# SIS 2017
# Statistics and Data Science: new challenges, new generations

28–30 June 2017
Florence (Italy)

# Proceedings of the Conference of the Italian Statistical Society

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.
(Proceedings e report ; 114)

http://digital.casalini.it/9788864535210

ISBN 978-88-6453-521-0 (online)

*Peer Review Process*
All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published on the website and in the online catalogue of the FUP (www.fupress.com).

*Firenze University Press Editorial Board*
A. Dolfi (Editor-in-Chief), M. Boddi, A. Bucelli, R. Casalbuoni, M. Garzaniti, M.C. Grisolia, P. Guarnieri, R. Lanfredini, A. Lenzi, P. Lo Nostro, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, G. Nigro, A. Perulli, M.C. Torricelli.

This work is licensed under a Creative Commons Attribution 4.0 International License
(CC BY 4.0: https://creativecommons.org/licenses/by/4.0/legalcode)

CC 2017 Firenze University Press
Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

**SOCIETÀ ITALIANA DI STATISTICA**
Sede: Salita de' Crescenzi 26 - 00186 Roma
Tel +39-06-6869845 - Fax +39-06-68806742
email: sis@caspur.it web:http://www.sis-statistica.it

La Società Italiana di Statistica (SIS), fondata nel 1939, è una società scientifica eretta ad Ente morale ed inclusa tra gli Enti di particolare rilevanza scientifica. La SIS promuove lo sviluppo delle scienze statistiche e la loro applicazione in campo economico, sociale, sanitario, demografico, produttivo ed in molti altri settori di ricerca.

**Organi della società:**

*Presidente:*
- Prof.ssa Monica Pratesi, Università di Pisa

*Segretario Generale:*
- Prof.ssa Filomena Racioppi, Sapienza Università di Roma

*Tesoriere*:
- Prof.ssa Maria Felice Arezzo, Sapienza Università di Roma

*Consiglieri:*
- Prof. Giuseppe Arbia, Università Cattolica del Sacro Cuore
- Prof.ssa Maria Maddalena Barbieri, Università Roma Tre
- Prof.ssa Francesca Bassi, Università di Padova
- Prof. Eugenio Brentari, Università di Brescia
- Dott. Stefano Falorsi, ISTAT
- Prof. Alessio Pollice, Università di Bari
- Prof.ssa Rosanna Verde, Seconda Università di Napoli
- Prof. Daniele Vignoli, Università di Firenze

*Collegio dei Revisori dei Conti:*
- Prof. Francesco Campobasso, Prof. Michele Gallo, Prof. Francesco Sanna, Prof. Umberto Salinas (supplente)

**SIS2017 Committees**

**Scientific Program Committee:**
Rosanna Verde (chair), Università della Campania "Luigi Vanvitelli"
Maria Felice Arezzo, Sapienza Università di Roma
Antonino Mazzeo, Università di Napoli Federico II
Emanuele Baldacci, Eurostat
Pierpaolo Brutti, Sapienza Università di Roma
Marcello Chiodi, Università di Palermo
Corrado Crocetta, Università di Foggia
Giovanni De Luca, Università di Napoli Parthenope
Viviana Egidi, Sapienza Università di Roma
Giulio Ghellini, Università degli Studi di Siena
Ippoliti Luigi, Università di Chieti-Pescara "G. D'Annunzio"
Matteo Mazziotta, ISTAT
Lucia Paci, Università Cattolica del Sacro Cuore
Alessandra Petrucci, Università degli Studi di Firenze
Filomena Racioppi, Sapienza Università di Roma
Laura M. Sangalli, Politecnico di Milano
Bruno Scarpa, Università degli Studi di Padova
Cinzia Viroli, Università di Bologna

**Local Organizing Committee:**
Alessandra Petrucci (chair), Università degli Studi di Firenze
Gianni Betti, Università degli Studi di Siena
Fabrizio Cipollini, Università degli Studi di Firenze
Emanuela Dreassi, Università degli Studi di Firenze
Caterina Giusti, Università di Pisa
Leonardo Grilli, Università degli Studi di Firenze
Alessandra Mattei, Università degli Studi di Firenze
Elena Pirani, Università degli Studi di Firenze
Emilia Rocco, Università degli Studi di Firenze
Maria Cecilia Verri, Università degli Studi di Firenze

**Supported by:**
Università degli Studi di Firenze
Università di Pisa
Università degli Studi di Siena
ISTAT
Regione Toscana
Comune di Firenze
BITBANG srl

# Index



Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

# A Bayesian semiparametric model for terrorist networks

## *Un modello Bayesiano semiparametrico per reti terroristiche*

Emanuele Aliverti

**Abstract** A recent field of research employs network-analysis' tools to the *dark network* framework, in which pairwise informations about terrorists' activities are available. In this work we focus on the "Noordin Mohamed Top" dataset, developing an asymmetric approach that treats one network as response and the remaining as covariates. The objective is to identify which information may be useful in predicting terrorists' collaboration in a bombing attack, identifying at the same time the most influential subjects involved in these dynamics. Such aim is addressed through an asymmetric Bayesian semi-parametric model for networks that, through a suitable prior specification, integrates a flexible regularization and the detection of leading nodes. Taking advantage of the Pólya-Gamma data augmentation scheme, we develop an efficient Gibbs sampler to make inference on the parameters involved.
**Abstract** *Un recente ambito di ricerca impiega strumenti tipici dell'analisi di reti nei contesti di* dark networks*, nei quali sono disponibili informazioni riguardanti attività terroristiche sotto forma di legami a coppie. In questo lavoro ci conentraimo sul dataset relativo a "Noordin Mohamed Top", sviluppando un approccio asimmetrico che considera una particolare rete come risposta, e le rimanenti come esplicative. L'obiettivo identificare quale informazione possa essere utile per predirre la collaborazione di diversi terroristi in un attentato, identificando contemporaneamente i pi influenti soggetti coinvolti in queste dinamiche. Il problema é affrontato tramite un modello Bayesiano semiparametrico per reti che, attraverso un opportuno specificazione delle distribuzioni a priori, incorpora al suo internouna regolazione flessibile e l'identificazioe dei nodi leader. Sfruttando lo schema Pólya-Gamma per dati aumentati, presentiamo un efficiente Gibbs* sampler *per fare inferenza sui parametri coinvolti.*

**Key words:** Terrorism, networks, Bayesian semiparametrics, latent space, spike-and slabs prior,matrix factorization

Emanuele Aliverti
Università di Padova, Dipartimento di Scienze Statistiche e-mail: aliverti@stat.unipd.it



Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

## 1 Introduction

After September 11th, intelligence agencies of different countries employed tools of the network science to serve in the fight against terroristic groups, often named *dark networks*. Great effort has been made to develop tools for identifying key players, that is actors within the network reporting high values in terms of some suitable network statistics. Since aggressive strategies encountered different failures, and the necessity of more sophisticated approaches became evident: [7] for example propose to focus on approaches less aggressive than direct military operations, involving a subtle application of informatics tools in order to gather different informations from various sources. The proper interpretation of retrieved data may provide a deeper description of terrorism, embracing at the same time social, economics and personal aspects, thus useful to develop strategies to defeat the roots of criminals associations.

Our motivating approach rises from the "Noordin Mohamed Top" dataset, drawn from a publication of the International Crisis Group; it consists of different ties among terrorists of the most ruthless group of the southwest Asia.
Data are coded into 10 symmetric relationships between network's leader, Noordin Mohamed Top, and 78 affiliates, thus naturally coded into a *multilayer simple graphs*, that is a structure $G = \{V, E_k\}$ where nodes (elements of $V$) represent terrorists and edges (unordered pairs situated in the set $E_k$) the presence of the particular $k$-th relationship among two generic subjects.

We expect a certain degree of association among different relationships, since they're defined over the same set of nodes. Therefore, we would like to propose an approach able to efficiently use the information held inside "simpler" network in order to predict and make inference on the most interesting one, which is the network referred to the co-participation at the same terroristic bombing..

## 2 Proposed approach

Our research objectives can be faced by setting up an asymmetric framework, that threats one network as response and the remaining as covariates. The proposal of [3] is the most appropriate, and hence we will adapt this approach to our purposes by including nodal random effects and a non-parametric matrix factorization that avoids the estimation of different models. Let $v$ the number of nodes of each network, $\mathbf{Y}$ the $v \times v$ adjacency matrix referred to the response network and $\underline{X}$ the $v \times v \times p$ array containing the $p$ adjacency matrices referred to the $p$ explanatory networks. We will consider only undirected and unweighted network (simple graphs), so adjacency matrices associated at are all dichotomous, symmetric and with non-defined elements on the main diagonal. Hence $y_{ij} = y_{ji} \in \{0,1\}$ and $x_{ijk} = x_{jik} \in \{0,1\}\ \forall i, j, k$. Since the response network can assume only two values (presence or absence of edges), it is reasonable to assume a conditional bernoulli distribution for the under-

lying generative mechanism. We parametrize $\pi_{ij}$, the probability of observing an edge between node $i$ and node $j$, through its log-odds $\theta_{ij}$. Formally:

$$y_{ij}|\pi_{ij} \overset{ind}{\sim} \text{Bin}(1,\pi_{ij}) \qquad \theta_{ij} = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)$$

Furthermore, we decompose the linear predictor $\theta_{ij}$ into two components: the first can be regarded as a parametric mixed model component, while the second as a non-parametric matrix factorization.

$$\theta_{ij} = \underbrace{\alpha + \sum_{k=1}^{p} [\beta_k + b_{ik} + b_{jk}] x_{ijk}}_{\substack{\text{Parametric component:} \\ \text{fixed and random effects}}} + \underbrace{\sum_{h=1}^{H} \lambda_h z_{ih} z_{jh}}_{\substack{\text{Non-parametric} \\ \text{component}}} \tag{1}$$

$$\alpha \in \mathbb{R}, \quad \beta_k \in \mathbb{R} \quad k = 1,\dots,p \quad z_{ih} \in \mathbb{R}, \lambda_h \in \mathbb{R}^+ \quad i = 1,\dots,v$$

$$\mathbf{b}_i = (b_{i1},\dots,b_{ip}) \sim \mathbf{G}, \quad i = 1,\dots,v$$

The parametric component describes the relationship between networks and detects potentially influential nodes, which in this application means subjects whose role in some relationships has been particularly different from the average one. The basic interpretation is the following: $\alpha$ provides an indication of the density of the response network, as an ordinary intercept in the binomial regression. Coefficient $\beta_k$ are *fixed effects* in a logistic regression, that is the mean variation in the log odds of the outcome attributable to the $k$-th explicative network. In order to take advantage of the explicative power of covariates networks, we introduce additive *random effects* referred to the generic nodes $i$ and $j$ involved in the $(i,j)$-th dyad. In 1 $b_{ki}$ represents the specific deviation of the $i$-th node from the main effect $\beta_k$, and so can account for his particular propensity in building ties in the response network. For each relationship, the purpose is to identify subjects more (or less) likely to commit an attack with, providing thus a brighter description of those dynamics. Furthermore, additive random effects can account for between-rows heterogeneity contained in the explanatory networks, allowing then a better estimation of the fixed counterpart.

The non-parametric component decompose the residual among response and explanatory networks in a flexible way, that is through a matrix factorization that allows the number of factors to vary adaptively. It can be interpreted as a latent space whose size is at most equal to $H$, in which $z_{ih}$ represents the $h$-th latent coordinate of the $i$-th node, while $\lambda_h$ defines the importance of the $h$-th dimension of the latent space in defining the final model. This strategy aims to adaptively account for the dependencies in the response not seized by explanatory networks, providing estimates for the parametric component deprived of potentially confounding factors.

## 3 Prior distribution and posterior simulation

For a complete Bayesian definition of the proposed model we need to specify proper prior distributions for the set of parameters involved.

### *3.1 Parametric component*

We specify zero-mean normal distributions over the fixed effects parameter. Formally,

$$\pi(\alpha) \sim N(\mu_{0_\alpha}, \sigma^2_{\alpha_0}) \qquad \pi(\beta) = \pi(\beta_1, \ldots, \beta_p) \sim N_p(\mu_{0_\beta}, \Sigma_{\beta_0}) \tag{2}$$

In our application, we expect a certain level of heterogeneity in nodes' behavior, both between different subject and within the same, when involved in different relationships. For example, it's reasonable that dealing directly with leaders may led to a higher propensity in participating at the same terroristic attack. However, certain subjects may have had a central role just in the some specific relationships, such as the school recruitment network, and a marginal position elsewhere; for that, we need a prior distribution able to differentiate particular subjects from standard ones, and hence we specify a *spike and slabs* prior distributions [4] independently for each $p$-dimensional vector referred to the generic $i$-th subject, $i = 1, \ldots, v$. Formally:

$$\begin{gathered}
\mathbf{G} \sim N(0, \Gamma_i), \quad \Gamma_i = \text{diag}(\gamma_{i1}, \ldots, \gamma_{ip}), \quad \gamma_{ik} = \theta_{ik}\, \tau^2_{ik}, \quad k = 1, \ldots, p \\
\pi(\theta_{ik}) \overset{iid}{\sim} (1 - w_i)\, \delta_{v_0}(\cdot) + w_i \delta_1(\cdot) \\
\pi(\tau^{-2}_{ik}) \sim \text{Gamma}(d_1, d_2), \quad \pi(w_i) \sim \text{Uniform}\,[0, 1]
\end{gathered} \tag{3}$$

In 3 $v_0$ is a value close to zero, and the hyper-parameters $d_1, d_2$ are chosen in order to obtain, for $\gamma_k = \theta_k\, \tau^2_k$, a continuous distribution characterized by a spike in $v_0$ and a continuous right tail; $\delta_{v_0}$ and $\delta_1$ are Formally, a Multiplicative Inverse Gamma (MIG) is specified as prior probability measure over the loading elements $\lambda_h$ in 1, and standard Gaussian distribution for the latent coordinates. See [2] for a recent discussion regarding the properties of the MIG prior. Formally:

$$\begin{gathered}
z_{ih} \overset{iid}{\sim} N(0, 1), \qquad i = 1, \ldots, v \\
\lambda_h = \prod_{m=1}^{h} \frac{1}{\theta_m}, \quad \theta_1 \sim \text{Gamma}(a_1, 1), \quad \theta_{h \geq 2} \overset{iid}{\sim} \text{Gamma}(a_2, 1)
\end{gathered} \tag{4}$$

with $a_1 > 0$ and $a_2 > 1$ fixed hyper parameters.

### *3.2 Posterior Simulation*

Adapting the Pólya-Gamma data augmentation strategy proposed by [6] in the logistic regression framework, we can obtain the full-conditional distributions for the parameters involved in our model, and hence implement a Gibbs sampling strategy.

## 4 Results

The effects of different network is heterogeneous: for example, a tie in the communication network increments, in mean, the log odds of collaborating in the same bombing operations of around 2 times; furthermore, if two terrorist had been in the same terroristic organization the log odds is lowered of an amount around 1.33 times, that is not so trivial. As for influential nodes, the spike and slabs strategy identify several terrorists, confirmed to be such in the Indonesian reports. The predictive performance recorded an an average area under the ROC curve equal to 0.864, a false positive rate equal to 0.225 and a total negative rate of 0.220, using as estimates for the missing edges the mean of the posterior predictive density and, where needed, the overall density of the response network as cutoff value.

## References

1. Bhattacharya, A and D B Dunson (2011). Sparse Bayesian infinite factor models. In: *Biometrika* 98.2, pp. 291-306.
2. Durante, Daniele (2017). A note on the multiplicative gamma process. In: *Statistics & Probability Letters* 122, pp. 198-204.
3. Hoff, Peter (2008). Modeling homophily and stochastic equivalence in symmetric rela- tional data. In: Advances in Neural Information Processing Systems, pp. 657-664.
4. Ishwaran, H., and Rao, J. S. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies". *Annals of Statistics*, 730-773.
5. Polson, Nicholas G., James G. Scott, and Jesse Windle (2013). Bayesian Inference for Logistic Models Using Pólya-Gamma Latent Variables. In: *Journal of the American Statistical Association* 108.504, pp. 1339-1349.
6. Roberts, Nancy and Sean F Everton (2011). Strategies for Combating Dark Networks. In: *Journal of Social Structure* 12,