PROJET 17

Estimation de la taille d'un graphe par marches aléatoires

sujet proposé par L. Massoulié

laurent.massoulie@inria.fr

Etant donné un graphe $G = (V, E)$ fini, non orienté et connexe, on cherche à estimer sa taille. On suppose qu'on a accès au graphe de la manière suivante: on peut accéder à un sommet particulier $i_0 \in V$ du graphe, et ayant accédé à un sommet i du graphe, on peut alors accéder à n'importe quel voisin j de i dans G. Cette situation se produit par exemple lorsque le graphe représente les individus participant à un réseau pair-à-pair et les connexions existant entre ces individus. On peut aussi considérer l'estimation de la taille du web, en supposant qu'on sait uniquement passer d'une page à une autre conectée à la précédente par un lien hypertexte.

On va considérer l'algorithme suivant pour l'estimation de cette taille. On fixe une longueur τ, et un nombre cible ℓ. A chaque étape $t \geq 1$, on construit une marche aléatoire de longueur τ, soit $X_0^t, \ldots, X_\tau^t$ sur le graphe, issue de $X_0^t = i_0$. Par définition, sachant les choix $X_0^t, \ldots, X_{s-1}^t$, X_s^t est sélectionné uniformément au hasard parmi les voisins de X_{s-1}^t dans G. On note $Y_t = X_\tau^t$ le dernier sommet obtenu à la t-ème marche aléatoire.

On dit qu'on a une **collision** à l'étape t si $Y_t \in \{Y_1, \ldots, Y_{t-1}\}$. On note C_ℓ l'instant de la ℓ-ème collision, soit:

$$C_\ell = \inf\{t \geq 1 : \sum_{s=1}^{t} \mathbf{1}_{Y_s \in \{Y_1, \ldots, Y_{s-1}\}} = \ell\}.$$

L'estimateur $\hat{N}$ du nombre de sommets $N := |V|$ du graphe est alors donné par

$$\hat{N} = \frac{C_\ell^2}{2\ell}.$$

La partie théorique établit des garanties théoriques sur cette procédure d'estimation. La partie simulation consiste à la mise en oeuvre de cet algorithme.

1 Partie théorique

On va supposer par la suite que le graphe G considéré est d-régulier, i.e. chaque sommet $i \in V$ a d voisins dans G. On notera par ailleurs A la matrice d'adjacence du graphe G, qui est par définition la matrice carrée symétrique, dont les lignes et les colonnes sont indicées par les sommets $i \in V$ du graphe, et telle que $A_{ij} = 1$ si (i, j) est un arc du graphe, et $A_{i,j} = 0$ sinon.

T1. Exprimer la loi de probabilité de chaque échantillon Y_t, soit $\pi := \{\mathbb{P}(Y_t = i)\}_{i \in V}$ au moyen de la matrice P^τ, où $P := d^{-1}A$.

T2. En supposant que le spectre de la matrice P est constitué de la valeur propre 1 et d'autres valeurs propres $\lambda_2, \ldots, \lambda_N$ dont les modules sont majorés par $1 - \epsilon$, pour $\epsilon > 0$, et en utilisant le théorème spectral pour P, en déduire que la loi π des Y_t converge pour la norme $\|\cdot\|_2$, lorsque τ tend vers l'infini, vers la loi uniforme sur V.

T3. On fait maintenant l'approximation suivante au vu de la question précédente: on suppose que τ est choisi suffisamment grand, de sorte qu'on peut supposer les Y_t i.i.d., et **uniformément distribués** sur V. Notant $C_1, \ldots, C_\ell$ les instants des ℓ collisions s'étant produites dans l'implémentation de l'algorithme, établir la formule suivante pour $n, m > 0$:

$$\mathbb{P}(C_\ell - C_{\ell-1} > n | C_{\ell-1} = m) = \frac{(N-m+\ell-1)(N-m+\ell-2)\cdots(N-m+\ell-n)}{N^n}.$$

T4. Etablir, pour tout ℓ fixé et tous $a, b > 0$ fixés, la convergence

$$\lim_{N\to\infty} \mathbb{P}(C_\ell - C_{\ell-1} > b\sqrt{N} | C_{\ell-1} = a\sqrt{N}) = e^{-ab-b^2/2}.$$

En déduire la convergence, pour tous $x, y > 0$ fixés:

$$\lim_{N\to\infty} \mathbb{P}\left([C_\ell^2 - C_{\ell-1}^2]/(2N) > y | C_{\ell-1}^2/(2N) = x\right) = e^{-y}.$$

Ce dernier résultat entraîne, par récurrence sur ℓ, la convergence en distribution $(2N)^{-1}C_\ell^2 \overset{N\to\infty}{\to} E_1 + \ldots + E_\ell$, où $E_1, \ldots E_\ell$ sont i.i.d. et exponentiellement distribuées. Avec quelques étapes supplémentaires (non demandées!), on peut en déduire que l'estimateur proposé $\hat{N} := C_\ell^2/(2\ell)$ vérifie $\mathbb{E}\hat{N}/N \sim 1$, et Var$(\hat{N}/N) = \Theta(1/\ell)$. Ces propriétés suggèrent que $\hat{N}$ est un estimateur raisonnable de N lorsque $N \gg 1$ et ℓ est suffisamment grand.

2 Partie simulation

S1. Implémenter l'algorithme proposé initialement, en remplaçant la marche aléatoire standard par la variante suivante: pour passer de X_s^t à X_{s+1}^t, avec probabilité 1/10 on prend $X_{s+1}^t = X_s^t$, et avec probabilité $9/10$ on prend pour X_{s+1}^t un voisin de X_s^t uniformément au hasard dans G. Cette variante est dite marche aléatoire paresseuse, elle a pour intérêt qu'elle vérifie la propriété spectrale de la question T2, ce qui n'est pas le cas pour la marche aléatoire standard sur un graphe bi-partite, pour lequel la valeur propre -1 appartient au spectre de la matrice P.

S2. Tester l'algorithme avec comme graphe $G = (V, E)$ l'hypercube à δ dimensions, i.e. $V = \{u = (u_1, \ldots, u_\delta) \in \{0,1\}^\delta\}$, et

$$E = \{(u, v) \in V^2 : \sum_{i=1}^{\delta} |u_i - v_i| = 1\}.$$

On prendra en particulier $\delta = 10$, et on dessinera les histogrammes des valeurs $\hat{N}$ de l'estimation de $N = 1024$ correspondant aux choix de paramètres $\tau = 5, 10, 50, 100$, et $\ell = 1, 10, 100$, avec pour chaque choix de valeurs (τ, ℓ) 100 réalisations de l'estimateur $\hat{N}$.

On commentera l'impact des deux paramètres τ, ℓ sur la qualité des estimations obtenues, mesurées par l'erreur quadratique relative $\sqrt{(1/100)\sum_{i=1}^{100}(\hat{N}_i/N - 1)^2}$ sur les 100 réalisations.

S3. Tester maintenant l'algorithme sur le cycle C_N, par définition constitué des sommets $1, \ldots, N$ et des arcs $(i, i+1)$, $i = 1, \ldots, N-1$, et de l'arc $(N, 1)$. On dessinera les histogrammes des valeurs $\hat{N}$ de l'estimation de $N = 50$ correspondant aux choix de paramètres $\tau = 10, 100, 1000, 10000$, et $\ell = 1, 10, 100$, avec pour chaque choix de valeurs (τ, ℓ) 100 réalisations de l'estimateur $\hat{N}$. Comme dans la question précédente on commentera l'impact de τ et ℓ sur la qualité des estimations obtenues.

On commentera la différence entre la performance de l'estimateur dans le cas précédent de l'hypercube et le cas précédent du cycle.
