

Modèles de graphes aléatoires et croissance logarithmique des distances

Résumé : On étudie comment, dans un espace métrique initialement grand, les distances peuvent être fortement diminuées si on ajoute quelques “raccourcis” aléatoires. Cela peut modéliser diverses situations concrètes (réseaux sociaux, internet, propagation d’épidémies etc.).

Mots clefs : distance, loi binomiale, inégalité de Markov

- *Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. Vous êtes laissé(e) libre d'organiser votre discussion comme vous l'entendez. Des suggestions de développement, largement indépendantes les unes des autres, vous sont proposées en fin de texte. Vous n'êtes pas tenu(e) de les suivre. Il vous est conseillé de mettre en lumière vos connaissances à partir du fil conducteur constitué par le texte. Le jury appréciera que la discussion soit accompagnée d'exemples traités sur ordinateur.*

Le monde est grand : plus de sept milliards d’êtres humains, dont un individu donné ne connaîtra jamais plus d’une infime partie. Et pourtant le monde est petit : si on considère la “distance des poignées de main” (on est à distance 1 de quelqu’un à qui on a serré la main, à distance 2 de quelqu’un qui a serré la main d’une personne à qui on a serré la main etc.), deux êtres humains quelconques sont, d’après certaines estimations, à distance au plus 7. Le présent texte étudie un modèle simplifié qui tente de rendre compte de ce phénomène.

Notations. Si x est un réel positif, $Ent(x)$ désigne la partie entière de x . Si E est un ensemble, $card(E)$ désigne le cardinal de E et $\mathcal{P}_2(E)$ est l’ensemble des paires d’éléments de E : $\mathcal{P}_2(E) = \{\{i, j\}, i \in E, j \in E, i \neq j\}$. Enfin si F est un événement d’un espace de probabilités Ω , $\mathbf{1}_F$ est la fonction indicatrice de F : pour $\omega \in \Omega$, $\mathbf{1}_F(\omega) = 1$ si $\omega \in F$ et $\mathbf{1}_F(\omega) = 0$ si $\omega \notin F$.

1. Modèles déterministes

1.1. Le modèle formel

On modélise la situation comme suit : l’ensemble des individus est noté S et deux individus $i, j \in S$ se connaissent si et seulement si la paire $\{i, j\}$ appartient à un certain sous-ensemble $A \subset \mathcal{P}_2(S)$. La donnée de S et A est appelée graphe et on utilisera la terminologie des graphes en appelant *sommets* les éléments de S et *arêtes* les éléments de A .

Si i et j sont deux sommets, un *chemin* de longueur $n \geq 1$ de i à j est une suite de sommets (i_0, \dots, i_n) telle que $i_0 = i$, $i_n = j$ et que pour tout $k \in [0, n - 1]$, on ait $\{i_k, i_{k+1}\} \in A$. La distance entre deux sommets i et j est la longueur minimale d’un chemin de i à j . Formellement, pour tout $i \in S$, $d(i, i) = 0$ et pour toute paire $\{i, j\} \in \mathcal{P}_2(S)$,

$$d(i, j) = \min\{n \in \mathbb{N}, \text{il existe un chemin de longueur } n \text{ de } i \text{ à } j\}$$

avec, par convention, $\min \emptyset = \infty$. On vérifie aisément que d est une distance.

1.2. Le graphe géographique

Soit $N \geq 2$ un entier. On considère le graphe circulaire \mathcal{C}_N à N points, dont l'ensemble des sommets est $S = \{1, 2, \dots, N\}$ et l'ensemble A des arêtes est l'ensemble des paires de la forme $\{i, j\}$ avec $j = i + 1$ modulo N .

Le graphe \mathcal{C}_N modélise, de manière simplifiée, un monde où les gens ne connaissent que leurs voisins géographiques. Si N est grand, la distance entre deux points pris au hasard de façon indépendante dans S est de l'ordre de $N/4$ et la distance maximale entre deux points est $Ent(N/2)$.

1.3. Graphes avec raccourcis réguliers

On veut modéliser le fait que les individus peuvent connaître des personnes géographiquement éloignées. Pour cela, on ajoute des "raccourcis", c'est-à-dire des arêtes supplémentaires. Pour être réaliste du point de vue de la modélisation, il faut que le nombre de connaissances éloignées d'un individu donné ne soit pas trop grand.

Un autre point de vue peut être le suivant : supposons que \mathcal{C}_N modélise un réseau informatique, les sommets étant les ordinateurs et les arêtes les connexions entre ces ordinateurs. Pour améliorer les performances du réseau, on peut ajouter des connexions à longue distance mais pour des raisons financières, on ne veut pas en ajouter trop. De plus, on ne veut pas ajouter trop de connexions à un ordinateur donné pour éviter des phénomènes de saturation.

Il peut sembler naturel d'ajouter des arêtes de manière régulière, c'est-à-dire avec invariance par rotation. Fixons donc un entier $k \geq 2$ et supposons que pour tout $i \in S$, on ajoute une arête entre les sommets i et $i + k$ modulo N . Alors il est facile de voir que les longues distances vont en gros être divisées par k . Quand N augmente, la distance maximale $D(N)$ entre deux points va croître linéairement en N , c'est-à-dire que $D(N)/N$ va converger quand $N \rightarrow \infty$.

Une solution plus astucieuse consiste à ajouter une arête entre i et $i + Ent(\sqrt{N})$ modulo N . Dans ce cas, on peut voir que $D(N)$ va croître comme \sqrt{N} quand N augmente. On va montrer dans la suite de ce texte qu'ajouter des arêtes de manière aléatoire donne de meilleurs résultats.

2. Le modèle aléatoire

2.1. Définition du graphe avec raccourcis aléatoires

On va construire un nouveau graphe \mathcal{C}'_N en ajoutant au graphe \mathcal{C}_N décrit en 1.2 des raccourcis aléatoires. Informellement, on ajoute un certain nombre d'arêtes, chaque arête $\{i, j\}$ étant ajoutée indépendamment avec probabilité c/N .

Formellement, l'ensemble des sommets de \mathcal{C}'_N est S et l'ensemble des arêtes est défini comme suit. Tout d'abord, on choisit un réel c avec $1 < c \leq N$. Soit Q l'ensemble des paires $\{i, j\}$ de sommets de \mathcal{C}'_N telles que $\{i, j\} \notin A$. Soit $(\xi_{\{i,j\}}, \{i, j\} \in Q)$ une famille indépendante de variables aléatoires de Bernoulli de paramètre c/N . Alors on pose $A' = \{\{i, j\} \in Q, \xi_{i,j} = 1\}$: A' est donc l'ensemble des "raccourcis". L'ensemble des arêtes de \mathcal{C}'_N est défini comme $A \cup A'$.

On dit que $j \in S$ est une *connaissance lointaine* de $i \in S$ si $\{i, j\} \in A'$. Le nombre de connaissances lointaines d'un sommet donné suit une loi binomiale de paramètres $(N-3, c/N)$. En particulier, le nombre maximal de connaissances lointaines n'est pas trop grand :

Proposition 1. *Pour tout entier $m \in [c, N]$, soit $Z(c, m, N)$ la probabilité qu'il existe un sommet $j \in S$ ayant au moins m connaissances lointaines. Alors*

$$Z(c, m, N) \leq N \left(\frac{ec}{m} \right)^m$$

Démonstration. Tout d'abord, $Z(c, m, N)$ est majorée par

$$\sum_{i=1}^N P(i \text{ a au moins } m \text{ connaissances lointaines}) = NP(1 \text{ a au moins } m \text{ connaissances lointaines})$$

Soit B le nombre de connaissances lointaines du sommet 1. On cherche donc à majorer $P(B \geq m)$. On utilise pour cela l'inégalité de Markov qui donne, pour tout $t > 1$,

$$(1) \quad P(B \geq m) t^m \leq E(t^B)$$

Or pour tout réel $t > 0$,

$$E(t^B) = \left(1 + \frac{c(t-1)}{N} \right)^{N-3}$$

Or, pour tous $x \geq 0$ et $k > 0$, on a $(1 + x/k)^k \leq e^x$, de sorte que pour tout $t > 1$,

$$\left(1 + \frac{c(t-1)}{N} \right)^{N-3} \leq \left(1 + \frac{c(t-1)}{N} \right)^N \leq \exp(c(t-1))$$

En prenant $t = m/c$ dans l'inégalité (1), on démontre la proposition 1. □

Notons que d'après la loi des événements rares, on a :

Proposition 2. *Le nombre de connaissances lointaines du sommet 1 converge en loi, quand N tend vers l'infini, vers la loi de Poisson de paramètre c .*

2.2. Taille des petits voisinages d'un sommet donné

Dans notre modèle, le sommet 1 a en moyenne environ $c+2$ voisins, chacun de ces voisins à son tour a en moyenne environ $c+1$ voisins si on ne compte pas le sommet 1 etc. Donc intuitivement, on peut penser qu'en faisant n pas, on peut atteindre environ $(c+1)^n$ personnes différentes ; c'est le principe du système pyramidal à la Madoff.

Il est bien connu que les systèmes pyramidaux ont des limites même si, au début, ils marchent. De même, le raisonnement du paragraphe précédent est évidemment faux dès que n est trop grand (par exemple si $(c+1)^n > N$) mais il est presque vrai pour n assez petit.

Plus précisément, pour n pas trop grand, le nombre de sommets à distance n du sommet 1 est au moins de l'ordre de c^n . On va donner une démonstration partielle de ce fait.

Pour $n \geq 1$, soit W_n l'ensemble des sommets k tels qu'il existe un chemin de longueur n de 1 à k (de tels sommets k sont donc à distance au plus n du sommet 1). On a

$$E(\text{card}(W_n)) = E\left(\sum_{k \in S} \mathbf{1}_{\{k \in W_n\}}\right) = \sum_{k \in S} P(k \in W_n)$$

Pour estimer $P(k \in W_n)$, on va considérer $C_n(k)$ le nombre de chemins de longueur n de 1 à k . Comme on a $P(k \in W_n) = P(C_n(k) > 0)$, on est ramené à estimer $P(C_n(k) > 0)$. Pour cela, une méthode consiste à minorer $E(C_n(k))$, majorer $E(C_n(k)^2)$ et utiliser le résultat suivant :

Proposition 3. Soit X une variable aléatoire à valeurs dans \mathbb{R}_+ , de variance finie. Alors

$$P(X > 0) \geq \frac{[E(X)]^2}{E(X^2)}$$

Démonstration. On démontre la proposition 3 en écrivant :

$$E(X) = E(X\mathbf{1}_{\{X>0\}}) \leq \sqrt{E(X^2)}\sqrt{E(\mathbf{1}_{\{X>0\}})}$$

□

Minorons $E(C_n(k))$. Soit $D_n(k)$, l'ensemble des suites de la forme (i_0, \dots, i_n) telles que les i_j soient des sommets de S , que $i_0 = 1$, $i_n = k$ et que pour tout $j \in [0, n-1]$, $i_j \neq i_{j+1}$. Une telle suite est un chemin de 1 à k si et seulement si pour tout $j \in [0, n-1]$, $\{i_j, i_{j+1}\}$ est une arête, c'est-à-dire $\{i_j, i_{j+1}\} \in A \cup A'$. Or pour tout $j \in [0, n-1]$, on a :

- si $\{i_j, i_{j+1}\} \in A$, alors $\{i_j, i_{j+1}\}$ est une arête,
- sinon, $P(\{i_j, i_{j+1}\} \in A') = c/N$ et donc $\{i_j, i_{j+1}\}$ est une arête avec probabilité c/N .

Donc pour tout $D \in D_n(k)$, la probabilité que D soit un chemin de 1 à k est au moins égale à $(c/N)^n$. Par ailleurs le cardinal de $D_n(k)$ satisfait

$$\text{card}(D_n(k)) \geq (N-1)^{n-1} - (N-1)^{n-2}$$

Ainsi

$$E(C_n(k)) = \sum_{D \in D_n(k)} P(D \text{ est un chemin}) \geq [(N-1)^{n-1} - (N-1)^{n-2}] \left(\frac{c}{N}\right)^n$$

Précision. La suite de ce texte énonce des résultats et donne des éléments de preuve mais pas de démonstration complète, ce qui serait trop long dans le cadre de l'épreuve.

La majoration de $E(C_n(k)^2)$ se fait par un raisonnement similaire, mais la combinatoire des chemins est plus compliquée. On admettra le résultat suivant : il existe une constante $K > 0$ telle que pour tout entier $N > 2$ et tout entier positif $n \leq \text{Ent}((2 \log N)/(3 \log c))$,

$$E(\text{card}(W_n)) \geq Kc^n$$

Plus généralement, on peut démontrer le résultat suivant :

Fait (admis). Il existe des réels $q \in]0, 1[$ et $a > 0$ tels que pour tout entier $N > 0$ et pour tout $n \in [1, \text{Ent}((2 \log N)/(3 \log c))]$, avec probabilité supérieure à q ,

$$\text{card}(W_n) \geq ac^n$$

D'après ce résultat, partant du sommet 1, on peut atteindre beaucoup de sommets différents par des chemins assez courts, de longueur environ $\log(N)$. Mais à cause de la condition $n \leq \text{Ent}((2 \log N)/(3 \log c))$, condition liée à la majoration de $E(C_n(k)^2)$, le fait admis établit seulement que le nombre de ces sommets accessibles par des chemins assez courts est au moins de l'ordre de $N^{2/3}$. On voudrait montrer qu'il est de l'ordre de N . Cela revient à dire que n'importe quel sommet donné est accessible par un chemin assez court avec une probabilité non négligeable. La dernière partie du texte aborde cette question.

2.3. Croissance logarithmique des distances

Un "paradoxe" assez connu est que si on prend 23 personnes au hasard, la probabilité que deux d'entre elles aient le même anniversaire est plus grande que 1/2. Sur d'autres planètes où il y a plus de jours dans l'année, disons n , le nombre de personnes à réunir est plus grand mais il est de l'ordre de \sqrt{n} (et pas $n/2$ comme pourrait le croire un non-mathématicien). Une autre formulation du "paradoxe des anniversaires" est la suivante :

Proposition 4. Soit $N > 1$ un entier et soit $E = \{1, 2, \dots, N\}$. Soient E_1 et E_2 deux sous-ensembles aléatoires, indépendants et uniformément répartis sur l'ensemble des sous-ensembles de E de cardinal $\text{Ent}(\sqrt{N}) + 1$. Alors

$$P(E_1 \cap E_2 \neq \emptyset) \geq 1 - e^{-7/10}$$

C'est ce phénomène qui permet de résoudre la difficulté soulevée à la fin de la partie précédente. Soit en effet j un sommet quelconque de \mathcal{C}'_N . Pour n entier positif, soit $B_1(N, n)$ (resp. $B_2(N, n)$) l'ensemble des sommets k tels qu'il existe un chemin de longueur n de 1 à k (resp. de j à k). Si $k \in B_1(N, n)$, alors k est à distance au plus n de 1. De plus, si $B_1(N, n)$ et $B_2(N, n)$ ne sont pas vides et ne sont pas disjoints, alors les sommets 1 et j sont à distance $\leq 2n$.

Or si on prend $n = \text{Ent}[(\log N)/(2 \log c) - (\log a)/(\log c)] + 1$, d'après le fait admis dans la partie 2.2, pour N assez grand, avec probabilité supérieure à q , on a

$$\text{card}(B_1(N, n)) \geq \sqrt{N}$$

et de même pour $\text{card}(B_2(N, n))$. En utilisant la proposition 4, on a envie d'en déduire qu'avec probabilité supérieure à $q^2(1 - e^{-7/10})$, les sommets 1 et j sont à distance au plus $2n$.

Le problème est que les ensembles $B_1(N, n)$ et $B_2(N, n)$ ne sont pas indépendants et ne vérifient pas les hypothèses de la proposition 4. Cependant en affinant un peu les arguments, on peut montrer qu'il existe $q' > 0$ tel que pour tout $N > 1$ et pour tout sommet j du graphe \mathcal{C}_N , avec probabilité au moins q' , les sommets 1 et j sont à distance au plus $(\log N / \log c) + 1$.

On voit donc que les distances du modèle étudié croissent en $\log(N)$, donc très lentement, alors que le nombre moyen de connaissances d'un individu donné est borné par $c + 2$. En particulier, les raccourcis aléatoires diminuent plus les distances que les raccourcis réguliers déterministes vus en 1.3.

Suggestions pour le développement

- ▶ *Soulignons qu'il s'agit d'un menu à la carte et que vous pouvez choisir d'étudier certains points, pas tous, pas nécessairement dans l'ordre, et de façon plus ou moins fouillée. Vous pouvez aussi vous poser d'autres questions que celles indiquées plus bas. Il est très vivement souhaité que vos investigations comportent une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats.*
- *Modélisation.*
 - On pourra analyser comment les raccourcis réguliers décrits en 1.3 rétrécissent le graphe.
 - On pourra discuter de la pertinence du modèle aléatoire et éventuellement suggérer des aménagements qui le rendraient plus réaliste.
 - Dans la pratique, on observe un réseau sans connaître *a priori* la valeur de c . Le candidat pourra indiquer comment on estime le paramètre c à partir de l'observation du réseau.
- *Applications informatiques.*
 - On pourra essayer de simuler le réseau aléatoire pour des valeurs de N d'environ 50, et $1 < c \leq 4$. On comptera alors le nombre de sommets à distance 1, 2 ou 3 du sommet 1.
 - On pourra aussi illustrer le paradoxe des anniversaires. Sous sa forme "classique", on tirera au hasard, indépendamment, des dates uniformes sur $\{1, 2, \dots, 365\}$ et on vérifiera si certaines d'entre elles coïncident. Sinon, on pourra illustrer ce paradoxe sous sa forme de la proposition 4.
- *Développements mathématiques.*
 - On peut étudier en détail les preuves des résultats énoncés dans le texte, notamment :
 - la majoration de la proposition 1 ;
 - l'inégalité de la proposition 3 ;
 - la minoration de $E(C_n(k))$.
 - Les candidats qui le souhaitent pourront démontrer la proposition 4, en remarquant que cela revient à calculer la probabilité que E_2 soit disjoint de l'ensemble $\{1, 2, \dots, Ent(\sqrt{N}) + 1\}$, le logarithme de cette probabilité se minorant par des estimations usuelles.