chapitre 2

Modèles de GRI

Section 1

graphes aléatoires G(n,p) d'Erdős-Rényi

Fabien de Montgolfier
fm@irif.fr

11 fevrier 2022

Plan du cours

1. **Description**
   Comprendre les propriétés structurelles des GRI
   en informatique, sociologie, biologie, physique, linguistique...
   - **en TP** : Calcul de divers paramètres.
     Big data → structures de données et algorithmes efficaces
2. **Modélisation**
   Classifier ces réseaux d'interaction.
   Modèles aléatoires.
   Si même propriétés que le réel, alors même génération ?
   - **en TP** : Génération aléatoire
3. **Deux applications**
   Dissémination (contagion)
   P2P (DHT ; diffusion temps réel)

Introduction

Un modèle c'est quoi ?

1. Pour l'informaticien : un processus de **génération aléatoire**
2. Pour le mathématicien : une **distribution de probabilités** sur la famille des graphes

Des modèles pour comprendre le monde

si le modèle à les mêmes propriétés que le réel alors peut-être le réel suit des **lois** semblables au modèle

- *Exemple* Erdős-Rényi «explique» les petites distances
- *Exemple* l'attachement préférentiel «explique» les power laws

Un modèle peut **valider** (en fait, non-infirmer) une hypothèse : **démarche scientifique** expérimentale (1.) ou mathématique (2.) !

Introduction

Des modèles pour démontrer

Le modèle d'Erdős-Rényi a été inventé comme outil de preuve : la **méthode probabiliste**. On y reviendra avec l'exemple du couplage.

Des modèles pour tester

Si les données réelles sont dures à obtenir, générer des données factices pour tester des protocoles ou algorithmes dessus.

Des modèles pour comparer

On peut comparer un graphe réel par rapport à un modèle aléatoire ("null model").

Attention à l'aléatoire

Beaucoup de confusion entre «génération faisant intervenir de l'aléatoire» et «génération aléatoire»

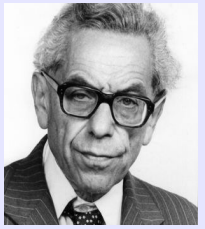
Le vrai sens de *génération aléatoire* est **équiprobable** ou uniforme : tout élément de la famille doit avoir la même probabilité de sortir. C'est très fort !

Par exemple tirer **un dé** est une «bonne» génération aléatoire d'un nombre entre 1 et 6. Mais tirer **deux dés** n'est pas une «bonne» génération aléatoire d'un nombre entre 2 et 12.

Beaucoup de gens en ingéniérie informatique ou en sciences expérimentales oublient de prouver que leur génération «avec de l'aléatoire» est uniforme. Ils ont souvent des **biais** importants...

Le modèle d'Erdős-Rényi [1959, 1960]

Paul Erdős (1913–1996)



Ce modèle a été inventé en même temps que *the probabilistic method*. Il suffit de montrer qu'un objet existe avec une probabilité > 0 pour être sûr qu'il existe ($\neq$ savoir le construire).

Quand on parle de «graphe aléatoire» sans plus de précision il s'agit de ce modèle, car il tire uniformément dans les graphes à densité fixée.

Attention, deux modèles

Erdős-Rényi $\mathcal{G}(n,p)$

n sommets. Pour deux sommets x et y l'arête xy existe avec probabilité p. $\binom{n}{2}$ tirages uniformes indépendants.

Erdős-Rényi $\mathcal{G}(n,m)$

n sommets, m arêtes. Chaque arête est placée uniformément au hasard entre deux sommets non-reliés. m tirages dépendants.

Quand n tend vers l'infini et $m = p.\binom{n}{2}$, ces deux modèles se confondent.

Attention choisir m fois de suite deux sommets au hasard ne produit pas un graphe mais un multigraphe (arêtes parallèles)

$\mathcal{G}(n,m)$ contre $\mathcal{G}(n,p)$

- $\mathcal{G}(n,m)$ tire avec probabilité **uniforme** sur tous les graphes à n sommets et m arêtes.
- On peut voir le modèle $\mathcal{G}(n,m)$ comme un processus de construction aléatoire en m étapes.
- En partant du graphe vide (sans arête), à chaque étape une nouvelle arête est choisie uniformément parmi les arêtes manquantes.
- Processus incrémental, comme par exemple un algorithme glouton
- Mais dans la suite on parlera de $\mathcal{G}(n,p)$ qui est beaucoup plus facile pour les calculs à cause de l'**indépendance** entre les arêtes
- La probabilité qu'il y ait une arête entre deux sommets ne dépend pas du reste du graphe, dans $\mathcal{G}(n,p)$.

- Plusieurs mathématiciens tels Gilbert en 1957, avaient eu l'idée de de ce modèle assez simple durant leurs travaux. Mais seuls Erdős et Rényi en ont fait un objet d'études systématiques en ayant compris l'importance du modèle.
- Si G est un graphe de $\mathcal{G}(n,p)$ et $m(G)$ sont nombre d'arêtes, l'espérance du nombre d'arêtes est :

$$E(m(G)) = \binom{n}{2}p \simeq \frac{n^2 p}{2}$$

- Dans $\mathcal{G}(n,p)$ un graphe particulier à n sommets, ayant m arêtes a une probabilité d'apparition égale à :

$$P(G) = p^m.(1-p)^{\binom{n}{2}-m}$$

On appelle ce modèle aussi : graphe aléatoire binomial.

On peut calculer plein de choses dans $\mathcal{G}(n,p)$

- La distribution des degrés (donc, le degré moyen)
- La distribution des distances (donc, distances moyennes)
- La cardialité espérée (plus grand k-cœur)
- Le coefficient de clustering
- La taille de la plus grande composante connexe

Tous les sommets jouants le même rôle on voit bien qu'il n'y a pas de communautés ni de navigabilité particulière

Loi binomiale

Problème

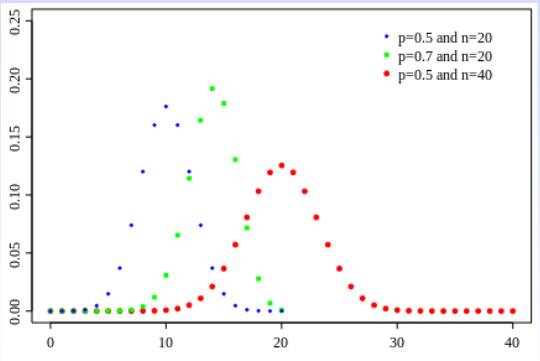
- On considères des évenements **binaires** : l'évenement se produit (avec probabilité p) ou pas (avec probabilité $1-p$).
- Les évenements sont indépendants les uns des autres
- On veut savoir sur n évenements **combien** se sont produit

Loi binomiale $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$

Exemple

Tirer $n = 100$ fois à pile ou face ($p = 1/2$). Proba d'avoir $k = 42$ face ? Réponse : $\frac{100!}{42! \times (100-42)!} \left(\frac{1}{2}\right)^{42} \left(\frac{1}{2}\right)^{100-42} \simeq 2,23\%$. Et $k = 50$ face ? Moins de 8% de chances

Loi binomiale



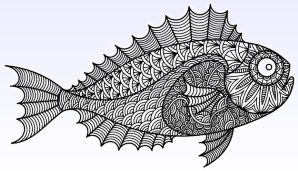
Degrés de $\mathcal{G}(n,p)$

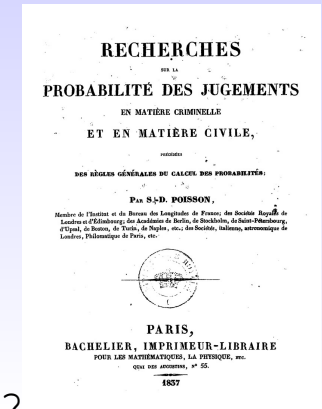
- La distribution des degrés ne suit pas une loi de puissance !
- Pour avoir un degré k, il faut k fois tirer un lien issu de x avec la probabilité p, et $n-1-k$ non liens avec probabilité $1-p$.
- Nb : $n-1$ en tout car pas de boucle
- Donc $P(d(x)=k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$

Théorème

Les **degrés** d'un graphe de $\mathcal{G}(n,p)$ suivent une loi binomiale

Loi de Poisson




RECHERCHES
SUR LA
PROBABILITÉ DES JUGEMENTS
EN MATIÈRE CRIMINELLE
ET EN MATIÈRE CIVILE,
PRÉCÉDÉES
DES RÈGLES GÉNÉRALES DU CALCUL DES PROBABILITÉS;
PAR S.-D. POISSON,
PARIS,
BACHELIER, IMPRIMEUR-LIBRAIRE
POUR LES MATHÉMATIQUES, LA PHYSIQUE, etc.
1837

?

Loi de Poisson

Problèmes modélisés

La loi de Poisson est une loi de probabilité discrète qui décrit le comportement du **nombre d'évènements** quand

- ils se produisent **indépendemment** les uns des autres
- chaque évenement arrivera **uniformément** dans un temps donné (*peu importe combien de temps on l'a déjà attendu*)
- en moyenne on en attend N dans l'intervalle de durée I

On pose $\lambda = N/I$ le taux d'arrivée.

Loi de poisson

$$P(X=k) = \frac{\lambda^k}{k!}.e^{-\lambda}$$

Loi de Poisson

Loi de poisson

$$P(X=k) = \frac{\lambda^k}{k!}.e^{-\lambda}$$

Exemple

Il arrive sur mon routeur un paquet par seconde en moyenne. Je suppose toutes ces arrivées indépendantes. En une minute, quelle est la probabilité que j'aie reçu 42 paquets ? Et 60 paquets ? Réponse : on pose $\lambda = 60/1 = 60$ et la réponse est

$$\frac{60^{42}}{42!}e^{-60} \simeq 0,03\% \quad \text{et} \quad \frac{60^{60}}{60!}e^{-60} \simeq 5,1\%$$

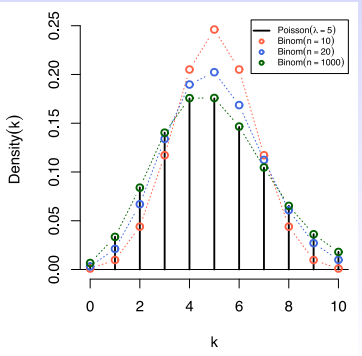
Lien entre ces deux lois

Limite d'une loi binomiale

En posant $\lambda = pn$, quand n tend vers l'infini, la loi binomiale de paramètre p **tend** vers la loi de Poisson de paramètre λ

Idée de la preuve :

$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$
$n - k \approx n$ donc :
$P(X=k) \approx \frac{n^k}{k!} p^k e^{n \log(1-p)}$
$P(X=k) \approx \frac{(pn)^k}{k!} e^{-pn}$
Avec $\lambda = pn$:
$P(X=k) \approx \frac{\lambda^k}{k!} e^{-\lambda}$



Revenons aux degrés de $\mathcal{G}(n,p)$

- Les degrés suivent des lois binomiales qui tendent vers une loi de Poisson...
- La queue est moins lourde qu'une loi de puissance ! En effet $\frac{\lambda^k}{k!}.e^{-\lambda}$, par la formule de Stirling $k! \simeq \sqrt{2\pi k} k^k e^{-k}$ est équivalent à $\left(\frac{\lambda e}{k}\right)^k$ qui tend de façon exponentielle vers 0 : la queue de Poisson est légère !
- Le modèle binomial aléatoire ne modélise donc pas très bien les graphes réels (réseaux sociaux) où on ne rencontre pas de telles distributions de degrés.
- De nombreux autres modèles aléatoires on été proposés récemment, cf suite du cours.
- Pour visualiser $\mathcal{G}(n,p)$ avec la percolation, cliquer ici !

Bornons le degré

- Degré moyen : $\overline{d} = pn = \lambda$.
- Pour $k > 3\lambda$ (avec la loi de Poisson) :
  $P(d(x) >= k) \approx \sum_{j\geq k} \frac{\lambda^j}{j!} e^{-\lambda} < \frac{\lambda^k}{k!} e^{-\lambda} \sum_{j\geq k} \left(\frac{\lambda}{k}\right)^j$
  $= P(d(x) = k) \frac{1}{1-\lambda/k} < \frac{3}{2} P(d(x) = k) < \frac{3}{2} \left(\frac{e}{3}\right)^k$.
- Pour $k = (1+\delta)\lambda$ (avec la loi binomiale) :
  $d(x) = X_1 + \cdots + X_n$ avec $X_i = 1$ si $xi \in E$, $X_i = 0$ sinon.
  Les X_i sont n variables indépendantes et $P(X_i = 1) = p$.
  Leur somme a moyenne $\mu = E(\sum_i X_i) = pn$.
  Borne de Chernoff : $P(\sum_i X_i \geq (1+\delta)\mu) < e^{-\min(\delta,\delta^2)\mu/4}$
  D'où : $P(d(x) \geq (1+\delta)\lambda) < e^{-\min(\delta,\delta^2)\lambda/4}$
  $< \frac{1}{n^\lambda}$ pour $\delta = 4 \ln n$.
  Si $\lambda \geq 2$ (i.e. $p \geq 2/n$), avec forte probabilité, $\mathcal{G}(n,p)$ a degré maximum $O(\lambda \ln n)$.

$\mathcal{G}(n,p)$ a petit diamètre

À p constant, quand n tend vers ∞, le diamètre tend vers 2.

1. En effet la proba que deux sommets ne soient *pas* liés est constante $1-p$
2. Le diamètre est donc > 1 avec forte probabilité
3. Proba que deux sommets x et y n'ont pas **un** voisin u en commun (pour trois sommets x, y, u donnés) : $1-p^2$
4. Proba que deux sommets x et y n'ont **aucun** voisin commun : $(1-p^2)^{(n-2)}$
5. En effet il y a indépendence entre «x et y ont un voisin commun u» et «x et y ont un voisin commun v»
6. La proba que le diamètre soit > 2 est inférieure à
$$\sum_{x\neq y} (1-p^2)^{(n-2)} < n^2 (1-p^2)^{(n-2)}$$
7. qui tend vers 0 (exponentielle $cste^n$ domine polynôme n^{cste})

- Pour savoir si **une propriété donnée est surprenante**, on peut regarder si un graphe aléatoire a la même propriété ou non. Si non, un autre processus est en jeu.
- Exemple : graphe de 10 000 sommets. On observe une clique de 100 sommets et les autres ont degré 0. Pour savoir si cela est étonnant, on peut le comparer à un graphe aléatoire avec le même n et $\overline{d}$ (donc $\overline{d} = 0.99$)
- Nous avons : $P(d(x) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$
- Donc la proba qu'un sommet ait degré 0 est $P(d(x) = 0) = \binom{n-1}{0} p^0 (1-p)^{n-1} = (1-p)^{n-1}$
- en sommant sur tous les sommets on a l'espérence du nombre de sommet isolés :
$$n(1-p)^{n-1} = 10000 \left(1 - \frac{2}{4950}\right)^{9999} \simeq 175 \neq 9900$$
- On devrait avoir 1,75% de sommets isolés au lieu de 99% : on a peu de chance d'obtenir ce graphe par tirage aléatoire à partir du modèle $\mathcal{G}(n,p)$.

3ème exemple typique de **méthode probabiliste** : le couplage parfait

La probabilité que $\mathcal{G}(2n, 1/2)$ admette un couplage parfait est supérieure à $1/3$

Preuve algorithmique

On construit un algorithme glouton qui calcule un couplage parfait.

Données : Une instance $G = (V, E)$ de $\mathcal{G}(2n, 1/2)$;
Résultat : Un couplage parfait M ;
$M \leftarrow \emptyset$; $S \leftarrow \{x_1, \ldots x_{2n}\} = V$;
pour $i \leftarrow 1$ **à** n **faire**
  Choisir un sommet x_i dans S;
  si $\exists y_i \in S$, *tel que* $x_i y_i \in E$ **alors**
    $M \leftarrow M + x_i y_i$;
    $S \leftarrow S\text{-}x_i\text{-}y_i$;
  sinon
    Erreur;

- La conditionelle revient à examiner toutes les arêtes possibles entre x_i et $S \setminus x_i$.
- Soit A_i l'événement qu'il n'existe pas d'arête entre x_i et $S \setminus x_i$.
- $P(A_i) = (1-p)^{|S|} = \left(\frac{1}{2}\right)^{2n-(2i+1)}$ indépendamment de ce qui s'est passé dans les itérations précédentes.
- La probabilité d'une erreur, c'est-à-dire qu'un des événements A_i soit vrai, est :

$$\Sigma_{i=1}^{n} \left(\frac{1}{2}\right)^{2n-(2i+1)} = \frac{1}{2}\Sigma_{j=0}^{j=n-1} \left(\frac{1}{4}\right)^j \leq \frac{1}{2}\Sigma_j \left(\frac{1}{4}\right)^j = \frac{1}{2}\frac{1}{1-1/4} = \frac{2}{3}$$

- Donc le glouton réussit (aucune erreur) avec probabilité $\geq 1/3$

- Nous avons montré qu'un algorithme glouton très simple avait une probabilité $1/3$ de trouver un couplage parfait.
- La probabilité que le $\mathcal{G}(2n, 1/2)$ admette un couplage parfait est peut-être très supérieure à $1/3$
- En fait on peut généraliser ce résultat et montrer que quand n tend vers l'infini, la probabilité tend vers 1.

Lois 0/1 sur ce modèle

Remarqué dès le début par Erdős :

Toute propriété de graphe est soit presque toujours vraie, soit presque toujours fausse après un certain seuil de taille ($n > n_0$).

Exemple

Dès que $m \geq n \ln n$ la probabilité d'être connexe est presque 1.

Logique du premier ordre (FOL)

Formule du premier ordre

Une formule de logique du premier ordre s'exprime à l'aide de variables, des connecteurs logiques classiques, et des quantificateurs $\exists x$, $\forall y$. On ne considère que des formules finies. Par exemple :

- Pas de sommet isolé : $\forall x \exists y [xRy]$
- Le graphe contient un triangle : $\exists x \exists y \exists z\ [xRy \wedge yRz \wedge xRz]$

Considérons l'ensemble $\mathcal{A}$ des propriétés FOL de graphes

- $\mathcal{G}(n,p)$ est un espace de probabilités sur les graphes.
- On peut donc s'intéresser à la probabilité que $\mathcal{G}(n,p)$ possède (ou **satisfait**) une propriété A de graphe, ce que l'on notera :
$$P[\mathcal{G}(n,p) \models A]$$

Lois 0-1

Théorème Glebski 1969, Fagin 1976

$\forall p \in [0,1], \forall A \in \mathcal{A}, \quad lim_{n\to\infty} P[\mathcal{G}(n,p) \models A] =$0 ou 1

Théorème vraiment important mais difficile à démontrer.

Conséquences

- Le terme random graph ou graphe aléatoire a vraiment été mal choisi !
- Comme on peut formaliser dans cette logique l'existence d'un couplage parfait, pour cet événement la probabilité tend vers 1 lorsqu'on choisit $p = 1/2$.

L'illusion des graphes aléatoires à la Erdős-Rényi

« Dès que l'on travaille sur ces graphes expérimentalement on remarque très vite qu'ils se ressemblent beaucoup (ce qui peut être contre-intuitif) : même densité de triangles, même taille de plus grande composante connexe ...

Ils ont tellement de propriétés communes que les appeler "aléatoires" en devient même bizarre. »

Extrait de : Large networks and graph limits, Lásló Lovász, AMS, 2012.

Conséquences

- Cela veut dire qu'à la limite le modèle d'Erdős-Rényi n'est plus vraiment aléatoire.
  A l'infini le graphe est unique, c'est le graphe de Rado !
- Il n'est pas judicieux pour tester un nouvel algorithme de graphes de le faire sur le modèle Erdős-Rényi.
  Car cela revient à tester son algorithme sur une classe restreinte de graphes.
- Par contre comme on connaît beaucoup de chose sur ces graphes, cela peut être intéressant de les comparer à des graphes existants en mesurant quelques paramètres.

Et la connexité ?

- Mais qu'en est-il de la connexité que l'on peut exprimer dans cette logique ?
- Là aussi la probabilité tend vers 1 ou 0, mais on peut en dire un peu plus (cf section suivante).

Lois 0-1 dans les autres modèles aléatoires

- Pas de loi 0-1 dans le cas de l'attachement préférentiel, d'après Kleinberg.
- Pour les modèles géométriques seulement en dimension 2.

Comparaison de $\mathcal{G}(n,p)$ et des GRI réels

	$\mathcal{G}(n,p)$, p petit	GRI
Degré moyen	np, petit	petit
Distribution des degrés	binomiale $\to$ Poisson	Power law
Distance moyenne	log ou constante	log (ou constante ?)
Diamètre	log ou constant	log
Cœur (selon définition)	non	oui
Composante géante	oui	oui
Transitivité (clustering)	non si p petit	oui
Navigabilité	non	oui, parfois
Communautés	non	oui, parfois

Les *petites distances* et la *connexité* des graphes réels ne sont donc **pas surprenantes**. Le reste, davantage.

Conclusion

Un modèle peu réaliste

$\mathcal{G}(n,p)$ et $\mathcal{G}(n,m)$ ne modélisent pas bien les réseaux d'interaction

On peut tout calculer sur les graphes aléatoires $\mathcal{G}(n,p)$

à condition d'être bon en calcul...
Certains en abusent !

Il existe d'autres modèles aléatoires de réseaux

Nous avons déjà vu l'anneau de Watts et Strogatz, nous verrons la semaine prochaine l'attachement préférentiel de Barabási-Albert et la grille de Kleinberg.

Generating Random Regular Graphs J. H. Kim & Van H. Vu [2003]