Overlapping mixture models for network data (`manet`) with covariates adjustment

Modelli di mistura a gruppi sovrapposti per dati network (`manet`) con covariate

Saverio Ranciati and Giuliano Galimberti and Ernst C. Wit and Veronica Vinciotti

Abstract Network data often come in the form of *actor-event* information, where two types of nodes comprise the very fabric of the network. Examples of such networks are: people voting in an election, users liking/disliking media content, or, more generally, individuals - *actors* - attending events. Interest lies in discovering communities among these actors, based on their patterns of attendance to the considered events. To achieve this goal, we propose an extension of the model introduced in [5]: our contribution injects covariates into the model, leveraging on parsimony for the parameters and giving insights about the influence of such characteristics on the attendances. We assess the performance of our approach in a simulated environment.

Abstract *I dati network vengono spesso strutturati sotto forma di informazioni* attore-evento*, ovvero network dove esistono due tipologie di nodi. Alcuni esempi sono: persone che votano durante le elezioni, utenti che esprimono preferenza o meno su contenuti multimediali, o, più in generale, individui* - attori - *che partecipano a eventi. L'interesse risiede nel rilevare la presenza di gruppi fra questi attori, cluster che si differenzino per la propensione nel partecipare agli eventi in esame. A tale scopo, proponiamo un'estensione del modello introdotto in [5]: il nostro contributo contempla la presenza di covariate nel modello, sfruttando quindi un*

Saverio Ranciati
Department of Statistical Sciences, University of Bologna, Via delle Belle Arti 41, Bologna, Italy, e-mail: saverio.ranciati2@unibo.it

Giuliano Galimberti
Department of Statistical Sciences, University of Bologna, Via delle Belle Arti 41, Bologna, Italy, e-mail: giuliano.galimberti@unibo.it

Ernst C. Wit
Johann Bernoulli Institute for Mathematics and Computer Sciences University of Groningen, 9747 AG Groningen, The Netherlands, e-mail: e.c.wit@rug.nl

Veronica Vinciotti
Department of Mathematics, Brunel University, London UB83PH, UK,
e-mail: veronica.vinciotti@brunel.ac.uk

approccio parsimonioso e dando potenziali informazioni sull'effetto delle caratteristiche considerate. Valutiamo la performance del nostro approccio in un ambiente simulato.

Key words: Bayesian inference, bimodal network, MCMC, probit regression

1 Introduction

Network data are becoming increasingly available and a propelling force in the pursuit of new methodological approaches devoted to analyze the complexity behind interactions among units in a system. A review about the methods and models adopted in this research area can be found in [3]. Some of these data come in the form of individuals attending events, or, more generally, a network structure where two different types of nodes exist: these are also called two-mode networks, bimodal networks, or affiliation networks [6, Chapter 8]. We focus on those data describing people's behavior with respect to attending or not a set of events, and we aim to discover if communities exist within the network itself, communities that differ in patterns of preferences to attend each event. A recent approach to do model-based clustering in this context was proposed by [5]: motivated by a dataset about terrorists participating to meetings and bombings, the authors introduced a mixture model for network data called `manet`, where each unit is allowed to potentially belong to more than one community. We build on their contribution and propose an extension of their model, in order to accomodate for external information about the network, in the form of covariates describing characteristics of the units or the events.

The main contributions of the paper are:

- extending `manet` [5] by introducing covariates into the model formulation;
- eliciting how some existing regression techniques can be used in the covariates-adjusted `manet` approach;
- providing results on a simulation study about the performances of our proposed model.

The remainder of the manuscript is organized as follows: in Section 2, first we outline `manet` original formulation, to familiarize the reader with the model's structure, and then we introduce the proposed extension; in Section 3, performance of `manet` with covariates adjustment is explored in a simulated environment; finally, in Section 4, we discuss the contribution and hint at future research trajectories.

2 Covariates adjustment for `manet`

Network data are organized in an $n \times d$ matrix Y of observations y_{ij}, collecting the attendances of $i = 1, \dots, n$ units - also called *actors* - to $j = 1, \dots, d$ events. Each realization y_{ij} comes from a binary random variable, with $y_{ij} = 1$ meaning individual i attends event j, and zero otherwise. We want to cluster these n actors based on their attendances via a model-based framework, and mixture models prove to be a suitable approach to achieve the task [1]. In the traditional setting, clusters are mutually exclusive and have (prior) sizes given by $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$. Usually, two conditions hold: (i) $\alpha_k \geq 0$, for each k; (ii) $\sum_{k=1}^{K} \alpha_k = 1$. Mixture models also have a hierarchical representation, attainable after introducing a unit-specific latent variable $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$: if actor i belongs to cluster k, the vector is full of zeros except for the k-th element $z_{ik} = 1$. Given the binary nature of response variables y_{ij}, for each actor $i = 1, \ldots, n$ we have

$$\begin{aligned} \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}; a_1, \ldots, a_K) \\ \mathbf{z}_i &\sim \text{Multinomial}(\mathbf{z}_i; \alpha_1, \ldots, \alpha_K) \\ \mathbf{y}_i|(\mathbf{z}_i, \boldsymbol{\pi}) &\sim \prod_{k=1}^{K} \left(\prod_{j=1}^{d} \pi_{kij}^{y_{ij}} (1 - \pi_{kij})^{1-y_{ij}} \right)^{z_{ik}} \end{aligned} \tag{1}$$

for some hyper-parameters $(a_1, \ldots, a_K)$; $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ij}, \ldots, y_{id})$ is the attendance profile of the i-th actor to the d events, which we assume - given $\mathbf{z}_i$ - to be independent for all $j, j' = 1, \ldots, d$ and $j \neq j'$. Vector $\boldsymbol{\pi}$ collects probabilities of attendance π_{kij} of a saturated model specification.

In many cases, one is interested in groups that are not mutually exclusive, allowing an *actor* to be allocated simultaneously to potentially more than a single cluster. We build on the approach suggested by [5], where a Bayesian **m**ultiple **a**llocation model for **net**work data (manet) is proposed. In manet, the hierarchical model in Equation 1 is modified by relaxing conditions on the cluster sizes $\boldsymbol{\alpha}$ and allocation vectors $\{\mathbf{z}_i\}$, allowing each actor to potentially belong to any number of the K clusters. The number of all possible group-allocating configurations is equal to $K^\star = 2^K$. Instead of working with the latent variables $\mathbf{z}_i$, a new $K^\star$-dimensional allocation vector $\mathbf{z}_i^\star$ is defined for each i. This vector satisfies $\sum_{h=1}^{K^\star} z_{ih}^\star = 1$, and a 1-to-1 correspondence exists between $\mathbf{z}_i$, which allocate actors into overlapping *parent* clusters, and $\mathbf{z}_i^\star$, which allocates actors into non-overlapping *heir* clusters. To this re-parametrization corresponds, in manet, the following hierarchical model

$$\begin{aligned} \boldsymbol{\alpha}^\star &\sim \text{Dir}(\boldsymbol{\alpha}^\star; a_1, \ldots, a_{K^\star}), \\ \mathbf{z}_i^\star|\boldsymbol{\alpha}^\star &\sim \text{Multinom}(\mathbf{z}_i^\star; \alpha_1^\star, \ldots, \alpha_{K^\star}^\star), \\ \mathbf{y}_i|\mathbf{z}_i^\star, \boldsymbol{\pi} &\sim \prod_{h=1}^{K^\star} \prod_{j=1}^{d} \left[\text{Ber}\big(y_{ij}; \pi_{hij}^\star\big) \right]^{z_{ih}^\star}. \end{aligned} \tag{2}$$

with prior $\pi_{kij} \sim \text{Beta}(\pi_{kij}; b_1, b_2)$, and (b_1, b_2) suitable hyper-parameters. For example, when $K = 2$, actor i may be assigned:

- to none of the two clusters, $\mathbf{z}_i = (0,0)$, corresponding to $\mathbf{z}_i^\star = (1,0,0,0)$;
- only to the first *parent* cluster, $\mathbf{z}_i = (1,0)$, corresponding to $\mathbf{z}_i^\star = (0,1,0,0)$;
- only to the second *parent* cluster, $\mathbf{z}_i = (0,1)$, corresponding to $\mathbf{z}_i^\star = (0,0,1,0)$;
- both of them $\mathbf{z}_i = (1,1)$, corresponding to $\mathbf{z}_i^\star = (0,0,0,1)$.

The advantage of working with re-parametrization in Equation 2 is that $\{\pi_{hij}\}$ are not additional parameters to be sampled, but probabilities of attendances produced by $\boldsymbol{\pi}$. For each actor i and event j, $\pi_{hij}^\star$ are computed via a function $\psi\big(\boldsymbol{\pi}_{\cdot ij}, \mathbf{z}_i\big)$, so that we obtain $\pi_{hij}^\star$ by looking at which parent clusters originated h, through the vector $\mathbf{z}_i$, and combining their corresponding probabilities $(\pi_{1ij}, \ldots, \pi_{Kij})$. We consider $\psi(\cdot) \equiv \min(\cdot)$. For the simple case where $K = 2$, an actor i belonging to both clusters, $\mathbf{z}_i = (1,1)$, deciding whether to attend an event j or not, will do so with probability $\pi^\star_{hij} = \psi(\pi_{1ij}, \pi_{2ij}) = \min(\pi_{1ij}, \pi_{2ij})$. When $\mathbf{z}_i = (0,0)$, $\pi^\star_{hij} = 0$. The saturated `manet` demands inference on $(K \times n \times d)$ probabilities of attendance, where each π_{kij} has only one observation to update the prior information with. In [5] authors prescribe a more feasible formulation for `manet` by setting π_{kij} to be only event- and cluster-specific, defining thus a quasi-saturated model with $\pi_{kij} \equiv \pi_{kj}$.

In this manuscript, we propose an extension of `manet` which introduces parsimony by exploiting covariates information. These covariates could be characteristic related to an actor, such as, gender, age, etc, or features of an event, i.e. type of event, date, duration, and so forth. We define $\mathbf{x}_{i\cdot} = (x_{i1}, \ldots, x_{il}, \ldots, x_{iL})$ to be the L-dimensional vector of covariates for actor i, and $\mathbf{w}_{j\cdot} = (w_{j1}, \ldots, w_{jq}, \ldots, w_{jQ})$ the Q-dimensional vector of covariates for event j. For simplicity, we assume the non-categorical covariates to be standardized, i.e. zero mean and unit variance. Covariates enter the model through a *link* function as in the generalized linear models context [4]. We add the following layers to the Bayesian hierarchical formulation

$$\begin{aligned}
&\boldsymbol{\mu}_k \sim \mathrm{N}(\mu_k; 0, \sigma^2_\mu), \quad \boldsymbol{\beta}_k \sim \mathrm{N}_L(\boldsymbol{\beta}_k; \mathbf{0}_L, \sigma^2_\beta I_L), \quad \boldsymbol{\gamma}_k \sim \mathrm{N}_Q(\boldsymbol{\gamma}_k; \mathbf{0}_Q, \sigma^2_\gamma I_Q),\\
&\eta_{kij} = \mu_k + \sum_{l=1}^{L} \beta_{kl} x_{il} + \sum_{q=1}^{Q} \gamma_{kq} w_{jq},\\
&\pi_{kij}(\mathbf{x}_{i\cdot}, \mathbf{w}_{j\cdot}) = g^{-1}(\eta_{kij}),
\end{aligned}$$

where: η_{kij} is the linear predictor; g^{-1} is the normal distribution's cumulative function $\Phi(\cdot)$, leading to a probit formulation; $\mathrm{N}(\cdot)$ is the normal distribution's density function, with subscripts denoting the dimension of vector or matrix; $(\sigma^2_\mu, \sigma^2_\beta, \sigma^2_\gamma)$ are hyper-parameters. Notice that, for g^{-1} set equal to the identity function and only $\{\mu_{kj}\}$ as parameters in the linear predictor, we revert back to the quasi-saturated `manet`. Once the probabilities π_{kij} are obtained from linear predictors η_{kij}, the corresponding heir parameters $\pi^\star_{hij}$ can be computed by means of the combining function $\psi(\cdot)$, exactly as in the formulation without covariates in [5]. Computing linear predictors requires regression coefficients and intercepts to be sampled, and this can be done separately for each cluster as they are independent. However, when an actor i belongs to multiple clusters, it is not univocally defined to which posterior distribution among the K sets of $(\mu_k, \beta_{k1}, \ldots, \gamma_{kQ})$ its likelihood term will contribute to. We follow the prescription of [5] to disentangle this issue. We introduce auxiliary variables $s(\boldsymbol{z}_i, \boldsymbol{\pi}) = s_{kij}$, such that, for a fixed (i,j) we have: $s_{kij} = z_{ik}$ if $\sum_{k=1}^{K} z_{ik} = 1$, whereas, if $\sum_{k=1}^{K} z_{ik} > 1$ then $\boldsymbol{s}_{ij\cdot}$ is a K-dimensional vector of zeros, except for $s_{ik_{\min},j} = 1$. Here $k_{\min}$ denotes the index corresponding to the *parent* cluster having the lowest value of η_{kij}, for a fixed event j and actor i. After introducing the auxiliary variables $\{s_{kij}\}$ into the model, the complete-data likelihood becomes

$$\mathscr{L}_{Y,Z}(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\gamma}; S) = \prod_{k=1}^{K} \prod_{j=1}^{d} \prod_{i=1}^{n} \left\{ \Big[\Phi(\eta_{kij})\Big]^{y_{ij}} \Big[1 - \Phi(\eta_{kij})\Big]^{1-y_{ij}} \right\}^{s_{kij}}. \tag{3}$$

Equation 3 is similar to the likelihood of a binary regression model with probit link function. Also, Equation 3 highlights that, in general, the model can be cast in a regression framework and thus, potentially, other regression techniques and/or extensions can be further exploited to refine manet. An outline of the general idea behind the MCMC implementation is: (i) first, all the observations Y are stacked into a vector $\tilde{\mathbf{y}}$ of length $(n \cdot d) \times 1$; (ii) for each k, a vector $\tilde{\mathbf{y}}_k$ is obtained by filtering $\tilde{\mathbf{y}}$ with rule $s_{kij} = 1$, and then a Bayesian probit regression is performed. For this last part, we refer to the work of [2], where the authors discuss a hierarchical Bayesian formulation of the probit model and provide technical details in the manuscript's Appendix. The MCMC algorithm for manet with covariates adjustment is implemented in an R script, and code is available upon request.

3 Simulation study

We generate 50 independent datasets from a probit model, with $K = 2$ overlapping groups; sample size and number of events are fixed to $n = 100$ and $d = 15$. Covariates are: (i) actor-specific categorical covariate with two levels, same proportions for both levels (50/100), coded with a single binary variable x_{i1}; (ii) event-specific categorical covariate with three levels, same proportions for the three levels (5/15), coded with two binary variables w_{j1} and w_{j2}. For each replicated dataset, we run: (i) our algorithm, labelled manet+cov; (ii) manet, with the homonymous R package. Number of MCMC iterations is set to 10000 with burn-in window equal to 4000. Results are numerically reported in terms of Adjusted Rand Index and misclassification error rate, averaged across the replicated datasets for both models, showed in percentage. Adjusted Rand Index (ARI) is a measure in the range $[0,1]$, with higher values indicating better performance. The misclassification error rate (MER) quantifies the proportion of wrongly allocated units, with smaller values indicating better performance. In terms of classification accuracy, manet and its covariates-adjusted extension attain comparable ARI and MER, with results slightly in favor of manet+cov: more specifically, average ARI is 79.06% for manet and 81.26% for manet+cov, while average MER is 8.92% for manet and 8.10% for manet+cov. This corroborates the idea of employing covariate information if available, as the performance of the parsimonious manet+cov is on par with the more flexible quasi-saturated manet. Results for manet+cov are also visualized (see Figure 1) through posterior distributions of the regression coefficients, plotted after aggregating chains from all the independent datasets. Despite the additional uncertainty introduced by combining MCMC samples from different datasets, the posterior distributions show good behavior in terms of location and scale: all the densities in Figure 1 are centered around the true values used to generate data, and exhibit limited dispersion.

4 Conclusions

We have proposed an extension of the model formulated in [5], in order to accomodate for additional information in the form of actor and/or event covariates. By cast-

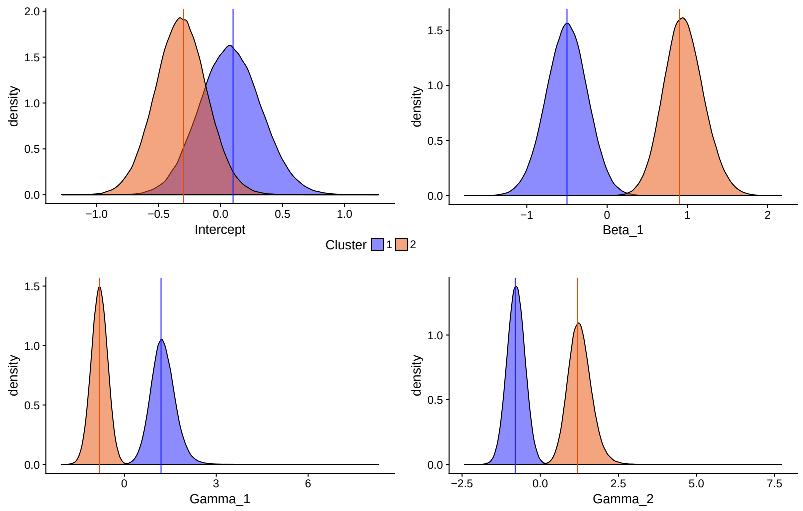

Fig. 1 Posterior distributions for the regression coefficients of `manet+cov`, computed after aggregating all the independent datasets' chains. True values are depicted as vertical lines in the plots, corresponding to parameters for: intercept, $\boldsymbol{\mu} = (0.1, -0.3)$; covariate x_{i1}, $\boldsymbol{\beta}_{\cdot 1} = (-0.5, 0.9)$; covariate w_{j1}, $\boldsymbol{\gamma}_{\cdot 1} = (1.2, -0.8)$; covariate w_{j2}, $\boldsymbol{\gamma}_{\cdot 2} = (-0.8, 1.2)$.

ing part of the inference problem into a binary regression framework, we highlighted the link between regression techniques and covariates-adjusted `manet`, paving the way for more interesting future refinements of the model - such as mixed effects, regularized regression, and so forth. We explored the performance of our proposed algorithm in a simulated environment, which shows appreciable preliminary outputs and encouraging results.

References

1. Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer Science & Business Media, New York (2006)
2. Holmes, C.C., Held, L.: Bayesian auxiliary variable models for binary and multinomial regression. Bayesian analysis **1**(1), 145–168 (2006)
3. Kolaczyk, E.D.: Statistical Analysis of Network Data: Methods and Models. Springer, New York (2009)
4. McCulloch, C.E., Neuhaus, J.M.: Generalized linear mixed models. Wiley Online Library (2001)
5. Ranciati, S., Vinciotti, V., Wit, E.C.: Identifying overlapping terrorist cells from the noordin top actor-event network. arXiv preprint arXiv:1710.10319 (2017)
6. Wasserman, S., Faust, K.: Social network analysis: Methods and applications, vol. 8. Cambridge university press (1994)