Bayesian statistics, AI, and networks for the analysis of the last French presidential election

Pierre Latouche

pierre.latouche@math.cnrs.fr
https://lmbp.uca.fr/~latouche/
@latouche_pierre

AI meeting, Aubière, 2024

Joint work

Joint work with: C. Bouveyron, R. Zreik, C. Ocanto, S. Petiot, R. Boutin

Outline

Outline

Introduction

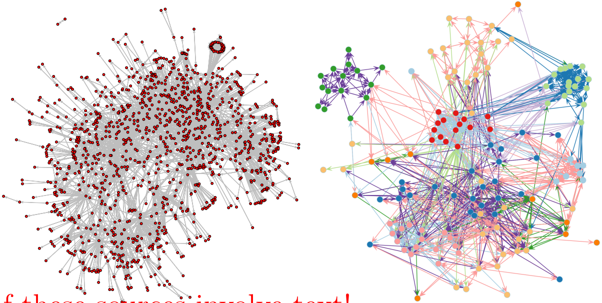
STBM

French presidential election

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, ...),
- mails 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!

Introduction

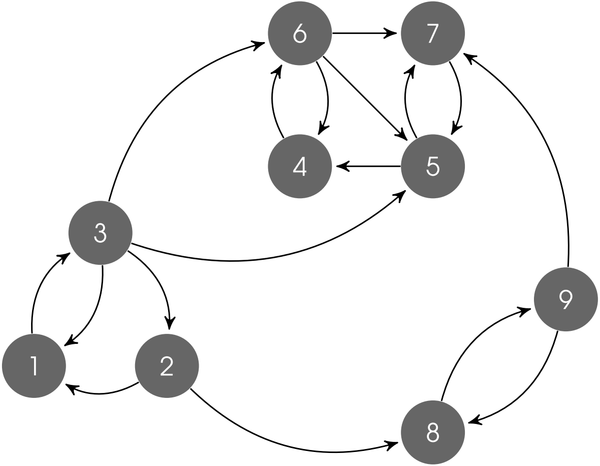


Figure: An (hypothetic) email network between a few individuals.

Introduction

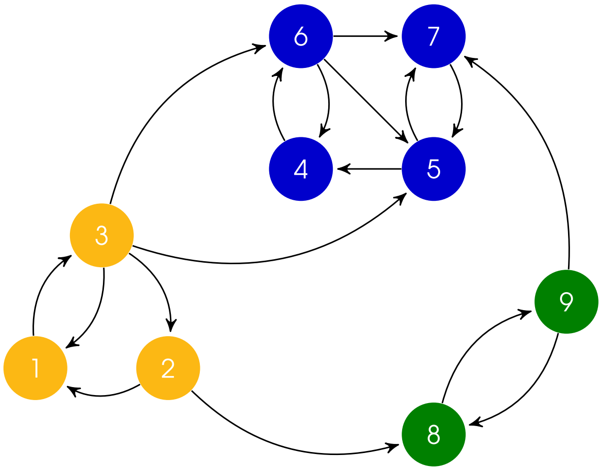


Figure: A typical clustering result for the (directed) binary network.

Introduction

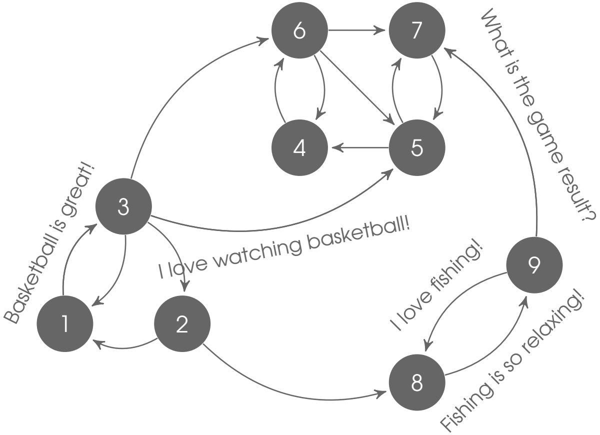


Figure: The (directed) network with textual edges.

Introduction

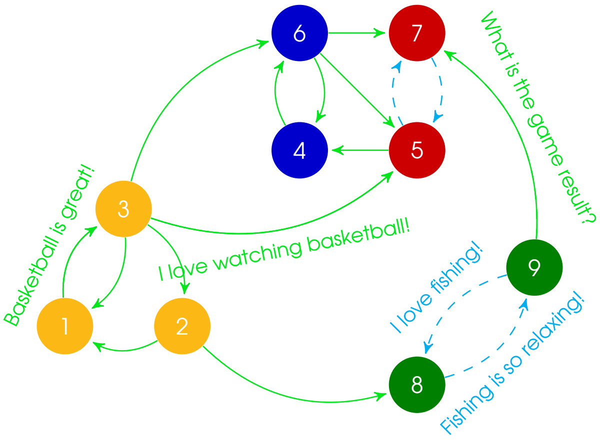


Figure: Expected clustering result for the (directed) network with textual edges.

The stochastic topic block model

the stochastic topic block model (STBM) [BLZ16]:

- Generalizes both SBM and LDA models
- Allows to analyze (directed and undirected) networks with textual edges.

Outline

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:

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

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:
$$Y_i \sim \mathcal{M}(1, \rho),$$
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.

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:

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

  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:

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

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

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

  where V is the vocabulary size.

STBM at a glance...

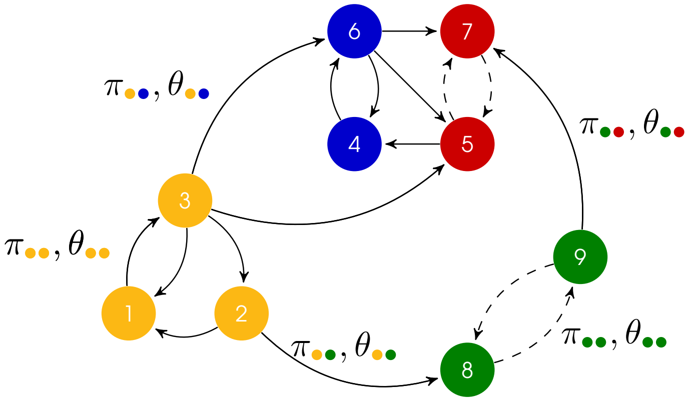


Figure: The stochastic topic block model.

Inference

The full joint distribution of the STBM model is given by:

$$p(A, W, Y, Z, \theta | \rho, \pi, \beta) = p(W, Z, \theta | A, Y, \beta) p(A, Y | \rho, \pi).$$

A key property of the STMB model:

- Let us assume that Y is observed (groups are known),
- It is then possible to reorganize the documents $D = \sum_{i,j} D_{ij}$ documents W such that:

$$W = (\tilde{W}_{qr})_{qr} \text{ where } \tilde{W}_{qr} = \left\{ W_{ij}^{d}, \forall (d, i, j), Y_{iq} Y_{jr} A_{ij} = 1 \right\},$$

- Since all words in $\tilde{W}_{qr}$ are associated with the same pair (q, r) of clusters, they share the same mixture distribution,
- and, simply seeing $\tilde{W}_{qr}$ as a document d, the sampling scheme then corresponds to the one of a LDA model with $D = Q^2$ documents.

Inference

Given the above property of the model, we propose for inference to maximize the complete data log-likelihood:

$$\log p(A, W, Y | \rho, \pi, \beta) = \log \sum_{Z} \int_{\theta} p(A, W, Y, Z, \theta | \rho, \pi, \beta) d\theta,$$

with respect to (ρ, π, β) and $Y = (Y_1, \ldots, Y_M)$.

Inference: the C-VEM algorithm

The C(-V)EM algorithm makes use of a variational decomposition:

$\log p(A, W, Y|\rho, \pi, \beta) = \mathcal{L}(R; Y, \rho, \pi, \beta) + \mathrm{KL}(R \parallel p(\cdot|A, W, Y, \rho, \pi, \beta))$,

where

$$\mathcal{L}(R(\cdot); Y, \rho, \pi, \beta) = \sum_{Z} \int_{\theta} R(Z, \theta) \log \frac{p(A, W, Y, Z, \theta|\rho, \pi, \beta)}{R(Z, \theta)} d\theta,$$

and $R(\cdot)$ is assumed to factorize as follows:

$$R(Z, \theta) = R(Z)R(\theta) = R(\theta) \prod_{i \neq j, A_{ij}=1}^{M} \prod_{d=1}^{D_{ij}} \prod_{n=1}^{N_{ij}^{d}} R(Z_{ij}^{dn}).$$

Inference: the C-VEM algorithm

The lower bound is given by:

$$\mathcal{L}\left(R(\cdot); Y, \rho, \pi, \beta\right) = \tilde{\mathcal{L}}\left(R(\cdot); Y, \beta\right) + \log p(A, Y | \rho, \pi),$$

where

$$\tilde{\mathcal{L}}\left(R(\cdot); Y, \beta\right) = \sum_{Z} \int_{\theta} R(Z, \theta) \log \frac{p(W, Z, \theta | A, Y, \beta)}{R(Z, \theta)} d\theta,$$

and $\log p(A, Y | \rho, \pi)$ is the complete data log-likelihood of the SBM model.
Algorithm: maximize the lower bound with respect to $R(\cdot), Y, \rho, \pi, \beta$, in turn

Model selection

- Need to estimate both Q and K

$$\log p(A, W, Y|K, Q) \approx BIC_{LDA|Y}(Y, K, Q) + ICL_{SBM}(Y, Q),$$

where

$$ICL_{SBM} = \max_{\rho,\pi} \log p(A, Y|\rho, \pi, Q) - \frac{Q^2}{2} \log M(M-1) - \frac{Q-1}{2} \log M,$$

and

$$BIC_{LDA|Y} = \max_{\beta} \tilde{\mathcal{L}} - \frac{K(V-1)}{2} \log Q^2.$$

- BIC here: Laplace (Schwarz, G., 1978) + variational approximation
- ICL: as in Biernacki et al. (2000): Stirling + Laplace

Outline

Introduction

STBM

French presidential election

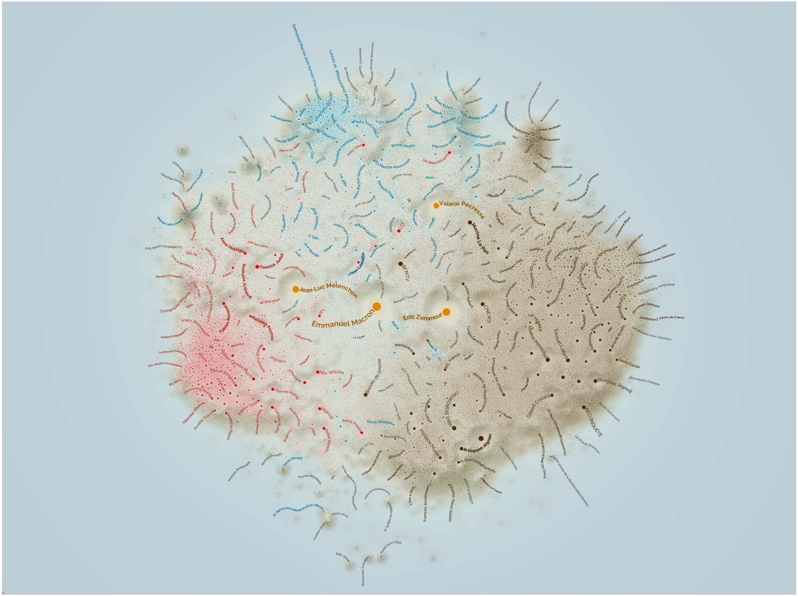
French presidential election

- Analysis of the last French presidential election
- Team:
  - Linkfluence a Meltwater company: G.Fouetillou
  - Linkage: CO, SP, CB, PL
  - M. Jacomy, TANT-lab, Copenhagen
  - 5 journalists of LeMonde : N. Chapuis, M. Goar, S. Auffret, S. Laurent, A. Mestre
- Data:
  - All tweets mentioning at least one of the twelve candidates between the 4th and the 21st of March 2022
  - 11.6 million tweets
  - 53.6%: negative
  - 31.5%: neutrals
  - 14.7%: positive

Data

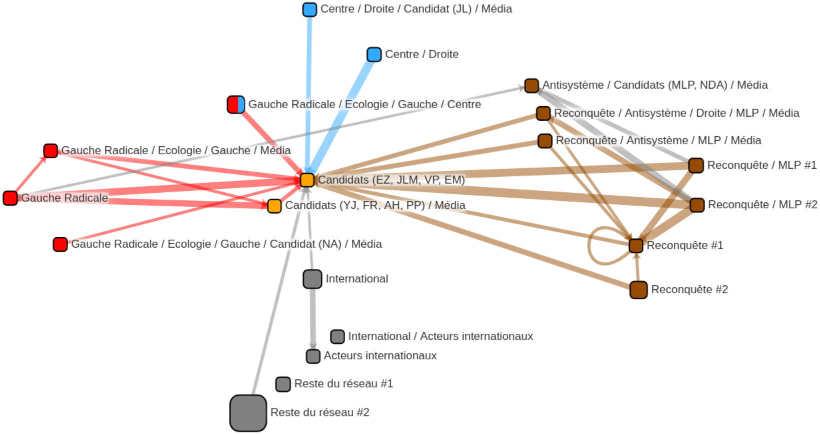
- Nodes = Twitter accounts
- An (directed) edge between two accounts is present if one mentioned or retweeted the other
- If an edge is present, we consider the full text (stacking) of all corresponding tweets
- Focus of the main connected component: 53774 nodes. 597484 edges
- NLP preprocessing
- Look for $Q = 20$ clusters and $K = 8$ with Linkage

The network

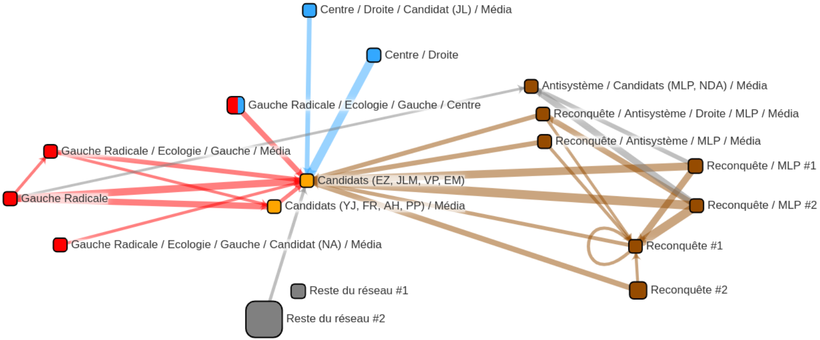


Jean-Luc Mélenchon
Emmanuel Macron
Eric Zemmour
Valérie Pécresse
Marine Le Pen
Yannick Jadot
Anne Hidalgo
Nathalie Arthaud
Philippe Poutou
Fabien Roussel
Adrien Quatennens
Manuel Bompard
Clémence Guetté
Mathilde Panot
Thomas Portes
Nils Wilcke
Libération
BFMTV
CNEWS
Brèves de presse
Samuel Lafont
Stanislas Rigault
Gilbert Collard
Marion Maréchal
Nicolas Dupont-Aignan
Florian Philippot
MIKE BOROWSKI
Denis Cieslik
Guillaume Peltier
RECONQUÊTE
Jean Lassalle
Christophe Castaner
Prisca THEVENOT
Le Monde
Les Républicains
Laurence Trochu
Bruno Jeudy
Duval Philippe
Alexis Poulin

The meta network



The meta network



Conclusions (in short)

- Main core: EZ, JLM, VP, EM
- Second main core: YJ, FR, AH, PP
- Left wing: LFI
- Right wing: Reconquête
- In grey: no connectivity clusters. No statistical decision. In particular: international actors
- Reconquête: use of political astroturfing techniques
- Winner: Reconquête
- Weights:
  - removing the international clusters: Ext Droite: 20%. Gauche radicale: 13%
  - removing all accounts (grey) with no connectivity structures: Ext Droite: 51%. Gauche radicale: 34%

Articles

Le Monde
VENDREDI 1ER AVRIL 2022

ÉLECTION PRÉSIDENTIELLE

Géopolitique de la twittosphère

POLITIQUE ET RÉSEAUX SOCIAUX 1|2 La carte des comptes Twitter liés à la campagne présidentielle montre les grandes zones d'influence d'un continent secoué par des éruptions « antisystème »

Après deux ans de pandémie qui ont changé profondément les rapports sociaux, jamais campagne présidentielle n'a autant reposé sur les outils numériques. Mais jamais l'impact de ces plates-formes sur la décision finale de l'électeur dans l'isoloir n'a été aussi difficile à cerner, à commencer par la plus politique d'entre elles, Twitter – 5,2 millions d'utilisateurs par jour en France selon Médiamétrie, pour plus de 48 millions d'inscrits sur les listes électorales. Comment se fabrique l'opinion des Français en ligne ? Quel camp politique remporte la bataille des réseaux ? Pour quels résultats ? Et, surtout, le leitmotiv des campagnes modernes : comment recréer en ligne cette « conversation » avec les électeurs dont rêvent tous les candidats ?

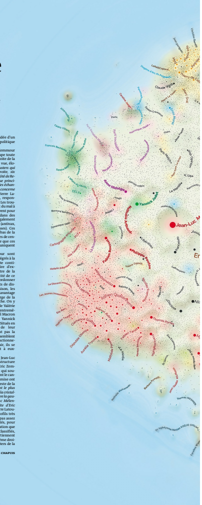
Pour essayer d'apporter des éléments de réponse, *Le Monde* a travaillé en partenariat avec Linkfluence-A Meltwater Company, […] ont travaillé avec l'aide d'un data scientist, Carlos Ocanto, et d'un ingénieur, Stéphane Petiot. Ces équipes ont développé un outil d'intelligence artificielle qui analyse les échanges de messages, mais aussi leur contenu, permettant de dessiner et de comprendre les réseaux créés à travers les interactions numériques.

Leurs résultats ont été exploités par Mathieu Jacomy, chercheur au TANT-Lab, à Copenhague, spécialisé dans l'analyse visuelle des réseaux, qui a pu établir une « cartographie » des interactions sur Twitter, entre le 3 et le 21 mars. Se forment naturellement, au gré des retweets, des mentions ou des mots employés, des « clusters » qui regroupent les comptes qui interagissent de la même façon. L'addition de ces clusters finit par dessiner, sous l'effet d'un algorithme de placement, des territoires militants.

Un instantané du Web politique sur Twitter passionnant, qui permet de confirmer certaines analyses sur les rapports de force politique en ligne […]

LE CAMP POLITIQUE DE ZEMMOUR A DÉMONTRÉ UNE CAPACITÉ À SE COORDONNER POUR PESER SUR LES SUJETS DE DISCUSSION

NICOLAS CHAPUIS



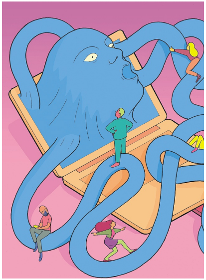
Le Monde
VENDREDI 1ER AVRIL 2022

ÉLECTION PRÉSIDENTIELLE

Faut-il se couper de Twitter, huis clos politique devenu hostile ?

Les équipes des candidats sont toujours rivées à ce réseau agité par les polémiques et les batailles militantes. Surfréquenté par les jeunes, les CSP+ et les populistes, son écosystème ne reflète pas les préoccupations des Français

ANALYSE



LE THÈME DU POUVOIR D'ACHAT, REDEVENU LA PREMIÈRE PRÉOCCUPATION DES FRANÇAIS DEPUIS L'ÉTÉ 2021, A EU BEAUCOUP DE MAL À ÉMERGER SUR TWITTER

SATURÉ PAR LES MILITANTS DE ZEMMOUR

EFFET DE LOUPE

Articles

- A. Mestre, "Eric Zemmour, nouveau président de la fachosphère ?". In: LeMonde (2022), p1. and p. 16-17 **[link]**.
- S. Laurent, "Comment la gauche sociale-démocrate a perdu la bataille des réseaux sociaux". In: LeMonde (2022), p. 16-17 **[link]**
- S. Auffret, "Brigitte Macron et Jean-Michel Trogneux, itiéraire d'une infox délirante". In: LeMonde (2022), p. 16-17 **[link]**
- M. Goar, N. Chapuis, "Présidentielle 2022 : faut-il se couper de Twitter, huis clos politique devenu hostile ?". In: LeMonde (2022), p. 1 and p. 16-19 **[link]**

Conclusion

- STBM: allows to model networks with textual edges
- Extension: ETSBM with ETM / variational auto-encoders
- C-VEM algorithm for inference
- Model selection criterion
- Find clusters of nodes and topics of discussions
- Analysis of the French presidential election

Biblio (1)

Charles Bouveyron, Pierre Latouche, and Rawya Zreik, *The stochastic topic block model for the clustering of vertices in networks with textual edges*, Statistics and Computing (2016), 1–21.