DÉTECTION DE COMMUNAUTÉS ET APPRENTISSAGE EN LIGNE

15 février 2022

Antoine Bak, Romain Farthoat

sous la supervision de Matthieu Lerasle

REMERCIEMENTS

Merci aux organisateurs du MODAL MAP472A pour avoir rendu possible ce projet, ainsi qu'à Matthieu Lerasle pour nous avoir guidés tout au long de ce travail.

TABLE DES MATIÈRES

1
PROBLÈME DU BANDIT MANCHOT

1.1 Problème posé

Soient $V = C_0 \sqcup C_1$ une partition de notre ensemble de sommets en deux communautés et un noeud $x \in V$ qu'on cherche à catégoriser. Ces deux communautés sont ici les « bras » du problème, selon une terminologie courante.

On notera $i^* \in \{0, 1\}$ l'indice de la communauté de notre nœud.

L'appartenance communautaire de tous les autres sommets est supposée connue dans la première partie, et seulement celle-ci.

On n'impose ici aucune contrainte sur la taille du graphe, qui peut être infini.

Notre but dans cette première partie sera de maximiser l'espérance du nombre d'arêtes découvertes en un nombre $t \in \mathbb{N}^*$ fixé de tours.

Étant donné un sommet autre que x quelconque, suppose que la probabilité qu'une arête relie x à ce sommet vaut p s'il appartient à la même communauté que x, q sinon, avec $p > q$. On veut donc d'après ce qui précède déterminer i^*, mais tout en restant efficace : on ne veut pas prendre le risque de moins gagner au final en jouant trop sur la mauvaise communauté.

On appellera μ_0 la probabilité que le sommet x soit relié à un autre sommet dans C_1. On définit similairement μ_1.

Conformément à ce qui précède :

$$\begin{cases} \mu_{i^*} = p \\ \mu_{1-i^*} = q \end{cases}$$

Au tour $s \in \mathbb{N}^*$ et compte tenu des informations acquises, on choisit une communauté $\pi(s)$, puis on procède à un tirage dans cette dernière, *i.e.* on y pioche un sommet — pas encore observé — pour voir s'il est relié à x ou non.

Soit, pour tout $s \in \mathbb{N}^*$, $X_{i^*,s} \sim \mathcal{B}(p)$ et $X_{1-i^*,s} \sim \mathcal{B}(q)$ avec $0 < q < p$. Ces variables représentent les tirages successifs effectués respectivement dans la communauté i^* et dans l'autre. On suppose toutes ces variables aléatoires **indépendantes.**

Posons, pour $s \in [\![1,t]\!]$: $Y_s = X_{\pi(s),N_{\pi(s)}(s)}$ et $N_i(s)$ le nombre de fois où le bras $i \in \{0,1\}$ a été choisi au cours des s premiers instants.

On cherche une stratégie $\pi : [\![1,t]\!] \longrightarrow \{0,1\}$ optimale, *i.e.* minimisant la fonction de regret suivante :

$$R_t = tp - \mathbb{E}\left[\sum_{s=1}^{t} Y_s\right]$$

$\pi(s)$ peut bien sûr dépendre des résultats passés et même être randomisé, *i.e.* défini comme suit :

$$\pi(s) = \psi_s\Big(U_1, Y_1, \ldots, U_{s-1}, Y_{s-1}, U_s\Big)$$

où les U_i sont i.i.d. de loi $\mathcal{U}([0,1])$ et indépendantes des $X_{i,s}$.

Remarque. Il n'y a pas de raison pour que $\pi(s)$ ne dépende pas de t, nous avons seulement allégé les notations.

On notera enfin à travers l'ensemble du présent rapport :

$$\Delta = p - q \qquad\qquad \tau = \frac{p+q}{2}$$

Revenons au risque. La formule de Wald permet de montrer qu'on a :

$$R_t = \Delta\mathbb{E}\Big[N_{1-i^*}(t)\Big]$$

et que de plus, en définissant :

$$\forall s \in [\![1,t]\!], \forall i \in \{0,1\}, \qquad \widehat{\mu}_i(s) = \frac{1}{N_i(s)} \sum_{k=1}^{N_i(s)} X_{i,k}$$

on trouve des estimateurs sans biais des μ_i :

$$\forall s \in [\![1,t]\!], \forall i \in \{0,1\}, \qquad \mathbb{E}[\widehat{\mu}_i(s)] = \mu_i$$

1.2 Stratégie *Upper Confidence Bound* (UCB)

Soit δ une fonction décroissante dont on déterminera les valeurs plus tard. On pose :

$$\Omega = \bigcap_{\substack{1\leqslant s\leqslant t\\ i\in\{0,1\}}} \Big(\mu_i \geqslant \widehat{\mu}_i(s) - \delta(s) \cap \mu_{i^*} \leqslant \widehat{\mu}_{i^*}(s) + \delta(s)\Big)$$

Remarque. Il n'y a pas non plus de raison pour que δ ne dépende pas de t.

L'évènement Ω correspond au cas où « tout va bien », *i.e.* où, peu importe le rang s auquel on s'intéresse, aucun des deux estimateurs n'est « trop grand », et où $\widehat{\mu}_{i^*}(s)$ n'est pas « trop petit ». $\delta(s)$ représente ici l'écart à l'espérance que l'on s'autorise avant de taxer un estimateur de « trop grand » ou « trop petit ».

On peut décider de choisir i^* constamment à compter du rang $s_0 \in \mathbb{N}^*$ si pour $s_0 \leqslant s \leqslant t$:

$$\widehat{\mu}_{1-i^*}(s) + \delta(s) < \widehat{\mu}_{i^*}(s) - \delta(s) \tag{1}$$

Or quitte à supposer Ω satisfait, on a pour $1 \leqslant s \leqslant t$:

$$\begin{cases} \mu_{i^*} - \delta(s) \leqslant \widehat{\mu}_{i^*}(s) \\ \widehat{\mu}_{1-i^*}(s) \leqslant \mu_{1-i^*} + \delta(s) \end{cases}$$

d'où la condition suffisante (par décroissance de δ) pour (1) :

$$\mu_{1-i^*} + 2\delta(s_0) \leqslant \mu_{i^*}$$

Cette condition est satisfaite si $s_0 \geqslant \delta^{-1}\left(\frac{\Delta}{2}\right)$, δ^{-1} étant la fonction inverse généralisée de δ.

Cela entraîne $\mathbb{E}\Big[N_{1-i^*}(t)\Big|\Omega\Big] \leqslant \delta^{-1}\left(\frac{\Delta}{2}\right)$ et :

$$R_t \leqslant \Delta\delta^{-1}\left(\frac{\Delta}{2}\right) + t\Delta\mathbb{P}\left(\Omega^C\right)$$

Restent à déterminer $\delta^{-1}\left(\frac{\Delta}{2}\right)$ et $\mathbb{P}\left(\Omega^C\right)$.

1.2.1 • Choix de δ

Comme :

$$\Omega \supset \bigcap_{1\leqslant s\leqslant t} \Big(|\widehat{\mu}_0(s) - \mu_0| > \delta(s) \cap |\widehat{\mu}_1(s) - \mu_1| > \delta(s)\Big)$$

on va utiliser l'inégalité de Hoeffding.

Pour $s \in [\![1, t]\!]$ et $N_0(s)$ **donnés :**

$$\begin{aligned}
&\mathbb{P}\Big(|\widehat{\mu}_0(s) - \mu_0| > \delta(s) \cap |\widehat{\mu}_1(s) - \mu_1| > \delta(s)\Big) \\
&\qquad \leqslant \mathbb{P}\Big(N_0(s)|\widehat{\mu}_0(s) - \mu_0| > N_0(s)\delta(s) \cap N_1(s)|\widehat{\mu}_1(s) - \mu_1| > N_1(s)\delta(s)\Big) \\
&\qquad \leqslant \mathbb{P}\left(\left|\sum_{k=1}^{N_0(s)} X_{0,k} - N_0(s)\mu_0\right| > N_0(s)\delta(s) \cap \left|\sum_{k=1}^{N_1(s)} X_{1,k} - N_1(s)\mu_1\right| > N_1(s)\delta(s)\right) \\
&\qquad \leqslant \mathbb{P}\left(\left|\sum_{k=1}^{N_0(s)} X_{0,k} + \sum_{k=1}^{N_1(s)} X_{1,k} - \Big(N_0(s)\mu_0 + N_1(s)\mu_1\Big)\right| > \Big(N_0(s) + N_1(s)\Big)\delta(s)\right) \\
&\qquad \leqslant 2\exp\Big(-2\Big(N_0(s) + N_1(s)\Big)\delta(s)^2\Big) \\
&\qquad \leqslant 2\exp\Big(-2s\delta(s)^2\Big)
\end{aligned}$$

On en déduit, indépendamment de la valeur de $N_0(s)$:

$$\mathbb{P}\Big(|\widehat{\mu}_0(s) - \mu_0| > \delta(s) \cap |\widehat{\mu}_1(s) - \mu_1| > \delta(s)\Big) \leqslant 2\exp\Big(-2s\delta(s)^2\Big)$$

Soit $\eta > 0$. Avec $\delta(s) = \sqrt{\dfrac{\log(2t/\eta)}{2s}}$ on obtient $\mathbb{P}\Big(|\widehat{\mu}_i(s) - \mu_i| > \delta(s)\Big) \leqslant \dfrac{\eta}{t}$ et :

$$\begin{aligned}
\mathbb{P}\left(\Omega^C\right) &\leqslant \mathbb{P}\left(\bigcup_{1\leqslant s\leqslant t}\Big(|\widehat{\mu}_0(s) - \mu_0| > \delta(s) \cap |\widehat{\mu}_1(s) - \mu_1| > \delta(s)\Big)\right) \\
&\leqslant \sum_{1\leqslant s\leqslant t} \mathbb{P}\Big(|\widehat{\mu}_0(s) - \mu_0| > \delta(s) \cap |\widehat{\mu}_1(s) - \mu_1| > \delta(s)\Big) \\
&\leqslant \eta
\end{aligned}$$

Enfin, $\eta = \dfrac{1}{t}$ donne la majoration :

$$R_t \leqslant \Delta + \frac{2\log(2t^2)}{\Delta}$$

1.3 Cas avec τ connu

Une idée naturelle consiste à observer la moyenne empirique des tirages pour chaque communauté et à la comparer avec la valeur de τ tant qu'aucune communauté n'a été épuisée :

- si les deux sont inférieures à τ, ou qu'aucun sommet n'a encore été tiré, on effectue un tirage supplémentaire dans chaque communauté ;
- sinon on tire un sommet dans la communauté ayant la moyenne empirique la plus grande.

On suppose $\Delta \leqslant 2q$ pour pouvoir plus tard appliquer l'inégalité de Chernov.

1.3.1 • Inégalité de Chernov

Théorème 1. *Soit $(A_i)_{i\geqslant 1}$ une suite de variables aléatoires i.i.d. de loi $\mathcal{B}(p)$ et $t \in \mathbb{N}^*$.*
Soit $S_t = \sum_{i=1}^{t} A_i$, de sorte que $\mathbb{E}[S_t] = tp$. On note également $\overline{A}_t = \frac{1}{t}\sum_{i=1}^{t} A_i$.
Si $\delta \in [0,1]$, alors :

$$\begin{cases} \mathbb{P}\Big(S_t \geqslant tp(1+\delta)\Big) \leqslant \exp\left(-\dfrac{tp\delta^2}{3}\right) \\ \mathbb{P}\Big(S_t \leqslant tp(1-\delta)\Big) \leqslant \exp\left(-\dfrac{tp\delta^2}{3}\right) \end{cases}$$

Si $\delta \geqslant 1$, on a quand même :

$$\mathbb{P}\Big(S_l \geqslant tp(1+\delta)\Big) \leqslant \exp\left(-\frac{tp\delta}{3}\right)$$

Remarque. Cette inégalité a pour intérêt d'obtenir de meilleures majorations que l'inégalité de Hoeffding dans les cas où p est faible :

Si $\varepsilon \leqslant p$:

$$\begin{aligned} \mathbb{P}\left(\left|\overline{A}_t - p\right| \geqslant \varepsilon\right) &\leqslant \mathbb{P}\left(S_t \geqslant pl\left(1+\frac{\varepsilon}{p}\right)\right) + \mathbb{P}\left(S_t \leqslant pl\left(1-\frac{\varepsilon}{p}\right)\right) \\ &\leqslant 2\exp\left(-\frac{l\varepsilon^2}{3p}\right) \end{aligned}$$

Démonstration. On applique l'inégalité de Markov à $e^{\lambda S_t}$ pour $\lambda > 0$:

$$\begin{aligned} \mathbb{P}\Big(S_t \geqslant tp(1+\delta)\Big) &= \mathbb{P}\left(e^{\lambda S_t} \geqslant e^{\lambda tp(1+\delta)}\right) \\ &\leqslant \mathbb{E}\left[e^{\lambda S_t}\right] e^{-\lambda tp(1+\delta)} \\ &= \exp\Big(t\log(1+p(e^{\lambda}-1)) - \lambda tp(1+\delta)\Big) \end{aligned}$$

D'une part $\lambda = \log(1+\delta)$ donne :

$$\begin{aligned}\mathbb{P}\Big(S_t \geqslant tp(1+\delta)\Big) &\leqslant \exp\Big(t\log(1+p\delta) - tp(1+\delta)\log(1+\delta)\Big)\\ &\leqslant \exp\Big(-tp\Big((1+\delta)\log(1+\delta) - \delta\Big)\Big)\end{aligned}$$

de même :

$$\begin{aligned}\mathbb{P}\Big(S_t \leqslant tp(1-\delta)\Big) &\leqslant \mathbb{E}\left[e^{-\lambda S_t}\right] e^{\lambda tp(1-\delta)}\\ &= \exp\Big(t\log(1+p(e^{-t}-1)) + \lambda tp(1-\delta)\Big)\end{aligned}$$

D'autre part $\lambda = -\log(1-\delta)$ donne :

$$\mathbb{P}\Big(S_t \leqslant tp(1-\delta)\Big) \leqslant \exp\Big(-tp\Big((1-\delta)\log(1-\delta) + \delta\Big)\Big)$$

Avec une rapide étude de fonction, on trouve que pour $\delta \in [-1, 1]$:

$$(1+\delta)\log(1+\delta) - \delta \geqslant \frac{\delta^2}{3}$$

ce qui achève la démonstration. □

1.3.2 • Calcul du regret

Intéressons-nous à présent à la probabilité de se tromper au s^{e} tour avec cette stratégie :

$$\mathbb{P}\Big(\pi(s) = 1 - i^*\Big) = \mathbb{P}\Big(\pi(s) = 1 - i^* \cap \widehat{\mu}_{1-i^*}(s) > \tau\Big) + \mathbb{P}\Big(\pi(s) = 1 - i^* \cap \widehat{\mu}_{1-i^*}(s) \leqslant \tau\Big)$$

Or :

$$\begin{aligned}
&\sum_{s=3}^{t} \mathbb{P}\Big(\pi(s) = 1 - i^* \cap \widehat{\mu}_{1-i^*}(s) > \tau\Big) \\
&\qquad = \sum_{s=3}^{t} \mathbb{P}\left(N_{1-i^*}(s) > N_{1-i^*}(s-1) \cap \widehat{\mu}_{1-i^*}(s) - \mu_{1-i^*} > \frac{\Delta}{2}\right) \\
&\qquad = \mathbb{E}\left[\sum_{s=3}^{t} \mathbf{1}_{N_{1-i^*}(s) > N_{1-i^*}(s-1) \cap \widehat{\mu}_{1-i^*}(s) - \mu_{1-i^*} > \frac{\Delta}{2}}\right] \\
&\qquad = \mathbb{E}\left[\sum_{\substack{3 \leqslant s \leqslant t \\ N_{1-i^*}(s) > N_{1-i^*}(s-1)}} \mathbf{1}_{\widehat{\mu}_{1-i^*}(s) - \mu_{1-i^*} > \frac{\Delta}{2}}\right] \\
&\qquad = \mathbb{E}\left[\sum_{l=1}^{N_{1-i^*}(t)} \mathbf{1}_{\frac{1}{l}\sum_{k=1}^{l} X_{1-i^*,k} - \mu_{1-i^*} > \frac{\Delta}{2}}\right] \\
&\qquad \leqslant \mathbb{E}\left[\sum_{l=1}^{t} \mathbf{1}_{\sum_{k=1}^{l} X_{1-i^*,k} - l\mu_{1-i^*} > l\frac{\Delta}{2}}\right] \\
&\qquad = \sum_{l=1}^{t} \mathbb{P}\left(\sum_{k=1}^{l} X_{1-i^*,k} > lq\left(1 + \frac{\Delta}{2q}\right)\right) \\
&\qquad \leqslant \sum_{l=1}^{t} \exp\left(-\frac{l\Delta^2}{12q}\right) \qquad \text{par Chernov} \\
&\qquad \leqslant \frac{1}{1 - \exp\left(\frac{\Delta^2}{12q}\right)} \\
&\qquad \leqslant \frac{12q}{\Delta^2}
\end{aligned}$$

De même, si :

$$\begin{cases} \pi(s) = 1 - i^* \\ \widehat{\mu}_{1-i^*}(s) \leqslant \tau \end{cases}$$

on a nécessairement $\widehat{\mu}_{i^*}(s) \leqslant \widehat{\mu}_{1-i^*}(s) \leqslant \tau$ car c'est $1 - i^*$ qui est choisi au s^{e} tour.

Il vient alors similairement que :

$$
\begin{aligned}
&\sum_{s=3}^{t} \mathbb{P}\Big(\pi(s) = 1 - i^* \cap \widehat{\mu}_{1-i^*}(s) \leqslant \tau\Big) \\
&\qquad \leqslant \sum_{s=3}^{t} \mathbb{P}\Big(\pi(s) = 1 - i^* \cap \widehat{\mu}_{i^*}(s) \leqslant \tau\Big) \\
&\qquad \leqslant \sum_{s=3}^{t} \mathbb{P}\left(N_{1-i^*}(s) > N_{1-i^*}(s-1) \cap \widehat{\mu}_{i^*}(s) - \mu_{i^*} \leqslant -\frac{\Delta}{2}\right) \\
&\qquad = \mathbb{E}\left[\sum_{s=3}^{t} \mathbf{1}_{N_{1-i^*}(s) > N_{1-i^*}(s-1) \cap \widehat{\mu}_{i^*}(s) - \mu_{i^*} < -\frac{\Delta}{2}}\right] \\
&\qquad = \mathbb{E}\left[\sum_{\substack{3 \leqslant s \leqslant t \\ N_{1-i^*}(s) > N_{1-i^*}(s-1)}} \mathbf{1}_{\widehat{\mu}_{i^*}(s) - \mu_{i^*} < -\frac{\Delta}{2}}\right] \\
&\qquad = \mathbb{E}\left[\sum_{l=1}^{N_{1-i^*}(t)} \mathbf{1}_{\frac{1}{l}\sum_{k=1}^{l} X_{i^*,k} - \mu_{i^*} < -\frac{\Delta}{2}}\right] \\
&\qquad \leqslant \mathbb{E}\left[\sum_{l=1}^{t} \mathbf{1}_{\sum_{k=1}^{l} X_{i^*,k} - l\mu_{i^*} < -l\frac{\Delta}{2}}\right] \\
&\qquad = \sum_{l=1}^{t} \mathbb{P}\left(\sum_{k=1}^{l} X_{i^*,k} < lp\left(1 - \frac{\Delta}{2p}\right)\right) \\
&\qquad \leqslant \sum_{l=1}^{t} \exp\left(-\frac{l\Delta^2}{12p}\right) \qquad \text{par Chernov} \\
&\qquad \leqslant \frac{1}{1 - \exp\left(\frac{\Delta^2}{12p}\right)} \\
&\qquad \leqslant \frac{12p}{\Delta^2}
\end{aligned}
$$

Par ailleurs comme $N_{1-i^*}(t) = \sum_{s=1}^{t} \mathbf{1}_{\pi(s)=1-i^*}$, on trouve :

$$\begin{aligned}
\mathbb{E}\Big[N_{1-i^*}(t)\Big] &= \mathbb{E}\left[\sum_{s=1}^{t} \mathbf{1}_{\pi(s)=1-i^*}\right] \\
&= \sum_{s=1}^{t} \mathbb{E}\left[\mathbf{1}_{\pi(s)=1-i^*}\right] \\
&= \sum_{s=1}^{t} \mathbb{P}\Big(\pi(s) = 1 - i^*\Big) \\
&\leqslant \frac{12(p+q)}{\Delta^2} \\
&= 24\frac{\tau}{\Delta^2}
\end{aligned}$$

et finalement :

$$R_t \leqslant \Delta + \frac{24\tau}{\Delta}$$

Le premier terme de la somme vient des deux tirages d'initialisation de l'algorithme.

Remarque. En utilisant l'inégalité de Hoeffding plutôt que celle de Chernov on peut obtenir :

$$R_t \leqslant \Delta + \frac{4}{\Delta}$$

Cette dernière inégalité a le mérite d'être valable même si $\Delta > 2q$.

2
APPRENTISSAGE EN LIGNE DANS UN GRAPHE

2.1 Problème posé

On reprend un formalisme similaire : soit $V = C_0 \sqcup C_1$ une partition des sommets d'un graphe en deux communautés telles que si i et j sont des sommets quelconque de la même communauté ils ont une probabilité p d'être reliés, contre q s'ils sont de communautés différentes.

On définit naturellement les variables aléatoires indépendantes :

$$X_{ij} \sim \begin{cases} \mathcal{B}(p) & \text{si } i \text{ et } j \text{ appartiennent à la même communauté} \\ \mathcal{B}(q) & \text{sinon} \end{cases}$$

Le problème posé est proche de celui du bandit manchot, où les « machines » sont les arêtes du graphe, et où il n'est pas possible de tirer deux fois le même levier.

Pour un nombre donné de tirages $T \in \mathbb{N}^*$, on veut jouer de manière à maximiser le nombre d'arêtes découvertes. Cette fois-ci en revanche, **on a *a priori* aucune connaissance sur l'appartenance communautaire de nos sommets.**

Le seul paramètre d'entrée de l'algorithme est l'entier $T \in \mathbb{N}^*$ qui désigne le nombre de tirages à effectuer au total.

2.2 Première solution

1. On tire un ensemble $\mathcal{N}$ de n sommets de manière équilibrée, numérotés de 1 à n, puis :
   - on tire les arêtes correspondant à chaque paire du noyau ;
   - on évalue $\widehat{\tau} = \binom{n}{2}^{-1} \sum_{1 \leqslant i < j \leqslant n} X_{ij}$;
   - on utilise un algorithme de détection de communautés dans le sous-graphe $\mathcal{N}$.

2. On se propose de classer un ensemble $\mathcal{S}$ de sommets de $G \backslash \mathcal{N}$.
Tant que nous n'avons pas effectué T tirages, on pioche $x \in V \backslash \mathcal{N}$, et on réutilise la stratégie avec τ connu en l'approximant par $\widehat{\tau}$, le tout sur les communautés $\widehat{C}_0$ et $\widehat{C}_1$:
   - On tire un sommet dans chaque $\widehat{C}$.
   - Tant que ni $\widehat{C}_0$ ni $\widehat{C}_1$ n'a été intégralement visitée :
     - Si les $\widehat{\mu}$ sont tous deux inférieurs à $\widehat{\tau}$, on tire un sommet supplémentaire dans chaque $\widehat{C}$.
     - Sinon, on tire un sommet de la communauté associée au $\widehat{\mu}$ le plus élevé.

2.2.1 • Schéma de preuve

Les notations utilisées ici seront introduites rigoureusement lors des développements suivants. On ne vise ici qu'à donner une idée globale de la direction des démonstrations qui suivront.

On montre tout d'abord que pourvu que notre noyau ait une composition équilibrée, *i.e.* qu'aucune communauté réelle n'y soit surreprésentée :

$$\left| n_0 - \frac{n}{2} \right| \leqslant n\varepsilon_n$$

on peut utiliser la méthode de classification du cas τ connu en l'estimant de manière satisfaisante, soit encore :

$$|\widehat{\tau} - \tau| \leqslant \Delta\varepsilon_\tau$$

On montre ensuite que pour chaque sommet donné, notre procédé de classification aboutit à un nombre borné d'erreurs, pourvu que la catégorisation du noyau soit suffisamment bonne, c'est-à-dire :

$$\left|\widehat{C}_0 \cap C_1\right| + \left|\widehat{C}_1 \cap C_0\right| \leqslant n\varepsilon_C$$

Enfin, le calcul du regret permet d'évaluer cet algorithme. On montre que la grande probabilité de la conjonction de ces trois conditions (notée Ω) rend possible un regret asymptotique en $O(T^{2/3})$ au lieu de $O(T)$ en tirant simplement des arêtes au hasard.

2.2.2 • Hypothèses

On suppose que $\Delta \geqslant \dfrac{C}{n\varepsilon_C^2}$

Pour pouvoir utiliser l'inégalité de Chernov, on suppose que $\sqrt{35}q \geqslant \Delta$.

2.2.3 • Évaluation de $\widehat{\tau}$

Posons $n_0 = |\mathcal{N} \cap C_0|$ et $n_1 = |\mathcal{N} \cap C_1| = n - n_0$.

Remarque. Noter que n_0 et n_1 dépendent du noyau, que l'on sélectionne aléatoirement. Ce sont donc **des variables aléatoires.**

On a quitte à renuméroter les sommets :

$$\binom{n}{2}\widehat{\tau} = \sum_{1\leqslant i<j\leqslant n_0} X_{ij} + \sum_{n_0+1\leqslant i<j\leqslant n} X_{ij} + \sum_{1\leqslant i\leqslant n_0, n_0+1\leqslant j\leqslant n} X_{ij}$$

où les X_{ij} des deux premières sommes suivent une loi $\mathcal{B}(p)$, alors que celles de la dernière somme suivent une loi $\mathcal{B}(q)$.

On peut donc écrire :

$$\binom{n}{2}\widehat{\tau} \sim B_1 + B_2$$

où

$$B_1 \sim \mathcal{B}\left(\binom{n_0}{2} + \binom{n_1}{2}, p\right)$$

et

$$B_2 \sim \mathcal{B}\big(n_0 n_1, q\big)$$

sont deux variables aléatoires indépendantes pour toute valeur de n_0 **donnée.**

On en déduit l'expression de l'espérance de $\widehat{\tau}$:

$$\binom{n}{2}\mathbb{E}[\widehat{\tau}|n_0] = p\left(\binom{n_0}{2} + \binom{n_1}{2}\right) + qn_0n_1$$

Soit $n\varepsilon_n = \sqrt{\dfrac{\Delta}{32\tau}\binom{n}{2}}$

Si $\left|n_0 - \frac{n}{2}\right| \leqslant n\varepsilon_n$, on a :

$$\begin{cases} \left|n_0 n_1 - \frac{1}{2}\binom{n}{2}\right| \leqslant \frac{\Delta}{32\tau}\binom{n}{2} \\ \left|\binom{n_0}{2} + \binom{n_1}{2} - \frac{1}{2}\binom{n}{2}\right| \leqslant \frac{\Delta}{32\tau}\binom{n}{2} \end{cases}$$

En utilisant l'inégalité de Serfling [1] :

$$\mathbb{P}\left(\left|n_0 - \frac{n}{2}\right| > n\varepsilon_n\right) \leqslant 2\exp\left(-2n\varepsilon_n^2\right)$$

On aimerait à présent avoir une estimation suffisamment précise de τ pour qu'utiliser les résultats de la première partie soit pertinent.

Plus concrètement, on désire avoir $|\widehat{\tau} - \tau| \leqslant \Delta\varepsilon_\tau$.

On peut appliquer l'inégalité de Hoeffding pour trouver :

$$\mathbb{P}\left(|\widehat{\tau} - \tau| > \Delta\varepsilon_\tau \,\middle|\, \left|n_0 - \frac{n}{2}\right| \leqslant n\varepsilon_n\right) \leqslant 2\exp\left(-2\binom{n}{2}\Delta^2(\varepsilon_\tau - \varepsilon_n)^2\right)$$

2.2.4 • Évaluation de l'erreur commise en fonction du nombre de sommets mal catégorisés

On peut montrer que s'il y a m sommets mal catégorisés dans $\widehat{C}_i$ alors :

$$\mathbb{E}\left[\widehat{\mu}_i(s)\middle|\widehat{C}_i\right] = \frac{\mu_i\left(\left|\widehat{C}_i\right| - m\right) + \mu_{1-i}m}{\left|\widehat{C}_i\right|}$$

Or avec une probabilité supérieure à $1 - e^{-N}$ on trouve des catégories estimées telles que :

$$\left|\widehat{C}_0 \cap C_1\right| + \left|\widehat{C}_1 \cap C_0\right| \leqslant n\varepsilon_C$$

On définit alors :

$$\Omega = \left(\left|n_0 - \frac{n}{2}\right| \leqslant n\varepsilon_n\right) \cap \left(|\widehat{\tau} - \tau| \leqslant \Delta\varepsilon_\tau\right) \cap \left(\left|\widehat{C}_0 \cap C_1\right| + \left|\widehat{C}_1 \cap C_0\right| \leqslant n\varepsilon_C\right)$$

On peut alors utiliser une démonstration analogue à celle du cas avec τ connu.

On reprend là encore des notations similaires. À supposer que l'on s'occupe d'un sommet $x \in V\backslash\mathcal{N}$ appartenant à C_{i^*}, $\pi(s)$ désigne l'indice de la communauté choisie lors du s^{e} tirage effectué pour classer x. Au cours des s premiers tours de cette classification, la communauté i a été choisie $N_i(s)$ fois.

Par la formule des probabilités totales :

$$\begin{aligned}\mathbb{P}\Big(\pi(s)=1-i^*\Big|\Omega\Big) &= \mathbb{P}\Big(\pi(s)=1-i^*\cap\widehat{\mu}_{1-i^*}(s)\leqslant\hat{\tau}\Big|\Omega\Big)\\ &\quad+\mathbb{P}\Big(\pi(s)=1-i^*\cap\widehat{\mu}_{1-i^*}(s)>\hat{\tau}\Big|\Omega\Big)\\ &= \mathbb{P}\Big(N_{1-i^*}(s)>N_{1-i^*}(s-1)\cap\widehat{\mu}_{1-i^*}(s)>\hat{\tau}\Big|\Omega\Big)\\ &\quad+\mathbb{P}\Big(N_{1-i^*}(s)>N_{1-i^*}(s-1)\cap\widehat{\mu}_{1-i^*}(s)\leqslant\hat{\tau}\Big|\Omega\Big)\end{aligned}$$

Puis comme dans la démonstration du cas τ connu, on note $(Y_{i,k})_{1\leqslant k\leqslant\left|\widehat{C}_i\right|}$ les variables aléatoires représentant les tirages d'arêtes entre x et les sommets de $\widehat{C}_i$ pris dans un ordre aléatoire et $\overline{Y}_{i,l}=\dfrac{1}{l}\displaystyle\sum_{k=1}^{l}Y_{i,k}$:

$$\begin{aligned}&\sum_{s=3}^{t}\mathbb{P}\Big(N_{1-i^*}(s)>N_{1-i^*}(s-1)\cap\widehat{\mu}_{1-i^*}(s)>\hat{\tau}\Big|\Omega\Big)\\ &\qquad\leqslant\sum_{l=1}^{t}\mathbb{P}\left(\overline{Y}_{1-i^*,l}-\mathbb{E}\left[\overline{Y}_{1-i^*,l}\right]>\hat{\tau}-\tau+\tau-\mathbb{E}\left[\overline{Y}_{1-i^*,l}\right]\Big|\Omega\right)\\ &\qquad\leqslant\sum_{l=1}^{t}\mathbb{P}\left(\overline{Y}_{1-i^*,l}-\mathbb{E}\left[\overline{Y}_{1-i^*,l}\right]>\tau-\mathbb{E}\left[\overline{Y}_{1-i^*,l}\right]-|\hat{\tau}-\tau|\Big|\Omega\right)\end{aligned}$$

On se convainc rapidement que le nombre de sommets mal catégorisés sur les l premiers choisis dans $\widehat{C}_{1-i^*}$ suit une loi hypergéométrique :

$$\mathcal{H}\left(l,\frac{\left|\widehat{C}_{1-i^*}\cap C_{i^*}\right|}{\left|\widehat{C}_{1-i^*}\right|},\left|\widehat{C}_{1-i^*}\right|\right)$$

Rappelons que sous Ω, $\left|\widehat{C}_{1-i^*}\cap C_{i^*}\right|\leqslant n\varepsilon_C$ et :

$$\begin{aligned}\left|\widehat{C}_{1-i^*}\right| &= n_{1-i^*}+\left|\widehat{C}_{1-i^*}\cap C_{i^*}\right|-\left|\widehat{C}_{i^*}\cap C_{1-i^*}\right|\\ &\geqslant n_{1-i^*}-\left|\widehat{C}_{i^*}\cap C_{1-i^*}\right|\\ &\geqslant\frac{n}{2}-n\varepsilon_n-n\varepsilon_C\\ &= n\frac{1-2\varepsilon_n-2\varepsilon_C}{2}\end{aligned}$$

La part $\dfrac{\left|\widehat{C}_{1-i^*}\cap C_{i^*}\right|}{\left|\widehat{C}_{1-i^*}\right|}$ de sommets mal classés dans $\widehat{C}_{1-i^*}$ valant au plus

$$\frac{n\varepsilon_C}{n\dfrac{1-2\varepsilon_n-2\varepsilon_C}{2}}=\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}$$

sous Ω, on a :

$$\mathbb{E}\left[\overline{Y}_{1-i^*,l}\middle|\,\Omega\right] \leqslant \mu_{i^*}\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C} + \mu_{1-i^*}\left(1-\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}\right)$$

D'où :

$$\tau - \mathbb{E}\left[\overline{Y}_{1-i^*,l}\middle|\,\Omega\right] \geqslant \Delta\left(\frac{1}{2}-\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}\right)$$

Et, en notant t le nombre de tours nécessaire à la classification de x :

$$\begin{aligned}
&\sum_{s=3}^{t}\mathbb{P}\Big(N_{1-i^*}(s) > N_{1-i^*}(s-1)\cap\widehat{\mu}_{1-i^*}(s)>\hat{\tau}\Big|\Omega\Big)\\
&\qquad\leqslant \sum_{l=1}^{t}\mathbb{P}\left(\overline{Y}_{1-i^*,l}-\mathbb{E}\left[\overline{Y}_{1-i^*,l}\middle|\,\Omega\right] > \tau-\mathbb{E}\left[\overline{Y}_{1-i^*,l}\middle|\,\Omega\right]-|\hat{\tau}-\tau|\middle|\Omega\right)\\
&\qquad\leqslant \sum_{l=1}^{t}\mathbb{P}\left(\overline{Y}_{1-i^*,l}-\mathbb{E}\left[\overline{Y}_{1-i^*,l}\middle|\,\Omega\right] > \Delta\left(\frac{1}{2}-\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}-\varepsilon_\tau\right)\middle|\Omega\right)
\end{aligned}$$

Par l'inégalité de Chernov pour $1\leqslant l\leqslant t$:

$$\begin{aligned}
&\mathbb{P}\left(\overline{Y}_{1-i^*,l}-\mathbb{E}\left[\overline{Y}_{1-i^*,l}\middle|\,\Omega\right] > \Delta\left(\frac{1}{2}-\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}-\varepsilon_\tau\right)\middle|\Omega\right)\\
&\qquad\leqslant \exp\left(-\frac{l\Delta^2\left(\frac{1}{2}-\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}-\varepsilon_\tau\right)^2}{3\mathbb{E}\left[\overline{Y}_{1-i^*,l}\middle|\,\Omega\right]}\right)
\end{aligned}$$

On obtient une majoration de la somme comme précédemment :

$$\begin{aligned}
&\sum_{l=1}^{t}\mathbb{P}\left(\overline{Y}_{1-i^*,l}-\mathbb{E}\left[\overline{Y}_{1-i^*,l}\middle|\,\Omega\right] > \Delta\left(\frac{1}{2}-\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}-\varepsilon_\tau\right)\middle|\Omega\right)\\
&\qquad\leqslant \frac{3p}{\Delta^2\left(\frac{1}{2}-\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}-\varepsilon_\tau\right)^2}
\end{aligned}$$

Similairement :

$$\frac{\left|\widehat{C}_{i^*}\cap C_{1-i^*}\right|}{\left|\widehat{C}_{i^*}\right|} \leqslant \frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}$$

Donc comme $\mu_{i^*} > \mu_{1-i^*}$:

$$\mathbb{E}\left[\overline{Y}_{i^*,l}\middle|\,\Omega\right] \geqslant \mu_{i^*}\left(1-\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}\right) + \mu_{1-i^*}\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}$$

On a :

$$\begin{aligned}
&\sum_{s=3}^{t}\mathbb{P}\Big(N_{1-i^*}(s) > N_{1-i^*}(s-1)\cap\widehat{\mu}_{i^*}(s)\leqslant\widehat{\tau}\Big|\Omega\Big)\\
&\qquad\leqslant\sum_{l=1}^{t}\mathbb{P}\left(\overline{Y}_{i^*,l}-\mathbb{E}\left[\overline{Y}_{i^*,l}\middle|\Omega\right]\leqslant\widehat{\tau}-\tau+\tau-\mathbb{E}\left[\overline{Y}_{i^*,l}\middle|\Omega\right]\middle|\Omega\right)\\
&\qquad\leqslant\sum_{l=1}^{t}\mathbb{P}\left(\overline{Y}_{i^*,l}-\mathbb{E}\left[\overline{Y}_{i^*,l}\middle|\Omega\right]\leqslant\tau-\mathbb{E}\left[\overline{Y}_{i^*,l}\middle|\Omega\right]+|\widehat{\tau}-\tau|\middle|\Omega\right)\\
&\qquad\leqslant\sum_{l=1}^{t}\mathbb{P}\left(\overline{Y}_{1-i^*,l}-\mathbb{E}\left[\overline{Y}_{1-i^*,l}\middle|\Omega\right]\leqslant-\Delta\left(\frac{1}{2}-\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}-\varepsilon_\tau\right)\middle|\Omega\right)\\
&\qquad\leqslant\frac{3p}{\Delta^2\left(\frac{1}{2}-\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}-\varepsilon_\tau\right)^2}
\end{aligned}$$

Et enfin le résultat désiré :

$$\begin{aligned}
&\sum_{s=3}^{t}\mathbb{P}\Big(\pi(s)=1-i^*\Big|\Omega\Big)\\
&\qquad=\sum_{s=3}^{t}\Big(\mathbb{P}\Big(\pi(s)=1-i^*,\widehat{\mu}_{1-i^*}(s)>\hat{\tau}\Big|\Omega\Big)\\
&\qquad\quad+\mathbb{P}\Big(\pi(s)=1-i^*,\widehat{\mu}_{1-i^*}(s)\leqslant\hat{\tau}\Big|\Omega\Big)\Big)\\
&\qquad\leqslant\sum_{s=3}^{t}\Big(\mathbb{P}\Big(N_{1-i^*}(s)>N_{1-i^*}(s-1)\cap\widehat{\mu}_{1-i^*}(s)>\hat{\tau}\Big|\Omega\Big)\\
&\qquad\quad+\mathbb{P}\Big(N_{1-i^*}(s)>N_{1-i^*}(s-1)\cap\widehat{\mu}_{i^*}(s)\leqslant\hat{\tau}\Big|\Omega\Big)\Big)\\
&\qquad\leqslant\sum_{l=1}^{t}\Big(\mathbb{P}\left(\overline{Y}_{1-i^*,l}>\hat{\tau}\middle|\Omega\right)+\mathbb{P}\left(\overline{Y}_{i^*,l}\leqslant\hat{\tau}\middle|\Omega\right)\Big)\\
&\qquad\leqslant\sum_{l=1}^{t}\left(\mathbb{P}\left(\overline{Y}_{1-i^*,l}-\mathbb{E}\left[\overline{Y}_{i^*,l}\middle|\Omega\right]>\Delta\left(\frac{1}{2}-\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}-\varepsilon_\tau\right)\middle|\Omega\right)\right.\\
&\qquad\quad\left.+\mathbb{P}\left(\overline{Y}_{i^*,l}-\mathbb{E}\left[\overline{Y}_{i^*,l}\middle|\Omega\right]\leqslant-\Delta\left(\frac{1}{2}-\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}-\varepsilon_\tau\right)\middle|\Omega\right)\right)\\
&\qquad\leqslant\frac{6p}{\Delta^2\left(\frac{1}{2}-\frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C}-\varepsilon_\tau\right)^2}
\end{aligned}$$

2.2.5 • Calcul du regret

Sur la première étape de l'algorithme avec la construction du graphe, on commet au plus $\binom{n}{2}$ erreurs, d'où un premier terme :

$$R_{T,1} = \Delta\binom{n}{2}$$

Par ailleurs le regret de la 2e étape $R_{T,2}$ vérifie :

$$\mathbb{E}[R_{T,2}] = \mathbb{E}[R_{T,2}|\Omega]\mathbb{P}(\Omega) + \mathbb{E}\left[R_{T,2}\middle|\Omega^C\right]\mathbb{P}(\Omega^C) \leqslant \mathbb{E}[R_{T,2}|\Omega] + \Delta T\mathbb{P}\left(\Omega^C\right)$$

Pour atteindre les T tirages, comme chaque sommet x choisi dans le graphe sera affecté à au moins $\min\left(\left|\widehat{C}_0\right|, \left|\widehat{C}_1\right|\right)$ sommets lors de la seconde étape.

On classe donc, en plus du noyau, un nombre de sommets vérifiant :

$$|\mathcal{S}| \leqslant \frac{T}{\min\left(\left|\widehat{C}_0\right|, \left|\widehat{C}_1\right|\right)}$$

Or sous Ω, on a :

$$\min\left(\left|\widehat{C}_0\right|, \left|\widehat{C}_1\right|\right) \geqslant \frac{n}{2} - n(\varepsilon_n + \varepsilon_C)$$

On peut donc majorer le nombre de sommets à classer en plus du noyau.

De plus, on a, pour chacun de ces sommets, une espérance du nombre d'erreurs majorée par :

$$1 + \frac{6p}{\Delta^2\left(\frac{1}{2} - \frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C} - \varepsilon_\tau\right)^2}$$

Cela donne :

$$\mathbb{E}[R_{T,2}|\Omega] \leqslant \left(\Delta + \frac{6p}{\Delta\left(\frac{1}{2} - \frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C} - \varepsilon_\tau\right)^2}\right)\frac{T}{n\left(\frac{1}{2} - (\varepsilon_n + \varepsilon_C)\right)}$$

Par ailleurs des résultats déjà établis donnent :

$$\mathbb{P}\left(\Omega^C\right) \leqslant 2\exp\left(-2\binom{n}{2}\Delta^2(\varepsilon_\tau - \varepsilon_n)^2\right) + e^{-n} + 2\exp\left(-2n\varepsilon_n^2\right)$$

On obtient enfin la borne suivante sur le regret global :

$$\begin{aligned} R_T \leqslant \Delta\binom{n}{2} &+ \left(\Delta + \frac{6p}{\Delta\left(\frac{1}{2} - \frac{2\varepsilon_C}{1-2\varepsilon_n-2\varepsilon_C} - \varepsilon_\tau\right)^2}\right)\frac{T}{n\left(\frac{1}{2} - (\varepsilon_n + \varepsilon_C)\right)} \\ &+ \Delta T\left(2\exp\left(-2\binom{n}{2}\Delta^2(\varepsilon_\tau - \varepsilon_n)^2\right) + e^{-n} + 2\exp\left(-2n\varepsilon_n^2\right)\right) \end{aligned}$$

On a posé :

$$n\varepsilon_n = \sqrt{\frac{\Delta}{32\tau}\binom{n}{2}}$$

En prenant par exemple : $\varepsilon_\tau = \frac{1}{4}$ et $\varepsilon_C = \frac{1}{16}$, on trouve :

$$R_T \leqslant \Delta\binom{n}{2} + \left(\Delta + \frac{210p}{\Delta}\right)\frac{4T}{n} + \Delta T\left(2\exp\left(-\binom{n}{2}\frac{\Delta^2}{32}\right) + e^{-n} + 2\exp\left(-\frac{\Delta(n-1)}{32\tau}\right)\right)$$

On choisit alors :

$$n = \frac{p^{1/3}}{\Delta^{2/3}}T^{1/3}$$

Ce qui donne :

$$R_T = O\left(\frac{p^{2/3}}{\Delta^{1/3}}T^{2/3}\right)$$

2.3 Amélioration possible

1. On tire un ensemble $\mathcal{N}$ de n sommets de manière équilibrée, numérotés de 1 à n, puis :
   - on tire les arêtes correspondant à chaque paire du noyau ;
   - on évalue $\widehat{\tau} = \binom{n}{2}^{-1} \sum_{1 \leqslant i < j \leqslant n} X_{ij}$;
   - on utilise un algorithme de détection de communautés dans le sous-graphe $\mathcal{N}$.

2. On se propose de classer un ensemble $\mathcal{S}$ de sommets de $V \backslash \mathcal{N}$.
Tant que nous n'avons pas effectué T tirages, on pioche $x \in \mathcal{S}$, et on réutilise la stratégie avec τ connu en l'approximant par $\widehat{\tau}$, le tout sur les communautés $\widehat{C}_0$ et $\widehat{C}_1$:
   - On tire un sommet dans chaque $\widehat{C}$.
   - Tant que ni $\widehat{C}_0$ ni $\widehat{C}_1$ n'a été intégralement visitée :
     - Si les $\widehat{\mu}$ sont tous deux inférieurs à $\widehat{\tau}$, on tire un sommet supplémentaire dans chaque $\widehat{C}$.
     - Sinon, on tire un sommet de la communauté associée au $\widehat{\mu}$ le plus élevé.

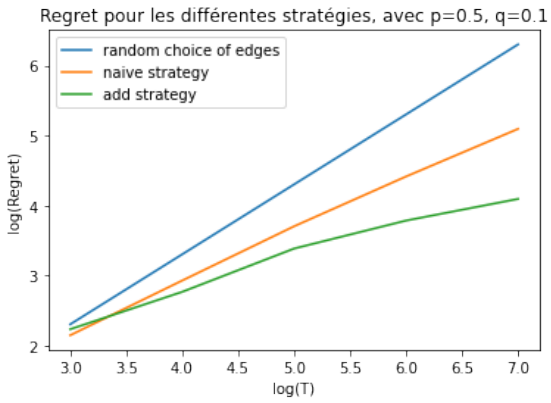
   - **On ajoute x à la communauté estimée dans laquelle on a effectué le dernier tirage.**

2.3.1 • Conjectures sur cet algorithme

On pense que cet algorithme, très similaire au premier à ceci près que le noyau croît après chaque nouvelle classification, donne de meilleurs résultats asymptotiques que l'algorithme précédent.

Plus précisément, on conjecture que cette méthode donne un regret en $O\left(\frac{p}{\Delta}\sqrt{T}\right)$, qui est la borne optimale pour ce modèle, *cf.* [2].

En effet, nos simulations suggèrent clairement que la deuxième stratégie proposée est la meilleure :



Donnons maintenant quelques arguments en faveur de ce résultat.

2.3.2 • Retour sur le choix de n

On a vu dans la partie précédente que le regret respectait une inégalité de la forme :

$$R_T \leqslant \Delta\left(\binom{n}{2} + \alpha\frac{p}{\Delta^2}|\mathcal{S}|\right) + \varepsilon(T)$$

avec $\varepsilon(T) \xrightarrow[T\to+\infty]{} 0$ et α une constante.

Le fait d'ajouter les sommets à nos communautés pourrait permettre de réduire drastiquement les valeurs de n et $|\mathcal{S}|$ optimales.

Dans l'algorithme précédent, la meilleure borne pour $|\mathcal{S}|$ était de la forme :

$$|\mathcal{S}| \leqslant \alpha\frac{T}{n}$$

avec α une constante, ce qui aboutissait à une valeur de n de l'ordre de $T^{1/3}$.

Nous pensons qu'ajouter des sommets au fur et à mesure permet d'arriver à une borne de la forme :

$$|\mathcal{S}| \leqslant \alpha\sqrt{T}$$

Pour peu que $n \geqslant T^{\varepsilon}$ pour un certain $\varepsilon > 0$, on aurait alors :

$$R_T \leqslant \Delta\left(\binom{n}{2} + \alpha\frac{p}{\Delta^2}\sqrt{T}\right) + \varepsilon(T)$$

En prenant $n = T^{1/4}$, on trouverait bien :

$$R_T = O\left(\frac{p}{\Delta}\sqrt{T}\right)$$

2.3.3 • Majoration de $|\mathcal{S}|$

On s'intéresse ici à $\mathcal{S}^*$, qu'on définit comme égal à $\mathcal{S}$ privé d'un éventuel dernier sommet problématique : rien n'indique que notre algorithme épuisera une classification au T^{e} tour.

En supposant que notre deuxième algorithme classifie **parfaitement** tous les sommets de $\mathcal{S}^*$, on se convainc rapidement que chaque sommet de $\mathcal{S}^* \cap C_i$ finit par voir toutes ses arêtes avec les sommets de $\widehat{C}_i$ et $\mathcal{S}^* \cap C_i$ tirées.

On aurait donc au moins

$$\binom{|\mathcal{S}^* \cap C_i|}{2} + \left|\widehat{C}_i\right| \cdot \left(|\mathcal{S}^* \cap C_i|\right)$$

tirages pour chaque communauté.

Cela donne au moins $\frac{|\mathcal{S}^*|^2}{4}$ tirages, toutes communautés confondues.

$|\mathcal{S}| = \left\lceil 2\sqrt{T} \right\rceil + 1$ sommets suffisent donc pour atteindre les T tirages.

2.3.4 • Résultats à prouver

Il reste à prouver que la catégorisation faite par ce deuxième algorithme est suffisament bonne pour garder une proportion suffisamment faible de sommets mal catégorisés dans les communautés estimées au fil des ajouts de sommets. C'est à dire qu'à tout instant et pour $i \in \{0, 1\}$ on ait l'existence d'$\varepsilon > 0$ tel que :

$$\frac{\left|\widehat{C}_i \cap C_{1-i}\right|}{\left|\widehat{C}_i\right|} \leqslant \varepsilon < \frac{1}{2} - \varepsilon_\tau$$

Cette inégalité serait suffisante à ce que la quantité

$$1 + \frac{6p}{\Delta^2 \left(\frac{1}{2} - \varepsilon - \varepsilon_\tau\right)^2}$$

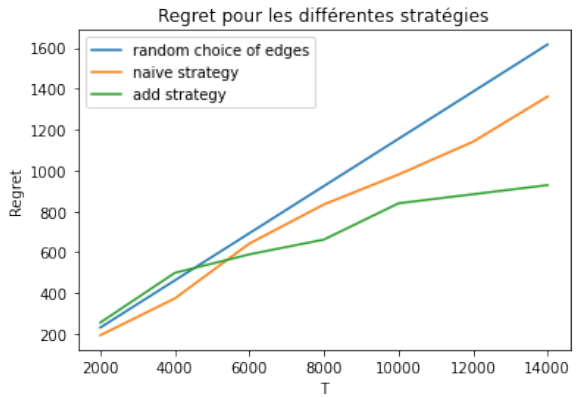
majore l'espérance du nombre de mauvais tirages pour un sommet donné, ce qui nous permettrait de conclure.

À ce stade, nous n'avons trouvé aucun moyen piste pour montrer que la qualité des communautés ne se détériore pas trop pour des valeurs de T plus grande, ce que nos simulations semblent pourtant suggérer.

2.3.5 • Remarques d'ordre pratique

Les détails techniques de notre implémentation sont consultables dans notre notebook Jupyter.

Voici les courbes obtenues pour nos données issues de Reddit :



Les données corroborent nos conclusions et aboutissent sensiblement aux mêmes résultats que nos simulations.

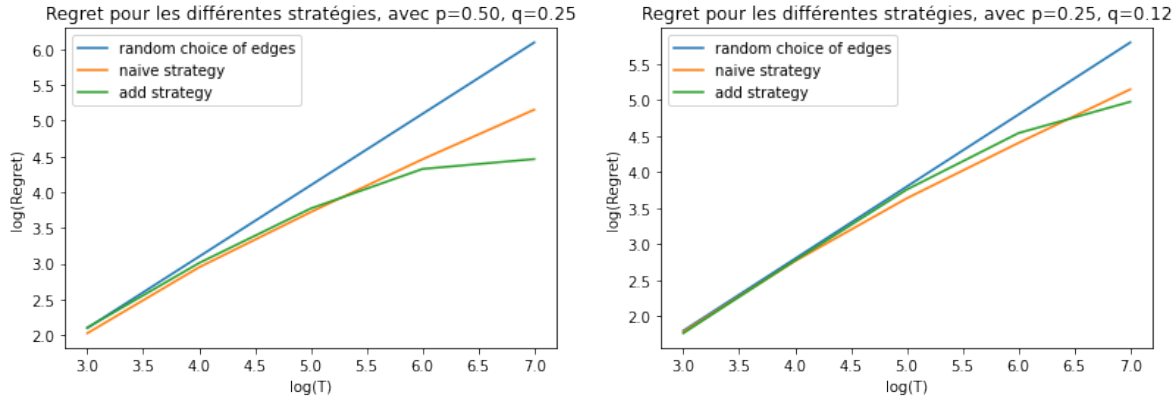
Cependant plusieurs remarques peuvent être faites, en sachant que nos valeurs de p et q sont à peu près similaires à celles choisies pour nos simulations : $p \simeq 0,32$ et $q \simeq 0,07$.

La première est que la deuxième stratégie n'a que très peu d'intérêt, voire est contre-productive pour des valeurs de T trop faibles.

La deuxième est que nos données n'ont *a priori* aucune raison de coller à notre modèle. En effet certains utilisateurs plus actifs sont à l'origine d'un plus grand nombre d'arêtes. Les valeurs estimées de p et q ne sont donc que des moyennes, et la variabilité dans le nombre de connexions aux autres sommets de la communauté est plus grande que dans nos simulations.

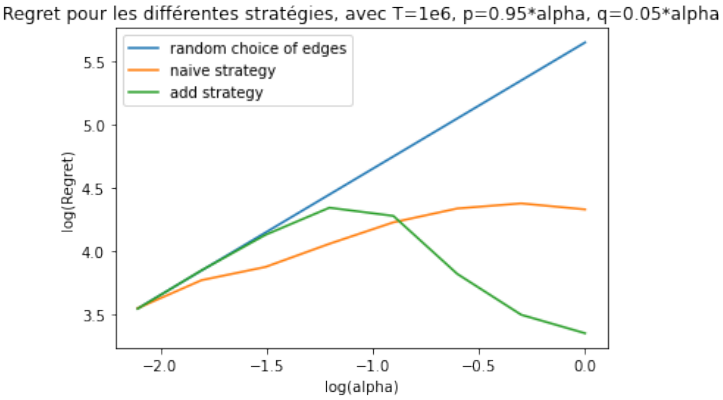
La troisième est que malgré tout ce que l'on a dit, on parvient à diminuer significativement notre regret par rapport à une sélection aléatoire d'arêtes.

On peut ajouter les résultats suivants :



Ainsi, le fait de réduire nos paramètres p et q d'un même facteur pousse vers le premier régime, celui où nos stratégies ne font guère mieux que le hasard.

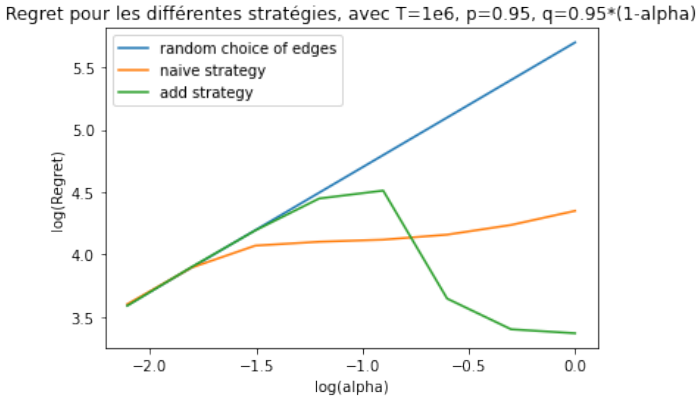
On peut encore mieux observer ce phénomène dans le graphique suivant :



On choisit ici d'étudier, pour $T = 10^6$, ce qu'il se passe pour des paramètres $p = \alpha p_0$ et $q = \alpha q_0$, avec $p_0 = 0,95$, $q_0 = 0,05$ et $\alpha \in]0, 1]$.

Il nous permet notamment de voir que pour un T de l'ordre du million, si nos valeurs de p et q sont trop faibles, il vaut parfois mieux utiliser notre premier algorithme, et ceux même si l'écart relatif entre p et q est important (il est ici constant et égal à 20).

Similairement, on peut fixer $p = p_0$ et poser $q = (1 - \alpha)p_0$, ce qui revient à faire varier Δ. On trouve alors, pour $p_0 = 0,95$:



On retrouve une dynamique très similaire, avec là encore une performance meilleure pour le premier algorithme lorsque p et q sont trop proches compte tenu de la valeur de T.

Que cela soit à p fixé ou non, cette meilleure performance du premier algorithme était attendue : pour $T = 10^6$, le noyau sur lequel se base le deuxième algorithme comporte $\sqrt[4]{10^6} < 32$ éléments. Si p et q sont trop proches, notre marge d'erreur sur la classification de ce noyau sera trop grande, et les erreurs commises se répercuteront sur toutes les classifications opérées durant la deuxième phase.

RÉFÉRENCES

[1] R. J. Serfling. Probability Inequalities for the Sum in Sampling without Replacement. *The Annals of Statistics*, 2(1) : 39 – 48, 1974.

[2] Christophe Giraud, Yann Issartel, Luc Lehéricy, and Matthieu Lerasle. Pair-matching : Links prediction with adaptive queries, 2020.