Proceedings

17th Applied Stochastic Models and Data Analysis
International Conference with
Demographics Workshop

ASMDA2017

Editor

Christos H Skiadas

6 - 9 June 2017

De Morgan House, London, UK

Imprint
Proceedings of the 17th Applied Stochastic Models and Data Analysis International Conference with the 6th Demographics Workshop
London, UK: 6-9 June, 2017

Published by: ISAST: International Society for the Advancement of Science and Technology.
Editor: Christos H Skiadas

ISBN (e-book) 978-618-5180-23-2

Preface

It is our pleasure to welcome the guests, participants and contributors to the International Conference (ASMDA 2017) on Applied Stochastic Models and Data Analysis and (DEMOGRAPHICS2017) Demographic Analysis and Research Workshop.

The main goal of the conference is to promote new methods and techniques for analyzing data, in fields like stochastic modeling, optimization techniques, statistical methods and inference, data mining and knowledge systems, computing-aided decision supports, neural networks, chaotic data analysis, demography and life table data analysis.

ASMDA Conference and DEMOGRAPHICS Workshop aim at bringing together people from both stochastic, data analysis and demography areas. Special attention is given to applications or to new theoretical results having potential of solving real life problems.

ASMDA 2017 and DEMOGRAPHICS 2017 focus in expanding the development of the theories, the methods and the empirical data and computer techniques, and the best theoretical achievements of the Applied Stochastic Models and Data Analysis field, bringing together various working groups for exchanging views and reporting research findings.

We thank all the contributors to the success of these events and especially the authors of this Proceedings Book. Many thanks to the honorary guest Gilbert Saporta and the Colleagues contributed in his special session on data analysis. Special thanks to the Plenary, Keynote and Invited Speakers, the Session Organisers, the Scientific Committee, the ISAST Committee, Yiannis Dimotikalis, Aristeidis Meletiou, the Conference Secretary Mary Karadima, and all the members of the Secretariat.

November 2017

Christos H. Skiadas
Conference Chair

ASMDA Conferences and Organizers

1st ASMDA 1981 Brussels, Belgium. Jacques Janssen

2nd ASMDA 1983 Brussels, Belgium. Jacques Janssen

3rd ASMDA 1985 Brussels, Belgium. Jacques Janssen

4th ASMDA 1988 Nancy, France. J. Janssen and Jean-Marie Proth

5th ASMDA 1991 Granada, Spain. Mariano J. Valderrama

6th ASMDA 1993 Chania, Crete, Greece. Christos H Skiadas

7th ASMDA 1995 Dublin, Ireland. Sally McClean

8th ASMDA 1997 Anacapry, Italy. Carlo Lauro

9th ASMDA 1999 Lisbon, Portugal. Helena Bacelar-Nicolau

10th ASMDA 2001 Compiègne, France. Nikolaos Limnios

11th ASMDA 2005 Brest, France. Philippe Lenca

12th ASMDA 2007 Chania, Crete, Greece. Christos H Skiadas

13th ASMDA 2009 Vilnious,Lithouania. Leonidas Sakalauskas

14th ASMDA 2011 Rome, Italy. Raimondo Manca

15th ASMDA 2013 Mataró (Barcelona), Spain. Vladimir Zaiats

16th ASMDA 2015 Piraeus, Greece. Sotiris Bersimis

17th ASMDA 2017 London, UK. Christos H Skiadas

SCIENTIFIC COMMITTEE

Claude Lefevre, Université Libre de Bruxelles, Belgium
Mei-Ling Ting Lee, University of Maryland, USA
Philippe Lenca, Telecom Bretagne, France
Nikolaos Limnios, Université de Techonlogie de Compiègne, France
Bo H. Lindquvist, Norvegian Institute of Technology, Norway
Brunero Liseo, University of Rome, Italy
Fabio Maccheroni, Università Bocconi, Italy
Claudio Macci, University of Rome "Tor Vergata", Italy
P. Mahanti. Dept. of Comp. Sci. and Appl. Statistics, Univ. of New Brunswick, Canada
Raimondo Manca, University of Rome "La Sapienza", Italy
Domenico Marinucci, University of Rome "Tor Vergata", Italy
Laszlo Markus, Eötvös Loránd University – Budapest, Hungary
Sally McClean, University of Ulster
Gilbert MacKenzie, Univerity of Limerick, Ireland
Terry Mills, Bendigo Health and La Trobe University, Australia
Leda Minkova, Dept. of Prob., Oper. Res. and Stat. Univ. of Sofia, Bulgaria
Ilya Molchanov, University of Berne, Switzerland
Karl Mosler, University of Koeln, Germany
Amílcar Oliveira, UAb-Open University in Lisbon, Dept. of Sciences and Technology and CEAUL-University of Lisbon, Portugal
Teresa A Oliveira, UAb-Open University in Lisbon, Dept. of Sciences and Technology and CEAUL-University of Lisbon, Portugal
Annamaria Olivieri, University of Parma, Italy
Enzo Orsingher, University of Rome "La Sapienza", Italy
T. Papaioannou, Universities of Pireaus and Ioannina, Greece
Valentin Patilea, ENSAI, France
Mauro Piccioni, University of Rome "La Sapienza", Italy
Ermanno Pitacco, University of Trieste, Italy
Flavio Pressacco University of Udine, Italy
Pere Puig, Dept of Math., Group of Math. Stat., Universitat Autonoma de Barcelona, Spain
Yosi Rinott, The Hebrew University of Jerusalem, Israel
Jean-Marie Robine, Head of the res. team Biodemography of Longevity and Vitality, INSERM U710, Montpellier, France
Leonidas Sakalauskas, Inst. of Math. and Informatics, Vilnius, Lithuania
Werner Sandmann, Dept. of Math., Clausthal Univ. of Tech., Germany
Gilbert Saporta, Conservatoire National des Arts et Métiers, Paris, France
W. Sandmann, Dept. of Mathematics, Clausthal University of Technology, Germany
Lino Sant, University of Malta, Valletta, Malta
José M. Sarabia, Department of Economics, University of Cantabria, Spain
Sergio Scarlatti, University of Rome "Tor Vergata", Italy
Hanspeter Schmidli, University of Cologne, Germany
Dmitrii Silvestrov, University of Stockholm, Sweden
P. Sirirangsi, Chulalongkorn University, Thailand
Christos H. Skiadas, Technical University of Crete, Greece (Co-Chair)
Charilaos Skiadas, Hanover College, Indiana, USA
Dimitrios Sotiropoulos, Techn. Univ. of Crete, Chania, Greece
Fabio Spizzichino, University of Rome "La Sapienza", Italy
Gabriele Stabile, University of Rome "La Sapienza", Italy
Valeri Stefanov, The University of Western Australia
Anatoly Swishchuk, University of Calgary, Canada
R. Szekli, University of Wroclaw, Poland
T. Takine, Osaka University, Japan
Andrea Tancredi, University of Rome "La Sapienza", Italy
P. Taylor, University of Melbourne, Australia
Cleon Tsimbos, University of Piraeus, Greece
Mariano Valderrama, University of Granada, Spain
Panos Vassiliou, Department of Statistical Sciences, University College London, UK
Larry Wasserman, Carnegie Mellon University, USA
Wolfgang Wefelmeyer, Math. Institute, University of Cologne, Germany
Shelly Zacks, Binghamton University, State University of New York, USA
Vladimir Zaiats, Universitat de Vic, Spain
K. Zografos, Department of Mathematics, University of Ioannina, Greece

Plenary/Keynote Talks
For ASMDA Conference

In celebration of Gilbert Saporta's 70th birthday and in honour of his contributions to Applied Statistics and Data Analysis and his support to ASMDA activities

Gilbert Saporta
Emeritus Professor of Applied Statistics
Conservatoire National des Arts et Métiers (CNAM)
Paris, France

N. Balakrishnan
Department of Mathematics and Statistics
McMaster University
Hamilton, Ontario, Canada

Robert J. Elliott
Haskayne School of Business,
University of Calgary, Canada and
Centre for Applied Financial Studies,
University of South Australia,
Adelaide, Australia

Sally McClean
School of Computing and Information Engineering
Ulster University
Coleraine
Northern Ireland

Fabrizio Ruggeri
CNR IMATI
Via Bassini 15
Milano, Italy

Anatoliy Swishchuk
Department of Mathematics and Statistics
University of Calgary, Canada

P.-C.G. VASSILIOU
Department of Statistical Sciences,
University College London, UK

For Demographics Workshop

Jean-Marie Robine
Université Montpellier 2, Place Eugène Bataillon
Montpellier, France

Rebecca Kippen
Rural Health,
Monash University
Victoria, Australia

Contents **Page**

Saporta at Seventy

Pieter M. Kroonenberg,

Emeritus Professor at the Department of Education and Child Studies, Leiden University and The Three-Mode Company, Leiden

Abstract. This paper is an introduction to the Keynote lecture by Prof. Gilbert Saporta at the occasion of his seventieth birthday. An overview of this major publications, his citation record, his academic non-statistical interests is presented as well as a pictorial overview.

1. Introduction

The *Applied Stochastic Models and Data Analysis International Society* (ASMDA) decided to pay a special tribute to Prof. Gilbert Saporta of the *Centre National des Arts et Métiers,* Paris at the occasion of his 70[th] birthday. Clearly such a tribute is not bestowed upon just any septuagenarian. If his contributions to applied statistics and data analysis and his support to ASMDA activities themselves were not already enough for such a tribute, his nomination as Président d'Honneur de la Société Française de Statistique, made just before the conference, is additional proof that Prof. Saporta is not an average man.
.

In his keynote lecture entitled "50 Years of data analysis: from EDA to predictive modelling and machine learning" Prof Saporta sketches what has taken place in data analysis during his academic career, but this introduction will concentrate on some of the highlights of his publishing career, looking at his key publications, his citation record and his presence at various statistical gremia. A full curriculum vitae of Prof. Saporta can be found at the CNAM site: *http://cedric.cnam.fr/~saporta/CVSaporta_english_April2017.pdf*.

2. Publication records and their citations

There are at present several organisations, publishers and individuals who provide citation records of individual academics and academic groups. Two of the older ones are the ISI *Web of Science* and *Google Scholar*. Given that the latter includes books and more publications in languages other than English, I have taken Google Scholar as the basis for the information presented in this article -- although its use is not without difficulty. Anne-Wil Harzing has created a program *Publish or Perish,* which uses Google Scholar as its data base. In this program she calculates various statistics about publications, satisfying specific search terms (authors, subjects, research groups, etc.). One unfortunate circumstance is that academics are human, too, and not uncommonly references to their colleagues' work are not completely accurate. Given the automated character of data gathering by Google Scholar, such inaccuracies are generally not detected, so that multiple variants of the same publications can be found in the data base, and hence also in that of Harzing's *Publish or Perish* database. Therefore, this article contains such inaccuracies as well, but they would be too time-consuming and too difficult to rectify. I have tried to eliminate some of the more glaring ones, but more will have remained.

ResearchGate indicates that Prof. Saporta obtains a (albeit somewhat ResearchGate-specific) score which exceeds the scores of 70% of other researchers on its site. I would imagine that if all his publications were uploaded on this site he would easily score in the 90s.

Incidentally, it turns out that references to Prof. Saporta's work also appear under "S. Gilbert" (see Table 1). The probable reason is that algorithms gathering information on a person need to allocate publications of "G. Saporta", "Gilbert Saporta", "Saporta, Gilbert", "Saporta, G" to the same person, but "Saporta Gilbert" (without

17[th] *ASMDA Conference Proceedings, 6 - 9 June 2017, London, UK*

the “,”) also occurs. How is the algorithm to know what which is the first name and which is the family name? Note that on the same line Jean-Marie Bourouche has been reduced to a mere Mr. B.

Table 1. Citations to publications by S. Gilbert (Source: *Publish or Perish*, 18/6/2017)

Cites	Per year	Rank	Authors	Title	Year	Publication	Publisher
h 19	0.70	2	S Gilbert	Probabilités, analyse des données et statistique	1990	Paris, Éditions Technip	
h 8	1.00	1	H Wang, M Ye, S Gilbert	Classification for Multiple Linear Regression Methods [J]	2009	Journal of System Simulation	en.cnki.com.cn
h 7	0.30	4	B Jean-Marie, S Gilbert	L'analyse des données	1994	Que Saisje	
2	0.06	3	S Gilbert	Multidimensional data analysis for categorical variables	1985		Matrtinus Nujholf Publishe
2	1.00	7	D Jean-Jacques, S Gilbert, TA Christine	Méthodes robustes en statistique	2015		books.google.com

An additional aspect is that Prof. Saporta has published in both French and English and that for the casual investigator such as me it is unclear whether some English publications are straightforward translations of the French ones or vice versa. Finally, do we count various editions of the same book as different publications, or as the same publication? I have merged the results of the citation analysis so that in these cases all references were to the same publication. This leads to higher citation counts for those books, but I think this is only proper.

3. Saporta’s productivity

Let us first look at Prof. Saporta’s productivity as found in *Publish or Perish* (Fig. 1), but only counting those publications which have been cited at least once.

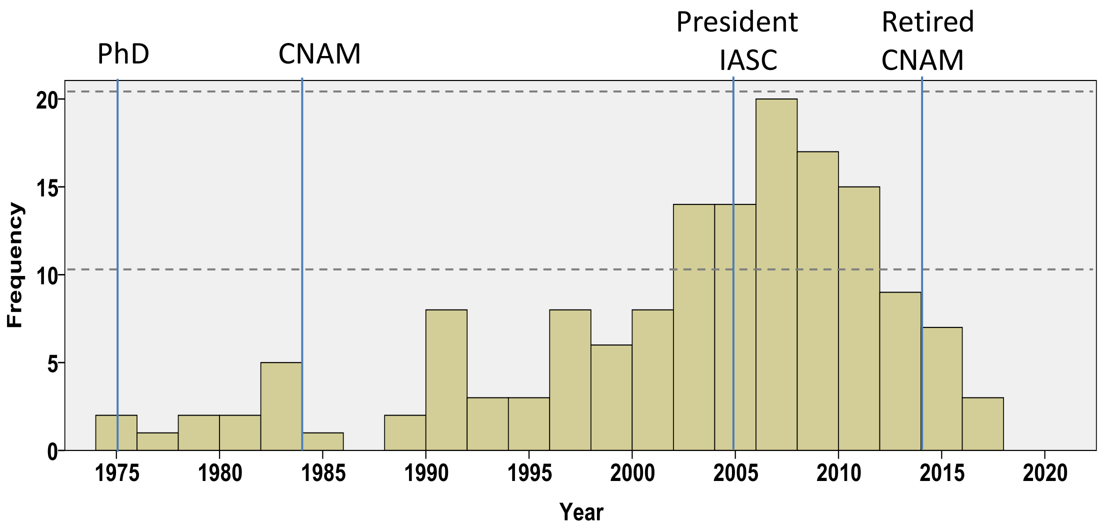


Figure 1. Number of publications cited at least once, arranged per year.

Figure 1 clearly shows that Prof. Saporta’s peak productivity was in his sixties. His publications included not only cited journal papers but also several books, including textbooks from which many generations of French students were taught (and hopefully learned) statistics; in particular *Probabilités, analyse des données et statistique*, which so far has known three editions (2006, 1991, 2011).

As a slightly frivolous exercise I asked Google to produce images of the covers of his books, which resulted in Figure 2. I have not edited the results, so there are some rogue and fantasy ‘covers’ included here as well. The one I loved best was the second from the right on the top row. It reads "*L’Analyse des données* (French Edition)". Why ‘French Edition’? Who would have been surprised that this book was not written in English? The solution to this riddle is that it is actually not a real cover (as stated almost illegibly in this figure), but a place holder for the real one, as is the first one of the same row. The actual covers of the two books from the *Que sais-je* series are given in Figure 3.

Figure 2. Covers of books (co)authored and/or (co)edited by Gilbert Saporta.

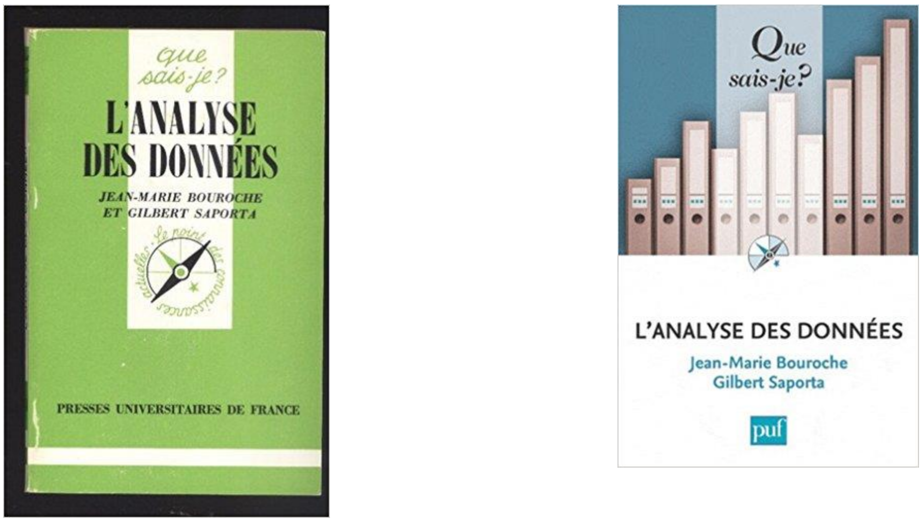


Figure 3. The real covers of the first and last editions of *L'Analyse des Données* from the *Que sais-je* series.

4. Saporta's prominent publications

In Table 2 below I present the results of a search for "Gilbert Saporta" in *Publish or Perish*. The outcomes are ranked according to frequency of citation. Prof. Saporta has an *h* index of 24, which means that on 17 June 2017, 24 of his publications had 24 citations or more, and it is those publications which are included in the list. The number of citations is a lower bound, because incorrect referencing has created new entries in the database. However, these citations should be part of the record of the correctly referenced publications.

The results in Table 2 make very clear that Prof. Saporta's books have been widely used, and one could even wonder what their citation count would have been had they also been available in English, the *lingua franca* of the scientific world. Finally, it is interesting to note how widely read and cited his two academic *thèses* have been. Not many scholars have that honour; of course it may be that this more usual in France than in the English-speaking world, but this does not diminish the acknowledged importance of these theses.

Measuring Latent Variables is space and/or time: A Gender Statistics exercise

Gaia Bertarelli[1], Franca Crippa[2], and Fulvia Mecatti[3]

[1] University of Perugia, Perugia, Italy
(E-mail: gaia.bertarelli@unipg.it)
[2] University of Milano-Bicocca, piazza dell'Ateneo Nuovo 1, 20126, Milano, Italy
(E-mail: franca.crippa@unimib.it)
[3] University of Milano-Bicocca, via Bicocca degli Arcimboldi 8, 20126, Milano, Italy
(E-mail: fulvia.mecatti@unimib.it)

Abstract. This paper concerns a Multivariate Latent Markov Model recently introduced in the literature for estimating latent traits in social sciences. Based on its ability of simultaneously dealing with longitudinal and spacial data, the model is proposed when the latent response variable is expected to have a time and space dynamic of its own, as an innovative alternative to popular methodologies such as the construction of composite indicators and structural equation modeling. The potentials of the proposed model and the added value with respect to the traditional weighted composition methodology, are illustrated via an empirical Gender Statistics exercise, focused on gender gap as the latent status to be measured and based on supranational official statistics for 30 European countries in the period 2010-2015.

Keywords: Latent clustering, Longitudinal data, Spatial ordering, Gender Gap.

1 Introduction

Composite indicators have the advantage of synthesizing a latent, multidimensional construct in a single number, usually included in the interval (0; 1). They can be derived as a weighted sum of simple indexes, as it is often the case in social statistics, specially when the set of indexes needs to stay unchanged in several geographic areas and/or time periods. In complex settings, the synthetic indicator is conceivable as a latent variable, typically estimated applying Structural Equation Models (SEM) in order to obtain a single measure.
When the latent variable is thought to have a time and-or space dynamic of its own, Multivariate Latent Markov Models (LMMs) may represent a valuable innovation to the construction of composite indicators. LMMs are a particular class of statistical models for the analysis of longitudinal data which assume the existence of a latent process affecting the distribution of the response variables [2] for a review). The rationale of this methodology considers the latent process as fully explained by the observable behaviour of some items, together with available covariates. The main assumption is conditional independence of the response variables given the latent process, which follow a first order discrete Markov chain with a finite number of states. The model is composed of two parts, analogously to SEM: the *measurement model*, concerning the conditional distribution of the response variables given the latent process, and the *latent*

17^{th} *ASMDA Conference Proceedings, 6 - 9 June 2017, London, UK*

model, pertaining the distribution of the latent process. LMMs can account for measurement errors or unobserved heterogeneity between areas in the analysis. LMMs main advantage is that the unobservable variable is allowed to have its own dynamics and it is not constrained to be time constant. In addition, when the latent states are identified as different subpopulations, LMMs can identify a latent clustering of the population of interest, with areas in the same subpopulation having a common distribution for the response variables. Under this respect, a LMM may be seen as an extension of the latent class (LC) model, in which areas are allowed to move between the latent classes during the observational period. Available covariates can be included in the latent model and then they may affect the initial and transition probabilities of the Markov chain. When covariates are included in the measurement model, the latent variables are used to account for the unobserved heterogeneity and the main interest is on a latent variable which is measured through the observable response variables (e.g., health status or gender inequalities) and on the evaluation of this latent variable depending on covariates. We focus on an extended model of the second type, as we are interested in ordinal latent states.
Very recently, Markov models for latent variables have contributed to in-depth investigations in highly specific and therefore narrow topics [**?**]. Extensive analyses of LMMs, both methodological and applicative, have been performed in the case of small area estimation, taking also into account several points in time [**?**]. Our viewpoint aims to adjust the LMMs approach to a wider area of synthetic social indicators in different geographical areas and in time, namely for national gender gap between countries. Gender statistics are defined as statistics that adequately reflect differences and inequalities in the situation of women and men in all areas of life [8]. Composite gender indicators are usually computed as weighted sum of simple indexes reflecting the multidimensionality of the phenomena and they are periodically released by supranational agencies (see for instance [6] for a comparative review.
We focus on gender gap as the latent status, since this construct is actually a latent trait, measurable only indirectly through a collection of observable variables and indicators purposively selected as micro-aspects that contribute to the latent macrodimension, aiming to add sensitiveness and discrimination power with respect to current indicators.

2 The proposed model

In this paper we use an extension of LMM proposed by Bertarelli [**?**]. The existence of two process is assumed: an observed process ca be expressed as:

$$Y_{jit}, \quad j = 1, \dots, J, \quad i = 1, \dots, n \text{ and } t = 1, \dots, T \tag{1}$$

where Y_{itj} denote the response variable j for unit i at time t, and an unobservable finite-state first-order Markov Chain

$$U_{it}, \quad i = 1, \dots, n \text{ and } t = 1, \dots, T \text{ with state space } \{1, \dots, m\}. \tag{2}$$

We assume that the distribution of Y_{jit} depends only on U_{it}; specifically the Y_{jit} are conditionally independent given U_{it}.

We also denote by $\tilde{\boldsymbol{U}}_{it} = \{U_{jt}, j \in \mathcal{G}_\rangle\}$, where $\mathcal{G}_i$ is the set of the neighbours, the latent states realisations in the neighborhood units.

In the *measurement model* we consider two Gaussian state-dependent distributions:

$$\begin{aligned} Y_{1it}|U_{it} &\sim N(\mu_1, \nu_1), \\ Y_{2it}|U_{it} &\sim N(\mu_2, \nu_2). \end{aligned} \tag{3}$$

The set of parameters of the *structural model*, corresponding to the latent Markov chain, includes the vector of initial probabilities

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_u, \ldots, \pi_m)^{'}, \tag{4}$$

where

$$\pi_u = P(U_{i1} = u)$$

is the probability of being in state u at the initial time for $u = 1, \ldots, m$ and the elements of the transition probability matrix

$$\boldsymbol{\Pi} = \{\pi_{u|\bar{u}},\ \bar{u}, u = 1, \ldots, m\}, \tag{5}$$

where

$$\pi_{u|\bar{u}} = P(U_{it} = u|U_{i,t-1} = \bar{u})$$

is the probability that unit i visits state u at time t given that at time $t-1$ it was in state $\bar{u}$.

Considering spatial dependence is a crucial point in our field of application [**?**]. As in [**?**], we propose to handle spatial dependence introducing a covariate in the structural model based on the information from a neighboring matrix and depending on the latent structure itself. In this way, the influence of spatial structure depends on the latent process, therefore it is not fixed during the observation period.

For each unit i we know the number of neighbouring units, g_i and their corresponding labels which are collected in the sets G_i. Let $\tilde{\boldsymbol{U}}_{it}$ be the vector of latent states at occasion t for the neighbours of unit i. We suppose to handle ordinal latent states in order to model the severity of the gender gap. Let us consider a function $\boldsymbol{\eta}(\cdot)$ that maps the g_i-dimensional vector $\tilde{\boldsymbol{U}}_{it}$ onto a $d-$dimensional covariate, the choice of $\boldsymbol{\eta}$ depending on the nature of latent states (ordinal or not). Due to our application context, we decide to work with the mean of neighbourhood latent states. Then, this time-varying covariate affects the initial and transition probabilities through the following multinomial logit parametrization:

$$\log \frac{p(U_{i1} = u|\tilde{\boldsymbol{U}}_{i1} = \tilde{\boldsymbol{u}}_{i1})}{p(U_{i1} = 1|\tilde{\boldsymbol{U}}_{i1} = \tilde{\boldsymbol{u}}_{i1})} = \beta_{0u} + \boldsymbol{\eta}(\tilde{\boldsymbol{u}}_{i1})'\boldsymbol{\beta}_{1u} \quad \text{for } u \geq 2, \tag{6}$$

$$\log \frac{p(U_{it} = u|U_{i,t-1} = \bar{u}, \tilde{\boldsymbol{U}}_{it} = \tilde{\boldsymbol{u}}_{it})}{p(U_{it} = \bar{u}|U_{i,t-1} = \bar{u}, \tilde{\boldsymbol{U}}_{it} = \tilde{\boldsymbol{u}}_{it})} = \gamma_{0u\bar{u}} + \boldsymbol{\eta}(\tilde{\boldsymbol{u}}_{it})'\boldsymbol{\gamma}_{1u\bar{u}}, \quad \text{for } t \geq 2 \text{ and } u \neq \bar{u}, \tag{7}$$

where $\boldsymbol{\beta_u} = (\beta_{0u}, \boldsymbol{\beta}_{1u}^{'})'$ and $\boldsymbol{\gamma}_{u\bar{u}} = (\gamma_{0u\bar{u}}, \boldsymbol{\gamma}_{1u\bar{u}}^{'})'$ are vectors of parameters to be estimated. An individual covariate has been introduced, accordingly both the assumptions of local independence and of a first order latent process still hold.

3 Estimation and Inference

To estimate the proposed model. we adopt the principle of data augmentation (Tanner et al, 1987) in which the latent states are introduced as missing data and augmented to the state of the sampler [**?**]. In this way we can simplify the process of sampling from the posterior distribution: we can use a Gibbs sampler for the parameters of the measurement model and we can estimate the initial and the transition probabilities by means of a Random Walk Metropolis-Hastings step. We then need to introduce a system of priors for the unknown model parameters. In particular, a system of Dirichlet priors is set on the initial and on the transition probabilities, while for the vectors $\boldsymbol{\beta_u}$ and $\boldsymbol{\gamma}_{u\bar{u}}$ we assume that they are a priori independent with distribution $N(0, \sigma^2_\beta \boldsymbol{I})$ and $N(0, \sigma^2_\gamma \boldsymbol{I})$, respectively. The choice for σ^2_β and σ^2_γ depends on the context of the application, typically $5 \leq \sigma^2_\beta = \sigma^2_\gamma \leq 10$. The prior distribution for the parameters of the measurement model depends on the distribution assumed for the state-dependent distribution. We choose a Gaussian distribution for the priors of μ_1 and μ_2 and inverse gamma distributions for the variances ν_1 and ν_2.

The choice of the number of latent states of the unobserved Markov chain, underlying the observed data, is part of the model selection procedure and is a very important step of the estimation process. We adopt the Bayesian information criterion (BIC) [**?**] among a restricted set of models ($m = 3, 4, 5$).

4 LMMs Composite Indicators. A Gender Statistics exercise

Gender inequality - both in space and time - is indirectly measurable through a collection of observable variables. Gender composite indicators are commonly constructed as statistics indicators, i.e. linear combinations of a collection of simple indexes, such as means and proportions, which represent observable items, aggregated by means of a weighing system. The choice of both indexes and weight introduce a certain level of arbitrariness. Their case-specific technical limitations [12],[6] often lead to internal inconsistency since the ranking of a single country can vary in relation to the indicator considered. Moreover, few simple indexes, as well as the weighing system, can outweigh the overall results..

LMMs is liable to offer a sound methodology for estimating the latent trait, i.e. the gender gap, in time and in space, resulting in a synthetic indicator. We move from existing source, namely from supranational official statistics, providing different indicators for all nations worldwide. In particular, we take into account the Gender Inequality Index (GII)[9] and the Global Gender Gap Index

(GGGI)[10]. The GII was introduced by UNDP in 2010 and it measures gender inequalities in three aspects of human development: reproductive health, empowerment and economic status. It focus on inequality, therefore a balanced women/man situation is represented by a zero value. The Global Gender Gap Index (GGGI)was introduced by the World Economic Forum in 2006 with the aim of capturing the magnitude of gender-based disparities. It comprises four dimensions: economic participation and opportunity, educational attainment, health and survival, political empowerment. Perfect parity leads to the value 1. Our applicative viewpoint intends to adapt the LMM approach to Gender synthetic index. Gender Inequality Index (GII) and Global Gender Gap Index (GGGI) are composite indicators which aim to capture differences between man and woman in several areas of life. In our case, we focus on gender gap as the latent status, both in space and time. The gap is in fact a latent trait, namely only indirectly measurable through a collection of observable variables and indicators purposively selected as micro-aspects contributing to the latent macro-dimension. To make the interpretation of results easier and more accessible to non-statisticians, we transformed the value of $\boldsymbol{\beta_u} = (\beta_{0u}, \boldsymbol{\beta}'_{1u})'$ and $\boldsymbol{\gamma}_{u\bar{u}} = (\gamma_{0u\bar{u}}, \boldsymbol{\gamma}'_{1u\bar{u}})'$ in order to obtain an unique set of initial and transition probabilities for all the countries and time occasion. That is, our values represent a cross-national, inter-temporal synthesis.
Applying LMMs to $n = 30$ European countries, with respect to $T = 6$ time points (from 2010 to 2015), we investigate the unobservable latent gender gap summarizing the GGGI and GII information in a single value and rearranging two distinct and rather different ranking into a single one, as the multivariate latent Markov model identifies latent statuses of countries. The model selects $k = 4$ latent states, allowing us to organize countries in 4 ordinal latent statuses through the proposed multivariate spatial Latent Markov model with multinomial logit parametrization, where 1 reflects a situation relatively closest to equality and 4 denotes the highest level of Gender Gap severity. The vector of estimated initial probabilities of latent states at the first measurement occasion is

$$\boldsymbol{\pi} = (0.212, 0.483, 0.139, 0.167).$$

These values can be interpreted as sort of relative frequency [1] in the first year of observation. On the whole, European countries under consideration are more likely to be in latent status 1 and 2, with a relatively low gender gap, with initial probability status of 0.212 and 0.483 respectively. The higher imparity condition, present in status 3 and 4 is less common, accounting for slightly more then 20%, i.e. 0.139 and 0.167 jointly considered.
The Transition Probabilities matrix Π for geographical areas is the following, where the identified latent status are denoted $S1 \cdots S4$

	to S1	to S2	to S3	to S4
from S1	0.98	0.02	0	0
from S2	0.1	0.9	0	0
from S3	0	0.14	0.85	0.01
from S4	0	0.3	0.2	0.4

(8)

It is noticeable that we obtained a matrix close to diagonality, with more sub-diagonal elements than over-diagonal. Such a matrix implies that on the whole countries did not undergo relevant changes in the ten-year observational periods. Probabilities of improving or worsening with respect to the gender gap are low, except for latent status 4, whose diagonal value is equal to 0.4, meaning that 60% or countries improved their gender gap since 2010. When moving, it is often to a better condition, the probability of joining a worse latent status being limited to the shift from latent status 1 to 2, with probability 0.02, and from latent status 2 to 3, with probability 0.02. This reflects, on the one side, a relatively high starting point in gender equality, under the constitutional rights perspective and under aspects such as educational opportunities. On the other side, in so called developed countries, gender disparities tend to stay, when not to worsen, even in the most advanced countries. To this respect, some remarks can be posed on the basis of spacial results.

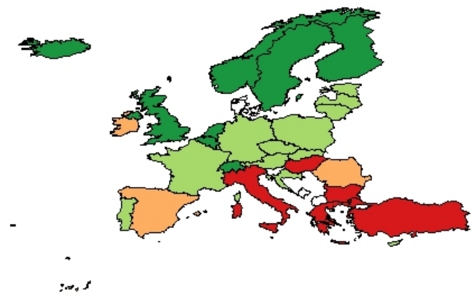
Figure 1 shows the geography of latent gap in Europe in 2010 and 2015 (at the beginning and at the end of the observational time period we considered for our exercise). The 4 latent statuses identified by our models are represented in darkening shades of gray from status $S1$ to $S4$, meaning a worsened gender gap situation.

In 2010 we obtain the following distribution: (i) Latent status 4: Bulgaria, Greece, Hungary, Italy, Malta, Turkey; (ii) Latent status 3: Ireland, Romania, Spain; (iii) Latent status 2: Austria, Cyprus, Croatia, Czech Republic, Germany, Estonia, France, Latvia, Lithuania, Luxembourg, Poland, Portugal, Slovenia; (iv) Latent status 1: Belgium, Finland, Island, Netherlands, Norway, Sweden, Switzerland, United Kingdom.

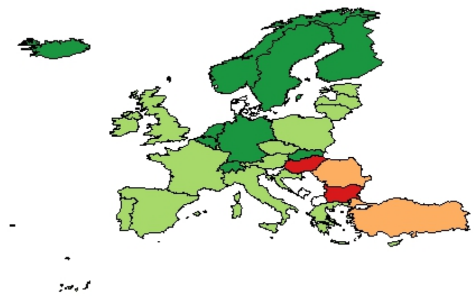
Despite the almost diagonal transition matrix, some changes in latent status structure are highlighted in 2015: (i) Latent status 4: Bulgaria, Hungary, Malta; (ii) Latent status 3: Romania, Turkey; (iii) Latent status 2: Austria, Cyprus, Croatia, Czech Republic, Estonia, France, Greece, Ireland, Italy, Latvia, Lithuania, Luxembourg, Poland, Portugal, Spain, United Kingdom; (iv) Latent status 1: Belgium, Finland, Germany, Island, Netherlands, Norway, Slovenia, Sweden, Swiss.

Latent status 2 becomes the most crowded. The ten-year span appears to have allowed some countries, like Italy, Greece, Spain, to narrow the gap especially in the educational and, to a lesser extent, in political representation. In the case of Slovenia, the upward shift was impressive. The downward shift experimented by the United Kingdom seems to reflect a general trend in economic conditions that cuts across all European countries, even the ones that are regarded as the most socially fair, like Norway, for instance. The overall change in time signals this aspect in a more concise and sharp form by the transition matrix in time, as discussed below.

Under a spacial point of view, then, a first relevant LMMs contribution can be identified in the synthetic single ranking from the information in two different preexisting ones, GGGI and GII respectively. The LMMs ranking establishes relations of equivalence and order that make a complex situation more accessible and readable to the public. For instance, with reference to 2015, the first latent status establishes that the relative best situation in terms of gender par-



(a) 2010



(b) 2015

Fig. 1. Latent Gender Gap Classification in 2010 and 2015

ity is reached with GGGI values in the interval [0.861; 0.947] and GII values in [0.044; 0.076]. Within this general framework, we gain a better understanding of individual countries changes or stability. As aforementioned, Slovenia upward shift from latent status 2 in 2010 to latent status 1 in 2015 relates to a remarkable increase in GGGI, from .698 to .874, as well as in GI, from .139 to .057. Table 1 shows values for countries that changed their ordinal clustering ranking in the five-year period.

Official statistics provide the two measure annually. With reference to time latent states, LMMs estimation showed an overall stability of the gender gap in the observational time, since the indicators transitional matrix (8) is almost diagonal. On the first hand, the widespread, general access to education and health has been experimented with different times and speed. Therefore, at the initial time point of our investigation (2010) some countries see slower, if not almost nonexistent, progress rates after 2010. On the other hand, GII has being decreasing far more slowly since 2010 not only in countries with a longer record of low GII values, like Switzerland, but also for countries that reached these goals more recently, like Greece. Furthermore, GGGI trend is generally very modest (fig.2) and it has often come to a halt after 2008 in

Country	2010 GGGI	2010 GII	2015 GGGI	2015 GII	2010 status	2015 status
Germany	0,7449	0,117	0,7790	0,073	2	1
Greece	0,6662	0,179	0,6850	0,121	4	2
Ireland	0,7597	0,192	0,8070	0,135	3	2
Italy	0,6798	0,175	0,7260	0,085	4	2
Slovenia	0,6982	0,139	0,7840	0,057	x	1
Spain	0,7345	0,118	0,7420	0,087	3	2
Turkey	0,5828	0,564	0,6240	0,340	4	3
United Kingdom	0,7402	0,206	0,7580	0,149	1	2

Table 1. GGGI, GII and latent status for countries with an upward shift in ordinal clustering

a specific dimension, Economic Opportunity and Political Empowerment, as signalled by the World Economic Forum's Global Gender Gap Report 2016, that states that the gap in the economic pillar is currently larger since 2008 [11]. Besides the disparities in opportunities and salary, a major critical issue is posed by the perspective need for women to acquire Stem (Science, Technology, Engineering and Mathematics) skills, with several implications for everyday social and personal lives.

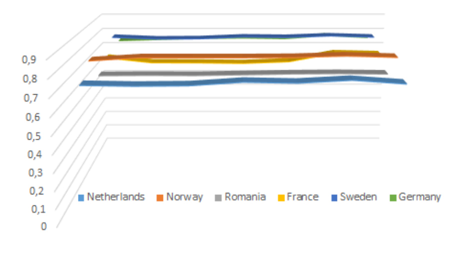


Fig. 2. GGGI trend from 2010 to 2016 in some European countries

5 Conclusion

LMMS have been recently applied to estimate latent traits in time and/or space in social sciences, mainly to highly specific research areas that did not respond adequately to other techniques. Adapting the model in [**?**] to a wider context of social sciences, our proposal consist in the application of LMMS to a more extensive and explored field, Gender Statistics. By means of an empirical exercise, we showed how these models can provide a relevant contribution, since they produced a latent ordinal classification of gender gap between 30 European countries from 2010 to 2015 using two different social composite indicators. They allowed us to obtain synthetic information from the transition matrix that, when diagonal, expresses absence of change. In our exercise, the matrix was nearly diagonal, with reduced margins of improvement for several

countries and in time, especially in the economic sector.
Given the complexity and the multidimensionality of social phenomena, LMMs can contribute highly to a unitarian view. Their latent approach, both in space and in time, can summarise information from different sources. As a matter of fact, both space and time components proved valuable in our application. As far as the former component is concerned, LMMs allowed to identify at a glance areas that are homogeneous or different with respect to gender equality and, in case of differences, permitted to set and order of such a divergence. With respect to the time component, LMMS returned a valuable, concise measure the trend to stagnation that gender parity is experimenting in western countries, due to the rigidness of the economic sector, in particular of the labour market. These models provided also information of national changes in time, i.e. if, how fast and how well some countries were able to set women and men more equal.
Further developments can focus on covariates, especially when expressing opportunities in everyday routines. The persistence of disparities in economic treatment, in fact, can rarely be attributed to explicit law discriminations in western countries, but they can be more often retrieved in availability and in simplification of services to the person and to parenthood, as well as in customs and in mental habits.

References

1. D.J. Bartholomew. *Stochastic Models for Social Processes* 2nd Edition, Wiley, New York, 1973.
2. F. Bartolucci, F. Pennoni and B. Francis. A latent Markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society, Series A*, 170, 151-132, 2007
3. G. Bertarelli. Latent Markov models for aggregate data: application to disease mapping and small area estimation. *Ph.D Thesis* `https://boa.unimib.it/handle/10281/96252`, 2015.
4. H. Crane. A hidden Markov model for latent temporal clustering with application to ideological alignment in the US Supreme Court. *Computational Statistics Data Analysis*, 110, 19-36, 2017.
5. B. Fisher and R. Naidoo. The Geography of Gender Inequality, *PlosOne*, 11(3), 0145778, 2016.
6. S.H. Germain. Bayesian spatio-temporal modelling of rainfall through non-homogenous hidden Markov models, *PhD thesis*, University of Newcastle Upon Tyne, 2010.
7. P.F. Lazarsfeld and N.W. Henry. *Latent Structure Analysis*, Houghton Mifflin, Boston, 1968.
8. F. Mecatti, F.Crippa and P. Farina. A special gen(d)re of statistics: roots, development and methodological prospects of gender statistics. *International Statistical Review*, 80(3), 452–467, 2012.
9. M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation, *Journal of the American statistical Association*, 82, 528–540, 1987.
10. I. ˜Permanyer. The measurement of multidimensional gender inequality: continuing the debate. *Social Indicators Research*, 95(2):181–198, 2010.

11. G.E. Schwarz. Estimating the dimension of a Model, *Annals of Statistics*, 6(2), 461-464, 1978.
12. United Nation. *What are Gender Statistics?*, http://unstats.un.org/unsd/genderstatmanual/What-are-gender-stats.ashx.
13. United Nations Development Programme. *Human Development Reports*, http://hdr.undp.org/en/content/gender-inequality-index-gii/, 2016.
14. World Economic Forum. *The Global Gender Index*, https://www.weforum.org, 2015.
15. World Economic Forum. *The Global Gender Gap Report 2016* http://reports.weforum.org/global-gender-gap-report-2016/, 2016.
16. W. Zucchini and I. MacDonald. *Hidden Markov models for time series*, Springer-Verlag, New York, 2009

PageRank, connecting a line of nodes with multiple complete graphs

Pitos Seleka Biganda[1,2], Benard Abola[2], Christopher Engström[2], and Sergei Silvestrov[2]

[1] Department of Mathematics, College of Natural and Applied Sciences, University of Dar es Salaam, Box 35062 Dar es Salaam, Tanzania
(E-mail: pitos.biganda@mdh.se)

[2] Division of Applied Mathematics, The School of Education, Culture and Communication (UKK), Mälardalen University, Box 883, 721 23, Västerås, Sweden
(E-mails: benard.abola@mdh.se, christopher.engstrom@mdh.se, sergei.silvestrov@mdh.se)

Abstract. PageRank was initially defined by S. Brin and L. Page for the purpose of ranking homepages (nodes) based on the structure of links between these pages. Studies has shown that PageRank of a graph changes with changes in the structure of the graph. In this article we examine how the PageRank changes when two or more outside nodes are connected to a line directed graph. We also look at the PageRank of a graph resulting from connecting a line graph to two complete graphs. In this paper we demonstrate that both the probability (or random walk on a graph) and blockwise matrix inversion approaches can be used to determine explicit formulas for the PageRanks of simple networks.
Keywords: Graph, PageRank, Random walk.

1 Introduction

PageRank was first introduced by Brin and Page [1] to rank homepages (nodes) on the Internet, based on the structure of links between these pages. When a person is interested in getting a certain information from the internet, he is most likely going to use a search engine (eg. Google search engine) to look for such information. Moreover, he will be interested in getting the most relevant ones. What PageRank aims to do, is to sort out and place the most relevant pages first in the list of all information displayed after the search.

It is known that the number of pages on the internet is very large and keeps on increasing over time. For this reason, the PageRank algorithm need to be very fast to accommodate the increasing number of pages and at the same time retaining the requirement for quality of the ranking results as one carries out an internet search [1].

Algorithms similar to PageRank are available, for instance, EigenTrust algorithm, by Kamvar *et al.*[2], applied to reputation management in peer-to-peer networks, and DeptRank algorithm, which is used to evaluate risk in financial networks (Battiston *et al.*[10]). These imply that PageRank concept can be adopted to various networks problem.

Usually PageRank is calculated using power method. The method has been found to be efficient for both small and large systems. The convergence speed

17^{th} *ASMDA Conference Proceedings, 6 - 9 June 2017, London, UK*

of the method on a webpage structure depends on the parameter c, where c is a real number such that $0 \leq c \leq 1$ (Haveliwala and Kamvar[12]), and the problem is well conditioned unless c is very close to 1 (Kamvar and Haveliwala[4]). However, many methods have been developed for speeding up the calculations of PageRank in order to meet the increasing number of pages on the internet. Some of these methods include aggregating webpages that are close and are expected to have similar PageRank (Ishii *et al.*[7]), partitioning the graph into components as in (Engström and Silvestrov[14]), removing the dangling nodes before computing PageRank and then calculate their ranks at the end or use a power series formulation of PageRank (Anderson and Silvestrov[8]), and not computing the PageRank of pages that have already converged in every iteration as suggested by Sepander *et al.*[13].

There are also studies on a large scale using PageRank and other measure in order to learn more about the Web. One of them is looking at the theoretical and experimental perspective of the distribution of PageRank as by Dyani *et al.*[11].

The theory behind PageRank is built from Perron-Frobenius theory (Berman and Plemmons[9]) and the study of Markov chains (Norris [3]). But how PageRank changes with changes in the system or parameters is not well known. Engström and Silvestrov[5,6] investigated the changes of PageRank of the nodes in the system consisting of a line of nodes and an outside node and/or a complete graph connected to the line of nodes in different ways. In this article, we will extend their work by looking at a line graph connected to multiple outsides nodes, and a line graph connected to two complete graphs. For instance, we will consider what happens when two or more nodes are linked to a line graph. Like in (Engström and Silvestrov[5]), we will consider PageRank as the solution to a linear system of equations as well as probabilities of a random walk through the graph. In the similar way, non-normalized PageRank will be considered.

2 Preliminaries

This section describes important notations and definitions. We start by giving some notations and thereafter essential definitions that are used throughout the article.

- S_G: The system of nodes and links for which we want to calculate PageRank. It contains both the system matrix A_G and a weight vector $\boldsymbol{v}_G$. A subindex G can be either a capital letter or a number in the case of multiple systems.
- n_G: The number of nodes in system S_G.
- A_G: A system matrix of size $n_G \times n_G$ where an element $a_{ij} = 0$ means there is no link from node i to node j. Non-zero elements are equal to $1/r_i$ where r_i is the number of links from node i.
- $\boldsymbol{u}_G$: Non-negative weight vector, not necessary with sum one. Its size is $n_G \times 1$.
- c: A parameter $0 < c < 1$ for calculating PageRank, usually $c = 0.85$.

- $\boldsymbol{g}_G$: A vector with elements equal to one for dangling nodes and zero otherwise in S_G. Its size is $n_G \times 1$.
- M_G: Modified system matrix, $\mathrm{M}_G = c(\mathrm{A}_G + \boldsymbol{g}_G \boldsymbol{u}_G^\top)^\top + (1-c)\boldsymbol{u}_G \boldsymbol{e}^\top$ used to calculate PageRank, where $\boldsymbol{e}$ is the unit vector. Size $n_G \times n_G$.
- S: Global system made up of multiple disjoint subsystems $S = S_1 \cup S_2 \ldots \cup S_N$, where N is the number of subsystems.

In the cases where there is only one possible system the subindex G is omitted. For the systems making up S we define disjoint systems in the following way.

Definition 1. Two systems S_1, S_2 are disjoint if there are no paths from any nodes in S_1 to S_2 or from any nodes in S_2 to S_1.

PageRank can be defined in various versions, for instance in [5] where two versions were presented. However, in this paper we will use the non-normalized PageRank, denoted as $\boldsymbol{R}_j$ for node j, and it is defined as

Definition 2. $\boldsymbol{R}_G$ for system S_G is defined as $\boldsymbol{R}_G = (\mathrm{I} - c\mathrm{A}_G^\top)^{-1} n_G \boldsymbol{u}_G$, where I is an identity matrix of same size as A_G.

Definition 3. Consider a random walk on a graph described by A_G, which is the adjacency matrix weighted such that the sum over every non-zero row is equal to one. In each step with probability $c \in (0,1)$, move to a new vertex from the current vertex by traversing a random outgoing edge from the current vertex with probability equal to the weight on the corresponding edge weight. With probability $1-c$ or if the current vertex have no outgoing edges, we stop the random walk. The PageRank $\boldsymbol{R}$ for a single vertex v_j can be written as

$$\boldsymbol{R}_j = \left(\sum_{v_i \in V, v_i \neq v_j} w_i P_{ij} + w_j \right) \left(\sum_{k=0}^{\infty} (P_{jj})^k \right), \tag{1}$$

where P_{ij} is the probability to hit node v_j in a random walk starting in node v_i described as above. This can be seen as the expected number of visits to v_j if we do multiple random walks, starting in every node once and weighting each of these random walks by $\boldsymbol{w}$ [5].

Next, let us define graph-structures we will encounter in the section that follows.

Definition 4. A simple line is a graph with n_L nodes where node n_L links to node n_{L-1} which in turn links to node n_{L-2} all the way until node n_2 link to node n_1.

Definition 5. A complete graph is a group of nodes in which all nodes in the group links to all other nodes in the group.

The following well known lemma for blockwise inversion will be used in this article. A proof can be found, for example in Bernstein [15].

Lemma 1.

$$\begin{bmatrix} \mathrm{B} & \mathrm{C} \\ \mathrm{D} & \mathrm{E} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathrm{B}-\mathrm{C}\mathrm{E}^{-1}\mathrm{D})^{-1} & -(\mathrm{B}-\mathrm{C}\mathrm{E}^{-1}\mathrm{D})^{-1}\mathrm{C}\mathrm{E}^{-1} \\ -\mathrm{E}^{-1}\mathrm{D}(\mathrm{B}-\mathrm{C}\mathrm{E}^{-1}\mathrm{D})^{-1} & \mathrm{E}^{-1} + \mathrm{E}^{-1}\mathrm{D}(\mathrm{B}-\mathrm{C}\mathrm{E}^{-1}\mathrm{D})^{-1}\mathrm{C}\mathrm{E}^{-1} \end{bmatrix} \tag{2}$$

where B, E *is square and* $\mathrm{E}, (\mathrm{B}-\mathrm{C}\mathrm{E}^{-1}\mathrm{D})$ *are nonsingular.*

3 Changes in PageRank when connecting the simple line graph with multiple outside nodes

In this section, we presents four graphs and associated PageRanks lemma and theorem. We will start with a lemma from where explicit PageRank for each vertex of the graph considered can be determined.

3.1 Connecting the simple line with multiple links from m outside nodes to one node in the line

Consider a simple line graph that has L vertices. Suppose vertex $n_j, j \in [1, L]$ is linked to m outside vertices as shown in Figure 1. It can be seen that if $j = 1$, then the node is said to be an authority node.

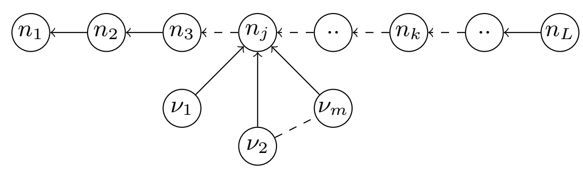


Fig. 1. A simple line directed graph with m outside vertices

Lemma 2. *The PageRank of a node e_i belonging to the line in a system containing a simple line with m outside nodes linking to one node j in the line when using uniform weight vector $\boldsymbol{u}$ can be expressed as*

$$\begin{aligned} \boldsymbol{R}_i &= \sum_{k=0}^{n_L - i} c^k + b_{ij} = \frac{1 - c^{n_L - i + 1}}{1 - c} + b_{ij} \\ b_{ij} &= \begin{cases} mc^{j-i+1}, & \text{if } i \le j \\ 0, & \text{if } i > j \end{cases} \end{aligned} \tag{3}$$

where $m \ge 1$ and n_L is the number of nodes in the line. The new nodes each have rank 1.

Proof. Applying the notion of probability, the PageRank for a node when a uniform $\boldsymbol{u}$ is used can be written in the form Equation (1). Let e_i and e_j be the nodes on the line. Suppose that P_{ji} is the probability of hitting node e_i starting at node e_j. Considering a random walk on a graph described by $c\mathrm{A}_G$, i.e. we walk to the new node with probability c and stop with probability $1-c$, therefore P_{ji} becomes

$$P_{ji} = c^{j-i}, \quad j > i$$

and zero, otherwise. It follows that the expected numbers of visits to e_i if multiple random walks is performed starting at any node e_j, for $j > i$ is expressed as

$$\sum_{\text{all}j:e_j \neq e_i} P_{ji} + 1 = \sum_{j=i+1}^{n_L} c^{j-i} + 1 = \frac{1 - c^{n_L - i+1}}{1-c},$$

where n_L is the number of nodes in the line. Next we show that the m outside nodes linking to node e_j on the line adds $b_{ij} = mc^{j-i+1}$ for $j \geq i$. The proof of this part is similar to Theorem 2 in [14], only that we need to show that it is generally true for m nodes. By induction; for $m = 1$, it is exactly the same as in [14]. Next, assume that it is true for $m = k$, then

$$b_{ij}(k) = \underbrace{c^{j-i+1} + c^{j-i+1} + \cdots + c^{j-i+1}}_{k \text{ times}} = kc^{j-i+1}.$$

It follows that for $m = k + 1$,

$$b_{ij}(k+1) = b_{ij}(k) + c^{j-i+1} = (k+1)c^{j-i+1}.$$

Finally, it is obvious that the PageRank of the m nodes is 1 each since no node links to each of the nodes.

Remark It is essential to note that we are dealing with simple line graph as given in Definition 4 thus it is not possible to hit node i from the left, that is., $i - 1$ if one takes a random walk from any node j such that $j < i$ as shown in Figure 1.

3.2 Connecting a simple line with multiple links from multiple outside nodes to the line

Assume that the nodes $n_1, n_2, \cdots, n_5$ on the line are linked to outside nodes $m_1, m_2, \cdots, m_5$ respectively, where $m_j \geq 0$ (the number of outside nodes linked to node j on the line graph). Suppose $m_j = 1$ for all $j \in \{1, 2, \cdots, 5\}$ as shown in the Figure 2. To gain a better understanding of how to obtain the

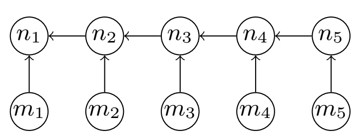


Fig. 2. A simple line graph with one outside vertex linked to one vertex on the line

PageRanks of Figure 2, let us have a look at $\boldsymbol{R}_4$ and $\boldsymbol{R}_5$ on the line graph which correspond to nodes n_4 and n_5 respectively. Using Definition 2, the Pagerank $\boldsymbol{R}_5 = 1 + m_5 c$. Similarly, we get $\boldsymbol{R}_4 = 1 + m_4 c + c\boldsymbol{R}_5$ and substituting for $\boldsymbol{R}_5$ yields $\boldsymbol{R}_4 = \frac{1-c^2}{1-c} + m_4 c + m_5 c^2 = \frac{1-c^2}{1-c} + \sum_{j=4}^{5} m_j c^{j-3}$. In overall PageRank

$\boldsymbol{R}_{S_L}$ on the line graph before substituting for m_j is

$$\begin{pmatrix} \boldsymbol{R}_1 \\ \boldsymbol{R}_2 \\ \boldsymbol{R}_3 \\ \boldsymbol{R}_4 \\ \boldsymbol{R}_5 \end{pmatrix} = \begin{pmatrix} 1 + c + c^2 + c^3 + c^4 + m_1 c + m_2 c^2 + m_3 c^3 + m_4 c^4 + m_5 c^5 \\ 1 + c + c^2 + c^3 + m_2 c + m_3 c^2 + m_4 c^3 + m_5 c^4 \\ 1 + c + c^2 + m_3 c + m_4 c^2 + m_5 c^3 \\ 1 + c + m_4 c + m_5 c^2 \\ 1 + m_5 c \end{pmatrix} \tag{4}$$

and the PageRank of each of the outside node is equal to 1.
It can be seen that a better approach to find the PageRank would be to start with $\boldsymbol{R}_5$, $\boldsymbol{R}_4$ and so on, that is, recursively then generalization can easily be made. In the theorem that follows, the PageRanks for such general network is proposed for $m_j \geq 0$.

Theorem 1. *The PageRank of a node e_i belonging to the line in a system containing a simple line with multiple outside nodes, $m_1, m_2, \cdots, m_i, \cdots, m_L$ linking to every nodes $n_1, n_2, \cdots, n_i, \cdots, n_L$ in that order respectively in the line when using uniform weight vector $\boldsymbol{u}$ can be written as*

$$\begin{aligned} \boldsymbol{R}_i &= \frac{1 - c^{n_L - i + 1}}{1 - c} + b_i, \quad \textit{where} \\ b_i &= \begin{cases} \sum_{j=i}^{n_L} m_j c^{j-i+1}, & \textit{if } j \geq i \\ 0 & \textit{if } i < j. \end{cases}, \end{aligned} \tag{5}$$

The outside nodes each have rank 1.

Proof. We start by calculating the PageRank of the nodes i on the directed line graph, we have partially shown how to achieve this in Lemma 2. However, the Pagerank $\boldsymbol{R}_i$ on the line graph is obtained by dividing the overall nodes of the graph into two: along the line and outside. Then writing the PageRank using Definition 3 while taking into account the weight $w_i = 1$. Hence, $\frac{1 - c^{n_L - i + 1}}{1 - c}$ is the expected number of visit to node i when arbitrary random walks are performed starting from any node j. The term b_i is the expected number of visits to node i starting from each outside nodes e_j, for $j \geq i$. Recall that if you are along the line, you can hit node $L - 1$ while starting from node L but not the vice verse. Now, without loss of generality, take the node L on the line, then

$$\begin{aligned} R_L = 1 + m_L c = \frac{1 - c}{1 - c} + m_L c &= \frac{1 - c^{L-L+1}}{1 - c} + m_L c^{L-L+1}, \\ = \frac{1 - c^{L-L+1}}{1 - c} &+ \sum_{j=L}^{L} m_j c^{j-L+1}. \end{aligned} \tag{6}$$

This proves that the formula is correct for the last node L in the line.

Next we prove that if the formula is correct for R_k then it is correct for R_{k-1} as well, which by induction proves that it is correct for all vertices in the line.