\neq j}$ and $\log(\cdot)$ is applied element-wise. Dropping terms that are not a function of Θ_j , expanding the norm term and completing the square within the integral, we obtain that the approximate marginal log-likelihood of α satisfies

$$-l_j(\alpha) \propto \log |C_\alpha| + \log(y_j^*)^\top C_\alpha^{-1} \log(y_j^*) \quad (20)$$

where $C_\alpha = \sigma_j I^2 + y_{\neq j} V^j y_{\neq j}^\top$.

Funding Not applicable.

Availability of data and materials The data used in Sect. 5 is part of the Six Degrees of Francis Bacon project (SDFB). Warren et al. (2016) can be contacted for the data.

Code availability All code corresponding to the paper is available upon request.

Declarations

Conflict of interest Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aitchison J, Ho C (1989) DThe multivariate Poisson-log normal distribution. *Biometrika* 76(4):643–653
- Allen GI, Liu Z (2012) A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In: 2012 IEEE international conference on bioinformatics and biomedicine. IEEE, pp 1–6
- Almqvist ZW, Bagozzi BE (2019) Using radical environmentalist texts to uncover network structure and network features. *Sociol Methods Res* 48(4):905–960
- Bartlett MS, Kendall D (1946) The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Suppl J R Stat Soc* 8(1):128–138

- Bonato A, D'Angelo DR, Elenberg ER, Gleich DF, Hou Y (2016) Mining and modeling character networks. In: International workshop on algorithms and models for the web-graph. Springer, pp 100–114
- Boulesteix AL, De Bin R, Jiang X, Fuchs M (2017) IPF-LASSO: integrative-penalized regression with penalty factors for prediction based on multi-omics data. *Comput Math Methods Med*, pp 1–14
- Calvo-Armengol A, Jackson MO (2004) The effects of social networks on employment and inequality. *Am Econ Rev* 94(3):426–454
- Chan AB, Vasconcelos N (2009) Bayesian Poisson regression for crowd counting. In: 2009 IEEE 12th international conference on computer vision. IEEE, pp 545–551
- Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3):432–441
- Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33(1):1
- Harvard University, Academia Sinica, Peking University (2021) China biographical database (CBDB). <https://projects.iq.harvard.edu/cbdb>. Accessed 15 May 2021
- Kossinets G, Watts DJ (2009) Origins of homophily in an evolving social network. *Am J Sociol* 115(2):405–450
- Lauritzen SL (1996) Graphical models, vol 17. Clarendon Press, Oxford
- Li Y, Nan B, Zhu J (2015) Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics* 71(2):354–363
- Marsden PV (1990) Network data and measurement. *Annu Rev Sociol* 16(1):435–463
- Matthew HCG, Harrison B, Goldman L et al (2014) Oxford dictionary of national biography. Oxford University Press, Oxford
- McPherson JM, Smith-Lovin L (1982) Women and weak ties: differences by sex in the size of voluntary organizations. *Am J Sociol* 87(4):883–904
- McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. *Annu Rev Sociol* 27(1):415–444
- Meinshausen N, Bühlmann P et al (2006) High-dimensional graphs and variable selection with the lasso. *Ann Stat* 34(3):1436–1462
- Mohamed ZT (2020) Studies in early modern social networks, 1400–1750. Ph.D. thesis, Harvard University
- Newman ME (2004) Coauthorship networks and patterns of scientific collaboration. *Proc Natl Acad Sci* 101:5200–5205
- Prentice RL (1974) A log gamma model and its maximum likelihood estimation. *Biometrika* 61(3):539–544
- Robins G (2015) Doing social network research: network-based research design for social scientists. Sage, London
- SDFB (2021) Six Degrees of Francis Bacon. <http://www.sixdegreesoffrancisbacon.com/>. Accessed 15 May 2021
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B (Methodol)* 58(1):267–288
- Üsdiken B, Pasadeos Y (1995) Organizational analysis in North America and Europe: a comparison of co-citation networks. *Organ Stud* 16(3):503–526
- Warren CN, Shore D, Otis J, Wang L, Finegold M, Shalizi C (2016) Six degrees of Francis Bacon: a statistical method for reconstructing large historical social networks. *Digit Humanit Q* 10(3):1–16
- Winkler WE (1990) String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods, J Am Stat Assoc* 354–359
- Yang E, Ravikumar P, Allen GI, Liu Z (2012) Graphical models via generalized linear models. *NIPS* 25:1367–1375
- Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. *J R Stat Soc Ser B (Stat Methodol)* 68(1):49–67
- Zeng C, Thomas DC, Lewinger JP (2021) Incorporating prior knowledge into regularized regression. *Bioinformatics* 37(4):514–521
- Zou H (2006) The adaptive lasso and its oracle properties. *J Am Stat Assoc* 101(476):1418–1429

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Xiaoyi Yang¹ · Nynke M. D. Niezink¹ · Rebecca Nugent¹

✉ Xiaoyi Yang
xya32436@creighton.edu

¹ Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15206, USA