Analyse spectrale de graphes

Thomas Soubiran
Ceraps (UMR 8026 CNRS - Université de Lille)
https://numa.hypotheses.org

Séminaire Quanti

Grenoble, 5 juin 2020

Introduction

- la présentation portera sur **l'analyse spectrale de graphes**
- $\sim$ l'utilisation de la décomposition **en valeurs propres** pour l'analyse de **matrices d'adjacence**
- et, plus particulièrement, différentes variantes **du laplacien d'un graphe**
- avec **deux applications**
  - les flux domicile–travail
  - le pétitionnement en ligne

Préambule

Quelques évidences ?

- postulat de l'existence **de partitions bien définies** dans les données
- postulat que les données sont **cohérentes**
- postulat que **l'information** (au sens mathématique) contenue par les données est **suffisante**

Quelques évidences ?

- postulat de l'existence **de partitions bien définies** dans les données
  - les méthodes de classification sont plus ou moins robustes **au flou des partitions**

    elles supposent souvent des coupures assez nettes
  - selon **des définitions variables** de la netteté de la partition

    qui ne donneront pas nécessairement les mêmes partitions (autre postulat d'équivalence des méthodes de classification)
  - d'où l'importance de déterminer **les critères d'homogénéité** qu'elles mettent en œuvre
  - ainsi que leurs conditions de **félicité et leurs limites**
  - pour jauger leur **adéquation** à un problème spécifique
- postulat que les données sont **cohérentes**
- postulat que **l'information** (au sens mathématique) contenue par les données est **suffisante**

Quelques évidences ?

- postulat de l'existence **de partitions bien définies** dans les données
- postulat que les données sont **cohérentes**
  - et ça, parce qu'il y a **un principe générateur** sous-jacent
  - habitus, rationalité, attitudes, valeurs,. . .ou appartenance à une classe (dans tous les sens du terme)
  - ou, plus particulièrement, à **une « communauté » (pour les graphes)**
  - particulièrement pour les réseaux **issus du web**
  - mais les arêtes d'un graphe **ne manifestent pas** toujours une communauté

    en tout cas pas de façon aussi immédiate
  - et peuvent être le résultat d'aspects **plus contingents**

    ne pas négliger le dispositif (interactions médiatisées par un tiers : l'application) qui fait que le graphe est biparti
  - comment analyser **un site non-communautaire** ?
- postulat que **l'information** (au sens mathématique) contenue par les données est **suffisante**

Quelques évidences ?

- postulat de l'existence **de partitions bien définies** dans les données
- postulat que les données sont **cohérentes**
- postulat que **l'information** (au sens mathématique) contenue par les données est **suffisante**

  - question évoquée dans une perspective différente dans **une autre présentation** à Grenoble en septembre 2019 au colloque inaugurale de la PUD-GA SOUBIRAN[2019b]
  - trouver des partitions nettes rendant compte d'un ou plusieurs principes de cohérence **nécessite de l'information**
  - ce qui renvoie à la question fondamentale de **la collecte des données**
  - disposer de beaucoup d'observations **ne garantit en rien** une grande quantité d'information

    indexer la quantité d'information sur le nombre d'observation peut-être trompeur

Quelques évidences

Relativement à l'analyse de graphes,

- importance **des caractéristiques des sommets du graphes** dans l'analyse de réseau

  sauf graphes structurés ou approche topologique (internet, routage de paquets)

- écho à une critique ancienne, l'analyse de réseaux sociaux ne peut être **un structuralisme stricte**
- dans les faits, les analyses de réseaux sociaux **reviennent aux sommets** (personnes, . . . pour l'interprétation
  - les paramètres sont souvent **interprétés** relativement à d'autres observations sur les caractéristiques des sommets (souvent de nature plus qualitative)

    novices dans un monastère, membres d'un club de karaté, familles florentines, *socialites* d'une ville des États-Unis ou associés d'un cabinet d'avocats

  - comme un nom ou identifiant
- ce type d'aide à l'interprétation **de plus en plus difficile** à mettre en œuvre lorsque le nombre de sommets augmente
  - comme dans le cas de signatures à des pétitions

    3,7 M de signatures, $>$ 12 000 pétitions

  - l'analyse de graphe des co-signatures aux pétitions a été réalisée à partir d'une **classification thématique** du texte des pétitions

Laplacien de graphes

Le spectre des graphes

- l'analyse spectrale des graphes est **ancienne**

  remonte au moins jusqu'au milieu du XIX^{e} siècle

- recouvre **de nombreuses méthodes**

  centralité des vecteurs propres, centralité de Katz, pagerank

- la présentation se concentrera sur l'utilisation **du laplacien d'un graphe** pour la classification

- **à grands traits**, l'analyse spectrale du laplacien d'un graphe consiste à

  1. transformer la matrice d'adjacence
  2. décomposer à la matrice ainsi obtenue en valeur propres
  3. appliquer une méthode classification sur les vecteurs propres

- ce qui nécessite de **formuler le partitionnement** comme un problème dont la solution peut être trouvée par **une décomposition en valeurs propres**

Les laplaciens de graphe

- la transformation de la matrice d'adjacence $\mathbf{A}$ peut prendre plusieurs formes telles que :
  - **le laplacien (combinatoire) :**

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \text{ avec } l_{ij} = \begin{cases} d_{v_i} & \text{si } i = j \\ -a_{ij} & \text{si } i \neq j \end{cases} \quad (1)$$

    où $\mathbf{D}$ est une matrice avec les degrés d_i des sommets v_i dans la diagonale
  - **le laplacien normalisé :**

$$\mathbf{L}_{\text{norm}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} \quad (2)$$

    avec

$$l_{ij} = \begin{cases} 1 & \text{si } i = j \\ -\frac{a_{ij}}{\sqrt{d_{v_i}}\sqrt{d_{v_j}}} & \text{si } i \neq j \end{cases}$$

- **Note :** la matrice binaire $\mathbf{A}$ peut être remplacée par une matrice $\mathbf{W}$ de poids w_{ij}

Critères d'homogénéité

- dans ce qui suit, on se limitera au partitionnement du graphe en **deux composants**

  mais le raisonnement peut être étendu à plus de deux

- classifier revient généralement **à optimiser** (min-max) un ou plusieurs critères
- **critères envisageables** pour partitionner un graphe en deux
  - minimiser **le nombre de liens entre partitions**
  - tout en maximisant **le nombre de lien à l'intérieur** des partitions
- de même que certains algorithmes de classification visent à
  - maximiser **la variance inter-groupes**
  - tout en minimisant **la variance intra**

Coupure et conductivité d'un graphe

- **le nombre de liens** entre deux partitions A et A^c a pour expression

$$vol(A, A^c) = \sum_{i \in A, j \in A^c} a_{ij} \tag{3}$$

On compte le nombre de liens qui ont un sommet dans un partition et dans l'autre

- et porte le nom de **coupure** (cut)
- la coupure est liée à **la conductivité** entre deux partitions d'un graphe

$$\varphi(A) = \frac{vol(A, A^c)}{\min(vol(A), vol(A^c)} \tag{4}$$

- le laplacien du graphe permet de trouver **les partitions qui minimisent la coupure du graphe**

$$\phi(G) = \min_{A \in V} \varphi(A) \tag{5}$$

soit la coupure du graphe avec la plus petite conductivité

Décomposition en valeurs propres et singulières

- décomposition **en valeurs propres** a pour expression :

$$\mathbf{A} = \mathbf{x}^\top \lambda \mathbf{x} \tag{6}$$

$$\lambda = \mathbf{x}^\top \mathbf{A}^\top \mathbf{x} \tag{7}$$

- **interprétation géométrique** intuitive

  ACP : directions principales du nuages de points suivant une distribution normale multivariée

- les valeurs propres peuvent être vue comme la solution du problème d'optimisation suivant (**caractérisation variationnelle** des valeurs propres) :

$$\max_{\mathbf{x}\in\mathbb{R},\ \mathbf{x}\neq 0} \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \max_{\mathbf{x}\in\mathbb{R},\ ||\mathbf{x}||_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} \tag{8}$$

- dans ce qui suit, on va chercher **à exprimer le problème du partitionnement** sous cette forme

Propriétés du laplacien combinatoire

- en remplaçant la matrice **A** par **L** dans la partie droite de (7) ($\mathbf{x}^\top \lambda \mathbf{x}$) et en substituant provisoirement **f** à **x**, **on obtient après développement** Luxburg[2007] :

$$\begin{aligned}\mathbf{f}^\top \mathbf{L}\mathbf{f} = \mathbf{f}^\top(\mathbf{D}-\mathbf{W})\mathbf{f} = \mathbf{f}^\top \mathbf{D}\mathbf{f} - \mathbf{f}^\top \mathbf{W}\mathbf{f} &= \sum_{i=1}^{N} d_i^2 - \sum_{i,j=1}^{N} w_{ij} f_i f_j \\ &= \frac{1}{2}\left(\sum_{i=1}^{N} d_i^2 - 2\sum_{i,j=1}^{N} w_{ij} f_i f_j + \sum_{j=1}^{N} d_j^2\right) \\ &= \frac{1}{2}\sum_{i,j=1}^{N} w_{ij}(f_i - f_j)^2 \qquad (9)\end{aligned}$$

- on a donc :
  - $\mathbf{f}^\top \mathbf{L}\mathbf{f} \geq 0 \; \forall f \in \mathbb{R}$, L est donc une matrice définie semi-positive

    quelque soit la matrice **A**, *on peut appliquer une décomposition en valeur propre à* **L**
  - sa plus petite valeur propre vaut zéro et le vecteur propre correspondant, $\mathbb{1}$
  - en conséquence, **L** a N valeurs propres non négatives $0 \leq \lambda_1, \ldots, \lambda_n$
- de plus, **la multiplicité des zéros** donne le nombre de composants connectés

Propriétés du laplacien normalisé

- pour $\mathbf{L}_{\text{norm}}$

$$\mathbf{L}_{\text{norm}} = \frac{1}{2} \sum_{i,j=1}^{N} w_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2$$

- $\mathbf{L}_{\text{norm}}$ est aussi définie semi-positive
- sa plus petite valeur propre vaut aussi zéro et le vecteur propre correspondant, $\mathbf{D}^{1/2}\mathbb{1}$
- $\mathbf{L}_{\text{norm}}$ a N valeurs propres non négatives $0 \leq \lambda_1, \ldots, \lambda_n$
- de même, la multiplicité des zéros donne le nombre de composants connectés
- $\mathbf{L}_{\text{norm}}$ est de plus liée à une autre variante du laplacien qui s'apparente à une marche aléatoire sur un graphe

Interprétation de f

- le vecteur $\mathbf{f}$ peut être interprété comme **une variable indicatrice** de l'appartenance à une grappe $\mathbf{f} \in \{a, b\}$
- toutefois a et b **ne peuvent pas prendre** n'importe quelle valeur
  - car $\mathbf{f}$ doit **être orthogonal** à $\mathbb{1}$ pour le laplacien combinatoire

    pour $\mathbf{L}$ on a doit avoir $\mathbf{Df} \perp \mathbb{1}$
  - donc $\mathbf{f} \notin \{0, 1\}$
- par contre,
  - $\mathbf{f} \in \{-1, 1\}$, entre autres possibilités
  - dans ce cas, $\mathbf{f}^\top \mathbf{Lf}$ permet de calculer la coupure du graphe
- **Exemple :** : *dumbbell graph*

Exemple

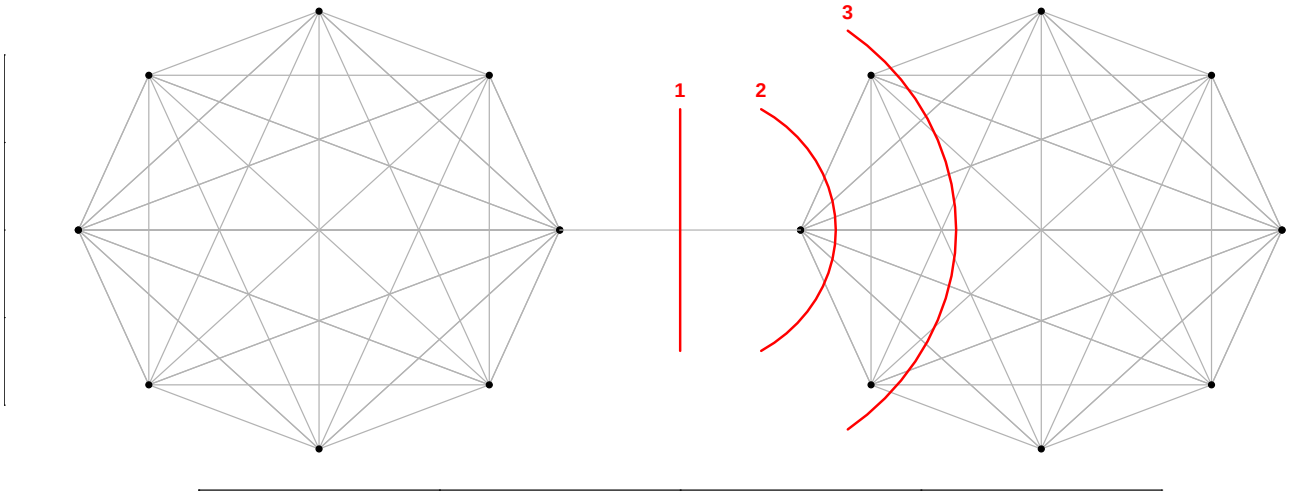


FIGURE 1 – Coupures de graphes

Pour les trois différentes partitions du graphe , $1/4\,\mathbf{f}^\top \mathbf{L}\mathbf{f}$ vaut respectivement

1. 1 ($\mathbf{f} = \{1,1,1,1,1,1,1,1,-1,-1,-1,-1,-1,-1,-1,-1\}$)
2. 7 ($\mathbf{f} = \{1,1,1,1,1,1,1,1,1,-1,-1,-1,-1,-1,-1,-1\}$)
3. 15 ($\mathbf{f} = \{1,1,1,1,1,1,1,1,1,1,1,-1,-1,-1,-1,-1\}$)

soit **le nombre de liens ayant des sommets** dans les deux partitions pour chaque cas.
En effet,

$$(f_i - f_j)^2 = \begin{cases} (1-(-1))^2 = 4 & \text{si } i \in A \text{ et } j \in B \\ (-1-(1))^2 = 4 & \text{si } i \in B \text{ et } j \in A \\ 0 & \text{sinon} \end{cases}$$

Critères de partionnement

- **L** et $\mathbf{L}_{\text{norm}}$ visent tous les deux à **minimiser la coupure** mais de façon différente
- **L** et $\mathbf{L}_{\text{norm}}$ minimisent en effet respectivement :
  - ***Ratio cut*** :

$$\texttt{RCut}\ (A, A^c) = \left(\frac{1}{|A|} + \frac{1}{|A^c|}\right) cut(A, A^c)$$

$$\texttt{RCut}\ (A_1, \ldots, A_K) = \sum_{k=1}^{K} \frac{cut(A_i, A_i^c)}{|A_i|} \tag{10}$$

  - ***Normalized cut*** :

$$\texttt{NCut}\ (A, A^c) = \left(\frac{1}{vol(A)} + \frac{1}{vol(A^c)}\right) cut(A, A^c)$$

$$\texttt{Ncut}\ (A_1, \ldots, A_K) = \sum_{k=1}^{K} \frac{cut(A_i, A_i^c)}{vol(A_i)} \tag{11}$$

avec $vol(A_i) = \sum_{v_i \in A_i} d_i$

Critères de partitionnement

- (10) et (11) correspondent à **deux normalisations** de la coupure
  - soit **par le nombre de sommets** dans chaque partition (10)

    $\mathbf{f}$ doit alors satisfaire à $\mathbf{f}^\top \mathbf{f} = N$
  - soit **par le nombre de liens** (éventuellement pondéré) entre les sommets de chaque partition (11)

    $\mathbf{f}$ doit alors satisfaire à $\mathbf{f}^\top \mathbf{f} = vol(G)$
- les laplaciens standardisent donc la coupure entre partitions
- intuitivement, le niveau de séparation induit par une coupure est **plus ou moins marquée** si une des partition contient un petit ou un grand nombre de sommet ou de liens
- en pratique, **si on minimise directement la coupure**
  - on risque d'obtenir des solutions qui isolent un (ou un petit nombre de) sommet du graphe
  - ce qui n'est pas l'objectif recherché ici

Laplacien combinatoire et Ncut

- si on choisit $\mathbf{f}$ tel que Luxburg[2007]

$$f_i = \begin{cases} \sqrt{|A^c|/|A|}, \text{si } v_i \in A \\ -\sqrt{|A|/|A^c|}, \text{si } v_i \in A^c \end{cases} \tag{12}$$

- $\mathbf{f}$ satisfait **les deux contraintes** $\mathbf{f} \perp \mathbb{1}$ et $\mathbf{f}^\top \mathbf{f} = N$

  Note : lorsque $|A| = |A^c|$, on a $\mathbf{f} \in \{-1, 1\}$

- alors,

$$\begin{aligned} \mathbf{f}^\top \mathbf{L}\mathbf{f} &= \frac{1}{2} \sum_{i,j=1}^{N} w_{ij}(f_i - f_j)^2 \\ &= 2|V| Rcut(A, A^c) \end{aligned} \tag{13}$$

- comme on souhaite **minimiser le Ncut**, le problème peut donc être formulé comme suit

$$\min \mathbf{f}^\top \mathbf{L}\mathbf{f} \qquad \mathbf{f} \perp \mathbb{1}, \ ||\mathbf{f}|| = \sqrt{N} \tag{14}$$

  cf. (8) (presque) mais en transformant le problème en minimisation

  on cherche les plus petites valeurs propres et non les plus grandes

Laplacien normalisé et Rcut

- si on choisit $\mathbf{f}$ tel que

$$f_i = \begin{cases} \sqrt{vol(A^c|/vol(A)}, \text{si } v_i \in A \\ -\sqrt{vol(A/vol(A^c)}, \text{si } v_i \in A^c \end{cases} \tag{15}$$

- $\mathbf{f}$ satisfait **les deux contraintes** $\mathbf{f} \perp \mathbb{1}$ et $\mathbf{f}^\top \mathbf{f} = vol(G)$
- alors,

$$\begin{aligned} \mathbf{f}^\top \mathbf{L}_{\text{norm}} \mathbf{f} &= \frac{1}{2} \sum_{i,j=1}^{N} w_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2 \\ &= 2vol(V) Ncut(A, A^c) \end{aligned} \tag{16}$$

- comme on souhaite **minimiser le Rcut**, le problème peut donc être formulé comme suit

$$\min \mathbf{f}^\top \mathbf{L}_{\text{norm}} \mathbf{f} \qquad \mathbf{Df} \perp \mathbb{1},\ \mathbf{f}^\top \mathbf{Df} = vol(V) \tag{17}$$

Note : le calcul du laplacien peut être transformé en maximisation

Décontraction du problème

- (17) (14) **ne permettent pas** de trouver une solution au problème dans un temps raisonnable

  NP-hard XXX

- **f** est **une variable indicatrice** et ne peut donc prendre que deux valeurs
- on **desserre donc les contraintes** en demandant seulement que $f_i \in R$

  approximation d'un problème discret

- f_i marque alors **le degré d'appartenance** du sommet v_i à une partition
- c'est pourquoi il faut **rajouter une étape supplémentaire** à la décomposition en valeurs propres
- pour partitionner le graphe

Quel laplacien ?

- le laplacien (combinatoire) est plutôt adapté pour **les graphes réguliers**
- ou dont le degré des sommets **varie peu**
- dans le cas contraire, il vaut mieux utiliser **le laplacien normalisé**
- sans pour autant que la distribution des degrés **soit trop hétérogène**

Laplacien régularisé

- comme souvent, la classification spectrale fonctionne bien sur des partitions **clairement séparées**
- ce qui n'est **pas forcément le cas** en pratique

  ce qui se manifeste par l'hétérogénéité des degrés des sommets

- alternative **pour tempérer** l'effet de l'hétérogénéité, **le laplacien régularisé** :

$$\mathbf{L}_\tau = \mathbf{D}_\tau^{-1/2}\mathbf{W}\mathbf{D}_\tau^{1/2} \tag{18}$$

  avec $\mathbf{D}_\tau = \mathbf{D} + \tau\mathbf{I}$

- le laplacien régularisé minimise le critère suivant Zhang et Rohe[2018]

$$\texttt{CoreCut}_\tau = \frac{cut(A, A^C) + \frac{\tau}{N}|A||A^c|}{vol(A, A^c) + \tau|V|} \tag{19}$$

- τ peut prendre **différentes valeurs**

  $\tau = \bar{d}_v$ marche bien en théorie et en pratique

- liens avec **le SBM** (*Stochastic Blockmodel*) Qin et Rohe[2013]

Applications

La base sur les flux de mobilité des « déplacements domicile-travail » de l'Insee

La base sur les flux de mobilité de l'INSEE

- données diffusées par l'INSEE issues **du recensement de population**
- renseignant **les flux** entre lieu de résidence avec le lieu de travail pour une année donnée
- la documentation indique qu'il s'agit **des déplacements domicile–travail** (navette)
- toutefois,
  - au regard des distances, semble plutôt renseigner les flux entre la résidence principale déclarée et le lieu de travail

    certains trajets traversent littéralement la France, voire proviennent de DOM
  - et agréger les mobilités professionnelles aux trajets quotidiens

    et ne semblent pas distinguer, p. ex., des personnes résident à proximité de leur lieu de travail en semaine mais qui habitent ailleurs le reste du temps
- les flux combinent sans doute **deux types de mobilité professionnelle**

  ce qui est susceptible de compliquer l'analyse en rendant les partitions plus floues

Le recensement général de population

- la particularité du fgurecensement français est d'être une enquête **par sondage probabiliste**
- en effet,
  - bon an, mal an, le recensement s'est déroulé en France **tous les cinq ans** de 1801 jusqu'aux années cinquante
  - l'intervalle inter-censitaire a ensuite progressivement **augmenté**
  - jusqu'à atteindre **neuf ans** en 1999
- l'INSEE a ensuite procédé à une **« rénovation »** radicale du RGP

  et ce, à l'invitation des pouvoirs publics
- qui a été effective **partir de 2004**

Les enquêtes annuelles de recensement de la population

- le recensement est une enquête tournante **par sondage** réalisée annuellement sur une période de six ans
- durant laquelle seule **une fraction** des ménages (logements ?) est interrogée
- en effet, chaque année :
  - seule une fraction des communes de moins de 10 000 hab.
  - seule une fraction des ménages des communes de plus de 10 000 hab.
  - sont interrogées
- ce changement a suscité de **nombreuses interrogations** dès sa mise en œuvre

  et continue d'en susciter, particulièrement de la part d'élus locaux
- les flux ne sont donc pas **comptabilisés** mais bien **inférés**

Note : pour plus de détails, voir notamment le n° spécial d'*Économie et statistiques* de 2016 sur le recensement rénové (n° 483-484-485 – 2016)

Les trajets entre communes

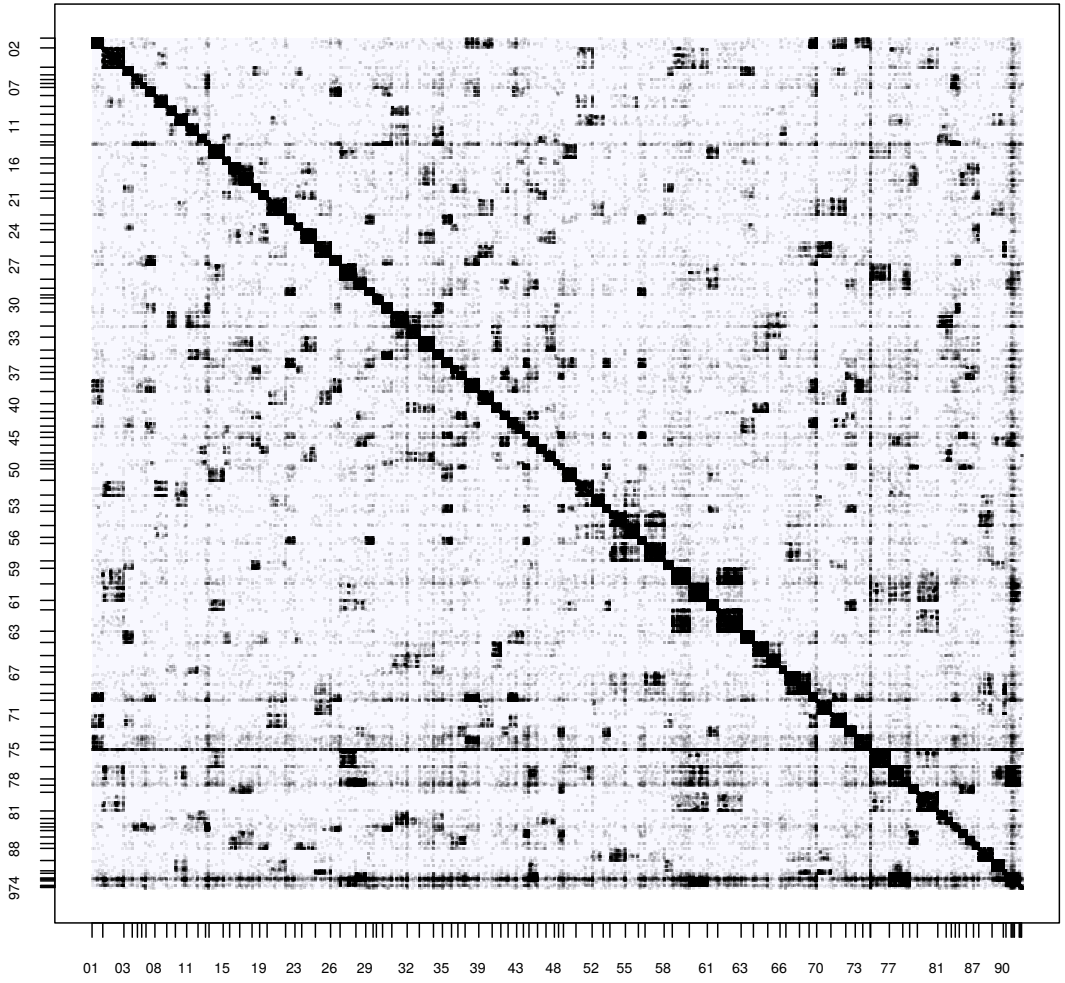
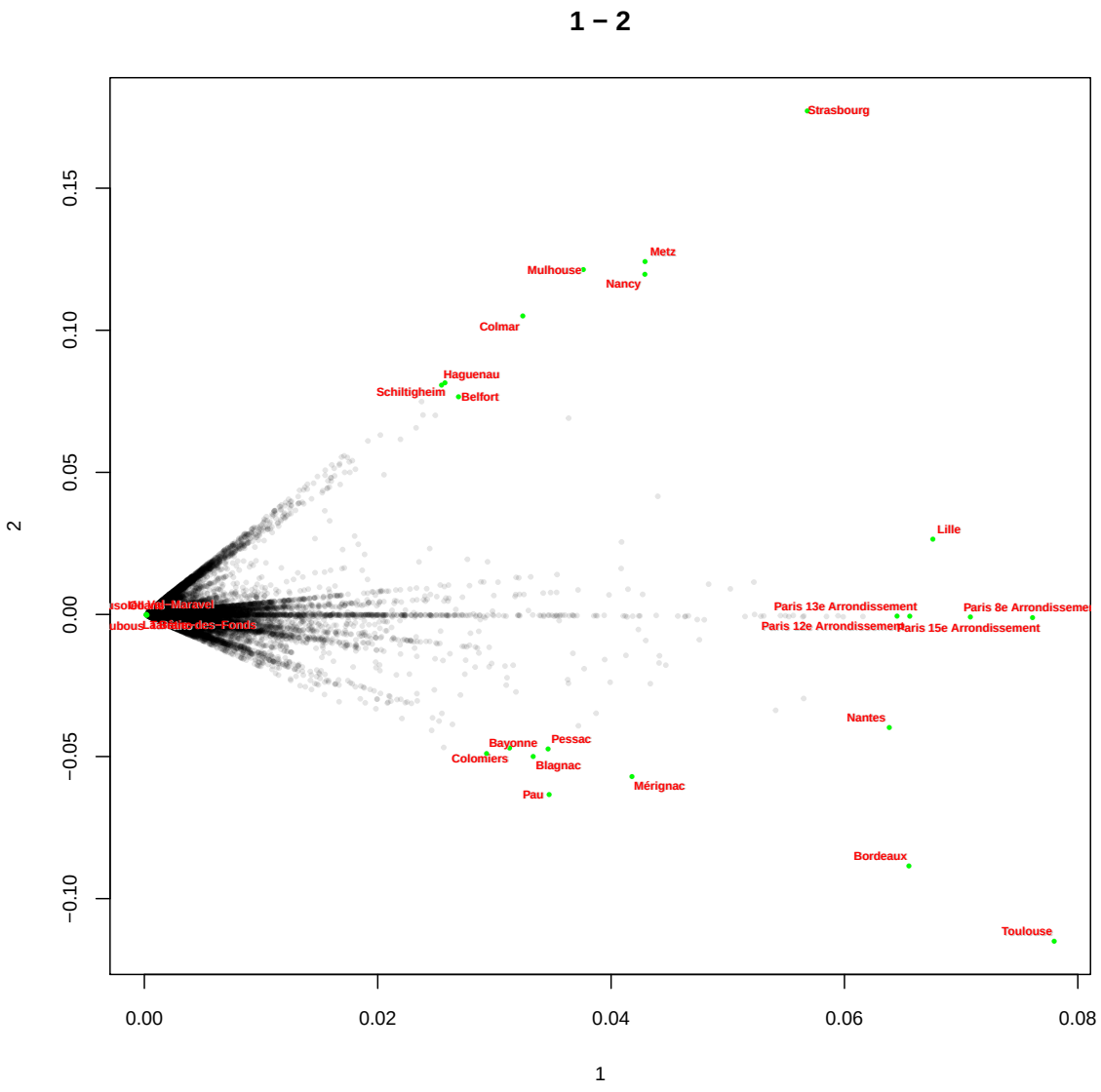
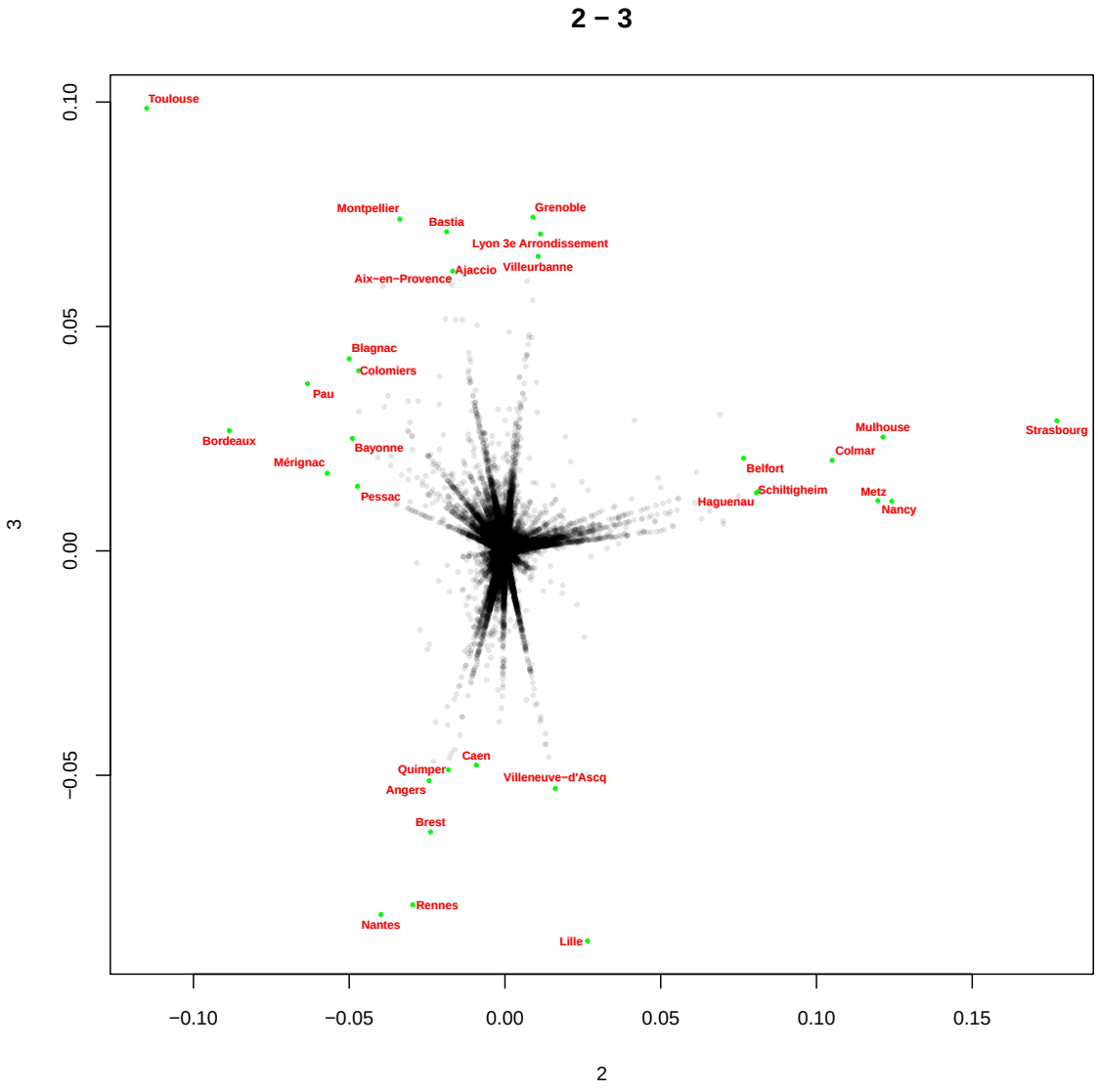


FIGURE 2 – Relations binaires entre les communes ($a_{ij} > 0$)), tri par cantons

Laplacien normalisé I



Laplacien normalisé II



Laplacien normalisé

- la première dimension correspond à $\lambda = 0$ et est fonction **du degré des sommets**

  -cf. (13)

- les premières dimensions dessinent **une coupure est-ouest et nord-sud**
- qui fait transparaître **« la diagonale du vide »** Oliveau et Doignon[2016]
- on obtient donc des **coupures nationales, loin des navettes**

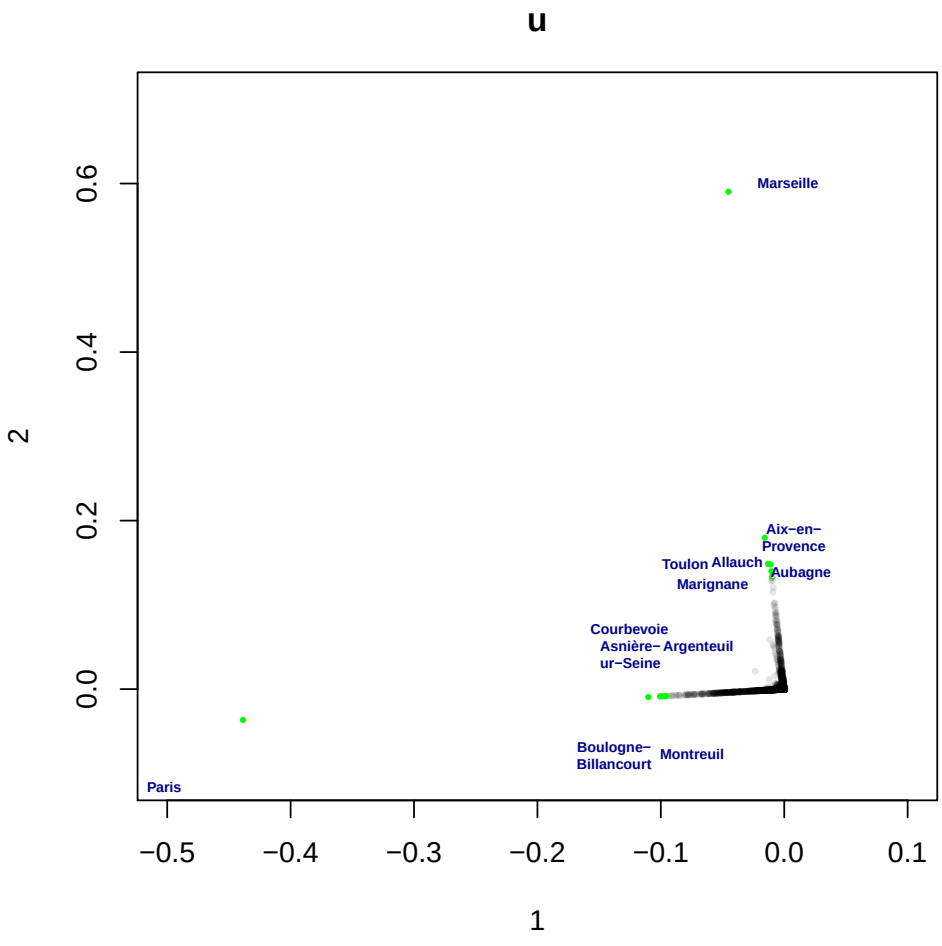
  sans doute en partie dû au fait que les flux mélangent différents types de mobilité professionnelles

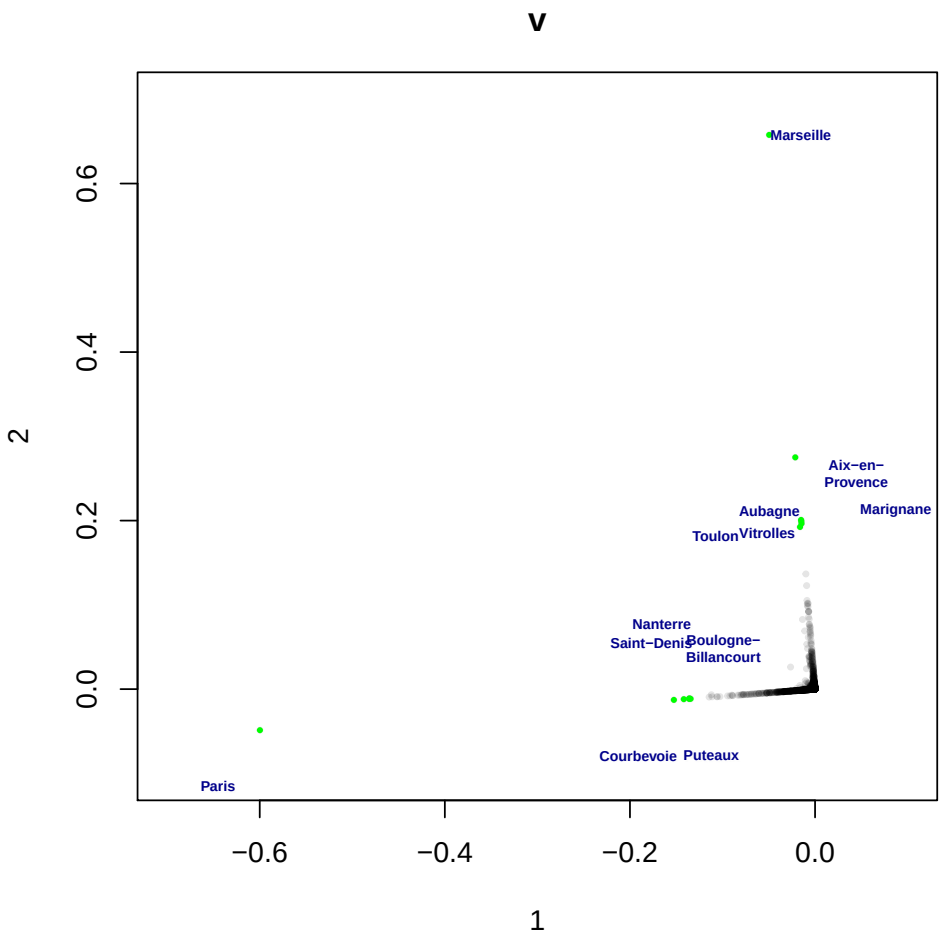
- effet de **la force des liens faible** ?

  petits diamètres sur le graphe qui sont un défi aux distances physiques

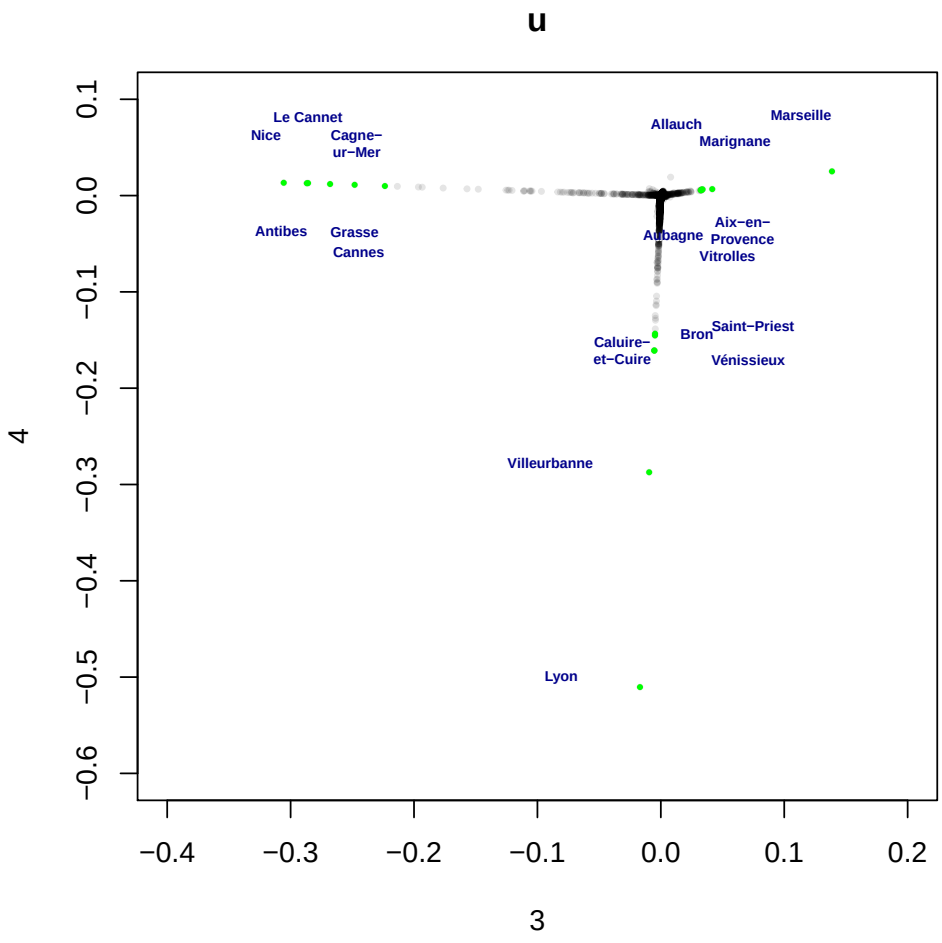
- qui conduisent à **flouter les partitions** ?

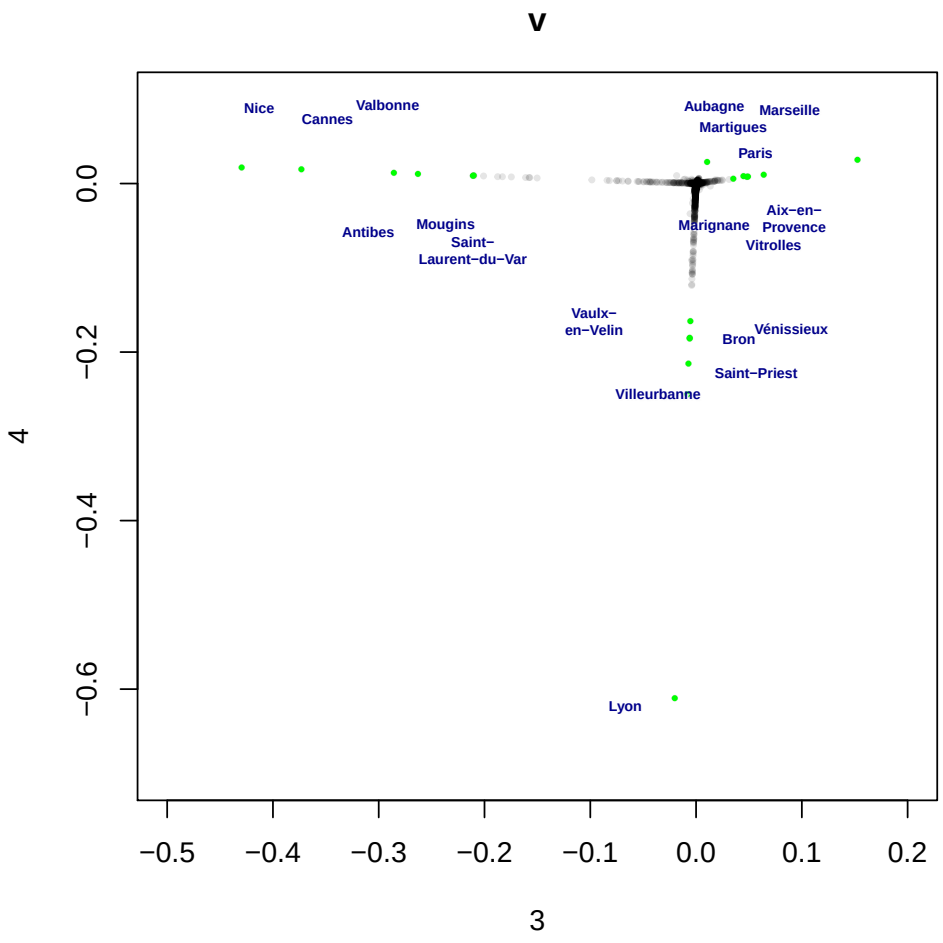
Laplacien régularisé I



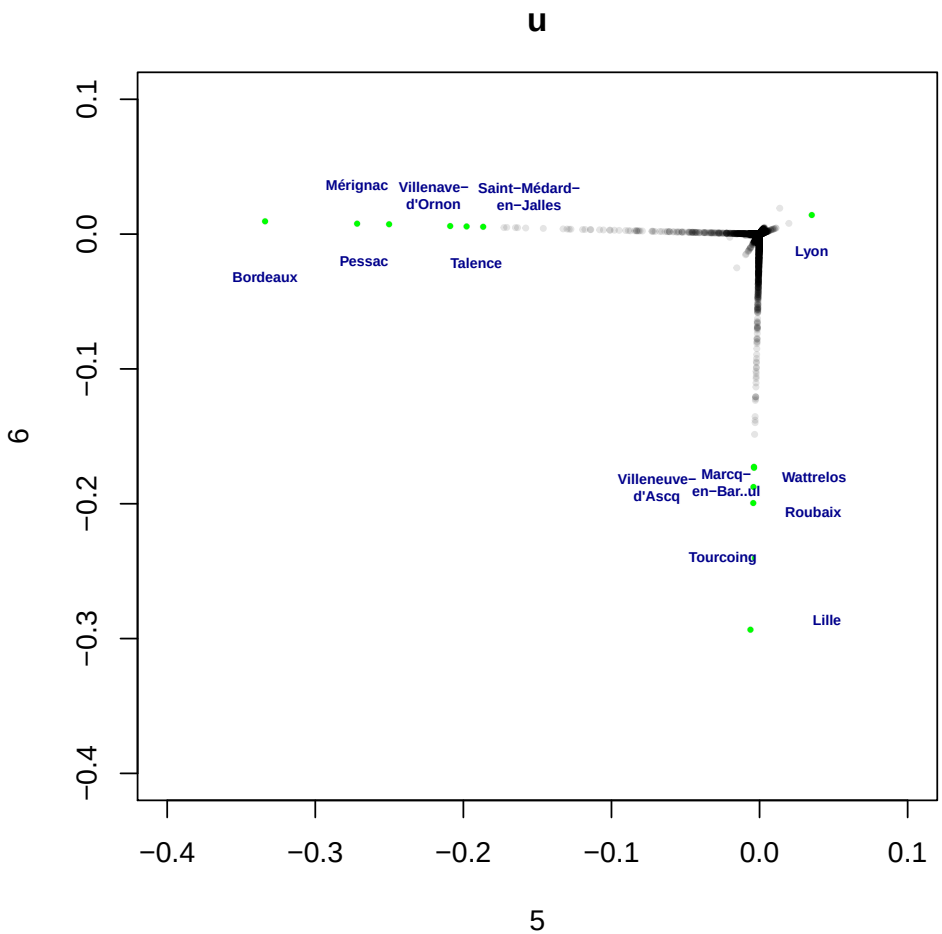


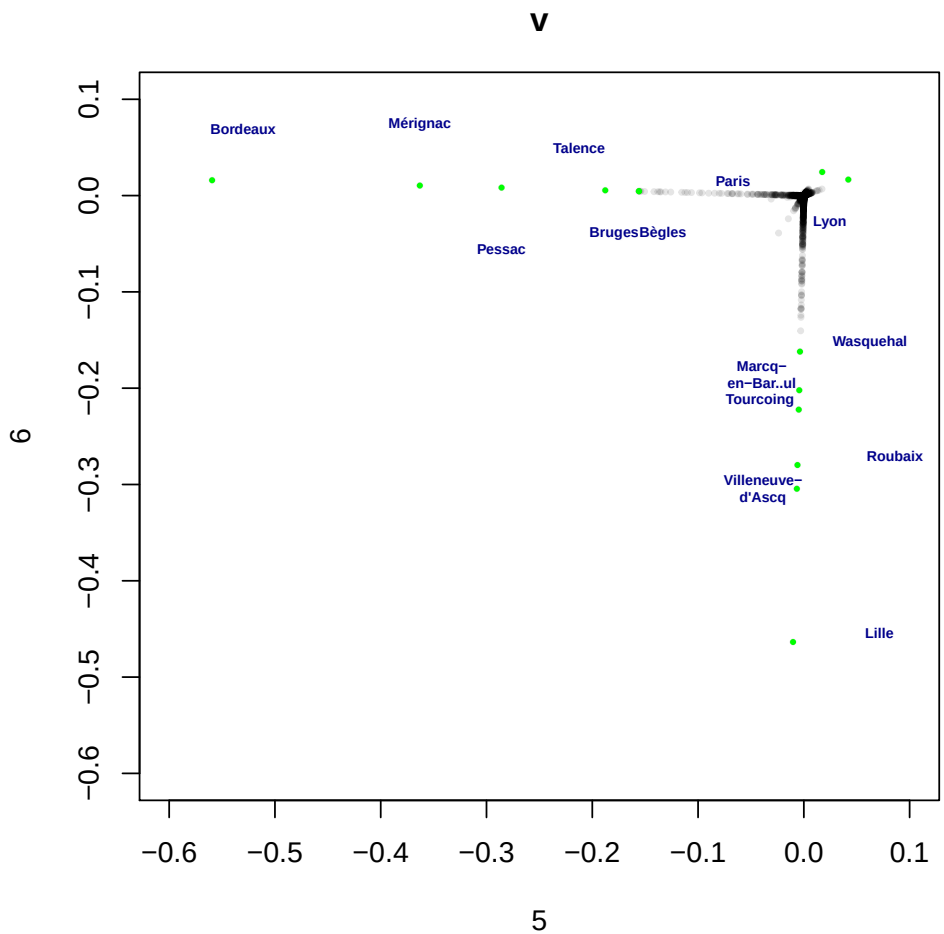
Laplacien régularisé II



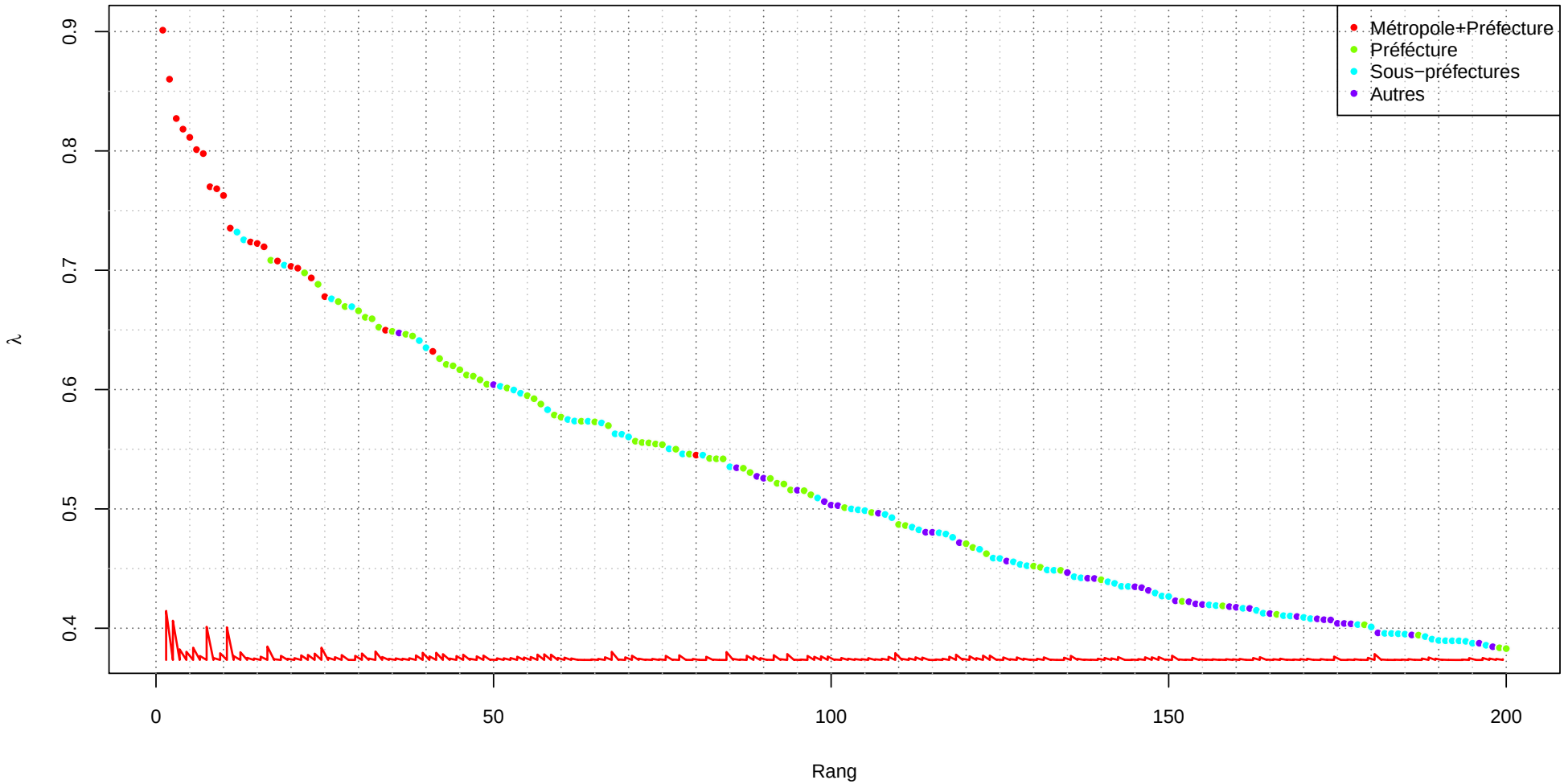


Laplacien régularisé III

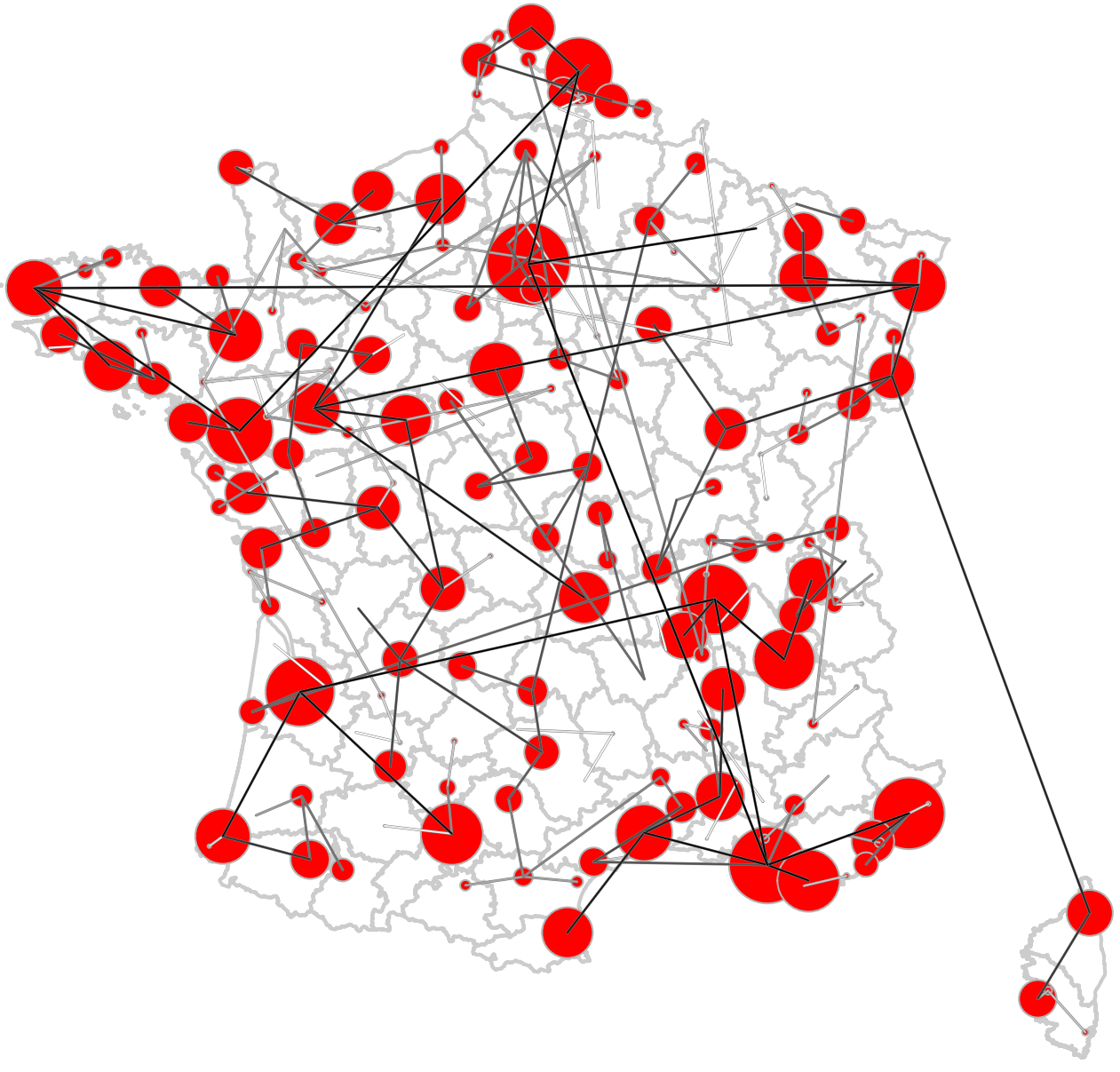




Valeurs propres du laplacien régularisé



Cartes des coupures du graphe



Laplacien régularisé

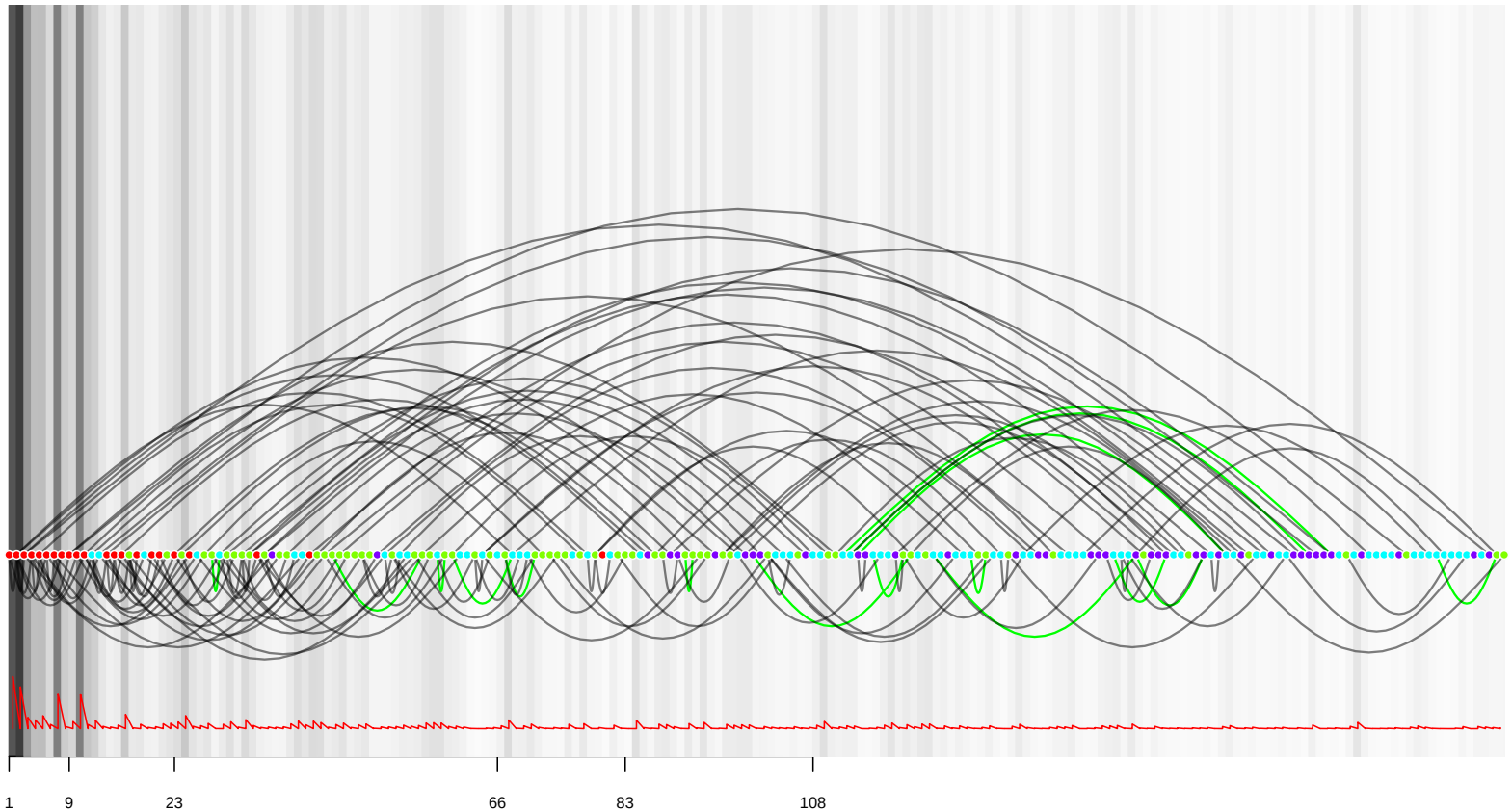
- par rapport au laplacien normalisé, le laplacien régularisé permet de mieux faire ressortir **les dimensions régionales**
- mais aussi **départementales des flux**
- toutefois, les axes apparaissent **très concentrés** autour de quelques communes
- un grand nombre de communes ne sont sans doute **pas bien caractérisées**

flou des données

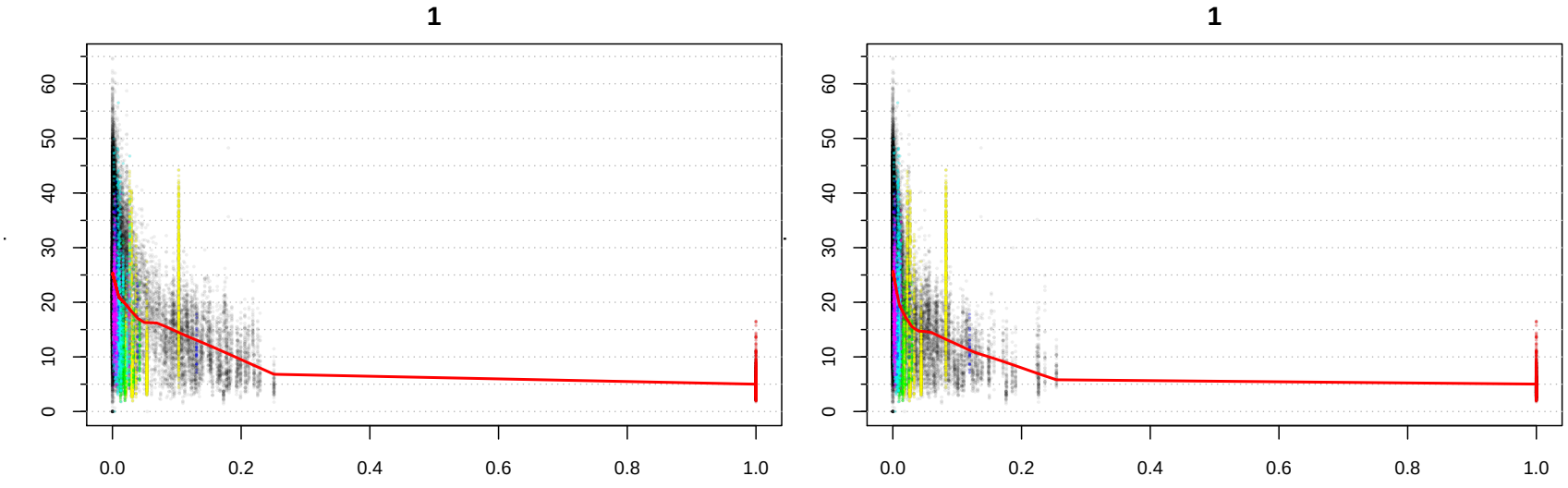
La position des communes dans les flux

- tentative de définition **d'un indicateur de position des communes dans les flux**
- en tenant compte de **la multipolarité des flux**
- approche exploratoire qui repose sur l'observation que
  - les valeurs propres permettent de rapprocher des communes dans **des positions similaires**
  - et peuvent donc servir **à discrétiser le spectre**
  - puis calcul de la position des communes dans chaque sous-espace ainsi défini
  - en standardisant la longueur des vecteurs propres
  - et en prenant la norme infinie
- **Application :** croisement avec les résultats au bureau de vote **au premier tour des élections présidentielles de 2017**

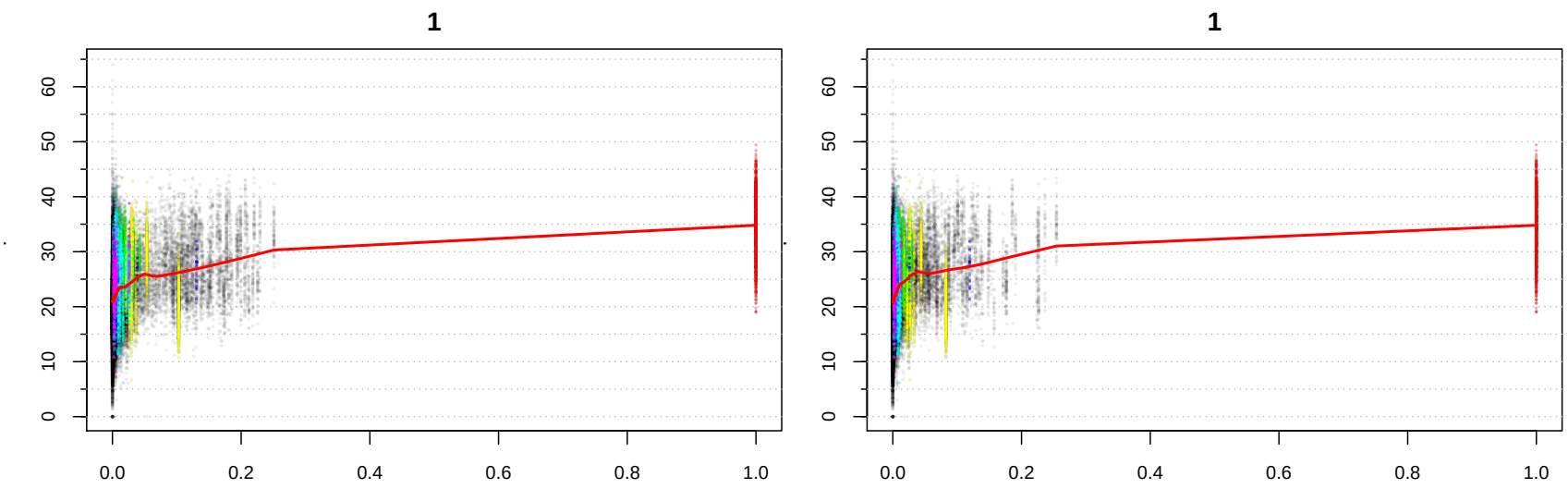
Discrtisation du spectre du graphe



Pourcentages exprimés au bureau de vote I

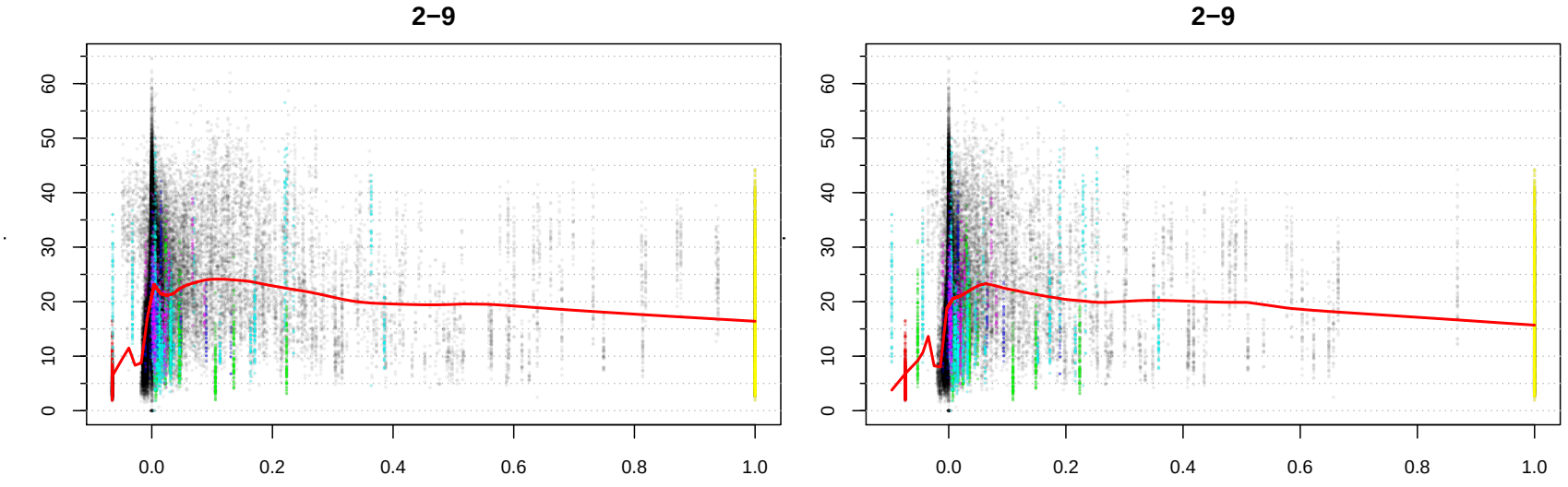


(a) Marine Le Pen

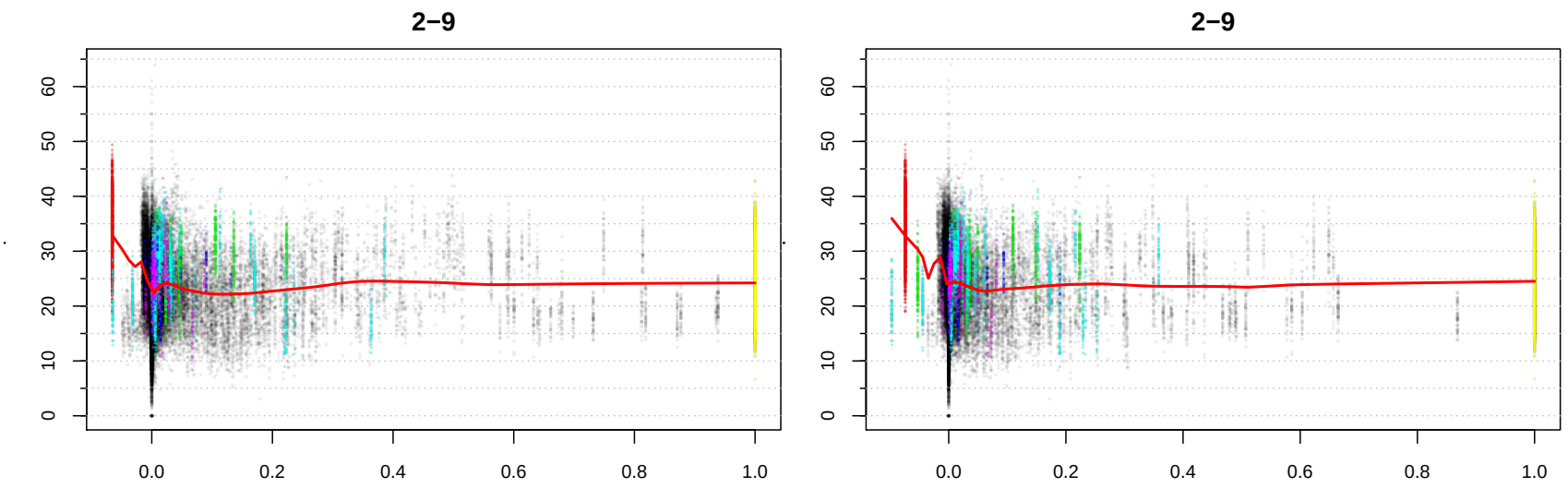


(b) Emmanuel Macron

Pourcentages exprimés au bureau de vote II



(a) Marine Le Pen



(b) Emmanuel Macron

Applications

lapetition.be

La base lapetition.be

- site lancé **fin 2006**
- extraction de la base SQL servant de *backend* au site en **février 2015**
- la base renseigne
  - plus de **15 000 pétitions**

    en fait, plutôt 12 4000, le nombre variant en fonction de la table considérée. . .
  - près de **3,8 millions** de signatures
- voir la présentation aux colloque Internet et les nouvelles formes de participation politique

Diagonalisation des signatures

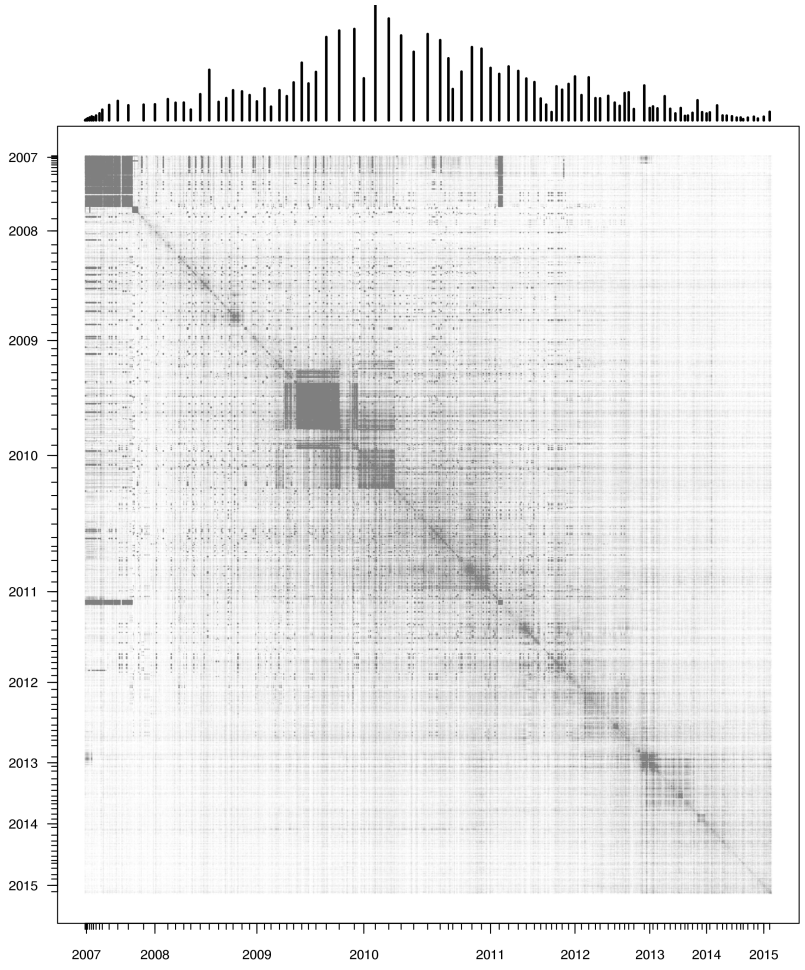


FIGURE 5 – Relations binaires des pétitions dans le temps ($n_{ij} > 0$)

Une site « non-communautaire »

- la plus part ($\sim$ 70%) des signataires ne signent **qu'une pétitions**
- lorsqu'ils en signent une autre, c'est généralement**dans l'heure qui suit**

  les co-signatures proviennent majoritairement de la navigation sur le site

- **difficile de trouver** des « communautés » dans ces conditions

  même si on peut trouver des traces soit de réseaux inter-personnels ou de mobilisations

- approche alternative : chercher **des relations thématiques** entre les signatures
- et analyser les co-signatures entre les thèmes des pétitions
- pour plus de détail, voir Contamin, Léonard et Soubiran[2017] ainsi que la présentation au colloque « Internet et les nouvelles formes de participations » **Soubiran[2019a]**

Analyse thématique des textes des pétitions

- un sous-ensemble de 8 159 pétitions

  « Droits de l'Homme », « Environnement, nature et écologie », « Politique », « Social »

- ont été regroupées en **108 classes**

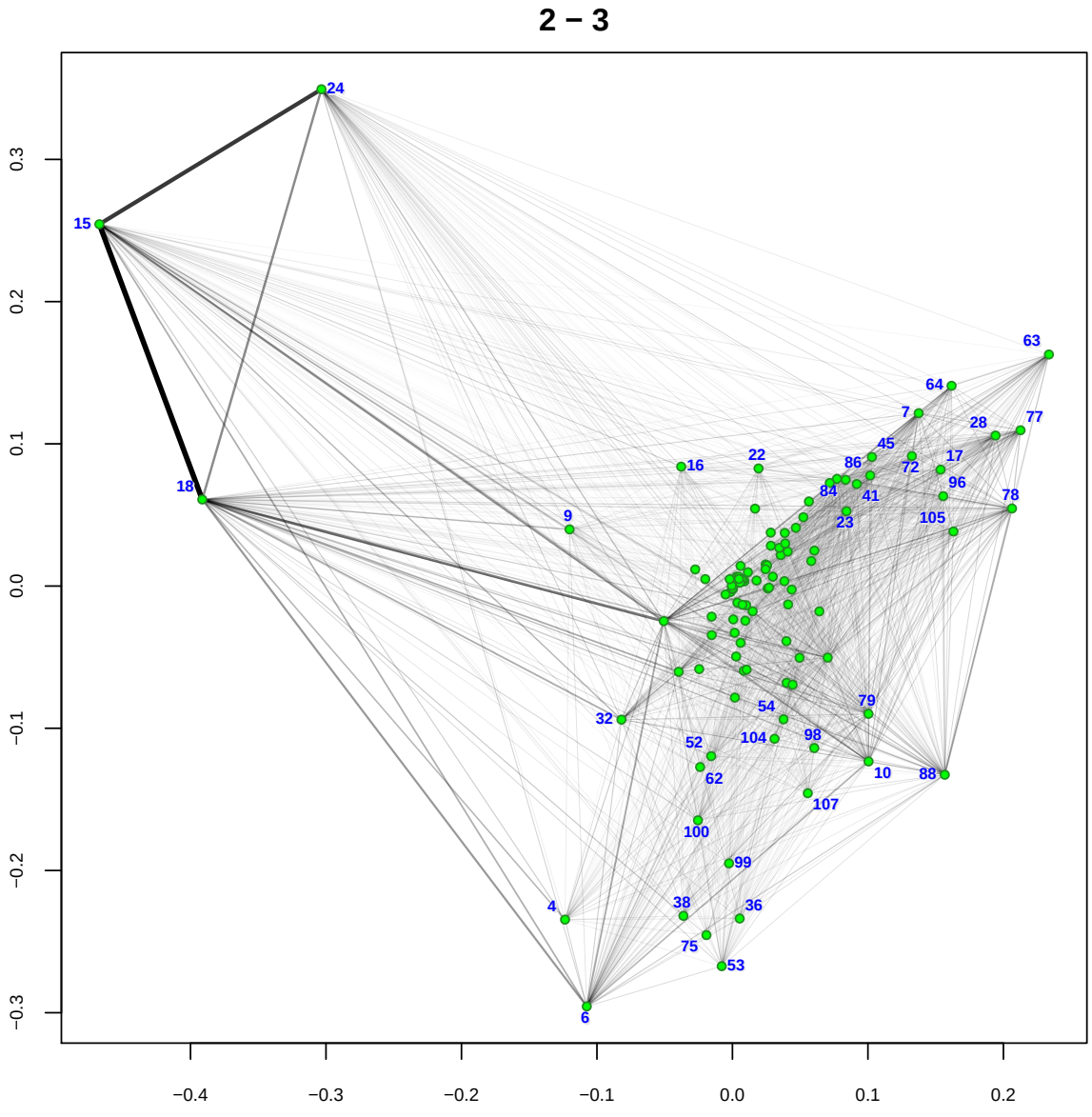
  dont un « blob » de $\sim$ 2 000 pétitions

- présentant **une homogénéité variable**
- qui renvoie elle-même à **l'hétérogénéité des textes**

  longueur, correction grammaticale, syntaxique,. . .

- plus plus de détails, voir Soubiran[2019a] ainsi que **Soubiran[2019c]** (pour une présentation des modèles thématiques)

Laplacien normalisé



- la première dimension oppose **le blob** aux autres pétition
- **la deuxième dimension** oppose d'une part :
  - les pétitions sur la fiscalité

    avec une classe spécifique sur la TVA
  - associées à la classe de pétitions relatives à la Belgique
  - et, d'autre part, un ensemble de pétitions relatives aux
    - conditions de travail, licenciement
    - la Palestine
    - la protection des espaces naturels
    - les expulsions et le droit au logement
    - la liberté d'expression dans les médias
    - pollution de sites (exploitations minière, déchets,. . .
    - les arrestations arbitraires et les violences policières

      cette classes regroupe plus généralement les pétitions sur les questions de sécurité et de police

Conclusion

Conclusion

- l'analyse spectrale permet d'analyser **un large éventail de graphes**

  pondérés, dirigés, bipartis

- **de grandes dimensions**

  car elle repose sur la décomposition en valeur singulière pour laquelle il existe des routines efficaces et numériquement stables

- avec **une théorie bien établie**

  et toujours en développement

- qui permet de faire ressortir différentes **propriétés des graphes** à partir de leur spectre

  connectivité, bipartition, marches aléatoires. . .

- parmi ses variantes, **le laplacien régularisé** permet de s'accommoder d'une situation fréquente où les partitions n'apparaissent pas de façon clairement définies

Merci pour votre attention

Bibliographie

Bibliographie I

Contamin, Jean-Gabriel, Thomas Léonard et Thomas Soubiran (2017), « Les transformations des comportements politiques au prisme de l'e-pétitionnement. Potentialités et limites d'un dispositif d'étude pluridisciplinaire », *Réseaux*, 204, 4, p. 97-131.

Luxburg, Ulrike von (2007), « A tutorial on spectral clustering », *Statistics and Computing*, 17, p. 395-416.

Oliveau, Sébastien et Yoann Doignon (2016), « La diagonale se vide ? Analyse spatiale exploratoire des décroissances démographiques en France métropolitaine depuis 50 ans », *Cybergeo : Revue européenne de géographie*, URL : https://journals.openedition.org/cybergeo/27439?lang=fr.

Qin, Tai et Karl Rohe (sept. 2013), « Regularized Spectral Clustering under the Degree-Corrected Stochastic Blockmodel », *arXiv e-prints*, arXiv :1309.4111, URL : https://ui.adsabs.harvard.edu/%5C#abs/2013arXiv1309.4111Q.

Soubiran, Thomas (2019a), « Analyse thématique du graphe des signatures à un site de pétitions en ligne », colloque du projet ANR Appel *« Internet et les nouvelles formes de participations »*, Lille, 29 mar. 2019, https://pro.univ-lille.fr/fileadmin/user_upload/pages_pros/thomas_soubiran/analyses/appel-2019.pdf.

– (2019b), « La mise en conformité des traitements de données personnelles en SHS : bases épistémologiques pour la négociation », colloque inaugural *PUD-GA*, Grenoble, 13 sept. 2019, https://pro.univ-lille.fr/fileadmin/user_upload/pages_pros/thomas_soubiran/dcp/pudga2019-dcp.pdf.

Bibliographie II

Soubiran, Thomas (2019c), « Quarante ans de délibérations de la Cnil. Analyse thématique de l'évolution d'un corps de doctrine », colloque du PIREH *Histoire, langues et textométrie*, Paris, 1er jan. 2019, https://pro.univ-lille.fr/fileadmin/user_upload/pages_pros/thomas_soubiran/dcp/pireh2019--doctrine-presentation.pdf.

Zhang, Yilin et Karl Rohe (2018), "Understanding Regularized Spectral Clustering via Graph Conductance", sous la dir. de S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi et R. Garnett, Curran Associates, Inc., p. 10631-10640, url : http://papers.nips.cc/paper/8262-understanding-regularized-spectral-clustering-via-graph-conductance.pdf.