2019

Missing Data Augmentation for Bayesian Multiplex ERGMs

Robert Krause
University of Groningen

Alberto Caimo
Technological University Dublin, alberto.caimo@tudublin.ie

Recommended Citation

Krause, R. W. and Caimo, A. (2019). Missing Data Augmentation for Bayesian Exponential Random Multi-Graph Models.*International Workshop on Complex Networks,* vol. 221, pg. 63–72. doi:10.21427/PME8-MT48

Missing Data Augmentation for Bayesian Multiplex ERGMs

Robert W. Krause[1][0000−0003−4288−4732]
Alberto Caimo[2][0000−0001−8956−7166]

[1] University of Groningen, 9712 TS Groningen, The Netherlands
r.w.krause@rug.nl
[2] Dublin Institute of Technology, Dublin D08 X622, Ireland
alberto.caimo@dit.ie

Abstract. In this paper we present an estimation algorithm for Bayesian multiplex exponential random graphs (BmERGMs) under missing network data. Social actors are often connected with more than one type of relation, thus forming a multiplex network. It is important to consider these multiplex structures simultaneously when analyzing a multiplex network. The importance of proper models of multiplex network structures is even more pronounced under the issue of missing network data. The proposed algorithm is able to estimate BmERGMs under missing data and can be used to obtain proper multiple imputations for multiplex network structures. It is an extension of Bayesian exponential random graphs (BERGMs) as implemented in the Bergm package in R. We demonstrate the algorithm on a well known example, with and without artificially simulated missing data.

Keywords: Social Networks · Missing Data · Multiple Imputation · ERGM · Bayesian Computation · Multiplex Networks.

1 Introduction

In recent years, it is becoming more and more apparent that the understanding of social structure often requires to take more than just one type of social relation into account, so called multiplex networks. Notable examples include the important interrelations between friendships and advice seeking behavior [17], the importance of antipathi-ties in the maintenance of friendship group structures [18], and the relationships between joined drug use, sexual relations, and co-visitation of social venues [5]. However, increasing the number of network items collected from the participants is likely to increase missing data. First, social network data collection is often time intensive and adding more items is likely to increase drop out. Second, social network data is by nature interpersonal and thus often sensitive data, thus more likely to not be fully reported by participants. While missing data is a problem for all (social) sciences, network models suffer particularly under missing data, because of the strong dependencies within the data. Non-response by one participant does not only mean we

know less about this participant, but we also know less about the social network of all other participants, after all, the missing participant could have nominated any of the other participants thus potentially changing the network structure drastically.

Statistical tools have been developed to handle missing social network data, both to obtain more reliable model estimates [6, 16, 9, 10], as well as reliable descriptive statistics [11]. In this paper we propose an extension of this work to the context of multiplex exponential random graph models (mERGMs). We advance the literature twofold, first, by proposing an estimation procedure for Bayesian mERGMs, and second by proving proper multiple imputation of missing multiplex network data.

2 Bayesian Multiplex ERGMs

2.1 Bayesian inference for ERGMs

The exponential random graph model family (ERGMs; [12]) is most commonly used to analyze cross-sectional network data. ERGMs model an observed network, or graph, as a function of sufficient network statistics (primarily counts of subgraph configurations, e.g., number of ties, number of reciprocated ties or number of transitive triplets). A network graph is expressed as a random $n \times n$ adjacency matrix Y with $Y_{ij} = 1$ when there is tie from node i to node j and $Y_{ij} = 0$ when there is no tie. Usually, edges connecting a node with itself are not allowed ($Y_{ii} = 0$). Networks can be directed or undirected ($Y_{ij} = Y_{ji}$). Let $\mathcal{Y}$ denote the set of all possible networks on n nodes and let y be a realization of $Y \in \mathcal{Y}$. Then, in Bayesian ERGMs (BERGMs) the posterior probability of the parameters conditional on the data is given by

$$p(\theta|y) = \frac{\exp\left[\theta^T s(y)\right]}{z(\theta)} \frac{p(\theta)}{p(y)}, \tag{1}$$

with θ being a vector of model parameters, $s(y)$ a vector of corresponding sufficient network statistics, $z(\theta)$ the normalizing constant, $p(\theta)$ the prior distribution of the parameters and $p(y)$ is the marginal probability. See Lusher et al. for an introduction to ERGMs [12].

2.2 Multiplexity

Multiplex networks are structures with multiple different types of relations on the same set of nodes. Multiplex networks can thus be expressed as a random $n \times n \times m$ adjacency array or cube Y with $Y_{ijm} = 1$ when there is tie from node i to node j on network m and $Y_{ijm} = 0$ when there is no such tie on network m. Each layer m of the multiplex network can be either directed or undirected. Multiplex ERGMs were first introduced by Pattison and Wasserman [14] and later extended by Wang [20]. Multiplexity increases the complexity of network models by an additional factor, while a single layer directed network has $2^{n \times n - n}$

possible configurations (e.g., a network of 20 nodes has $\sim 2.5 \times 10^{114}$ possible configurations), this number increases exponentially to the number of layers, $2^{(n\times n-n)\times m}$ (e.g., a multiplex network of 20 nodes with 2 layers has $\sim 6.1\times 10^{228}$ possible configurations).

2.3 Posterior Parameter Estimation for BmERGMs

The MCMC estimation algorithm of the posterior $p(\theta|y)$ is an extension of the approximate exchange algorithm introduced by Caimo and Friel [1] and currently implemented in the Bergm package in R [3]. The algorithm samples from the following distribution:

$$p(\theta', y', \theta|y) \propto p(y|\theta)\ p(\theta)\ \epsilon(\theta'|\theta)\ p(y'|\theta'), \tag{2}$$

with $p(y'|\theta')$ being the likelihood on which the simulated data y' are defined and belongs to the same exponential family of densities as $p(y|\theta)$, $\epsilon(\theta'|\theta)$ is any arbitrary proposal distribution for the parameter θ'. This proposal distribution is set to be a normal centred at θ.

At each MCMC iteration, the exchange algorithm consists of three main steps: First, proposing a Gibbs update of θ'. Followed by a Gibbs update of y', a drawn from $p(\cdot|\theta')$ with an MCMC algorithm [8]. Third an exchange, or swap, from the current state θ to the proposed new parameter θ' is taken. This deterministic proposal is accepted with the following probability:

$$\min\left(1, \frac{q_\theta(y')\ p(\theta')\ \epsilon(\theta|\theta')\ q_{\theta'}(y)}{q_\theta(y)\ p(\theta)\ \epsilon(\theta'|\theta)\ q_{\theta'}(y')} \times \frac{z(\theta)\ z(\theta')}{z(\theta')\ z(\theta)}\right) \tag{3}$$

where q_θ and $q_{\theta'}$ indicate the unnormalized likelihoods for parameters θ and θ', respectively. The intractable normalizing constants cancel each other out in this equation, thus avoiding the problem of calculating them.

Concretely, the algorithm is implemented in the following way:

Algorithm 1 Approximate exchange algorithm for BmERGMs

Initialise θ
for $i = 1, \ldots, N$ **do**
  Generate θ' from $\epsilon(\cdot|\theta)$
  loop
    for $m = 1, \ldots, M$ **do**
      Simulate one (or a few) tie swap y'_m from $p(\cdot|\theta', y')$
    end for
  end loop
  Update $\theta \to \theta'$ with the log of the probability:

$$\min\left(0, [\theta - \theta']^T[s(y') - s(y)] + \log\left[\frac{p(\theta')}{p(\theta)}\right]\right) \tag{4}$$

end for

The key change to the regular Bergm algorithm is in the network simulation loop which is here sampling from a multiplex network space. Instead of directly simulating a new multiplex network y' with the proposed parameter θ', the simulation is done iteratively for each of the M layers of ties by proposing one (or a few) tie swap on each layer, conditional on the proposed parameter vector θ' and on the current state of y', thus including all tie swaps simulated on all layers of the network in this and previous iterations. This is repeated until convergence is reached and a sample is drawn from $p(\cdot|\theta')$. Adaptive procedures such as the adaptive direction sampling [1, 19] or the delayed rejection sampling [2] can be adopted.

2.4 Cross Network Effects

Currently three fundamental dyadic cross network effects are implemented for the algorithm. These effects are: 1) co-occurrence, 2) entrainment, and 3) cross network reciprocity. Co-occurrence expresses the tendency of edges on one layer to occur with edges on another layer in an undirected graph and entrainment is its directed counterpart. The corresponding sufficient statistic can thus be calculated similarly for both:

$$s_{co|ent}(y) = \sum_{i<j} y_{ij1} y_{ij2}. \tag{5}$$

Cross network reciprocity models the co-occurrence of outgoing ties of one type with incoming ties of another type on the same dyad.

$$s_{cross-recip.}(y) = \sum_{i<j} y_{ij1} y_{ji2}. \tag{6}$$

3 Missing Data Imputation

The proposed missing data augmentation procedure is an extension of the work by Koskinen et al. [9]. In short, every time a new θ' is accepted in the algorithm outlined above, the missing network data is imputed conditional on the observed data and θ'. The imputation follows a similar simulation procedure as the parameter estimation. However, only tie-swaps for missing ties are proposed. The obtained multiplex network y^* is then used as the starting point for the next iteration and treated as new baseline. Thus equation (4) optimizes $[s(y') - s(y^*)]$, and not $[s(y') - s(y)]$.

The imputed networks y^* can be retained after the estimation of the posterior $p(\theta|y)$ and used for additional analyses, because they constitute proper multiple imputations of y, assuming a well fitting model.

This algorithm has been shown to provide reliable estimates of $p(\theta|y)$ [9], and low biases in descriptive statistics [11] in the single-layer network setting. However, if y is a multiplex network, it is important to impute missing data with a multiplex network model to guarantee that the observed relationships

Algorithm 2 Approximate exchange algorithm for BmERGMs under missing data

Use naive imputation obtain starting values for $s(y^*)$
Initialise θ
for $i = 1, \ldots, N$ **do**
  Generate θ' from $\epsilon(\cdot|\theta)$
  loop
    for $m = 1, \ldots, M$ **do**
      Simulate one (or a few) tie swap y'_m from $p(\cdot|\theta', y')$
    end for
  end loop
  Update $\theta \rightarrow \theta'$ with the log of the probability:

$$\min\left(0, [\theta - \theta']^T [s(y') - s(y*)] + \log\left[\frac{p(\theta')}{p(\theta)}\right]\right) \tag{7}$$

  if $\theta' \in p(\theta|y)$ **then**
    loop
      for $m = 1, \ldots, M$ **do**
        Simulate one (or a few) tie swap of missing ties y^*_m from $p(\cdot|\theta', y^*)$
      end for
    end loop
  end if
end for

between the layers are maintained in the imputation process. Thus, this algorithms provides an important advancement in the treatment of missing network data.

4 Illustration - Florentine Families

As a simple illustration we present Padgett's network of the Florentine banking families, a classical example of network analysis [13]. The network consists of 16 nodes (the banking families), their business relations, and their martial connections (fig. 1). We present only a simple model for the multiplex graph for illustrative purposes. The within network layer effects are similar for both business and marriage network. The model consists of a set of parameters for edges (modeling the density), geometrically weighted degree (GWDEGREE - modeling the degree distribution) and geometrically weighted edgewise shared partners (GWESP - modeling triadic closure) [7, 15]. Additionally, the model includes a parameter for the co-occurrence of ties between the layers.

Missing data was created by randomly selecting three (∼20%) of the families and setting their outgoing and incoming ties and no-ties for both layers to missing. The posterior distributions for the complete data model as well as for the missing data model are presented jointly in fig. 2. The missing data augmentation algorithm performs well in approximating the posterior of the full data model.

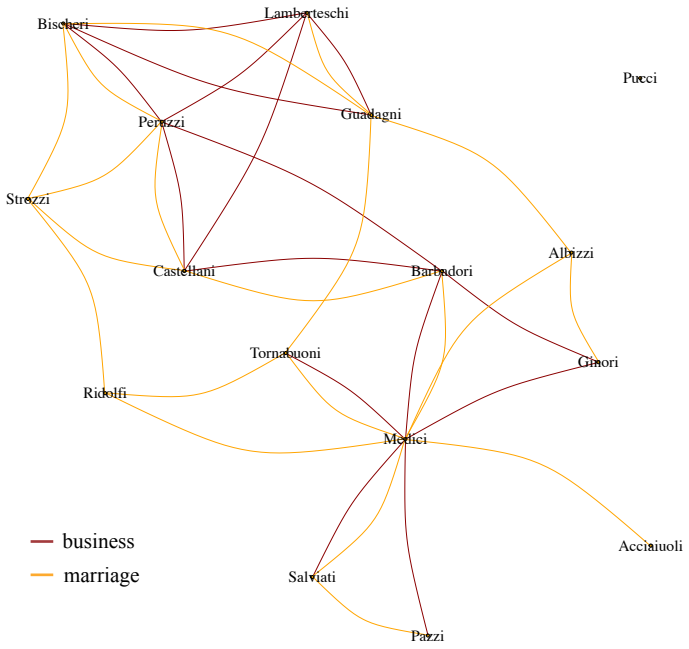

Fig. 1. Business and marriage relations of the 16 Florentine families

5 Discussion

In this paper, we present a Bayesian computational algorithm for the estimation of multiplex exponential random graphs under missing data. The code implementing the methodology is currently available on GitHub and in future will be part of the Bergm package in R. It is thus far the only implementation of multiplex random graphs in R. Currenly, Bergm, and by extension the proposed algorithm, are heavily reliant on the ergm package. Unfortunately, ergm does not facilitate estimation of multiplex ERGMs, which limits the availability of cross network effects. The proposed algorithm can be easily adapted to estimate Bayesian (multiplex) exponential random network models [4], an extension of the ERG-family models where also nodal attributes are random and dependent on the connectivity structure of the network. The estimation of this joint network and attribute distribution can be implemented similarly to the estimation of the multiplex structure.

References

1. Caimo, A., Friel, N.: Bayesian inference for exponential random graph models. Social Networks **33**(1), 41 – 55 (2011)

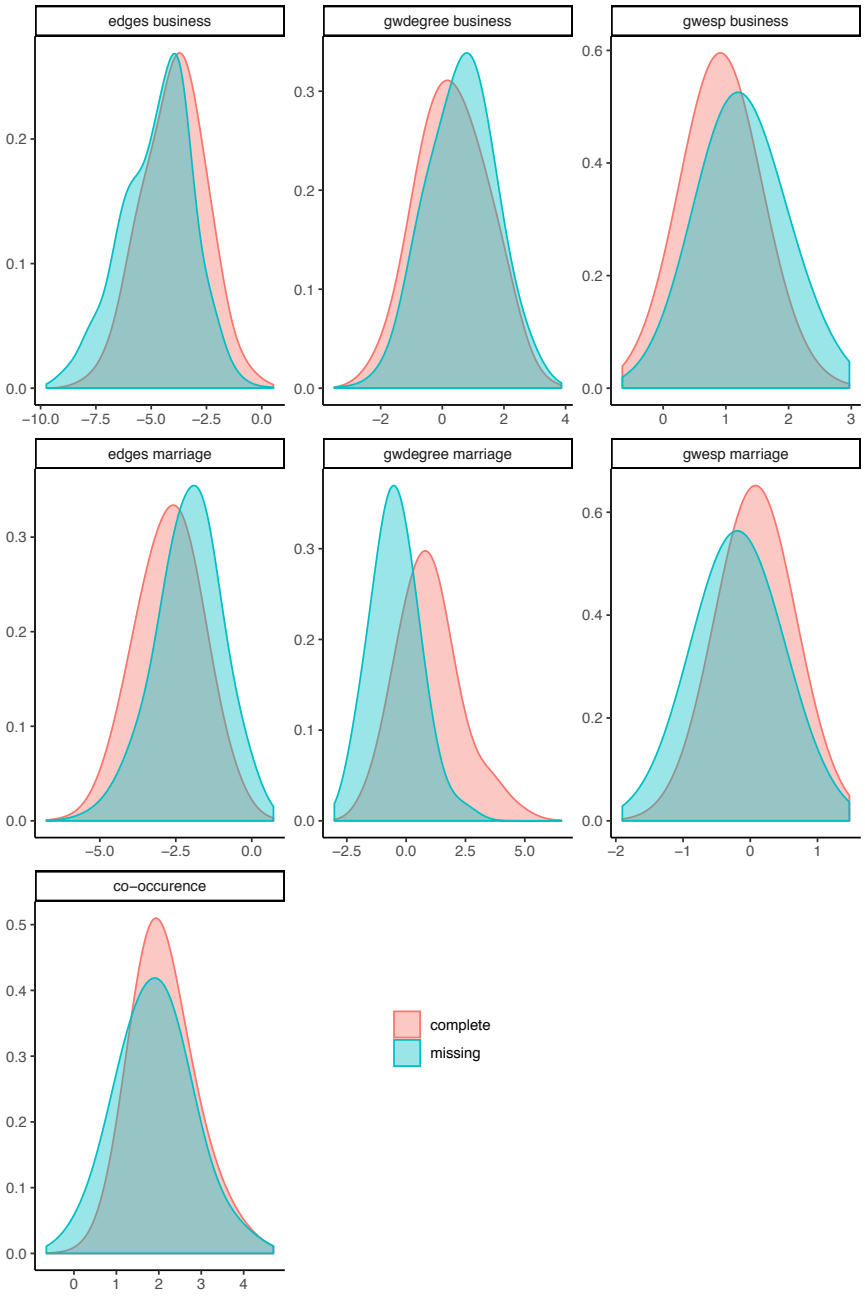

Fig. 2. Posterior Distributions of BmERGM for complete and missing data

2. Caimo, A., Mira, A.: Efficient computational strategies for doubly intractable problems with applications to Bayesian social networks. Statistics and Computing **25**, 113–125 (2015)
3. Caimo, A., Friel, N.: Bergm: Bayesian exponential random graphs in R. Journal of Statistical Software **61**(2), 1–25 (2014), http://www.jstatsoft.org/v61/i02/
4. Fellows, I., Handcock, M.S.: Exponential-family random network models. arXiv preprint arXiv:1208.0121 (2012)
5. Fujimoto, K., Wang, P., Ross, M.W., Williams, M.L.: Venue-mediated weak ties in multiplex hiv transmission risk networks among drug-using male sex workers and associates. American Journal of Public Health **105**(6), 1128–1135 (2015)
6. Handcock, M.S., Gile, K.J.: Modeling social networks from sampled data. The Annals of Applied Statistics **4**(1), 5 (2010)
7. Hunter, D.R., Handcock, M.S.: Inference in curved exponential family models for networks. Journal of Computational and Graphical Statistics **15**, 565–583 (2006)
8. Hunter, D.R., Handcock, M.S., Butts, C.T., Goodreau, S.M., Morris, M.: ergm: A package to fit, simulate and diagnose exponential-family models for networks. Journal of Statistical Software **24**(3), 1–29 (2008), http://www.jstatsoft.org/v24/i03
9. Koskinen, J.H., Robins, G.L., Pattison, P.E.: Analysing exponential random graph (p-star) models with missing data using bayesian data augmentation. Statistical Methodology **7**(3), 366–384 (2010)
10. Krause, R.W., Huisman, M., Snijders, T.A.: Multiple imputation for longitudinal network data. Italian Journal of Applied Statistics **30**, 33–57 (2018)
11. Krause, R.W., Huisman, M., Steglich, C., Snijders, T.A.: Missing network data a comparison of different imputation methods. Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (2018)
12. Lusher, D., Koskinen, J., Robins, G.: Exponential random graph models for social networks: Theory, methods, and applications. Cambridge University Press (2013)
13. Padgett, J.F., Ansell, C.K.: Robust action and the rise of the medici, 1400-1434. American Journal of Sociology **98**(6), 1259–1319 (1993)
14. Pattison, P., Wasserman, S.: Logit models and logistic regressions for social networks: Ii. multivariate relations. British Journal of Mathematical and Statistical Psychology **52**(2), 169–193 (1999)
15. Snijders, T.A.B., Pattison, P.E., Robins, G.L., S., H.M.: New specifications for exponential random graph models. Sociological Methodology **36**, 99–153 (2006)
16. Snijders, T.A., Koskinen, J., Schweinberger, M.: Maximum likelihood estimation for social network dynamics. The Annals of Applied Statistics **4**(2), 567 (2010)
17. Snijders, T.A., Lomi, A., Torló, V.J.: A model for the multiplex dynamics of two-mode and one-mode networks, with an application to employment preference, friendship, and advice. Social networks **35**(2), 265–276 (2013)
18. Stadtfeldt, C., Takács, K., Vörös, A.: The emergence and stability of groups in social networks. SSRN, https://ssrn.com/abstract=3232958 or http://dx.doi.org/10.2139/ssrn.3232958 (2018)
19. Thiemichen, S., Friel, N., Caimo, A., Kauermann, G.: Bayesian exponential random graph models with nodal random effects. Social Networks **46**, 11–28 (2016)
20. Wang, P.: Ergm extensions: models for multiple networks and bipartite networks. Exponential Random Graph Models for Social Networks: Theory, Methods, Applications pp. 115–129 (2012)