Statistical Learning with Networks and Texts

Charles BOUVEYRON

Professor of Statistics
Chair of Excellence Inria on "Data Science"

Laboratoire LJAD, UMR CNRS 7351
Equipe Asclepios, Inria Sophia-Antipolis
Université Côte d'Azur

charles.bouveyron@unice.fr
@cbouveyron

Preamble

"Essentially, all models are wrong but some are useful"

George E.P. Box

Outline

Introduction

In statistical learning, the challenge nowadays is to learn from data which are:

- high-dimensional (p large),
- big or as stream (n large),
- evolutive (evolving phenomenon),
- heterogeneous (categorical, functional, networks, texts, ...)

Introduction

In statistical learning, the challenge nowadays is to learn from data which are:

- high-dimensional (p large),
- big or as stream (n large),
- evolutive (evolving phenomenon),
- heterogeneous (categorical, functional, networks, texts, ...)

In any case, the understanding of the results is essential :

- the practitioners are interested in visualizing or clustering their data,
- to have a selection of the relevant original variables for interpretation,
- and to have a probabilistic model supposed to have generated the data.

Introduction

Statistical analysis of (social) networks has become a strong discipline:

- description and comparison of networks,
- network visualization,
- clustering of network nodes.

Introduction

Statistical analysis of (social) networks has become a strong discipline:

- description and comparison of networks,
- network visualization,
- clustering of network nodes.

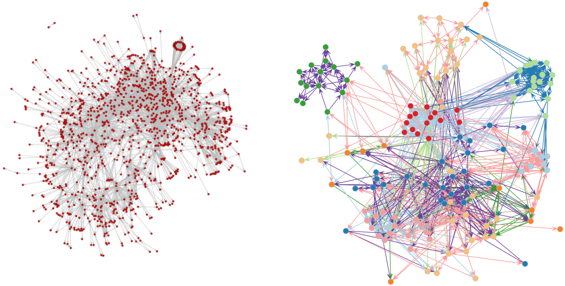
with applications in domains ranging from biology to historical sciences:

- biology: analysis of gene regulation processes,
- social sciences: analysis of political blogs,
- historical sciences: clustering and comparison of medieval social networks
  - Bouveyron, Lamassé et al., *The random subgraph model for the analysis of an ecclesiastical network in merovingian Gaul*, The Annals of Applied Statistics, vol. 8(1), pp. 377-405, 2014.

Introduction

Networks can be observed directly or indirectly from a variety of sources:

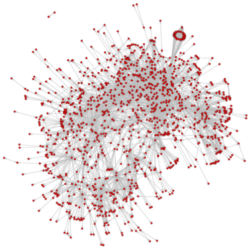
- social websites (Facebook, Twitter, ...),
- personal emails (from your Gmail, Clinton's mails, ...),
- emails of a company (Enron Email data),
- digital/numeric documents (Panama papers, co-authorships, ...),
- and even archived documents in libraries (digital humanities).

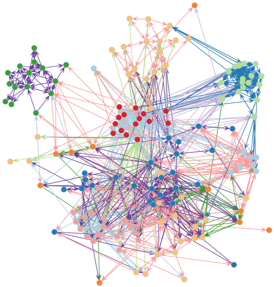


Introduction

Networks can be observed directly or indirectly from a variety of sources:

- social websites (Facebook, Twitter, ...),
- personal emails (from your Gmail, Clinton's mails, ...),
- emails of a company (Enron Email data),
- digital/numeric documents (Panama papers, co-authorships, ...),
- and even archived documents in libraries (digital humanities).





⇒ most of these sources involve text!

An introductory example

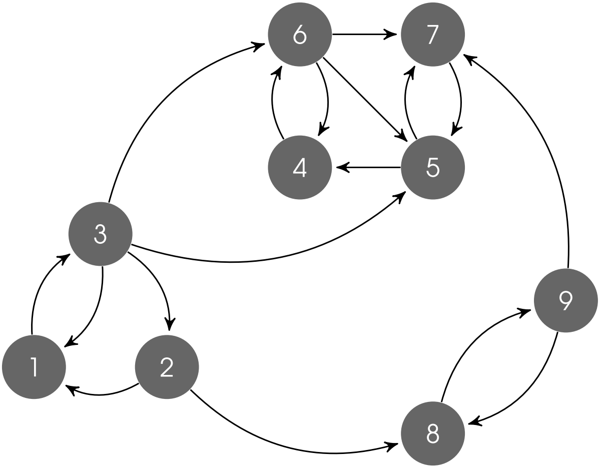


Figure: An (hypothetic) email network between a few individuals.

An introductory example

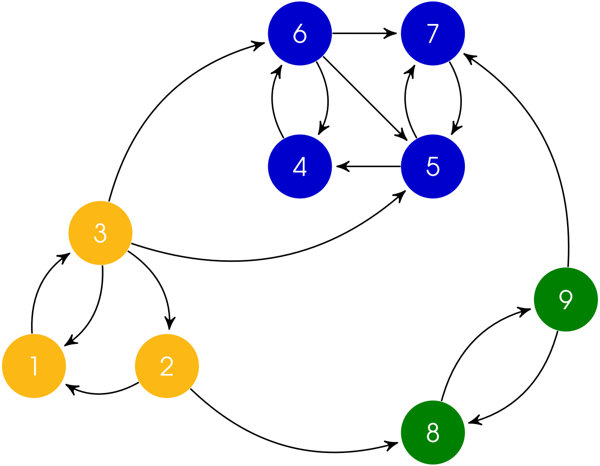


Figure: A typical clustering result for the (directed) binary network.

An introductory example

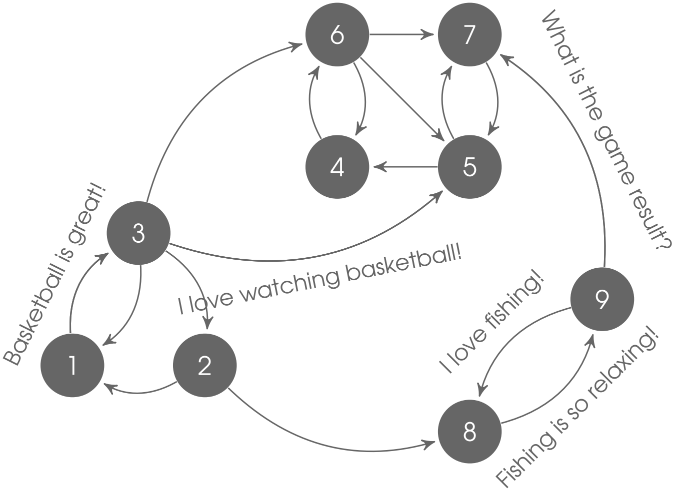


Figure: The (directed) network with textual edges.

An introductory example

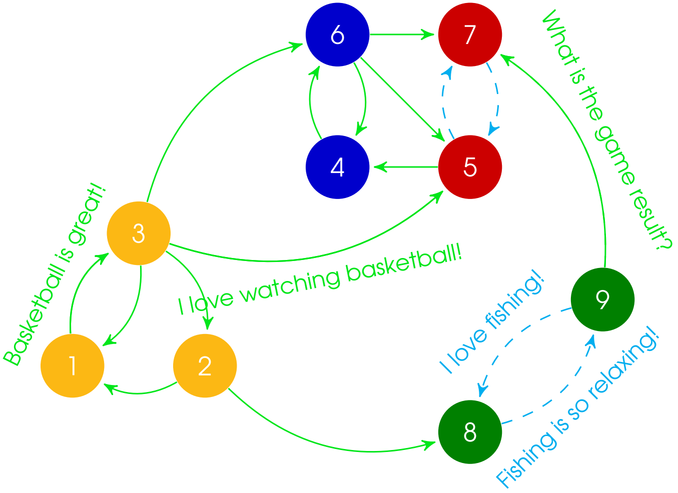


Figure: Expected clustering result for the (directed) network with textual edges.

Outline

STBM : Context and notations

We are interesting in clustering the nodes of a (directed) network of M vertices into Q groups:

- the network is represented by its $M \times M$ adjacency matrix A:

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between i and j} \\ 0 & \text{otherwise} \end{cases}$$

- if $A_{ij} = 1$, the textual edge is characterized by a set of D_{ij} documents:

$$W_{ij} = (W_{ij}^1, ..., W_{ij}^d, ..., W_{ij}^{D_{ij}}),$$

- each document W_{ij}^d is made of N_{ij}^d words:

$$W_{ij}^d = (W_{ij}^{d1}, ..., W_{ij}^{dn}, ..., W_{ij}^{dN_{ij}^d}).$$

STBM : Modeling of the edges

Let us assume that edges are generated according to a SBM model:

- each node i is associated with an (unobserved) group among Q according to:

$$Y_i \sim \mathcal{M}(\rho),$$

  where $\rho \in [0, 1]^Q$ is the vector of group proportions,

STBM : Modeling of the edges

Let us assume that edges are generated according to a SBM model:

- each node i is associated with an (unobserved) group among Q according to:

$$Y_i \sim \mathcal{M}(\rho),$$

  where $\rho \in [0,1]^Q$ is the vector of group proportions,

- the presence of an edge A_{ij} between i and j is drawn according to:

$$A_{ij} | Y_{iq} Y_{jr} = 1 \sim \mathcal{B}(\pi_{qr}),$$

  where $\pi_{qr} \in [0,1]$ is the connection probability between clusters q and r.

STBM : Modeling of the documents

The generative model for the documents is as follows:

- each pair of clusters (q, r) is first associated to a vector of topic proportions $\theta_{qr} = (\theta_{qrk})_k$ sampled from a Dirichlet distribution:

$$\theta_{qr} \sim \mathrm{Dir}\left(\alpha\right),$$

such that $\sum_{k=1}^{K} \theta_{qrk} = 1, \forall (q, r)$.

STBM : Modeling of the documents

The generative model for the documents is as follows:

- each pair of clusters (q, r) is first associated to a vector of topic proportions $\theta_{qr} = (\theta_{qrk})_k$ sampled from a Dirichlet distribution:

$$\theta_{qr} \sim \mathrm{Dir}(\alpha),$$

  such that $\sum_{k=1}^{K} \theta_{qrk} = 1, \forall (q, r)$.

- the nth word W_{ij}^{dn} of documents d in W_{ij} is then associated to a latent topic vector Z_{ij}^{dn} according to:

$$Z_{ij}^{dn} | \{A_{ij} Y_{iq} Y_{jr} = 1, \theta\} \sim \mathcal{M}(1, \theta_{qr}).$$

STBM : Modeling of the documents

The generative model for the documents is as follows:

- each pair of clusters (q, r) is first associated to a vector of topic proportions $\theta_{qr} = (\theta_{qrk})_k$ sampled from a Dirichlet distribution:

$$\theta_{qr} \sim \mathrm{Dir}(\alpha),$$

  such that $\sum_{k=1}^{K} \theta_{qrk} = 1, \forall (q, r)$.

- the nth word W_{ij}^{dn} of documents d in W_{ij} is then associated to a latent topic vector Z_{ij}^{dn} according to:

$$Z_{ij}^{dn} | \{A_{ij} Y_{iq} Y_{jr} = 1, \theta\} \sim \mathcal{M}(1, \theta_{qr}).$$

- then, given Z_{ij}^{dn}, the word W_{ij}^{dn} is assumed to be drawn from a multinomial distribution:

$$W_{ij}^{dn} | Z_{ij}^{dnk} = 1 \sim \mathcal{M}(1, \beta_k = (\beta_{k1}, \ldots, \beta_{kV})),$$

  where V is the vocabulary size.

STBM at a glance...

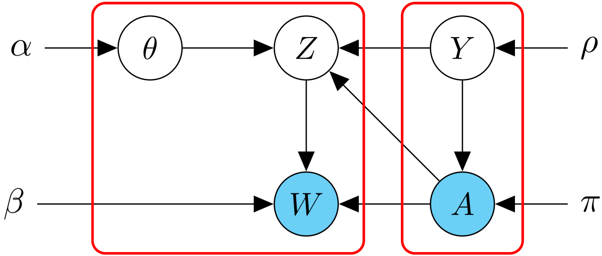


Figure: The stochastic topic block model.

The C-VEM algorithm for inference

The C-VEM algorithm is a follows:

- we use a VEM algorithm to maximize $\tilde{\mathcal{L}}$ with respect β and $R(Z, \theta)$, which *essentially* corresponds to the VEM algorithm of Blei et al. (2003),
- then, $\log p(A, Y|\rho, \pi)$ is maximized with respect to ρ and π to provide estimates,
- finally, $\mathcal{L}\left(R(\cdot); Y, \rho, \pi, \beta\right)$ is maximized with respect to Y, which is the only term involved in both $\tilde{\mathcal{L}}$ and the SBM complete data log-likelihood.

The C-VEM algorithm for inference

The C-VEM algorithm is a follows:

- we use a VEM algorithm to maximize $\tilde{\mathcal{L}}$ with respect β and $R(Z, \theta)$, which *essentially* corresponds to the VEM algorithm of Blei et al. (2003),
- then, $\log p(A, Y|\rho, \pi)$ is maximized with respect to ρ and π to provide estimates,
- finally, $\mathcal{L}\left(R(\cdot); Y, \rho, \pi, \beta\right)$ is maximized with respect to Y, which is the only term involved in both $\tilde{\mathcal{L}}$ and the SBM complete data log-likelihood.

Optimization over Y:

- we propose an online approach which cycles randomly through the vertices,
- at each step, a single vertex i is considered and all membership vectors $Y_{j \neq i}$ are held fixed,
- for vertex i, we look for every possible cluster assignment Y_i and the one which maximizes $\mathcal{L}\left(R(\cdot); Y, \rho, \pi, \beta\right)$ is kept.

Outline

Analysis of the Enron Emails

The Enron data set:

- all emails between 149 Enron employees,
- from 1999 to the bankrupt in late 2001,
- almost 253 000 emails in the whole data base.

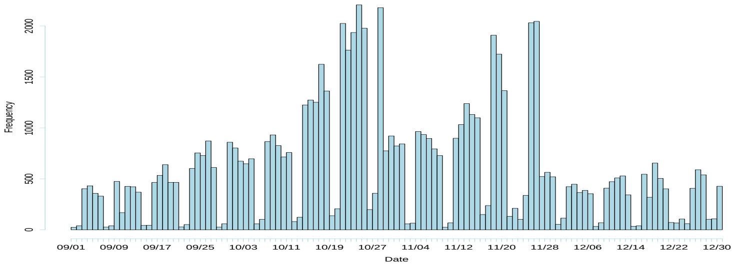


Figure: Temporal distribution of Enron emails.

Analysis of the Enron Emails

Model selection criterion

	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7	K = 8	K = 9	K = 10	K = 11	K = 12	K = 13	K = 14	K = 15	K = 16	K = 17	K = 18	K = 19	K = 20
Q = 14	−1853	−1840	−1835	−1824	−1834	−1823	−1824	−1855	−1845	−1859	−1865	−1868	−1893	−1901	−1915	−1938	−1947	−1955	−1973
Q = 13	−1856	−1841	−1829	−1826	−1827	−1840	−1837	−1839	−1874	−1863	−1874	−1873	−1907	−1911	−1931	−1935	−1956	−1973	−1977
Q = 12	−1853	−1834	−1834	−1828	−1838	−1827	−1851	−1847	−1854	−1879	−1878	−1880	−1901	−1912	−1930	−1948	−1955	−1978	−1998
Q = 11	−1856	−1838	−1836	−1845	−1844	−1831	−1834	−1863	−1877	−1886	−1884	−1900	−1910	−1927	−1968	−1958	−1969	−1991	−1995
Q = 10	−1858	−1840	−1826	$\boxed{-1822}$	−1841	−1837	−1835	−1864	−1857	−1883	−1897	−1912	−1917	−1938	−1945	−1951	−1981	−1975	−1995
Q = 9	−1852	−1835	−1841	−1843	−1825	−1845	−1854	−1863	−1879	−1877	−1894	−1903	−1940	−1936	−1976	−1986	−1982	−2014	−2004
Q = 8	−1858	−1839	−1847	−1836	−1842	−1862	−1845	−1847	−1869	−1873	−1902	−1909	−1927	−1966	−1947	−1988	−2003	−2009	−2013
Q = 7	−1853	−1846	−1840	−1838	−1840	−1854	−1853	−1858	−1899	−1897	−1916	−1918	−1944	−1945	−1958	−1972	−2030	−2027	−2038
Q = 6	−1855	−1842	−1849	−1831	−1842	−1845	−1854	−1864	−1875	−1901	−1921	−1943	−1957	−1974	−1960	−2021	−2015	−2034	−2064
Q = 5	−1857	−1870	−1851	−1860	−1866	−1864	−1902	−1898	−1899	−1919	−1939	−1970	−1970	−1984	−1996	−2013	−2031	−2068	−2085
Q = 4	−1860	−1870	−1870	−1870	−1878	−1891	−1902	−1919	−1895	−1906	−1954	−1954	−1992	−2003	−2018	−2047	−2051	−2064	−2084
Q = 3	−1868	−1876	−1887	−1865	−1915	−1895	−1909	−1926	−1924	−1951	−1964	−1983	−2006	−2014	−2013	−2044	−2062	−2089	−2104
Q = 2	−1876	−1867	−1889	−1912	−1924	−1939	−1957	−1975	−1989	−2009	−2023	−2041	−2054	−2075	−2092	−2106	−2122	−2139	−2157
Q = 1	−1904	−1921	−1938	−1955	−1971	−1988	−2005	−2022	−2038	−2055	−2072	−2089	−2106	−2122	−2139	−2156	−2173	−2189	−2206

Figure: Model selection for STBM on the Enron network.

Analysis of the Enron Emails

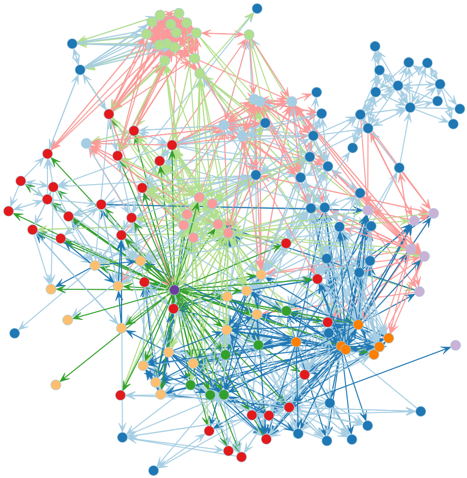


Figure: Clustering of the Enron network.

Analysis of the Enron Emails

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
cycle	grigsby	edison	backup	mmbtud
ofo	afghanistan	puc	seat	harris
usage	viewing	interview	test	watson
select	desk	state	location	capacity
prorata	phillip	interviewers	building	transwestern
storage	ground	dwr	supplies	deliveries
interruptible	park	davis	computer	hayslett
declared	taleban	fantastic	announcement	master
equal	forces	dinner	notified	lynn
ridge	named	said	phones	socalgas
forecast	sheppard	saturday	seats	shackleton
windows	fundamental	super	locations	donoho
wheeler	tori	california	regular	lindy
nom	allen	mara	assignments	kay
injections	ermis	dasovich	rely	sara
elapsed	ina	phase	assignment	geaccone
limits	kuykendall	governor	numbers	pkgs
gas	gaskill	steffes	equipment	juan
receipt	bin	rto	aside	kilmer
clock	heizenrader	contracts	floors	netting

Figure: Most specific terms in the found topics for the Enron data.

Analysis of the Enron Emails

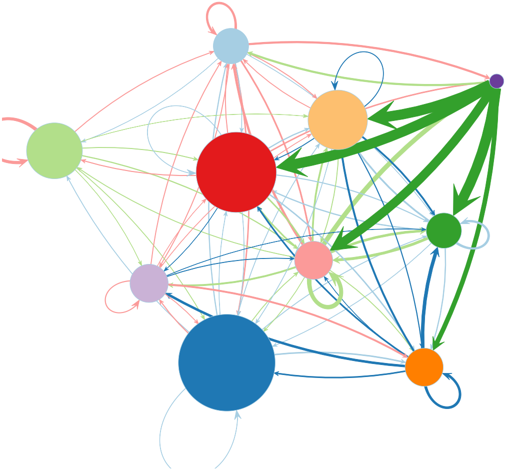


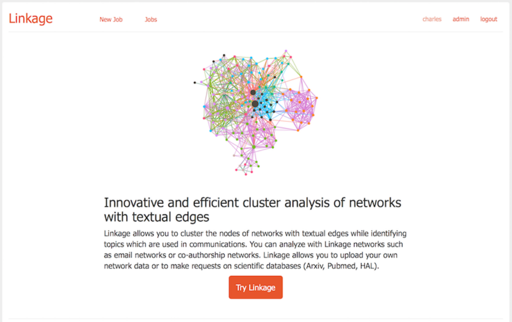
Figure: Meta-network for the Enron data set.

Outline

Innovation: the linkage project

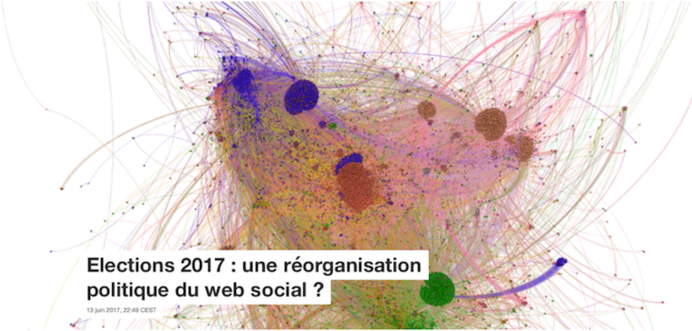
From research to Innovation:

- the project is supported by SATT IDFInnov,
- 50 k€ for SaaS plateform www.linkage.fr
- 200k€ for further work (dynamic, sparsity)



www.linkage.fr

Analysis of the 2017 French presidential election



Elections 2017 : une réorganisation politique du web social ?

13 juin 2017, 22:49 CEST

Réseaux des tweets des français liés à la politique des 17 et 18 avril. P. Latouche, CC BY

Adresse électronique
Twitter 13
Facebook 48
Linkedin 11
Imprimer

Emmanuel Macron vient d'être élu à la présidence de la République sur un programme dont une des priorités est la recomposition de la vie politique. La période que nous traversons, entre les deux tours des législatives, est donc sujette à de fortes interrogations quant à la réorganisation à venir des partis politiques.

Afin d'apporter un éclairage sur ce point, nous avons étudié pendant les semaines qui ont précédé le second tour de l'élection présidentielle les mouvements et transferts entre les partis, avec un prisme particulier,

Auteurs

Pierre Latouche
Maître de conférences en Mathématiques Appliquées, Université Paris 1 Panthéon-Sorbonne

Charles Bouveyron
Professeur des Universités en Mathématiques Appliquées, , Université Paris Descartes – USPC

up5.fr/presid2017

Outline

Conclusion

We proposed a new statistical model, called STBM, for :

- the clustering of the nodes of networks with textual edges,
- which also "clusters" the messages into general topics,
- it provides an effective summary of the whole data (network + texts).

STBM can be applied to:

- communication networks (emails, web forums, twitters, ...),
- co-authorship networks (scientific publications, patents, ...),
- and can even applied to networks with images (Instagram, ...).

Reference:

C. Bouveyron, P. Latouche and R. Zreik, The Stochastic Topic Block Model for the Clustering of Networks with Textual Edges, Statistics & Computing, in press, 2017.