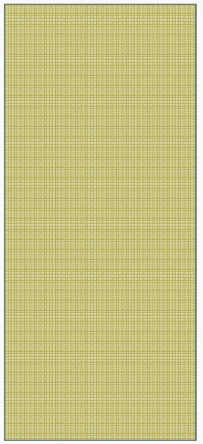
VII JORNADAS DE USUARIOS DE R

“ CLASIFICACIÓN DE TEXTOS CIENTÍCOS CON R ”

UNIVERSIDAD COMPLUTENSE DE MADRID

Departamento de Arquitectura de Computadores y Automática (DACyA)
Grupo ABSys (Adaptative and Bioinspired Systems group)

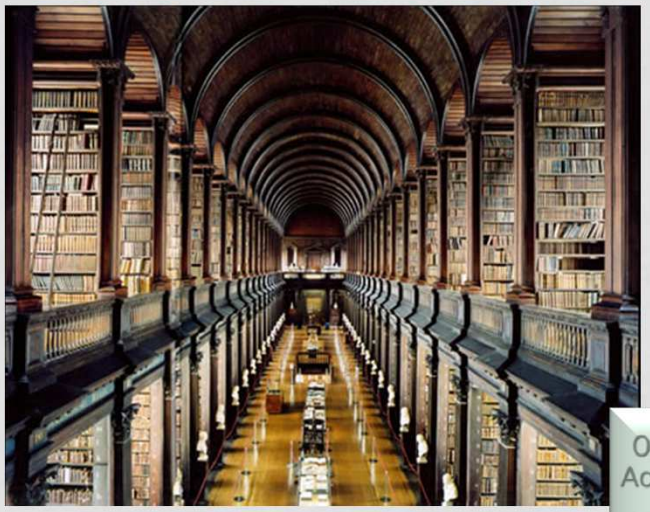


noviembre 2015

AUTOR: SERGIO CONTADOR PACHÓN

MOTIVACIÓN

- Crecimiento continuo de la capacidad de generar, coleccionar y almacenar datos.
- Fuentes cada vez más grandes con abundantes datos:



Open Access

Necesidad de sistemas de análisis masivos de información automáticos y escalables

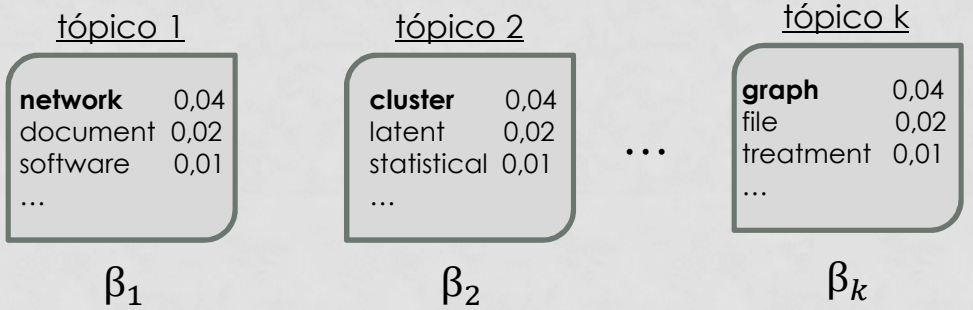
ÍNDICE

TOPIC MODELING

- "Topic modeling" (TM) son modelos probabilísticos de textos:
  - TM descubren los tópicos (temas) en largas colecciones de documentos como un problema de inferencia a posteriori.
  - TM calculan la estructura temática oculta, los tópicos.

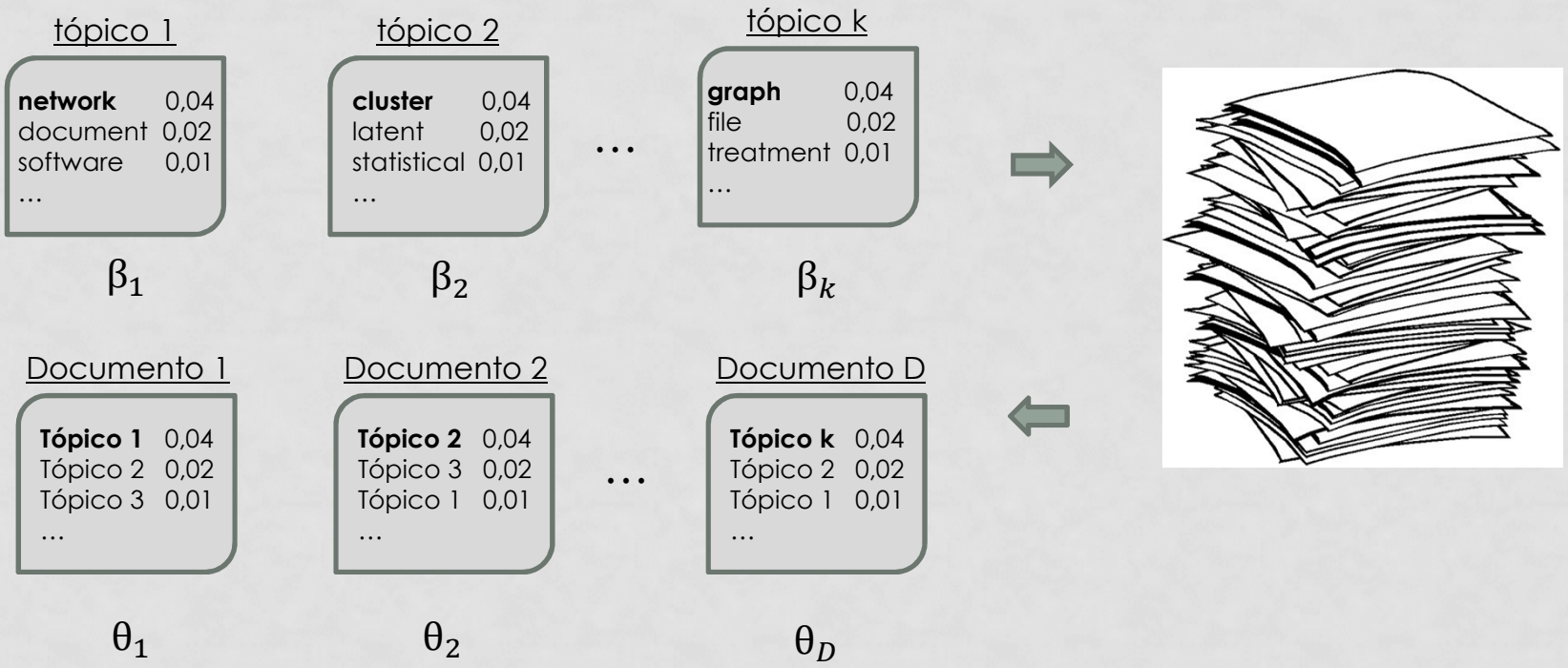
TM: CREAR TÓPICOS

- Los tópicos se crean calculando las distribuciones $\beta_{1:k}$ de palabras.



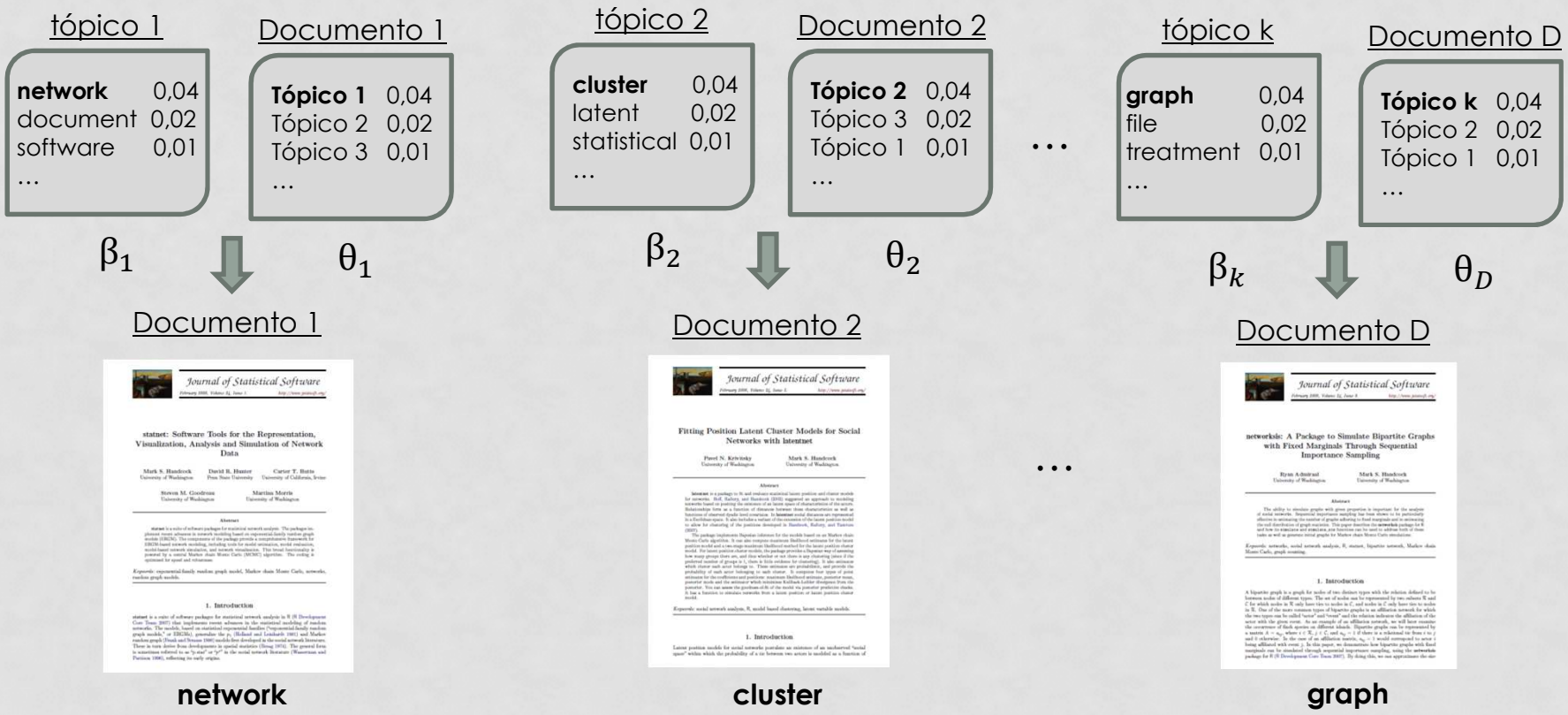
TM: ASIGNAR TÓPICOS

- Los tópicos se asignan calculando la distribución $\theta_{1:D}$ de tópicos.



TM: CLASIFICAR DOCUMENTOS

- La palabra mas probable del tópico mas probable es el tema principal del documento.



TM: ALGORITMOS

- Probabilistic Lantent Semantic Analysis (PLSA) [Hofmann.1999].
- Latent Dirichlet Allocation (LDA) [Blei et al. 2003].
- Dinamic Topic Model [Blei et al. 2006].
- Bigram Topical Mdel [Wallach, 2006].
- Correlated Topic Model (CTM) [Blei et al. 2007].
- Bayesian Non-Parametric Topic Model [Griffiths et al. 2007].
- Supervised Topic Model [Blei et al. 2007].
- Topical N-Gram Model (TNG) [Wang et al. 2007].
- Label Topic [Mei et al. 2007].
- TurboTopics (LDAPD) [Blei & Lafferty, 2009].
- Relational Topic Model [Chang et al. 2009].
- Ideal Topic Model [Sean Gerrish et al. 2010].
- Phrase discovering Topic Model (PDLDA) [Lindsey et al. 2012].
- KERT [Danilevsky et al. 2014].
- TopMine [El-kishky et al. 2015].

LATENT DIRICHLET ALLOCATION

- “Latent Dirichlet Allocation” (LDA) es el TM más simple.
- LDA es la versión Bayesiana de “ Probabilistic Latent Semantic Analysis ” (PLSA) [Deerwester et al., 1990; Hofmann, 1999].
- LDA modeliza las probabilidades prior/posterior con funciones dirichlet de hiperparámetros η y α.
- LDA necesita conocer los textos y el número de tópicos (temas) k a los que hacen referencia los textos.

LATENT DIRICHLET ALLOCATION

- LDA es la versión Bayesiana PLSA:

Probabilidad de los documentos dados los tópicos (prior):
asignar los tópicos a los documentos de la colección

$$p(\beta_{1:k}, \theta_{1:D}, Z_{1:D}, W_{1:D}) = \prod_{i=1}^{k} p(\beta_i/\eta) \prod_{d=1}^{D} p(\theta_d/\alpha) \prod_{n=1}^{N} p(Z_{d,n}/\theta_d)\, p(W_{d,n}/\beta_{1:k}, Z_{d,n})$$

Probabilidad de los tópicos dados los documentos (posterior):
crear los tópicos de la colección

$$p(\beta_{1:k}, \theta_{1:D}, Z_{1:D}/W_{1:D}) = \prod_{i=1}^{k} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N} p(Z_{d,n}/\theta_d)\, p(W_{d,n}/\beta_{1:k}, Z_{d,n})/\, p(W_{1:D})$$

$p(W_{1:D})$ *es la probabilidad marginal,*
requiere mucho tiempo de computo, no − viable

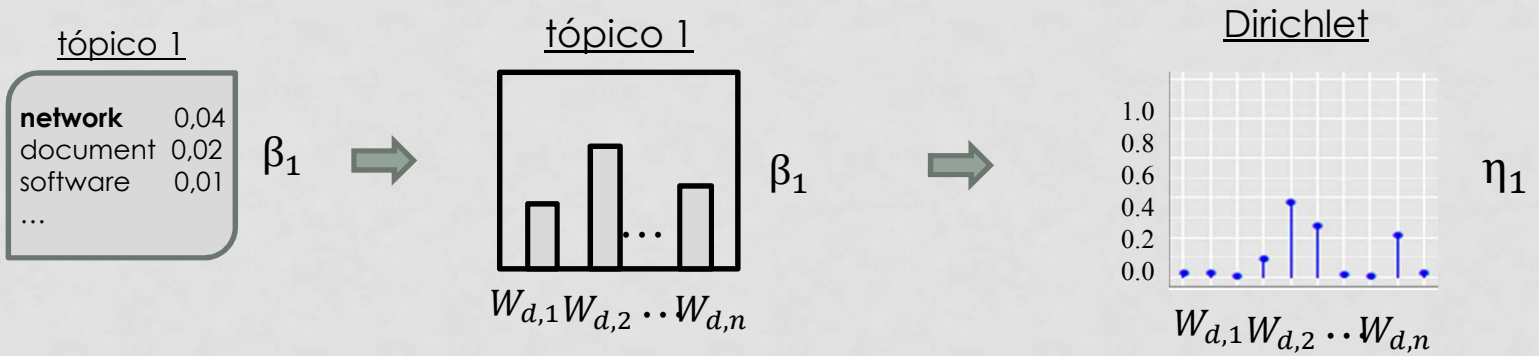
LDA: APROXIMANDO EL POSTERIOR

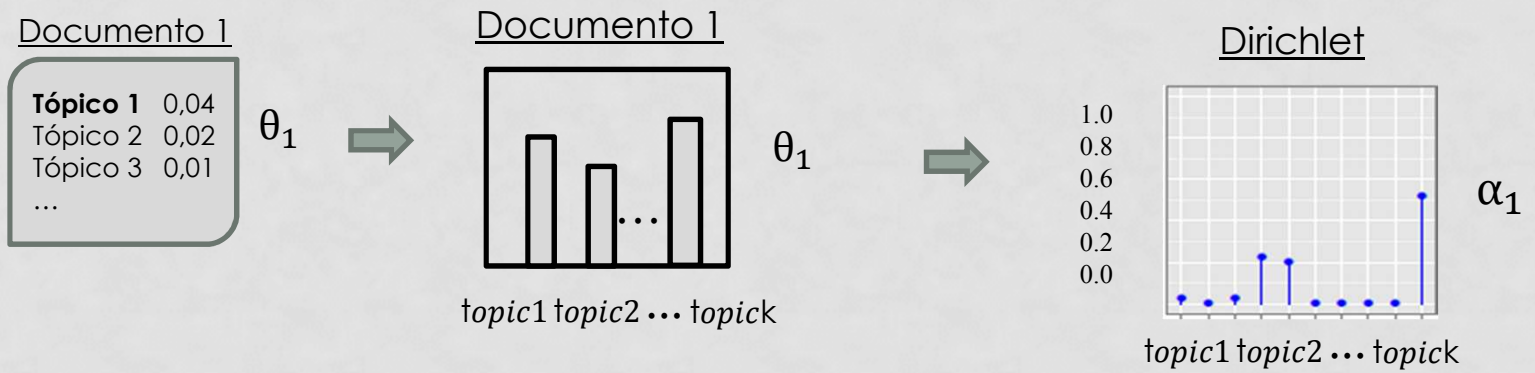
- Mean field variational methods [Blei et al.., 2001].
- Collapsed Gibbs sampling (CGS) [Griffiths and Steyvers, 2002].
- Expectation propagation [Minka and Lafferty, 2002].
- Collapse variational inference [The et al.., 2006].
- Online variational inference [Hoffman 2010].

LDA: MODELO

- LDA modeliza las probabilidades prior/posterior con funciones dirichlet:

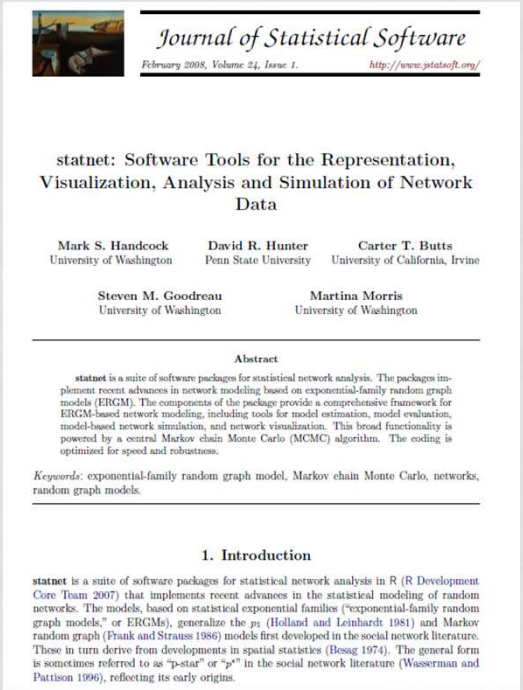
$$\eta = \eta(\eta_1, \eta_2, \ldots, \eta_k) \qquad \alpha = \alpha(\alpha_1, \alpha_2, \ldots, \alpha_D)$$





LDA INPUT

- LDA necesita conocer los textos y el número de tópicos k:



Journal of Statistical Software
February 2008, Volume 24, Issue 1. http://www.jstatsoft.org/

statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data

Mark S. Handcock, University of Washington
David R. Hunter, Penn State University
Carter T. Butts, University of California, Irvine
Steven M. Goodreau, University of Washington
Martina Morris, University of Washington

Abstract

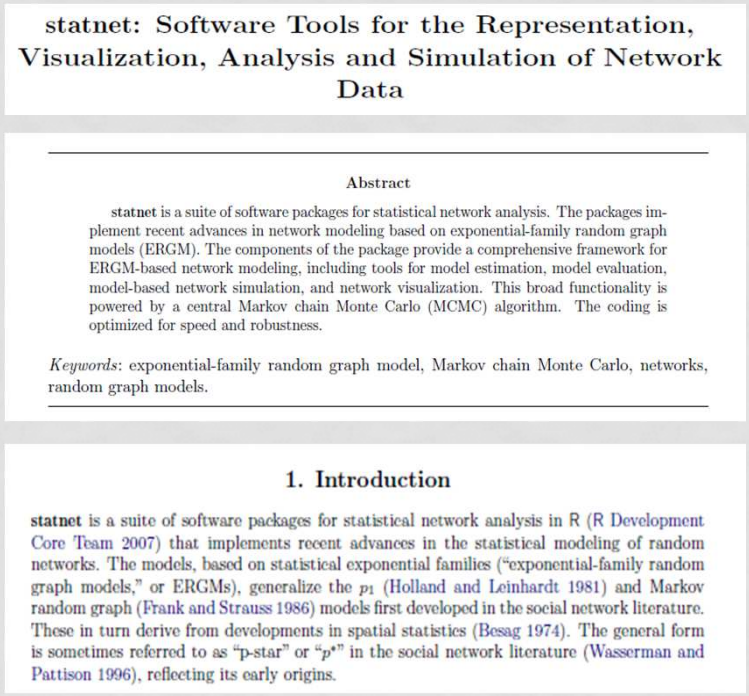
statnet is a suite of software packages for statistical network analysis. The packages implement recent advances in network modeling based on exponential-family random graph models (ERGM). The components of the package provide a comprehensive framework for ERGM-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm. The coding is optimized for speed and robustness.

Keywords: exponential-family random graph model, Markov chain Monte Carlo, networks, random graph models.

1. Introduction

statnet is a suite of software packages for statistical network analysis in R (R Development Core Team 2007) that implements recent advances in the statistical modeling of random networks. The models, based on statistical exponential families ("exponential-family random graph models," or ERGMs), generalize the p_1 (Holland and Leinhardt 1981) and Markov random graph (Frank and Strauss 1986) models first developed in the social network literature. These in turn derive from developments in spatial statistics (Besag 1974). The general form is sometimes referred to as "p-star" or "p^*" in the social network literature (Wasserman and Pattison 1996), reflecting its early origins.

Granularidad



statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data

Abstract

statnet is a suite of software packages for statistical network analysis. The packages implement recent advances in network modeling based on exponential-family random graph models (ERGM). The components of the package provide a comprehensive framework for ERGM-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm. The coding is optimized for speed and robustness.

Keywords: exponential-family random graph model, Markov chain Monte Carlo, networks, random graph models.

1. Introduction

statnet is a suite of software packages for statistical network analysis in R (R Development Core Team 2007) that implements recent advances in the statistical modeling of random networks. The models, based on statistical exponential families ("exponential-family random graph models," or ERGMs), generalize the p_1 (Holland and Leinhardt 1981) and Markov random graph (Frank and Strauss 1986) models first developed in the social network literature. These in turn derive from developments in spatial statistics (Besag 1974). The general form is sometimes referred to as "p-star" or "p^*" in the social network literature (Wasserman and Pattison 1996), reflecting its early origins.

⋮

LDA OUTPUT: CREAR TÓPICOS

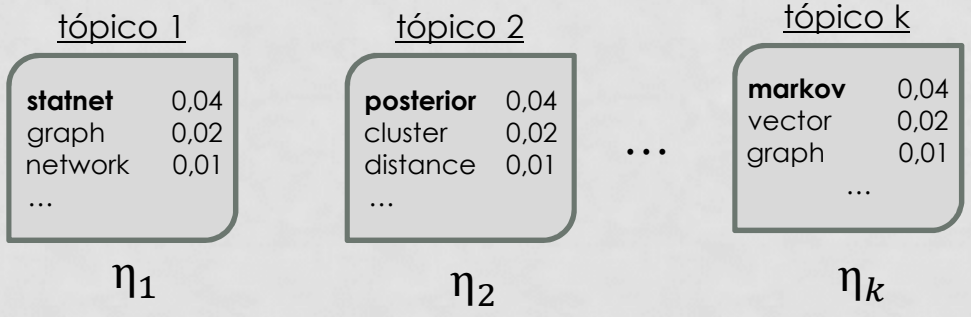
Abstract

statnet is a suite of software packages for statistical network analysis. The packages implement recent advances in network modeling based on exponential-family random graph models (ERGM). The components of the package provide a comprehensive framework for ERGM-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm. The coding is optimized for speed and robustness.

Keywords: exponential-family random graph model, Markov chain Monte Carlo, networks, random graph models.

- Inicializar los tópicos:

$$\eta = \eta(\eta_1, \eta_2, \ldots, \eta_k)$$



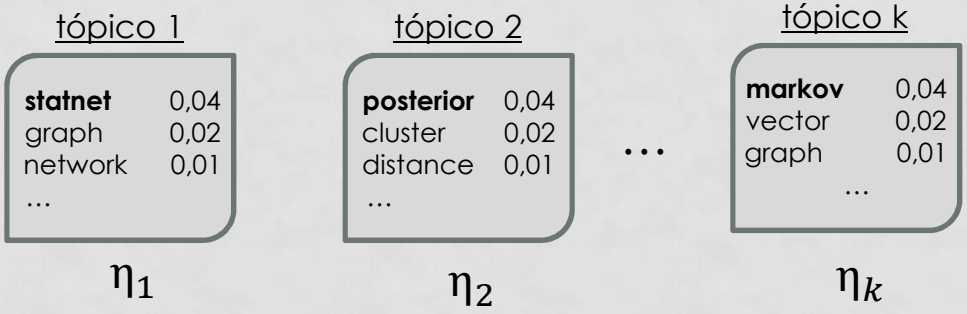
LDA OUTPUT: CREAR TÓPICOS

Abstract

statnet is a suite of software packages for statistical network analysis. The packages implement recent advances in network modeling based on exponential-family random graph models (ERGM). The components of the package provide a comprehensive framework for ERGM-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm. The coding is optimized for speed and robustness.

Keywords: exponential-family random graph model, Markov chain Monte Carlo, networks, random graph models.

- Seleccionar subespacio de busqueda: MCMC (Markov Chain Monte Carlo)



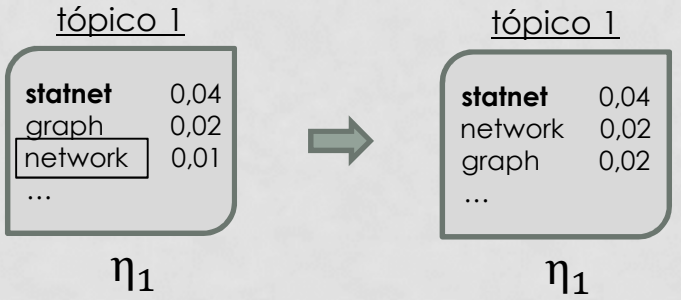
LDA OUTPUT: CREAR TÓPICOS

Abstract

statnet is a suite of software packages for statistical network analysis. The packages implement recent advances in network modeling based on exponential-family random graph models (ERGM). The components of the package provide a comprehensive framework for ERGM-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm. The coding is optimized for speed and robustness.

Keywords: exponential-family random graph model, Markov chain Monte Carlo, networks, random graph models.

- Modificar un parámetro del hiperparámetro η (manteniendo normalización)



LDA OUTPUT: CREAR TÓPICOS

Abstract

statnet is a suite of software packages for statistical network analysis. The packages implement recent advances in network modeling based on exponential-family random graph models (ERGM). The components of the package provide a comprehensive framework for ERGM-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm. The coding is optimized for speed and robustness.

Keywords: exponential-family random graph model, Markov chain Monte Carlo, networks, random graph models.

- Calcular la probabilidad de que el modelo explique los datos (**posterior):**

$$P(iter = q) = \prod_{i=1}^{k} p(\beta_i/\eta) \prod_{d=1}^{D} p(\theta_d/\alpha) \prod_{n=1}^{N} p(Z_{d,n}/\theta_d)\, p(W_{d,n}/\beta_{1:k}, Z_{d,n})$$

LDA OUTPUT: CREAR TÓPICOS

Abstract

statnet is a suite of software packages for statistical network analysis. The packages implement recent advances in network modeling based on exponential-family random graph models (ERGM). The components of the package provide a comprehensive framework for ERGM-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm. The coding is optimized for speed and robustness.

Keywords: exponential-family random graph model, Markov chain Monte Carlo, networks, random graph models.

- Comparar $P(iter = q)$ con la probabilidad de la iteración anterior, aplicando MLE:

Si $P(iter = q) > P(iter = (q - 1))$ → actualizar hiperparámetros β, η con β_i, η_j y continuar

Si $P(iter = q) \leq P(iter = (q - 1))$ → descartar β_i, η_j y continuar

LDA OUTPUT: CREAR TÓPICOS

Abstract

statnet is a suite of software packages for statistical network analysis. The packages implement recent advances in network modeling based on exponential-family random graph models (ERGM). The components of the package provide a comprehensive framework for ERGM-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm. The coding is optimized for speed and robustness.

Keywords: exponential-family random graph model, Markov chain Monte Carlo, networks, random graph models.

- Repetir hasta alcanzar resultado óptimo:

El criterio de convergencia seleccionado se ha alcanzado

ó

$$P(iter = q) \cong P(iter = (q - 1)) \; \forall \; q \in N$$

LDA OUTPUT: ASIGNAR TÓPICOS

Abstract

statnet is a suite of software packages for statistical network analysis. The packages implement recent advances in network modeling based on exponential-family random graph models (ERGM). The components of the package provide a comprehensive framework for ERGM-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm. The coding is optimized for speed and robustness.

Keywords: exponential-family random graph model, Markov chain Monte Carlo, networks, random graph models.

---

tópico 1		tópico 2		...	tópico k	
network	0,04	**cluster**	0,04		**graph**	0,04
document	0,02	latent	0,02		file	0,02
software	0,01	statistical	0,01		treatment	0,01
...		...			...	
η_1		η_2			η_k	

LDA OUTPUT: ASIGNAR TÓPICOS

Abstract

statnet is a suite of software packages for statistical network analysis. The packages implement recent advances in network modeling based on exponential-family random graph models (ERGM). The components of the package provide a comprehensive framework for ERGM-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm. The coding is optimized for speed and robustness.

Keywords: exponential-family random graph model, Markov chain Monte Carlo, networks, random graph models.

tópico 1		tópico 2			tópico k	
network	0,04	**cluster**	0,04		**graph**	0,04
document	0,02	latent	0,02		file	0,02
software	0,01	statistical	0,01	...	treatment	0,01
...		...			...	
η_1		η_2			η_k	

LDA OUTPUT: ASIGNAR TÓPICOS

Abstract

statnet is a suite of software packages for statistical network analysis. The packages implement recent advances in network modeling based on exponential-family random graph models (ERGM). The components of the package provide a comprehensive framework for ERGM-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm. The coding is optimized for speed and robustness.

Keywords: exponential-family random graph model, Markov chain Monte Carlo, networks, random graph models.

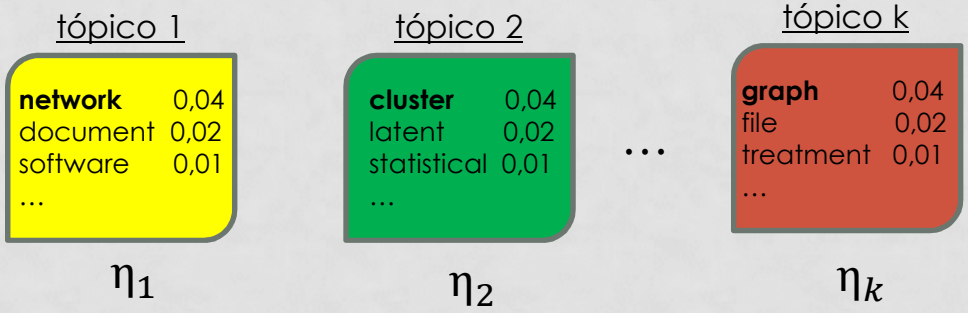
tópico 1		tópico 2			tópico k	
network	0,04	**cluster**	0,04		**graph**	0,04
document	0,02	latent	0,02		file	0,02
software	0,01	statistical	0,01		treatment	0,01
…		…			…	
η_1		η_2			η_k	

LDA OUTPUT: ASIGNAR TÓPICOS

Abstract

statnet is a suite of software packages for statistical network analysis. The packages implement recent advances in network modeling based on exponential-family random graph models (ERGM). The components of the package provide a comprehensive framework for ERGM-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm. The coding is optimized for speed and robustness.

Keywords: exponential-family random graph model, Markov chain Monte Carlo, networks, random graph models.



LDA OUTPUT: ASIGNAR TÓPICOS

Abstract

statnet is a suite of software packages for statistical network analysis. The packages implement recent advances in network modeling based on exponential-family random graph models (ERGM). The components of the package provide a comprehensive framework for ERGM-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm. The coding is optimized for speed and robustness.

Keywords: exponential-family random graph model, Markov chain Monte Carlo, networks, random graph models.

LDA OUTPUT: CLASIFICAR DOCUMENTOS

network

Abstract

statnet is a suite of software packages for statistical network analysis. The packages implement recent advances in network modeling based on exponential-family random graph models (ERGM). The components of the package provide a comprehensive framework for ERGM-based network modeling, including tools for model estimation, model evaluation, model-based network simulation, and network visualization. This broad functionality is powered by a central Markov chain Monte Carlo (MCMC) algorithm. The coding is optimized for speed and robustness.

Keywords: exponential-family random graph model, Markov chain Monte Carlo, networks, random graph models.

tópico 1

network	0,04
document	0,02
software	0,01
…	

η_1

tópico 2

cluster	0,04
latent	0,02
statistical	0,01
…	

η_2

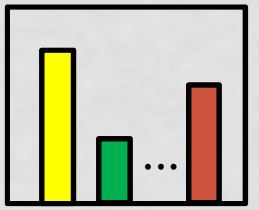
…

tópico k

graph	0,04
file	0,02
treatment	0,01
…	

η_k

Documento 1



θ_1

$topic1\ topic2 \cdots topick$

TM CON R

- Hay dos librerías en R para trabajar con Topic Modeling:
  "lda" y "topicmodels"

  - Funciones de la librería "lda":

    LDA_CGS (Collapse Gibbs Sampling)
    nubbi_CGS (networks uncovered by bayesian inference)
    rtm_CGS (relational topic models)

  - Funciones de la librería "topicmodels":

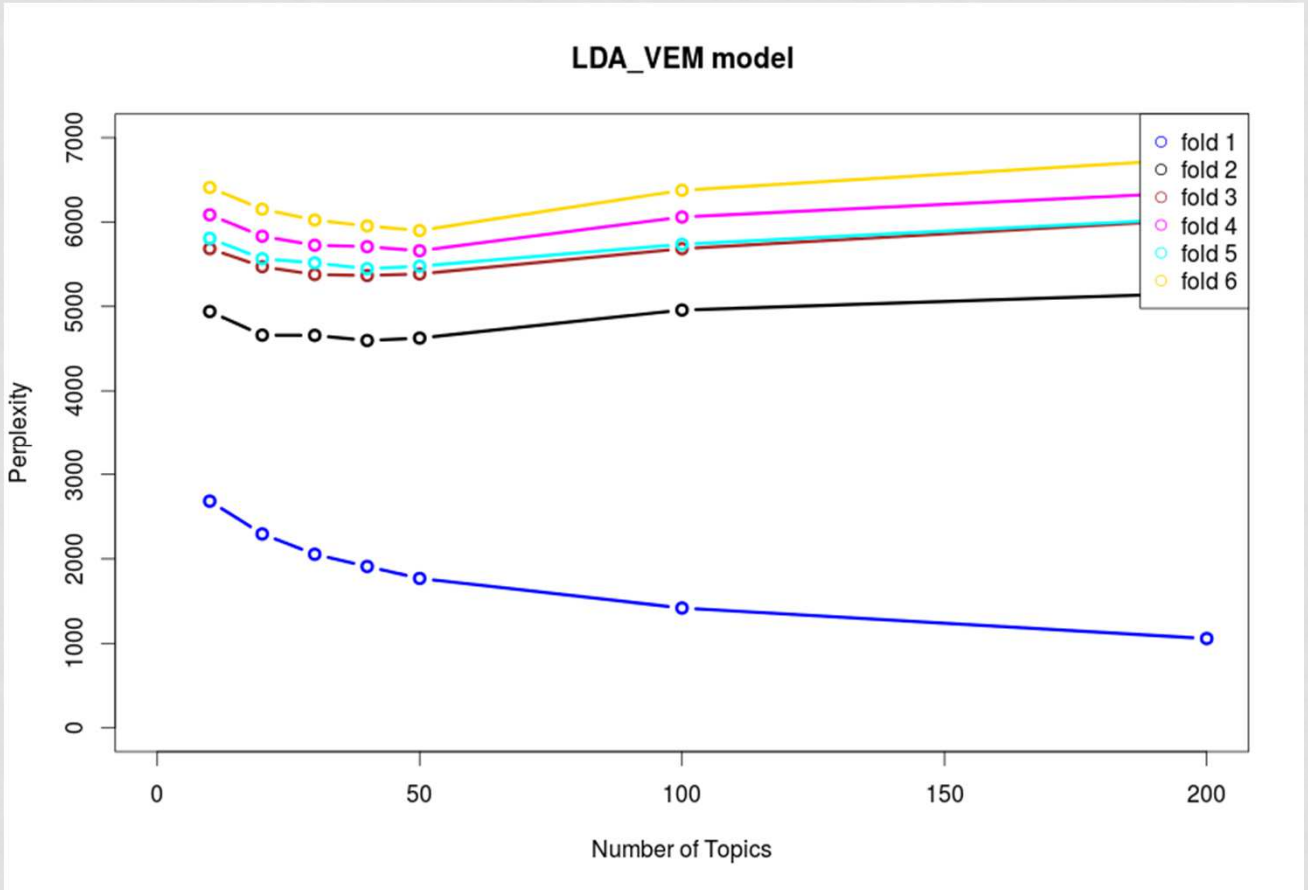
    LDA_CGS
    LDA_VEM (Variational Expectation Maximization)
    LDA_VEM_fixed
    CTM_VEM
    CTM_CGS (working on it)

EJEMPLO

- Programa: 
- Librería: “topicmodels”
- Datos: Associated Press (AP) (2246 documentos)
  Journal of Statistical Software (JSS) (636 documentos)
- Análisis:
  Modelos: LDA_VEM
  Evalución Modelo : 6-fold cross-validation
  Clasificación Volumen 24 modelo JSS (9 documentos)

RESULTADOS: EVALUACIÓN MODELO AP

- 40 tópicos explican mejor los datos:



RESULTADOS: CLASIFICACIÓN MODELO JSS









RESULTADOS: CLASIFICACIÓN MODELO JSS



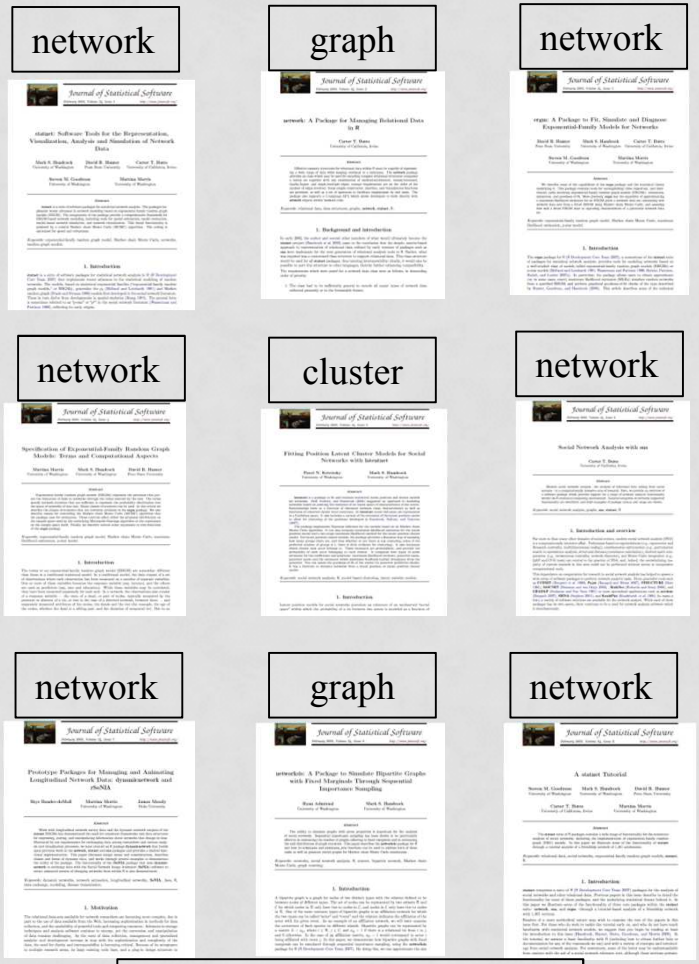






Tópico LDA_VEM	Tópico LDA_VEM	Tópico LDA_VEM
"network"	"cluster"	"graph"
"ergm"	"input"	"ordin"
"document"	"posit"	"file"
"statnet"	"latent"	"treatment"
"graph"	"nearest"	"balance"
"exponential"	"distance"	"boost"
"metabol"	"match"	"demograph"
"pdfs"	"miss"	"exercise"
"random"	"modelbas"	"human"
"polynomi"	"neighbor"	"impure"
"modern"	"posterior"	"machine"
"belief"	"search"	"misclassification"
"coin"	"space"	"ruby"
"flux"	"understand"	"vector"
"isotopom"	forest"	"agestructuedr"
"melt"	"actor"	"biology"

RESULTADOS: CLASIFICACIÓN MODELO JSS



Tópico LDA_VEM	Tópico LDA_VEM	Tópico LDA_VEM
"network"	"cluster"	"graph"
"ergm"	"input"	"ordin"
"document"	"posit"	"file"
"statnet"	"latent"	"treatment"
"graph"	"nearest"	"balance"
"exponential"	"distance"	"boost"
"metabol"	"match"	"demograph"
"pdfs"	"miss"	"exercise"
"random"	"modelbas"	"human"
"polynomi"	"neighbor"	"impure"
"modern"	"posterior"	"machine"
"belief"	"search"	"misclassification"
"coin"	"space"	"ruby"
"flux"	"understand"	"vector"
"isotopom"	forest"	"agestructuedr"
"melt"	"actor"	"biology"

CONCLUSIONES

- Hemos visto que es TM , como crea y asigna tópicos, y como los utiliza para clasificar documentos.
- Hemos utilizado 2246 documentos de la revista AP para crear un modelo LDA_VEM, validado utilizando 6-fold cross-validation, comprobando que el valor optimo de k es de 40 tópicos.
- Hemos utilizado 636 documentos de la revista JSS para crear un modelo LDA_VEM, clasificando el volumen 24 con el tema “network”, y los 9 documentos del volumen con los temas “network”, “cluster” y “graph”.


Mining
Learning
Information
Text
Models
Retrieval
Analysis
Web
Semantic
Classification
Link
mining
Latent
Variable
Extraction
Discovery
Introduction
Experience
School
Search
Data
Eric
Semi-Supervised
Holistic
Access
povzemanja
support
behaviour
Učenje
Active
Large
web
Open
data
Terrorism
SVM's
Document
mrežo
Personalized
Current
Images
Lexicon-Based
Automatic
Approach
Devices
Staff
Never
Probabilistic
Patterns
retrieval
Counter
Systems
Machine
Explanation
Challenges
Reports
classification
News
Summarization
Scale
text
library
Twitter
Art
Engine
Xing
Applications
Language
MEAD
Network
Building
Automated
Mobile
Autumn
Content
Analyzing
semantično
Java
Large-Scale
Semi-supervised
Social
besedil
State
Ending
pretvorbo
Textual
IMF
using
TMIFE
Opinion
Fact

End