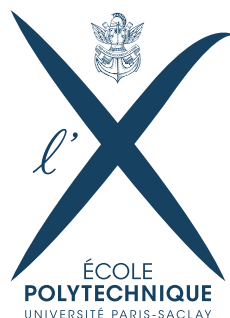


# **SIMULATION NUMÉRIQUE ALÉATOIRE - IDENTIFICATION DE COMMUNAUTÉS**

**Rapport de projet**

June 5, 2017

—  
Arnaud AUTF  
Ignacio MADRID CANALES



# CONTENTS

---

# 1

## INTRODUCTION

---

On considère le problème d'identification de communautés au sein d'une famille d'individus dont on connaît *a priori* leurs classes. On vise à pouvoir évaluer la fiabilité d'un algorithme de partitionnement - le clustering spectral - pour reconstruire ces relations d'appartenance. Pour ce faire on fait deux pas successives:

1. On commence avec  $n$  individus qui appartient à  $k$  classes en total. On construit un graphe de manière aléatoire suivant le Stochastic Block Model, i.e., deux individus de la même classe sont liés avec une probabilité  $p_{in}$ , et deux de classes différents avec probabilité  $p_{out}$ . On se pose au cas  $p_{in} = c_{in}/n$  et  $p_{out} = c_{out}/n$  pour deux paramètres  $c_{in}$  et  $c_{out}$  qu'on va adapter selon nos besoins. Les graphes construits sont non orientés (relations d'amitié).
2. Ensuite, on fait un clustering spectral sous le graphe généré. Le clustering spectral est un technique de partitionnement qui passe par un réduction de la dimension du problème. Cette méthode utilise les vecteurs propres associés au  $k$  plus grands vecteur propres de la matrice d'adjacence  $A$  du graphe (option 1) ou d'une matrice appelée laplacienne qui résout de faire  $M = \sqrt{D}^{-1}.A.\sqrt{D}^{-1}$ , où  $D$  est la matrice diagonale contenant les degrés de chaque sommet (option 2).

Dans [Ver16] il est montré que lorsque la condition

$$c_{in} - c_{out} \gg \sqrt{\log n(c_{in} + c_{out})} \quad (1)$$

est bien vérifié l'algorithme décrit au-dessus fonctionne bien.

Dans ce rapport on étudiera sa fiabilité, plus précisément, les probabilités associées à l'échec du partitionnement déduit par le clustering spectral, dans le cas particulière de  $k = 2$  classes.

Pour ce faire, on analyse d'abord le comportement du Stochastic Block Model et du clustering spectral, et particulièrement leurs sensibilités aux changements de paramètres ( $c_{in}$  et  $c_{out}$ ) aussi qu'à la distribution initiale de classes (combien des individus appartient à chaque cluster). Dans ce même esprit, on compare la performance des méthodes 1 et 2 du clustering spectral, sous différent jeux de paramètres et distributions initiales (Section 2).


Dans un deuxième temps (Section 3), on mesure la fiabilité du clustering. Pour cela on utilise la technique de l'*importance sampling* pour estimer deux probabilités : la probabilité que deux individus fixés soient mal classifiés par le clustering spectral lorsque la condition (??) est bien vérifié, ainsi que la probabilité qu'un grand nombre d'individus soient mal clusterisés lorsque (??) n'est presque plus satisfaite.

## 2

## SIMULATION DU MODÈLE

## 2.1 VISUALISATION DU STOCHASTIC BLOCK MODEL

Pour la visualisation des graphes et leurs clusters on a utilisé la librairie **igraph**. Dans une première instance, on a analysé la méthode du clustering spectral avant l'analyse probabiliste. Pour ce faire, on a d'abord visualisé les graphes résultants d'un Stochastic Block Model (SBM) dont les paramètres de probabilités ( $c_{in}, c_{out}$ ) variait (cf. figure ??). On voit *grosso modo* qu'au mesure que  $c_{in}$  est plus élevé et  $c_{out}$  est plus petit, la fiabilité du SBM est supérieure, sous une même méthode de clustering (ici, clustering spectral).



images/comparaison\_cinsetcouts.png

Figure 1: Graphes d'une population de 100 individus partitionnés initialement en deux classes de 75 et 25. Les trois graphes correspondent à simulations issues du SBM avec trois sets de probabilité : (a)  $c_{in} = 15$ ,  $c_{out} = 2$ , (b)  $c_{in} = 5$ ,  $c_{out} = 1$ , (c)  $c_{in} = 3$ ,  $c_{out} = 1$ .

De même, on peut aussi visualiser la qualité des clusterings. Dans la suite on a coloré les individus selon leurs classes initiales. Cela permet de constater les différences entre les clusters trouvés par les différentes méthodes et les classes originales. Un exemple est présenté dans la figure ??, qui permet de visualiser la qualité du partitionnement spectral fait à partir de la matrice d'adjacence (option 1) contre celui fait à partir de la matrice laplacienne du graphe (option 2). Une comparaison plus détaillée entre ces deux méthodes est traitée dans la section suivante.

images/comparaison\_option1et2.png

Figure 2: Graphes d'une population de 100 individus partitionnés en deux classes de 75 et 25. Les sommets sont distribués selon les classes définies par le clustering spectral fait a posteriori un SBM avec  $c_{in} = 20, c_{out} = 7.5$ , alors qu'ils sont colorés selon leurs classes initiales. À gauche: partitionnement obtenu par l'option 1. À droite : partitionnement obtenu par l'option 2.

## 2.2 COMPARAISON DES MÉTHODES

A fin de comparer la réussite ou l'échec des partitionnements on utilise l'indice de Rand. L'indice de Rand prend valeurs entre 0 et 1, 1 étant le cas de correspondance parfaite entre les classes originales et les nouvelles classes définies par le clustering spectral.

Le tableau ?? compare la réussite de différents clusterings spectrales sous différentes combinaisons de paramètres  $c_{in}$  et  $c_{out}$ , pour une même famille de 200 sommets, 75 % appartenant à la classe 0 et le reste à la classe 1. Pour chaque combinaison on a fait 100 stimulations. On peut vérifier que lors que la condition ?? est satisfaite, le clustering est effectivement mieux réussi. De même, de manière systématique on observe que l'option 2 arrive à partitionner le graphe avec indices de réussite supérieures à ceux de l'option 1. Par exemple, avec  $c_{in} = 40$  et  $c_{out} = 20$ , la condition est faiblement satisfaite et l'option 1 n'arrive qu'à un indice de Rand moyen de 0.549, tandis que l'option 2 arrive à une moyenne de 0.834.

$c_{in}$	$c_{out}$	$c_{in} - c_{out}$	$\sqrt{\log n(c_{in} + c_{out})}$	Rand Option 1	Rand Option 2
100.0	2.0	98.0	24.1202	1.0	1.0
50.0	10.0	40.0	18.4994	0.991	0.999
40.0	15.0	25.0	17.7118	0.718	0.959
35.0	15.0	20.0	16.8875	0.538	0.89
40.0	20.0	20.0	18.4994	0.549	0.834
35.0	20.0	15.0	17.7118	0.507	0.601
10.0	5.0	5.0	9.2497	0.504	0.546
10.0	1.0	9.0	7.921	0.507	0.797
20.0	100.0	-80.0	26.1621	0.499	0.5

Table 1: Comparaison du partitionnement spectral par l'option 1 et par l'option 2 dans différents scénarios de  $c_{in}$  et  $c_{out}$  pour 200 individus (partitionnement original en ratio 3:1). La troisième et la quatrième colonne servent à vérifier le respect de la condition ?. La cinquième et la sixième colonne montrent l'indice de Rand moyen des clusterings.


Ensuite, on a voulu tester la performance des deux options dans différents partitionnements pour les classes originales, *ceteris paribus* (200 individus,  $c_{in} = 40$ ,  $c_{out} = 15$ ). Les résultats sont montrés dans le tableau ??.

Proportion Classe 0	Proportion Classe 1	Rand Option 1	Rand Option 2
0.05	0.95	0.5	0.535
0.1	0.9	0.502	0.685
0.2	0.8	0.511	0.942
0.25	0.75	0.744	0.954
0.3	0.7	0.908	0.96
0.4	0.6	0.98	0.976
0.5	0.5	0.984	0.977
0.55	0.45	0.982	0.984

Table 2: Comparaison des indices de Rand du partitionnement spectral par l'option 1 et par l'option 2 pour différents partitionnement des deux classes originales. Pour chaque essai on a fait 100 simulations avec 200 individus, et  $c_{in}$  et  $c_{out}$  constantes, avec des valeurs de 40 et 15 respectivement.

On peut observer qu'à mesure que le partitionnement devienne plus régulier (ratio 1:1) le clustering est plus performant, pour l'option 2 aussi bien que pour l'option 1. Néanmoins, quand le partitionnement est déséquilibré l'option 2 s'avère être assez plus effective que l'option 1. Par exemple avec une distribution 80% - 20%, le clustering fait par la méthode 2 arrive à un indice de Rand moyen de 0.9, tandis que la méthode 1 reste aux alentours de 0.5. La figure ?? montre l'un des clusterings simulés sous ces paramètres.





images/comparaison\_1et2\_8020.png

Figure 3: Graphes d'une population de 200 individus partitionnés en deux classes du 80 % et 20% du total. Les sommets sont distribués selon les classes définies par le clustering spectral fait à  $c_{in} = 40$ ,  $c_{out} = 15$ , alors qu'ils sont colorés selon leurs classes initiales. À gauche: partitionnement obtenu par l'option 2. À droite : partitionnement obtenu par l'option 1.

## 3

## ETUDE DE LA FIABILITÉ DU CLUSTERING

Dans cette partie, nous allons mettre en oeuvre la technique de l'importance sampling dans le but d'étudier les limites du clustering spectral dans le cas particulier graphes générés par le Stochastic Block Model.

Dans un premier temps, nous allons estimer la probabilité que, lorsque le critère de clustering réussi est bien vérifié, un ensemble réduit et fixé de sommets soit mal clusterisé.

Dans un second temps, nous allons estimer la probabilité que, lorsque le critère de clustering réussi n'est presque plus satisfait, un grand nombre de sommets soient mal clusterisés.

### 3.1 CHOIX DES PARAMÈTRES

Avant de débiter notre estimation, il s'agit de déterminer quels paramètres il est pertinent de prendre pour générer nos graphes : le nombre d'individus  $n$ , le nombre de classes  $k$ , la répartition de ces classes entre les individus, la probabilité d'amitié intra-classe  $p_{in}$  et celle d'amitié inter-classe  $p_{out}$ . Pour simplifier l'étude, le sujet nous invite d'abord à fixer  $k = 2$ .

Pour  $n$ , nous avons décidé de prendre des valeurs assez élevées, tout en permettant à nos ordinateurs d'effectuer les simulations en un temps raisonnable. Ainsi, sachant que nous effectuons la plupart du temps 1000 simulations pour estimer une probabilité, nous avons pris des valeurs de  $n$  entre 200 et 300. En effet, en lisant quelques articles traitant du clustering spectral appliqué au Stochastic Block Model, nous avons eu le sentiment que la validité de celui-ci varie moins brutalement avec  $c_{in}$  et  $c_{out}$  lorsque  $n$  est grand, ce qui nous permettra par la suite d'être plus "grossiers" dans notre recherche d'un bon changement de probabilité sur  $p_{in}$  et  $p_{out}$ .

Enfin, il semble intuitivement plus facile d'observer un mauvais clustering d'individus d'une classe si celle-ci compte beaucoup moins d'individus que l'autre. Par conséquent, nous avons décidé de placer un tiers des individus dans la première classe et deux tiers dans la seconde, puis de chercher à observer les potentiels échecs du clustering parmi les individus de la première.

### 3.2 CRITÈRE SATISFAIT - PEU DE SOMMETS MAL CLUSTERISÉS

Dans cette partie nous allons donc présenter nos résultats pour le premier problème. Tout d'abord, étant donné que nous allons nous intéresser à la probabilité que deux individus fixés de la première classe de notre graphe soient mal clusterisés, le changement de probabilité ne sera effectué que sur ces deux individus. De fait, après avoir mené quelques tests infructueux en changeant la probabilité de tous les individus, nous avons remarqué qu'un changement de probabilité portant sur un nombre trop important de variables de Bernoulli conduit à la multiplication de nos résultats par un facteur correctif trop important qui dégrade fortement la qualité de notre estimation, alors qu'avec le choix des paramètres que nous avons fait nous pouvons espérer que de légers changements de probabilité sur ces deux individus nous permettront d'arriver à une bonne estimation de la probabilité recherchée.

L'enjeu est maintenant de choisir intelligemment comment modifier  $p_{in} = \frac{c_{in}}{n}$  et  $p_{out} = \frac{c_{out}}{n}$ . La lecture d'articles tels que [NN12] nous a alors encouragés à prendre une valeur fixe pour la quantité  $c_{in}+c_{out}$ , ce qui facilite la recherche du changement de probabilités idéal. Nous avons donc pris pour le modèle  $c_{in} = 40$  et  $c_{out} = 10$ .

Pour savoir dans quelle plage effectuer notre première vague de tests de changements de probabilités, nous avons ensuite fait un petit calcul "intuitif". En effet, nous avons essayé de comprendre comment fonctionnait le clustering spectral, et avons entre autres compris que l'existence d'un écart important entre les premières valeurs propres de la matrice clusterisée était un élément déterminant pour que celui-ci soit correct. En particulier, l'article [RCY11], en définissant des sommets comme "mal clusterisés" lorsqu'ils se trouvent à une distance trop importante de leur centroïde, parvient à montrer que ce nombre de sommets mal clusterisés est dominé par une quantité proportionnelle à  $\frac{1}{\lambda^4}$  où  $\lambda$  est la plus petite valeur propre d'une matrice proche de celle que nous utilisons effectivement dans notre cas pour mettre en oeuvre le clustering spectral, qui (dans notre cas où les individus sont répartis équitablement entre les  $k$  classes et où la probabilité inter-classe et la probabilité intra-classe peuvent s'écrire respectivement  $p_{in} = p+r$  et  $p_{out} = r$ ) s'exprime alors  $\lambda = \frac{1}{k(\frac{r}{p})+1}$ .

Alors, dans notre cas où  $k=2$ ,  $p_{in} = \frac{c_{in}}{n}$ ,  $p_{out} = \frac{c_{out}}{n}$ , le calcul permet d'obtenir :

$$\lambda = \frac{c_{in}-c_{out}}{c_{in}+c_{out}}$$

Par conséquent, pour obtenir un nombre de sommets mal clusterisés environ deux fois supérieur à la normale, on peut penser intuitivement à prendre des paramètres modifiés  $nc_{in}$  et  $nc_{out}$  tels que :

$$\left(\frac{nc_{in}+nc_{out}}{nc_{in}-nc_{out}}\right)^4 = 2 \times \left(\frac{c_{in}+c_{out}}{c_{in}-c_{out}}\right)^4$$

Ce qui donne ici (avec  $c_{in} = 40$  et  $c_{out} = 10$ ,  $c_{in}+c_{out}$  fixe) :

$$nc_{in} + nc_{out} = c_{in} - c_{out} - 4.77$$

Donc, des paramètres "intuitivement intéressants" pour le changement de probabilité seraient  $nc_{in} = 37.75$  et  $c_{out} = 32.75$ .

Cependant, étant donné que nous ne changeons les probabilités que pour deux individus, nous pouvions nous attendre à ce que le changement de probabilité optimal modifie plus fortement  $c_{in}$  et  $c_{out}$ .

Ainsi, après ces nombreuses observations, nous avons donc lancé une première vague de simulations avec les paramètres suivants :

- $n = 300$
- $k = 2$
- $n1 = 102$  et  $n2 = 198$
- $c_{in} = 40$
- $c_{out} = 10$
- 1000 simulations par estimation

Le changement de probabilités est effectué sur deux individus fixés de la première classe (qui compte  $n_1=102$  individus), et la probabilité estimée est celle que ces deux individus soient mal clusterisés. Le but de ces premières simulations est de déterminer quel changement de probabilité conserver afin de produire la meilleure estimation de la probabilité recherchée. Leurs résultats sont présentés dans le tableau suivant, obtenus en faisant varier  $nc_{in}$  et  $nc_{out}$  à  $c_{in}+c_{out}$  constant.

Valeurs $c_{in}/c_{out}$	Probabilité estimée	Demi-largeur de l'intervalle de confiance à 95%
39.5/10.5	$13.9 \times 10^{-3}$	$6.7 \times 10^{-3}$
39/11	$11.3 \times 10^{-3}$	$5.8 \times 10^{-3}$
38.5/11.5	$17.5 \times 10^{-3}$	$6.7 \times 10^{-3}$
38/12	$13.8 \times 10^{-3}$	$6.3 \times 10^{-3}$
37.5/12.5	$17.4 \times 10^{-3}$	$7.2 \times 10^{-3}$
37/13	$13.3 \times 10^{-3}$	$5.1 \times 10^{-3}$
36.5/13.5	$8.3 \times 10^{-3}$	$3.4 \times 10^{-3}$
36/14	$14.9 \times 10^{-3}$	$5.9 \times 10^{-3}$
35.5/14.5	$14.6 \times 10^{-3}$	$7.0 \times 10^{-3}$
35/15	$11.9 \times 10^{-3}$	$5.0 \times 10^{-3}$
34.5/15.5	$23.5 \times 10^{-3}$	$12.6 \times 10^{-3}$
34/16	$10.2 \times 10^{-3}$	$5.9 \times 10^{-3}$
33.5/16.5	$5.4 \times 10^{-3}$	$2.4 \times 10^{-3}$
33/17	$7.8 \times 10^{-3}$	$3.8 \times 10^{-3}$
32.5/17.5	$9.9 \times 10^{-3}$	$7.9 \times 10^{-3}$
32/18	$3.7 \times 10^{-3}$	$1.5 \times 10^{-3}$
31.5/18.5	$16.2 \times 10^{-3}$	$17.7 \times 10^{-3}$
31/19	$22.3 \times 10^{-3}$	$23.5 \times 10^{-3}$
30.5/19.5	$17.8 \times 10^{-3}$	$17.6 \times 10^{-3}$
30/20	$8.0 \times 10^{-3}$	$5.8 \times 10^{-3}$

Ces premières simulations nous ont alors permis de restreindre la recherche à des changements de probabilités à  $c_{in}+c_{out}$  constant avec  $nc_{in}$  variant entre 37.5 et 36. Une deuxième vague de simulations a donc été menée, dont les résultats sont présentés ci-après.

Valeurs $c_{in}/c_{out}$	Probabilité estimée	Demi-largeur de l'intervalle de confiance à 95%
37.5/12.5	$8.9 \times 10^{-3}$	$4.1 \times 10^{-3}$
37.33/12.67	$14.8 \times 10^{-3}$	$7.1 \times 10^{-3}$
37.17/12.83	$12.8 \times 10^{-3}$	$5.0 \times 10^{-3}$
37/13	$16.8 \times 10^{-3}$	$13.1 \times 10^{-3}$
36.83/13.17	$9.7 \times 10^{-3}$	$3.6 \times 10^{-3}$
36.67/13.3	$11.2 \times 10^{-3}$	$4.1 \times 10^{-3}$
36.5/13.5	$11.2 \times 10^{-3}$	$4.68 \times 10^{-3}$
36.33/13.67	$12.3 \times 10^{-3}$	$5.7 \times 10^{-3}$
36.11/13.83	$9.99 \times 10^{-3}$	$4.78 \times 10^{-3}$
36/14	$13.6 \times 10^{-3}$	$7.5 \times 10^{-3}$

Cette deuxième vague de simulations nous amène enfin à effectuer une dernière recherche de  $nc_{in}$  entre 36.6 et 36.8 pour  $c_{in} + c_{out}$  constant. Dont voici les résultats :

Valeurs $c_{in}/c_{out}$	Probabilité estimée	Demi-largeur de l'intervalle de confiance à 95%
36.6/13.4	$12.3 \times 10^{-3}$	$4.5 \times 10^{-3}$
36.65/13.35	$17.8 \times 10^{-3}$	$8.0 \times 10^{-3}$
36.7/13.30	$10.0 \times 10^{-3}$	$4.2 \times 10^{-3}$
36.75/13.25	$14.1 \times 10^{-3}$	$5.8 \times 10^{-3}$
36.8/13.2	$21.5 \times 10^{-3}$	$10.1 \times 10^{-3}$

On décide finalement de choisir :  $nc_{in} = 36.6$  et  $nc_{out} = 13.4$ . On obtient donc les résultats, par rapport à la méthode de Monte-Carlo classique :

Méthode employée	Probabilité estimée (moyenne et variance sur 10 simulations)	Demi-largeur de l'intervalle de confiance à 95% (moyenne et variance sur 10 simulations)
Monte-Carlo classique	$15.0 \times 10^{-3}$ , var $3.5 \times 10^{-3}$	$7.5 \times 10^{-3}$ , var $1 \times 10^{-3}$
Importance sampling	$13.9 \times 10^{-3}$ , var $2.86 \times 10^{-3}$	$6.1 \times 10^{-3}$ , var $2.2 \times 10^{-3}$

La probabilité retenue, dans la configuration que nous avons choisie, pour que les deux premiers individus du groupe 1 soient mal clusterisés, est donc finalement de  $13.9 \times 10^{-3}$ , estimée par notre méthode à avec un intervalle de confiance à 95% de  $6.1 \times 10^{-3}$ . Les résultats obtenus par importance sampling sont donc meilleurs que par la méthode de Monte-Carlo classique. Néanmoins, le gain de précision sur la probabilité n'est pas significatif. Ceci s'explique sans doute par le fait qu'avec le jeu de paramètres que nous avons choisis, l'événement considéré n'est pas si rare et la méthode de Monte-Carlo classique est donc assez efficace dès 1000 simulations.

### 3.3 CRITÈRE NON SATISFAIT - BEAUCOUP DE SOMMETS MAL CLUSTERISÉS

Dans cette partie nous allons présenter nos résultats pour le second problème. L'objectif est de déterminer la probabilité que, lorsque le critère de clustering réussi est mal satisfait, la probabilité qu'un nombre important de sommets soit mal clusterisés.

Pour cette question, nous avons d'abord choisi une manière de juger qu'un "nombre élevé de sommets" sont mal clusterisés. Nous avons alors retenu l'indice de Rand, quantité inférieure 1 qui permet d'évaluer la proximité entre deux clusterings (1 correspondant à deux clusterings identiques). Le calcul de l'indice de Rand étant plus gourmand en calcul, nous limiterons pour la suite le nombre d'individus à 200.

Alors, pour avoir un critère de clustering réussi mal satisfait, nous avons pris  $c_{in} = 36$  et  $c_{out} = 14$ , ce qui donne pour le critère :

$$c_{in} - c_{out} - \sqrt{\log(n)(c_{in} + c_{out})} = 1.72$$

Il s'agit maintenant de choisir une valeur "seuil" de l'indice de Rand telle que, sous les paramètres choisis, l'événement "Indice de Rand < seuil" puisse raisonnablement être assimilé à l'événement "Un grand nombre de sommets mal clusterisés". Avant de calculer ce seuil, il est nécessaire de choisir la méthode de clustering spectral retenue : après avoir retenu la méthode 1 dans la partie précédente, nous retenons la méthode 2 dans celle-ci. La méthode 2, comme nous l'avons mis en évidence précédemment, se distingue par une meilleure performance de l'indice de Rand lorsque les classes sont très inégalement réparties.

En ayant choisi les paramètres suivants, avec un clustering réalisé selon la méthode 2 :

- $n = 200$
- $k = 2$
- $n_1 = 50$  et  $n_2 = 150$
- $c_{in} = 36$
- $c_{out} = 14$
- 1000 simulations par tentative

Quelques simulations permettent rapidement d'observer que l'événement "Indice de Rand < 0.54" survient moins d'une dizaine de fois sur 1000 essais. C'est donc cette valeur que nous retenons.

En nous appuyons sur les raisonnements exposés plus tôt, nous avons donc débuté nos simulations pour explorer les valeurs de  $c_{in}$  et  $c_{out}$  qui constitueront un bon changement de probabilités. Une difficulté supplémentaire apparaît dans ce problème : il s'agit de déterminer combien d'individus doivent voir leurs probabilités être modifiées par notre méthode d'importance sampling. En sachant que la modification des probabilités d'un nombre important d'individus limite exponentiellement l'amplitude des changements de probabilité que nous pouvons mettre en oeuvre. Nous avons donc décidé, après plusieurs tentatives, de nous limiter à 5 individus voyant leurs probabilités modifiées sous le changement.

Après de nombreux échecs, nous avons tentés de chercher des paramètres corrects pour nos probabilités modifiées en nous restreignant à des valeurs comprises entre 34 et 33 pour  $c_{in}$ . Les résultats, obtenus avec les paramètres précisés plus haut, sont présentés dans la table suivante :

Valeurs $c_{in}/c_{out}$	Probabilité estimée	Demi-largeur de l'intervalle de confiance à 95%
33.9/16.1	$6.8 \times 10^{-3}$	$5.0 \times 10^{-3}$
33.8/16.2	$9.1 \times 10^{-3}$	$5.7 \times 10^{-3}$
33.7/16.3	$9.1 \times 10^{-3}$	$5.7 \times 10^{-3}$
33.6/16.4	$9.2 \times 10^{-3}$	$7.5 \times 10^{-3}$
33.5/16.5	$4.3 \times 10^{-3}$	$3.5 \times 10^{-3}$
33.4/16.6	$9.0 \times 10^{-3}$	$6.3 \times 10^{-3}$
33.3/16.7	$4.8 \times 10^{-3}$	$4.0 \times 10^{-3}$
33.2/16.8	$3.1 \times 10^{-3}$	$3.8 \times 10^{-3}$
33.1/16.9	$9.0 \times 10^{-3}$	$6.4 \times 10^{-3}$
33.0/17.0	$10.0 \times 10^{-3}$	$6.7 \times 10^{-3}$

Après cette vague de simulations nous avons décidé une dernière recherche de  $nc_{in}$  entre 33.8 et 33.7 pour  $c_{in}+c_{out}$  constant. Dont voici les résultats :

Valeurs $c_{in}/c_{out}$	Probabilité estimée	Demi-largeur de l'intervalle de confiance à 95%
33.8/16.2	$9.6 \times 10^{-3}$	$6.0 \times 10^{-3}$
33.78/16.22	$3.0 \times 10^{-3}$	$3.4 \times 10^{-3}$
33.76/16.24	$8.0 \times 10^{-3}$	$5.2 \times 10^{-3}$
33.74/16.26	$6.5 \times 10^{-3}$	$4.8 \times 10^{-3}$
33.72/16.28	$6.7 \times 10^{-3}$	$5.0 \times 10^{-3}$

On décide finalement de choisir :  $nc_{in} = 33.76$  et  $nc_{out} = 16.24$ . On obtient donc les résultats, par rapport à la méthode de Monte-Carlo classique :

Méthode employée	Probabilité estimée (moyenne et variance sur 10 simulations)	Demi-largeur de l'intervalle de confiance à 95% (moyenne et variance sur 10 simulations)
Monte-Carlo classique	$6.4 \times 10^{-3}$ , var $3.4 \times 10^{-3}$	$4.8 \times 10^{-3}$ , var $1.3 \times 10^{-3}$
Importance sampling	$8.0 \times 10^{-3}$ , var $2.6 \times 10^{-3}$	$5.3 \times 10^{-3}$ , var $1 \times 10^{-3}$

La probabilité retenue, dans la configuration que nous avons choisie, pour que "beaucoup d'individus" soient mal clusterisés, est donc finalement de  $8.0 \times 10^{-3}$ , estimée par notre méthode à avec un intervalle de confiance à 95% de  $2.6 \times 10^{-3}$ . Les résultats obtenus par importance sampling sont donc presque identiques à ceux obtenus par la Méthode de Monte Carlo classique. La raison principale de cette similarité est que nous ne sommes pas parvenus à "pousser" le changement de probabilité, c'est à dire à changer de manière plus marquée  $c_{in}$  et  $c_{out}$ , ce qu'il aurait sans doute été possible de faire en ayant une meilleur connaissance du problème et en choisissant le cadre de nos simulations de manière plus précise.

Finalement, notre travail nous a tout de même permis d'obtenir une estimation correcte des probabilités recherchée, en particulier dans le cas où nous nous intéressions au problème d'un critère de clustering bien satisfait avec l'événement "deux individus fixés sont mal clusterisés". Si nous aurions apprécié de pouvoir présenter des résultats finaux plus éclatants, les nombreuses difficultés de simulation et de calibration de nos paramètres que nous avons rencontrées tout au long de ce projet nous ont beaucoup appris. Nous avons particulièrement pris conscience des liens subtils qui lient théorie et simulations pratiques : ici les possibilités d'un changement de probabilité étaient prometteuses *a priori*, difficiles à obtenir en réalité, et la clé de simulations plus réussies résidait sans doute dans une connaissance plus approfondie des tenants théoriques de la validité d'un clustering spectral.