Étude des Pandémies

Épreuve pratique d'algorithmique et de programmation
Concours commun des écoles normales supérieures

Durée de l'épreuve: 3 heures 30 minutes

Juillet 2009

ATTENTION !

N'oubliez en aucun cas de recopier votre u_0
à l'emplacement prévu sur votre fiche réponse

Important.

Sur votre table est indiqué un numéro u_0 qui servira d'entrée à vos programmes. Les réponses attendues sont généralement courtes et doivent être données sur la fiche réponse fournie à la fin du sujet. À la fin du sujet, vous trouverez en fait deux fiches réponses. La première est un exemple des réponses attendues pour un $\widetilde{u_0}$ particulier (précisé sur cette même fiche et que nous notons avec un tilde pour éviter toute confusion!). Cette fiche est destinée à vous aider à vérifier le résultat de vos programmes en les testant avec $\widetilde{u_0}$ au lieu de u_0. Vous indiquerez vos réponses (correspondant à votre u_0) sur la seconde et vous la remettrez à l'examinateur à la fin de l'épreuve.

En ce qui concerne la partie orale de l'examen, lorsque la description d'un algorithme est demandée, vous devez présenter son fonctionnement de façon schématique, courte et précise. Vous ne devez en aucun cas recopier le code de vos procédures!
Quand on demande la complexité en temps ou en mémoire d'un algorithme en fonction d'un paramètre n, on demande l'ordre de grandeur en fonction du paramètre, par exemple: $O(n^2)$, $O(n \log n)$,...
Il est recommandé de commencer par lancer vos programmes sur de petites valeurs des paramètres et de ***tester vos programmes sur des petits exemples que vous aurez résolus préalablement à la main ou bien à l'aide de la fiche réponse type fournie en annexe.*** Enfin, il est recommandé de lire l'intégralité du sujet avant de commencer afin d'effectuer les bons choix de structures de données dès le début.

1 Introduction

Comme l'actualité nous l'a encore montré récemment avec le virus H1N1, il est important de comprendre les mécanismes de propagation de virus dans les populations, lors des épisodes de pandémie.

Dans cette épreuve, nous allons tâcher de modéliser une telle propagation en affinant un modéle de représentation des contacts dans une population.

Nous allons tout d'abord modéliser la population par un graphe.

Définition 1 (Graphe) *Un graphe $G = (V, E)$ est un couple d'ensembles, V de sommets, et E d'arêtes. L'ensemble $V = \{0, \ldots, |V|-1\}$ représente les individus, et les arêtes (l'ensemble $E \subset V^2$) encodent les relations entre individus, qui sont les vecteurs susceptibles de propager le virus. Le graphe est naturellement non-orienté, c'est-à-dire que si $(i, j) \in E$, alors $(j, i) \in E$. Il n'y a pas d'arête de la forme (i, i). On dira qu'une arête $e = (u, v)$ est incidente à un sommet i si $i = u$ ou $i = v$.*

Définition 2 (Composante connexe) *La composante connexe d'un sommet s dans un graphe est l'ensemble des sommets s' du graphe qui peuvent être joints à s en passant par un nombre quelconque d'arêtes, c'est-à-dire qu'il existe un ensemble de sommets $s = s_1, \ldots, s_k = s'$ tels que $(s_i, s_{i+1}) \in E$ pour tout i dans $1 \ldots k-1$.*
Les composantes connexes forment une partition du graphe. Le graphe est dit connexe lorsqu'il n'y a qu'une seule composante connexe, et que donc elle contient tous les sommets du graphe.

Dans un premier temps, nous allons générer des graphes de population simples, en extraire quelques informations, et faire des simulations dessus. Puis, nous essayerons d'extraire le chemin préférentiel de propagation du virus.

2 Graphes pseudo-aléatoires

Considérons les suites d'entiers (u_n) et (v_n) définies pour $n \geq 0$ par :

$$u_n = \begin{cases} \textit{votre } \mathtt{u_0} \textit{ (à reporter sur votre fiche)} & \text{si } n = 0 \\ 15\,091 \times u_{n-1} \mod 64\,007 & \text{si } n \geq 1 \end{cases}$$

$$v_n = \begin{cases} \mathtt{u_0} & \text{si } n = 0 \\ 1\,129 \times v_{n-1} \mod 63\,997 & \text{si } n \geq 1 \end{cases}$$

Question 1 *Que valent :* ***a)*** u_{10} ***b)*** $u_{1\,000}$ ***c)*** $v_{1\,000}$

On s'assurera de précalculer et stocker suffisamment de valeurs de u_n et v_n de manière à pouvoir y accéder en temps constant par la suite.

On notera $G_{n,t}$ le graphe à n sommets dont les arêtes sont :

$$\{(u_i \mod n, v_i \mod n) \textit{ et } (v_i \mod n, u_i \mod n) \mid i \in \{0, \ldots, t-1\}\}$$

On ignorera les couples de la forme (i, i) générés, ainsi que les doublons.
Pour représenter le graphe, on choisira une structure de données compacte, sachant que chaque sommet aura très peu de voisins en comparaison de n.

3 Statistiques sur les graphes

Définition 3 (Degré) *Le degré d'un sommet dans un graphe est le nombre d'arêtes qui y sont incidentes. Le degré moyen d'un graphe est naturellement la moyenne des degrés des sommets du graphe.*

La distance entre deux sommets d'un graphe est le nombre minimum d'arêtes consécutives qu'il faut traverser pour joindre les deux sommets. Elle est considérée infinie s'il n'existe pas de chemin entre les deux sommets (ils sont alors dans des composantes connexes différentes).

Nous nous intéresserons également au sommet le plus éloigné d'un sommet s dans sa composante connexe, que nous noterons $f(s)$, et $d(s)$ la distance entre s et $f(s)$ (qui est donc finie).

Définition 4 (Diamètre) *Finalement, nous définissons le diamètre du graphe comme étant le maximum de $d(s)$ sur tous les sommets du graphe (en fait le maximum des diamètres de chaque composante connexe du graphe).*

Question 2 *Que valent les degrés moyens des graphes suivants :*
a) $G_{5,10}$ **b)** $G_{1000,2000}$ **c)** $G_{10000,40000}$?

Question 3 *Que valent $d(0)$ sur les graphes suivants :*
a) $G_{5,10}$ **b)** $G_{1000,2000}$ **c)** $G_{10000,40000}$?

Question 4 *Que valent les diamètres des graphes suivants :*
a) $G_{5,10}$ **b)** $G_{100,200}$ **c)** $G_{1000,2000}$?

4 Graphes connexes minimaux

Question 5 *Combien les graphes suivants ont-ils de composantes connexes :*
a) $G_{5,10}$ **b)** $G_{1000,2000}$ **c)** $G_{10000,40000}$?

Nous cherchons maintenant à déterminer la plus petite valeur de e telle que $G_{n,e}$ soit connexe, qu'on note e_n. On note H_n le graphe G_{n,e_n}.

Question à développer pendant l'oral : Proposer un algorithme qui détermine e_n, et donner sa complexité. On cherchera un algorithme de complexité sous-quadratique.

Indication : on cherchera une structure de donnée qui permette efficacement : (a) de savoir dans quelle composante connexe un sommet donné se trouve (on pourra identifier une composante connexe par un de ses sommets), et (b) d'unifier deux composantes connexes ensemble. On pourra par exemple utiliser un arbre, dont la racine connaitrait le représentant de la composante connexe, et les sommets de l'arbre représenteraient les sommets de la composante connexe. Unifier deux composantes connexes reviendrait alors à accrocher un arbre à l'autre de manière judicieuse.

Exemple : prenons un graphe à 5 sommets avec les arêtes $(1, 2)$ et $(0, 4)$. Ce graphe a 3 composantes connexes, et on peut choisir les sommets 1, 4 et 3 comme représentants de ces composantes. Nous obtenons donc 3 arbres :
- l'arbre ne contenant qu'une racine, 3,
- l'arbre de racine 1 ayant un fils qui est 2,
- l'arbre de racine 4 ayant un fils qui est 0,

Si on ajoute l'arête $(0, 2)$, dont les sommets sont dans des composantes connexes différentes, on pourra par exemple raccrocher les deux derniers arbres en faisant en sorte que 4 devienne un nouveau fils de 1. On n'a donc maintenant plus que 2 composantes connexes, donc 2 arbres.

Question à développer pendant l'oral : On remarque que le choix des accrochages implique un coût de parcours variable pour trouver la racine d'un arbre, qui est le représentant de la composante connexe. On proposera donc une méthode qui permet d'obtenir une bonne complexité, que l'on essayera d'analyser.

Question 6 *Déterminer les valeurs de e_n pour les valeurs de n suivantes :*
a) 5 ***b)*** 1000 ***c)*** 10000 *?*

Question 7 *Déterminer la moyenne du nombre de composantes connexes des graphes $G_{n,e}$ pour e évoluant de 0 à e_n (inclus) pour les valeurs de n suivantes :*
a) 5 ***b)*** 1000 ***c)*** 10000 *?*

Question 8 *Que valent les diamètres des graphes suivants :* ***a)*** H_5 ***b)*** H_{100} ***c)*** H_{1000} *?*

5 Propagation du virus

Modélisons maintenant la propagation d'un virus dans les graphes H_n. Deux paramètres sont considérés : le nombre de jours J où un individu malade est contaminant, et la probabilité r qu'il a de contaminer chaque jour ses connaissances.

À chaque pas de temps, un individu peut être dans 3 états : sain, malade (et contagieux), ou vacciné (donc non-contaminable à nouveau, et ne contaminant plus). Si un individu est contaminé à un instant t, il est contaminant aux instants $t+1, \ldots, t+J$, et devient vacciné à $t + J + 1$. Le virus se propage sur une arête (i, j) à un instant t si $u_{i+j+t}/64007 < r$.

Le virus apparait en contaminant l'individu 0 au temps $t = 0$.

Question 9 *Combien d'invididus sont contaminés dans les graphes H_n, pour les valeurs de n suivantes, avec les paramètres $J = 5$ et $r = 0.2$:* ***a)*** 5 ***b)*** 1000 ***c)*** 10000 *?*

6 Squelette de propagation préférentiel

L'OMS aimerait également analyser le graphe de population de manière statique, afin d'obtenir des informations sur les chemins préférentiels de propagation du virus. Dans cette partie, nous considérerons uniquement le cas où toute la population est connectée (graphe connexe).

Dans cette optique, on fait correspondre à chaque arête (i, j) du graphe une probabilité de propagation de virus, ou poids, $p(i, j) = u_{i+j}/64007$.

On s'intéresse au "squelette" principal de la propagation du virus dans le graphe, c'est-à-dire à un sous-graphe connectant toute la population, de taille minimale (donc $|V| - 1$ arêtes), mais dont la somme des $p(i, j)$ pour toutes les arêtes soit maximum. On remarquera qu'un tel sous-graphe ne contient pas de cycle, et est donc (isomorphe à) un arbre. On parle d'arbre couvrant de poids maximal.

Pour calculer un tel arbre, nous allons appliquer l'algorithme glouton suivant :
- créer une forêt : un ensemble d'arbres où chaque sommet du graphe est un arbre séparé,
- créer un ensemble S contenant toutes les arêtes du graphe,
- retirer l'arête de S qui a le poids maximum (en cas d'égalité on considèrera, pour des raisons de reproductibilité des réponses de l'épreuve, l'arête qui a un de ses sommets de plus petit indice, et s'il y a encore égalité celle dont l'autre sommet a le plus petit indice),
- si ses deux sommets sont dans des arbres disjoints, alors combiner les deux arbres en un, et sinon, passer à l'arête suivante.

L'arbre obtenu au final a la propriété recherchée.

Question à développer pendant l'oral : Quel est la complexité de cet algorithme ? Vous discuterez également vos choix d'implémentation.

Question à développer pendant l'oral : Pouvez-vous prouver que l'arbre obtenu maximise bien la somme des poids ?

Question 10 *Donnez le minimum, le maximum et la moyenne des $p(i, j)$ sur les arbres ainsi obtenus pour les graphes H_n, pour les valeurs de n suivantes :*
a) 5 ***b)*** 1000 ***c)*** 10000 ?

Question 11 *Même question pour les graphes $H'_n = G_{n,10e_n}$, pour les valeurs de n suivantes :* ***a)*** 5 ***b)*** 1000 ***c)*** 10000 ?

◇
◇ ◇
◇