JOURNÉE TÉLÉCOM-UPS 2015

en prélude de l'assemblée générale de l'UPS du 30 mai 2015

Conférence

Le Numérique Pour Tous

vendredi 29 mai 2015

Stage LIESSE de Télécom ParisTech

http://perso.telecom-paristech.fr/~rioul/liesse.html

Journée organisée par Olivier Rioul, enseignant-chercheur à Télécom ParisTech

Inscription en ligne : www.telecom-paristech.fr/liesse/
Contact : liesse@telecom-paristech.fr

ups Union des Professeurs de Spéciales

Liesse

Télécom ParisTech
46 rue Barrault
75013 Paris
www. telecom-paristech.fr

TELECOM ParisTech

Table des matières

Synopsis

« Le monde d'aujourd'hui est beaucoup plus mathématique qu'hier. Une étude Déloitte a montré que la recherche en mathématiques impacte directement 16% du PIB britannique » rappelait Cédric Villani, le 18 septembre 2014, au premier forum d'IncubAlliance intitulé : Résolution de l'équation « Mathématiques et entrepreneuriat ».

Venez découvrir, à Télécom ParisTech, comment les mathématiques fondent le monde numérique d'aujourd'hui, que ce soit dans les domaines de la cybersécurité, de la performance des communications que de la prédiction de tous types d'événements. Au travers des présentations de nos enseignants-chercheurs, vous identifierez des outils algébriques, des algorithmes déterministes ou probabilistes et des traitements statistiques utiles pour l'informatique, les réseaux et la science des données...

...et venez également à la rencontre des ingénieurs Télécom ParisTech, de vos anciens élèves, qui sont les architectes de ce monde numérique, qu'ils soient inventeurs, entrepreneurs, ou transformateurs des entreprises par le numérique !

Yves Poilane, directeur de Télécom ParisTech

David Madore

Maître de Conférences en Cryptographie au département Informatique et Réseaux de Télécom ParisTech

Application des réseaux euclidiens à la cryptographie

Les réseaux euclidiens (sous-groupes discrets de $\mathbb{R}^n$) ont trouvé des applications en cryptographie (chiffrement, signature électronique, hachage… mais aussi “chiffrement complètement homomorphe”) dont l’intérêt est notamment nourri par l’espoir qu’ils résistent aux attaques par les ordinateurs quantiques. La sécurité de ces schémas cryptographique réside dans la difficulté de problèmes tels que celui du plus court vecteur ou du plus proche vecteur, et la recherche de bases adaptées dans un réseau, ce qui conduit notamment à réexaminer l’algorithme LLL. Nous expliquerons quelques unes des idées dans cette direction.

Réseaux euclidiens et cryptographie

Journées Télécom-UPS
« Le numérique pour tous »

David A. Madore
Télécom ParisTech
david.madore@enst.fr

29 mai 2015

Plan

Réseaux euclidiens : définition

▶ Un **réseau** de $\mathbb{R}^m$ est un sous-groupe (additif) discret L de l'espace euclidien $\mathbb{R}^m$.

Un tel sous-groupe est nécessairement isomorphe à $\mathbb{Z}^n$ (où $n \leq m$) comme groupe abélien : il existe $b_1, \ldots, b_n \in L$ tels que $L = \mathbb{Z}b_1 \oplus \cdots \oplus \mathbb{Z}b_n$.

De plus, $b_1, \ldots, b_n$ sont $\mathbb{R}$-libres (=linéairement indép$^{\text{ts}}$).

On dit qu'ils sont une **base** de L, et que n est le **rang** de L.

Définition équivalente :

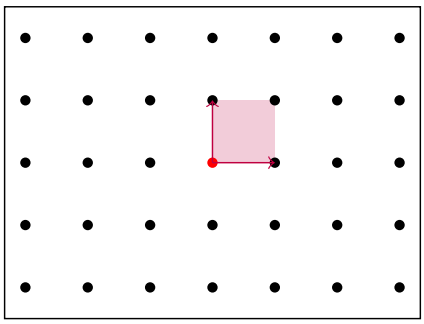
▶ Un **réseau** de $\mathbb{R}^m$ est un $\mathcal{L}(B) := \{uB : u \in \mathbb{Z}^n\}$ où $B \in \mathbb{R}^{n \times m}$ est une matrice de rang n.

(B est la matrice dont les b_i sont les lignes.)

▶ On suppose souvent $m = n$ (réseau de *rang plein*), quitte à se placer dans $\mathrm{Vect}_{\mathbb{R}}(L) = \mathbb{R}b_1 \oplus \cdots \oplus \mathbb{R}b_n$.

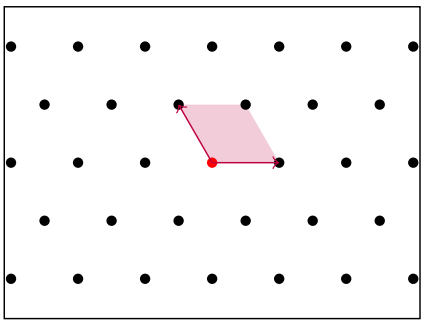
Exemples

Les deux réseaux de rang 2 admettant le plus grand groupe de symétries sont (à similitude près) :



$(A_1)^2$

$\mathbb{Z}^2 \subseteq \mathbb{R}^2$



A_2

$\{(x,y,z) \in \mathbb{Z}^3 : x+y+z=0\} \subseteq \mathbb{R}^3$

Bases et parallélotopes fondamentaux

Soit $\mathcal{L}(B) = \{uB : u \in \mathbb{Z}^n\} \subseteq \mathbb{R}^m$ (où $\operatorname{rg} B = n$).

- $\mathcal{P}(B) := \{uB : u \in [0;1[^n\}$ s'appelle **parallélotope fondamental** associé à la base B.
- On a $\mathcal{L}(B) = \mathcal{L}(B')$ ssi $B' = UB$ où $U \in GL_n(\mathbb{Z})$.

Ici, $GL_n(\mathbb{Z})$ est l'ensemble des matrices $n \times n$ à coefficients entiers, de déterminant ± 1 (**unimodulaires**).

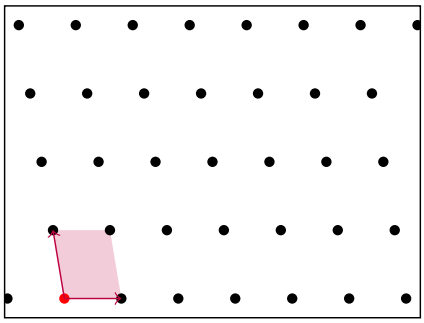
Dès que $n > 1$, un réseau admet une infinité de bases.

On peut voir l'ensemble des réseaux de rang plein dans $\mathbb{R}^n$ comme l'ensemble quotient $GL_n(\mathbb{Z})\backslash GL_n(\mathbb{R})$.

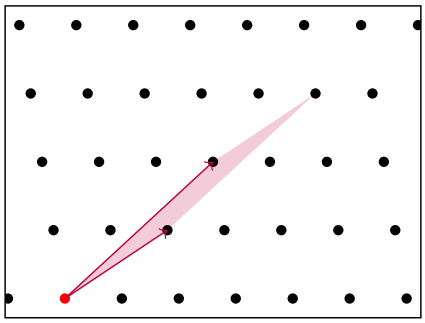
- $\operatorname{vol}(\mathcal{P}(B)) =: \operatorname{covol}(\mathcal{L}(B)) = |\det(B)|$ (lorsque $m = n$) : volume du parallélogramme fondamental : **(co)volume** ou **déterminant** du réseau. Ne dépend pas de B !

Toutes les bases ne se valent pas

Certaines bases sont plus « agréables » que d'autres :



« Bonne » base



Moins « bonne » base

Les deux parallélogrammes fondamentaux dessinés ont la même aire, mais pas la même forme / la même longueur des côtés.

« Bonne » ≈ constituée de petits vecteurs.

Thèmes : Comment construire de « bonnes » bases à partir de « mauvaises » ? (Par des opérations élémentaires entières sur les lignes de B.) Comment exploiter la difficulté de ce problème ?

Réseaux euclidiens et cryptographie

David Madore

Plan

Généralités sur les réseaux euclidiens

L'algorithme LLL

Réseaux et cryptographie

Similitudes

Soit $\mathcal{L}(B) = \{uB : u \in \mathbb{Z}^n\} \subseteq \mathbb{R}^m$ (où $\operatorname{rg} B = n$).

- Si $t \in \mathbb{R}^\times$, on a $t \cdot \mathcal{L}(B) = \mathcal{L}(tB)$ (*homothétie*).
Multiplie le covolume par t^n.
- Si $\Omega \in O_m$, on a $\mathcal{L}(B) \cdot \Omega = \mathcal{L}(B\Omega)$ (*isométrie*).
Ne change pas le covolume.
Si $\mathcal{L}(B) \cdot \Omega = \mathcal{L}(B)$, on dit que Ω est une **symétrie** de $\mathcal{L}(B)$.

On identifie souvent deux réseaux homothétiques, isométriques, ou les deux (semblables).
Ceci permet de **normaliser** $\operatorname{covol}(L) = 1$.

On peut considérer $SL_n^{\pm}(\mathbb{R})/O_n$ comme l'espace des *formes de parallélotopes* de dimension n [espace riemannien symétrique], et $GL_n(\mathbb{Z})\backslash SL_n^{\pm}(\mathbb{R})/O_n$ comme l'espace des *formes de réseaux* de rang plein.

- **Matrice de Gram :** $G := BB^{\mathrm{tr}}$ soit $G_{ij} = b_i \cdot b_j$, invariante par isométrie ($B\Omega(B\Omega)^{\mathrm{tr}} = BB^{\mathrm{tr}}$).

Matrice de Gram

Soit $\mathcal{L}(B) = \{uB : u \in \mathbb{Z}^n\} \subseteq \mathbb{R}^m$ (où $\operatorname{rg} B = n$).

Matrice de Gram : $G := BB^{\mathrm{tr}}$ soit $G_{ij} = b_i \cdot b_j$.

- Invariante par isométrie ($B\Omega(B\Omega)^{\mathrm{tr}} = BB^{\mathrm{tr}}$ si $\Omega \in O_m$).
- Est la matrice de la forme quadratique sur $\mathbb{Z}^n$ définie par $q(u) = \|uB\|^2$ (norme euclidienne transportée au réseau), donc définie positive. (⇒ Lien avec les f.q. sur les entiers.)
- Vérifie $\det(G) = \operatorname{covol}(L)^2$ (**discriminant** de $L = \mathcal{L}(B)$).
En effet, $\det(G) = \det(B)^2$ est évident si $m = n$.
- Réciproquement, si G est définie positive, on peut écrire $G = BB^{\mathrm{tr}}$ pour $B \in GL_n(\mathbb{R})$ (conséquence de Cholesky ou du théorème spectral), et B est unique à isométrie près.

L'espace $SL_n^{\pm}(\mathbb{R})/O_n$ s'identifie donc à l'ensemble des matrices définies positives de déterminant 1, et $GL_n(\mathbb{Z})\backslash SL_n^{\pm}(\mathbb{R})/O_n$ à l'ensemble des formes quadratiques définies positives sur un $\mathbb{Z}$-module de rang n.

Orthogonalisation de Gram-Schmidt

- Si $b_1, \ldots, b_n \in \mathbb{R}^m$ sont $\mathbb{R}$-libres, on définit par récurrence $b_i^\star := b_i - \sum_{j<i} \mu_{i,j}\, b_j^\star$ où $\mu_{i,j} := (b_i \cdot b_j^\star)/\|b_j^\star\|^2$ (i.e., $b_i^\star = \mathrm{proj}_{\mathrm{Vect}(b_j : j<i)^\perp}(b_i)$).

Les $(b_i^\star)_{i \le s}$ sont donc une base orthogonale de $\mathrm{Vect}(b_i^\star : i \le s) = \mathrm{Vect}(b_i : i \le s)$.

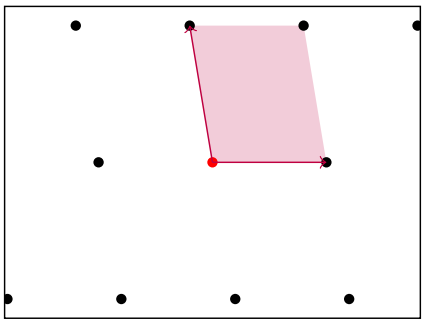
- Formulation matricielle (pour $m = n$) : $B = MDV$ avec M triangulaire inférieure de diagonale 1 (soit : $M_{ij} = \mu_{i,j}$ si $j < i$, 1 si $j = i$, et 0 si $j > i$), D diagonale de diagonale $\|b_i^\star\|$, et V orthogonale.

En particulier, $|\det(B)| = \det D = \prod_{i=1}^n \|b_i^\star\|$.

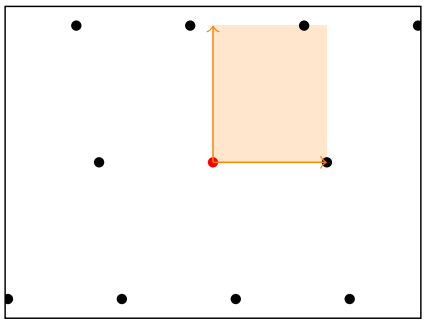
- Dépend de l'ordre : si on permute $b_i \leftrightarrow b_{i+1}$, alors $(b_i^\star, b_{i+1}^\star)$ devient $(b_{i+1}^\star + \mu_{i+1,i}\, b_i^\star,\ \frac{\|b_{i+1}^\star\|^2\, b_i^\star - \mu_{i+1,i}\, \|b_i^\star\|^2\, b_{i+1}^\star}{\|b_{i+1}^\star\|^2 + \mu_{i+1,i}^2\, \|b_i^\star\|^2})$.

Gram-Schmidt (suite)

Calcul de l'aire d'un parallélogramme :



(b_1, b_2)



$(b_1^\star, b_2^\star)$

La matrice (DV) des $b_i^\star$ définit un parallélotope rectangle ayant le même volume $\mathrm{covol}(L)$ que celui défini par les b_i.

Les $b_i^\star$ n'appartiennent pas à L en général.

Minima successifs d'un réseau

Soit L un réseau euclidien de rang n dans $\mathbb{R}^m$. On définit, pour $1 \leq i \leq n$:

$$\lambda_i(L) = \min\{r \in \mathbb{R}_+ : \dim \mathrm{Vect}(L \cap \mathrm{B_f}(0,r)) \geq i\}$$

où $\mathrm{B_f}(0,r) = \{x \in \mathbb{R}^m : \|x\| \leq r\}$.

Autrement dit, $\lambda_i(L)$ est le plus petit r tel qu'on puisse trouver i vecteurs $\mathbb{R}$-libres tous de norme $\leq r$ dans L.

Attention : $L \cap \mathrm{B_f}(0,\lambda_n)$ ne contient pas forcément une $\mathbb{Z}$-base de L.

En particulier, $\lambda_1(L) = \min\{\|x\| : x \in L \setminus \{0\}\}$ est la norme du plus petit vecteur non nul de L.

Exercice : Montrer que $\lambda_1(L) \geq \min\{\|b_i^\star\| : 1 \leq i \leq n\}$.
Indication : $\|uMDV\| = \|uMD\|$ avec MDV comme dans G-S.

Question : Peut-on borner $\lambda_1(L)\ \mathrm{covol}(L)^{-1/n}$?

Empilements de sphères

Ici, L est de rang plein.

Soit $\rho(L) := \frac{1}{2}\lambda_1(L)$. Il s'agit du plus grand rayon ρ tel que les boules ouvertes de rayon ρ centrées sur les points de L soient deux à deux disjointes.

La **densité** = fraction du volume occupé par les boules vaut alors $\mathscr{V}_n\ \rho(L)^n / \mathrm{covol}(L)$ où[†] $\mathscr{V}_n := \frac{\pi^{n/2}}{(n/2)!}$ est le volume de la n-boule unité.

Il est souvent plus commode de travailler avec $\rho(L)^n / \mathrm{covol}(L)$, ou encore $\lambda_1(L)\ \mathrm{covol}(L)^{-1/n}$.

Question : Quelles valeurs ces nombres peuvent-ils prendre ? (Quel réseau empile le mieux les boules en dimension n ?) Réponse connue pour $n \leq 8$ et $n = 24$.

Constante de Hermite :
$\gamma_n := \sup\{\lambda_1(L)^2 : L \text{ t.q. } \mathrm{covol}(L) = 1\}$ (atteint ; on a alors $\gamma_1 = 1$, $\gamma_2 = \frac{2}{3}\sqrt{3}$, $\gamma_3 = \sqrt[3]{2}$, $\gamma_8 = 2$, $\gamma_{24} = 4$).

[†]Où $(k+\frac{1}{2})! := \frac{(2k+1)!!}{2^{k+1}}\sqrt{\pi}$ (et $(2k+1)!! = \prod \text{impairs}$).

Réseaux euclidiens et cryptographie
David Madore
Plan
Généralités sur les réseaux euclidiens
L'algorithme LLL
Réseaux et cryptographie

Une borne de Minkowski

Explicitons le fait que la densité d'un empilement est ≤ 1.

Théorème (Blichfeld) : Si $L \subseteq \mathbb{R}^n$ de rg. pl., et $S \subseteq \mathbb{R}^n$ t.q. $\mathrm{vol}(S) > \mathrm{covol}(L)$, alors $\exists z_1 \neq z_2 \in S$ t.q. $z_1 - z_2 \in L$.

Preuve : sinon, les $S_z := (S + z) \cap \mathcal{P}$ sont disjoints (pour $z \in L$). Or $\sum_z \mathrm{vol}(S_z) = \sum \mathrm{vol}(S_z - z) = \mathrm{vol}\, S > \mathrm{vol}\, \mathcal{P}$, contradiction.

Théorème (Minkowski) : Si $L \subseteq \mathbb{R}^n$ de rg. pl., et S convexe sym$^{\text{que}}$ t.q. $\mathrm{vol}(S) > 2^n \,\mathrm{covol}(L)$, alors $S \cap (L \setminus \{0\}) \neq \varnothing$.

Preuve : $\mathrm{vol}(\frac{1}{2}S) = 2^{-n}\,\mathrm{vol}(S) > \mathrm{covol}(L)$ donc il existe $z_1 \neq z_2 \in \frac{1}{2}S$ t.q., $z_1 - z_2 \in L$, or $z_1 - z_2 = \frac{1}{2}(2z_1 - 2z_2) \in S$.

Corollaire : $\lambda_1(L) \leq \sqrt{n}\ \mathrm{covol}(L)^{1/n}$ (c'est-à-dire, $\gamma_n \leq n$).

Preuve : Appliquer le théorème à la boule ouverte de centre 0 et rayon λ_1, et utiliser la minoration $\mathscr{V}_n \geq (2/\sqrt{n})^n$ (car la boule unité contient un cube de côté $2/\sqrt{n}$).

Amélioration : $(\prod_{i=1}^n \lambda_i(L))^{1/n} \leq \sqrt{n}\ \mathrm{covol}(L)^{1/n}$.

Idée : Remplacer la boule par l'ellipsoïde de demi-axes $\lambda_1, \ldots, \lambda_n$ orientés selon le Gram-Schmidt des minima successifs.

←13/31→

Le réseau dual

Si $L \subseteq \mathbb{R}^n$ est un réseau de rang plein, son **dual** est

$$L^* := \{y \in \mathbb{R}^n : \forall x \in L,\ x \cdot y \in \mathbb{Z}\}$$

où $x \cdot y$ est le produit scalaire (euclidien).

Matriciellement, si les vecteurs sont vus comme des vecteurs-lignes :

$$\begin{aligned} L^* &= \{y \in \mathbb{R}^n : \forall x \in L,\ xy^{\mathrm{tr}} \in \mathbb{Z}\} \\ &= \{y \in \mathbb{R}^n : \forall u \in \mathbb{Z}^n,\ uBy^{\mathrm{tr}} \in \mathbb{Z}\} \\ &= \{y \in \mathbb{R}^n : yB^{\mathrm{tr}} \in \mathbb{Z}^n\} = \mathcal{L}(B^{-\mathrm{tr}}) \end{aligned}$$

C'est donc aussi un réseau, et $(L^*)^* = L$. Covolume : $\mathrm{covol}(L^*) = \mathrm{covol}(L)^{-1}$. Homothéties : $(t \cdot L)^* = \frac{1}{t} \cdot L^*$.

Inverse la matrice de Gram. Cas de rang non plein : on peut définir $\mathcal{L}(B)^* = \mathcal{L}((G^{-1}B)^{\mathrm{tr}}) \subseteq \mathrm{Vect}_{\mathbb{R}}(\mathcal{L}(B))$.

Symétrie sur l'espace riemannien symétrique $SL_n^{\pm}(\mathbb{R})/O_n$.

←14/31→

Réseaux euclidiens et cryptographie

David Madore

Plan

Généralités sur les réseaux euclidiens

L'algorithme LLL

Réseaux et cryptographie

Réseaux entiers

▶ Si $L \subseteq L^*$, i.e., si la matrice de Gram G est à coefficients entiers, on dit que L est **entier**.

Notamment, dans ce cas, le discriminant $\det G = \mathrm{covol}(L)^2$ est entier.

$\Rightarrow$ Lien avec les formes quadratiques entières ($q(u) = \|uB\|^2 = uGu^{\mathrm{tr}}$).

▶ On a $L = L^*$ ssi L est entier et $\mathrm{covol}(L) = 1$ (i.e., $G \in GL_n(\mathbb{Z})$). On dit alors que L est **unimodulaire**.

Si de plus $\|x\|^2 \in 2\mathbb{Z}$ pour tout $x \in L$ (i.e., q prend des valeurs paires), on dit que L est **pair** (=de type II), sinon **impair** (=de type I).

Le plus petit rang d'un réseau unimodulaire pair est 8, et ce réseau est unique à isométrie près (c'est E_8).

Quelques réseaux remarquables

▶ $\mathbb{Z}^n$ réseau entier de covolume 1, avec $\lambda_1 = \cdots = \lambda_n = 1$.

▶ $A_n := \{(x_0, \dots, x_n) \in \mathbb{Z}^{n+1} : \sum_{i=0}^n x_i = 0\}$ réseau entier de covolume $\sqrt{n+1}$, avec $\lambda_1 = \cdots = \lambda_n = \sqrt{2}$.
Note : A_1 est isométrique à $\sqrt{2}\mathbb{Z}$, et A_2 est le réseau hexagonal, A_3 le « cubique faces centrées ».

▶ $A_n^* = A_n + \mathbb{Z}(-\frac{n}{n+1}, \frac{1}{n+1}, \frac{1}{n+1}, \dots, \frac{1}{n+1})$ ici $\lambda_1 = \sqrt{\frac{n}{n+1}}$.
Note : A_1^* est isométrique à $\frac{1}{\sqrt{2}}\mathbb{Z}$ et A_2^* à $\frac{1}{\sqrt{3}}A_2$, et A_3^* est le « cubique centré ».

▶ $D_n := \{(x_1, \dots, x_n) \in \mathbb{Z}^n : \sum_{i=1}^n x_i \in 2\mathbb{Z}\}$ réseau entier de covolume 2, avec $\lambda_1 = \cdots = \lambda_n = \sqrt{2}$.
Note : D_2 est isométrique à $\sqrt{2}\mathbb{Z}^2$, et D_3 est isométrique à A_3.

▶ $D_n^* = \mathbb{Z}^n \cup (\mathbb{Z} + \frac{1}{2})^n$, avec $\lambda_1 = 1$ si $n \geq 4$.
Note : D_4^* est isométrique à $\frac{1}{\sqrt{2}}D_4$.

▶ $E_8 := \{(x_1, \dots, x_8) \in (\mathbb{Z}^8 \cup (\mathbb{Z} + \frac{1}{2})^8) : \sum_{i=1}^8 x_i \in 2\mathbb{Z}\}$ réseau entier de covolume 1, avec $\lambda_1 = \cdots = \lambda_8 = \sqrt{2}$.

Quelques problèmes algorithmiques

Algorithmiquement, on considère généralement des réseaux $L \subseteq \mathbb{Z}^n$ (ou en tout cas $L \subseteq \mathbb{Q}^n$). Parfois $N\mathbb{Z}^n \subseteq L \subseteq \mathbb{Z}^n$ (« N-modulaires »).

- **Problème SVP$_h$** (« Shortest Vector Problem ») : pour $h \geq 1$, donnée une base B de $L = \mathcal{L}(B)$, trouver $z \in L$ tel que $0 \neq \|z\| \leq h \cdot \lambda_1(L)$.

SVP$_h$ est NP-dur pour $h \lesssim \sqrt{n}$, polynomial (P) par LLL pour $h = 2^{n/2}$. SVP = SVP$_1$ est résoluble en complexité $2^{O(n)}$.

- **Problème CVP$_h$** (« Closest Vector Problem ») : pour $h \geq 1$, donnée une base B de $L = \mathcal{L}(B)$ et $t \in \mathbb{R}^n$, trouver $z \in L$ tel que $\|t - z\| \leq h \cdot \mathrm{dist}(t, L)$.

CVP$_h$ est au moins aussi dur que SVP$_h$, et polynomial (P) pour $h = 2^{n/2}$ par LLL+Babai.

Bases LLL-réduites

Gram-Schmidt : $b_i^\star := b_i - \sum_{j<i} \mu_{i,j}\, b_j^\star$ où $\mu_{i,j} := (b_i \cdot b_j^\star)/\|b_j^\star\|^2$.

La base $b_1, \ldots, b_n$ est dite LLL-δ-réduite ($\frac{1}{4} < \delta < 1$) si :

- pour tous $i > j$, on a $|\mu_{i,j}| \leq \frac{1}{2}$, et
- pour tout $i < n$, on a $\|b_{i+1}^\star + \mu_{i+1,i}\, b_i^\star\|^2 \geq \delta \cdot \|b_i^\star\|^2$.

Intuitivement, la première condition assure que les b_i ne sont pas trop loin d'être orthogonaux, et la seconde, qu'on ne gagne pas trop à échanger $b_i \leftrightarrow b_{i+1}$ avant d'appliquer G-S.

Notion de « bonne » base : on va voir que tout réseau a une base LLL-réduite, calculable en temps polynomial.

On déduit $\|b_{i+1}^\star\|^2 \geq (\delta - \mu_{i+1,i}^2)\|b_i^\star\|^2 \geq (\delta - \frac{1}{4})\|b_i^\star\|^2$.
Donc $\|b_i^\star\| \geq (\delta - \frac{1}{4})^{(i-1)/2}\|b_1\|$.
Comme $\lambda_1 \geq \min \|b_i^\star\|$, on a $\|b_1\| \leq (\delta - \frac{1}{4})^{-(n-1)/2}\lambda_1$.

Opérations élémentaires

- **Réduction** de la ligne b_i par b_j $(j < i)$: remplacer b_i par $b_i - cb_j$ (soit $B \leftarrow (1_n - cE_{ij})B$) où $c = \lceil \mu_{i,j} \rfloor$ (arrondi†).

Effet sur G-S : $\mu_{i,k} \leftarrow \mu_{i,k} - c\mu_{j,k}$, donc $\mu_{i,j} \leftarrow |\cdot| \leq \frac{1}{2}$.
Les $b_i^\star$ ne changent pas.

- **Réduction de taille** de la base :

pour i allant de 2 à n,
  pour j allant de $i-1$ à 1 (décroissant),
    réduire b_i par b_j (soit $b_i \leftarrow b_i - \lceil \mu_{i,j} \rfloor\, b_j$).

Assure la propriété $|\mu_{i,j}| \leq \frac{1}{2}$.

- **Échange** $b_i \leftrightarrow b_{i+1}$ [et recalculer / m.à.j. G-S !]

L'échange servira à assurer la propriété de Lovász
$\|b_{i+1}^\star + \mu_{i+1,i}\, b_i^\star\|^2 \geq \delta \cdot \|b_i^\star\|^2$.
Il faut refaire une réduction de taille après chaque échange !

†Soit $\lceil \xi \rfloor := \lfloor (\xi + \frac{1}{2}) \rfloor$ où $\lfloor \cdot \rfloor$ = partie entière.

L'algorithme LLL

Soit $\frac{1}{4} < \delta < 1$ (typiquement $\delta = \frac{3}{4}$ ou mieux $\frac{1}{4} + (\frac{3}{4})^{n/(n-1)}$).

Algorithme de Lenstra-Lenstra-Lovász donnés $b_1, \ldots, b_n$ base d'un réseau L de $\mathbb{R}^m$, calcule une base LLL-δ-réduite.

- (1) Calculer (ou m.à.j.) Gram-Schmidt.
- (2) Réduction de taille de la base :

pour i allant de 2 à n,
  pour j allant de $i-1$ à 1 (décroissant),
    réduire b_i par b_j (soit $b_i \leftarrow b_i - \lceil \mu_{i,j} \rfloor\, b_j$)
      (et $\mu_{i,k} \leftarrow \mu_{i,k} - \lceil \mu_{i,j} \rfloor\, \mu_{j,k}$).

- (3) S'il existe i tel que $\|b_{i+1}^\star + \mu_{i+1,i}\, b_i^\star\|^2 < \delta \cdot \|b_i^\star\|^2$: échanger $b_i \leftrightarrow b_{i+1}$, et retourner en (1).

Théorème : LLL termine en temps polynomial.

Idée : $\prod_{i=1}^n \|b_i^\star\|^{2(n-i+1)} = \prod_{i=1}^n \mathrm{covol}(\mathcal{L}(b_1, \ldots, b_i))^2$ décroît d'un facteur δ pour chaque échange.

Note : pour $n = 2$, LLL $\cong$ algo. de Lagrange|Gauß $\approx$ « Euclide centré ».

Réseaux euclidiens et cryptographie
David Madore
Plan
Généralités sur les réseaux euclidiens
L'algorithme LLL
Réseaux et cryptographie

Propriétés des bases LLL-réduites

Soit $\alpha = \frac{1}{\delta - \frac{1}{4}}$ et $b_1, \ldots, b_n$ une base LLL-δ-réduite.

On a vu $\|b_1\| \leq \alpha^{(n-1)/2} \lambda_1$, donc pour $\delta = \frac{3}{4}$ on a $\|b_1\| \leq 2^{(n-1)/2} \lambda_1$ et LLL résout SVP$|_h$ pour $h = 2^{(n-1)/2}$ (renvoyer b_1) en temps poly.

Plus généralement, on a :

- $\|b_i\| \leq \alpha^{(n-1)/2} \lambda_i$
- $\|b_1\| \leq \alpha^{(n-1)/4} \operatorname{covol}(L)^{1/n}$
- $\prod_{i=1}^{n} \|b_i\| \leq \alpha^{n(n-1)/4} \operatorname{covol}(L)$

Expérimentalement, sur des réseaux et bases aléatoires, on observe des inégalités meilleures (mais toujours exponentielles), par exemple $\|b_1\| \leq 1.022^n \operatorname{covol}(L)^{1/n}$.

Algorithme de Babai

Soit $L = \mathcal{L}(B)$ un réseau et $t \in \mathbb{R}^n$. On veut résoudre le problème CVP$_h$ avec $h = 2^{n/2}$, i.e., trouver $z \in L$ tel que $\|z - t\| \leq 2^{n/2} \operatorname{dist}(t, L)$.

- Appliquer LLL avec $\delta = \frac{3}{4}$ à B.
- Faire $x \leftarrow t$, puis
  pour j allant de n à 1 (décroissant),
  remplacer $x \leftarrow x - cb_j$
  où $c = \lceil (b \cdot b_j^\star)/\|b_j^\star\|^2 \rfloor$.
- Retourner $z = t - x$.

De façon équivalente : on choisit d'abord $c \in \mathbb{Z}$ tel que l'hyperplan affine $cb_n^\star + \operatorname{Vect}(b_1, \ldots, b_{n-1})$ soit aussi proche que possible de t, puis on applique récursivement pour trouver un élément proche de x dans $cb_n + \mathcal{L}(b_1, \ldots, b_{n-1})$ (i.e., proche de $x - cb_n$ dans $\mathcal{L}(b_1, \ldots, b_{n-1})$).

Approximation diophantienne simultanée

- Soient $(\xi_1, \ldots, \xi_r) \in \mathbb{R}$ irrationnels. On cherche à approcher les ξ_i par des rationnels p_i/q *de même dénominateur*, i.e., trouver $(p_1, \ldots, p_r) \in \mathbb{Z}^r$ et $q \in \mathbb{N}_{>0}$ tels que les $|q\xi_i - p_i|$ soient petits et q pas trop grand. Qualité prédite par :
- **Dirichlet :** Il existe des q arbitrairement grands tels que $|q\xi_i - p_i| \leq q^{-1/r}$ où $p_i = \lceil q\xi_i \rfloor$.

Preuve : Découper $(\mathbb{R}/\mathbb{Z})^r$ en N^r cubes de côté $1/N$, et considérer les $N^r + 1$ classes des points $q\vec{\xi}$ pour $0 \leq q \leq N^r$: il existe $0 \leq q_1 < q_2 \leq N^r$ tels que les classes tombent dans la même boîte, et si $q = q_2 - q_1$ alors on a $|q\xi_i - p_i| \leq \frac{1}{N} \leq q^{-1/r}$.

- **Réseau :** pour $N > 0$ réel, considérer l'image de $\mathbb{Z}^{r+1} \to \mathbb{R}^{r+1}$ envoyant $(p_1, \ldots, p_r, q)$ sur $(N(q\xi_1 - p_1), \ldots, N(q\xi_r - p_r), q/N^r)$. On vient de voir que ce réseau a des petits vecteurs non nuls.
- LLL donne $|q\xi_i - p_i| \leq 2^{r/2}/N$ avec $q \leq 2^{r/2}N^r$.

Le problème du sac à dos

Problème : Donnés $a_1, \ldots, a_r, s$ entiers > 0, on cherche un sous-ensemble P de $\{1, \ldots, r\}$ tel que $\sum_{i \in P} a_i = s$ (supposé exister).

Approche par LLL : soit B une constante bien choisie ($\lceil \sqrt{n2^n} \rceil$). considérer l'image de $\mathbb{Z}^{r+1} \to \mathbb{R}^{r+1}$ envoyant $(u_1, \ldots, u_r, v)$ sur $(u_1, \ldots, u_r, B \cdot (vs - \sum u_i a_i))$.

Avec les bonnes conditions sur les a_i (uniformément choisis sur un intervalle assez grand) et s (supérieur à $\frac{1}{2}\sum a_i$, ce qu'on peut toujours supposer), on montre qu'avec une probabilité extrêmement élevée, le plus court vecteur trouvé par LLL résout le problème du sac à dos.

Réseaux en cryptographie : principes

Utilisation pour le chiffrement à clé publique :

- La clé secrète sera typiquement une « bonne » base d'un réseau L (ou de son dual).

- La clé publique sera typiquement une « mauvaise » base du même réseau L.

Il est facile de générer la mauvaise base à partir de la bonne, difficile de faire l'opération inverse.

- Le chiffrement consiste à fabriquer un problème difficile à partir d'une mauvaise base, que la connaissance d'une bonne base permet de résoudre.

Par exemple : pour chiffrer, écrire le message sous forme d'un petit vecteur e, choisir z aléatoirement dans L, et renvoyer $x = z + e$. Déchiffrer demande de retrouver $z \in L$ proche de x.

Réseaux en cryptographie : espoirs et limitations

Espoirs de la cryptographie basée sur les réseaux :

- Résistance aux ordinateurs quantiques.

Contrairement aux problèmes de théorie des nombres (factorisation, pb. du log discret) utilisés comme source de difficulté en cryptographie à clé publique traditionnelle, et qui sont cassés par les ordinateurs quantiques[†], les problèmes de réseaux *paraissent* aussi difficiles pour les ordinateurs quantiques que pour les ordinateurs classiques.

- Outils plus puissants, p.ex., chiffrement complètement homomorphe (⇒calculs sur les chiffrés).

Limitations :

- Taille de clés/chiffrés beaucoup plus grande.
- Encore mal compris : pas de paramètres de sécurité standardisés.

[†]Si un jour ils existent vraiment...

Réseaux N-modulaires

Notation : $\mathbb{Z}_{/N} := \mathbb{Z}/N\mathbb{Z}$

Un réseau L tel que $N\mathbb{Z}^m \subseteq L \subseteq \mathbb{Z}^m$ est dit N-**modulaire**. Équivalent à la donnée d'un sous-groupe $L/N\mathbb{Z}^m \subseteq \mathbb{Z}_{/N}^m$ (si $N{=}q$ premier, d'un sous-$\mathbb{F}_q$-esp. vect. de $\mathbb{F}_q^m$).

Attention : Le rang du réseau ici est m, même si $L/N\mathbb{Z}^m$ est très petit.

Si $A \in (\mathbb{Z}_{/N})^{n\times m}$ (typiquement, $n \leq m \approx n \log n$), soient :

$$\Lambda(A) := \mathcal{L}(A) + N\mathbb{Z}^m = \{x \in \mathbb{Z}^m : \exists u \in \mathbb{Z}^n,\ x \equiv uA\ [N]\}$$
$$\Lambda^\perp(A) := \{v \in \mathbb{Z}^m : Av^{\mathrm{tr}} \equiv 0\ [N]\}$$

les réseaux N-modulaires (de rang m) engendré par les lignes de A, resp. orthogonal aux lignes de A.

- On a $\Lambda^\perp(A) = N \cdot \Lambda(A)^*$ et $\Lambda(A) = N \cdot \Lambda^\perp(A)^*$.
- Si $N{=}q$ premier, et A de rang n, on a $\Lambda^\perp(A) = \Lambda(B)$ où $B \in \mathbb{Z}_{/q}^{(m-n)\times m}$ de rang $m-n$ (lignes de B base du suppl. ortho. des lignes de A, soit $BA^{\mathrm{tr}} = 0$).
- Avec haute probabilité, $\mathrm{covol}(\Lambda(A)) = q^{m-n}$ et $\mathrm{covol}(\Lambda^\perp(A)) = q^n$.

« Learning With Errors » (LWE)

Soit q premier. Typiquement, $10^3 < q < 10^5$ ici, $10^2 < n < 10^3$ et $10^3 < m < 10^4$.

Soit $A \in \mathbb{Z}_{/q}^{n\times m}$ tiré au hasard uniformément. Le vecteur $x \in \mathbb{Z}_{/q}^m$ est défini par l'un des deux procédés suivants :

- tiré au hasard uniformément dans $\mathbb{Z}_{/q}^m$, ou bien
- calculé par $x = uA + e$ où $u \in \mathbb{Z}_{/q}^n$ est tiré au hasard uniformément, et $e \in \mathbb{Z}_{/q}^m$ selon une distribution gaussienne (arrondie aux entiers et réduite mod q).

Défi : distinguer ces deux cas avec probabilité $> \frac{1}{2} + \varepsilon$.

Si l'écart-type est assez petit, application du CVP à x pour le réseau $\Lambda(A)$. Correction de l'« erreur » e.

Théorème (informel[t]) : pour un écart-type assez élevé dans la gaussienne ($> \sqrt{\frac{2\pi}{n}}$), LWE est au moins aussi difficile que certains problèmes difficiles « standards » sur les réseaux.

Un chiffrement basé sur LWE (Regev / GPV)

- Paramètre : $A \in \mathbb{Z}_{/q}^{n \times m}$ tiré au hasard uniformément. Clé secrète : $s \in \mathbb{Z}_{/q}^{m}$ selon une distribution gaussienne (« petit vecteur » secret). Clé publique : $p := As^{\mathrm{tr}} \in \mathbb{Z}_{/q}^{n}$.
- Chiffrement d'un bit $b \in \{0,1\}$: tirer $u \in \mathbb{Z}_{/q}^{n}$ uniformément et $(e, e_0) \in \mathbb{Z}_{/q}^{m+1}$ selon une distribution gaussienne (« erreur »). Renvoyer $x = uA + e \in \mathbb{Z}_{/q}^{m}$ ainsi que $c = b\lfloor \frac{q}{2} \rfloor + u \cdot p + e_0 \in \mathbb{Z}_{/q}$.
- Déchiffrement : recevant $x \in \mathbb{Z}_{/q}^{m}$ et $c \in \mathbb{Z}_{/q}$, calculer $c - x \cdot s^{\mathrm{tr}}$, qui vaut $b\lfloor \frac{q}{2} \rfloor + e_0 - e \cdot s^{\mathrm{tr}}$: si ce nombre est plus proche de $\frac{q}{2}$, décoder 1, sinon, décoder 0. Validité : $e_0 - e \cdot s^{\mathrm{tr}}$ a une probabilité négligeable d'être $\gtrsim \frac{q}{2}$.

Le paramétrage de m, n, q et les écarts-types des gaussiennes doit être fait pour rendre le chiffrement difficile à casser et la probabilité d'erreur au décodage négligeable.

Un chiffrement basé sur LWE : explications

- Paramètre : $A \in \mathbb{Z}_{/q}^{n \times m}$. Clé secrète : $s \in \mathbb{Z}_{/q}^{m}$ (« petit vecteur »). Clé publique : $p := As^{\mathrm{tr}} \in \mathbb{Z}_{/q}^{n}$.

La clé publique est plutôt $(A|p) \in \mathbb{Z}_{/q}^{n \times (m+1)}$. Soit $L := \Lambda(A|p)$ le réseau engendré par ses lignes.

- Chiffrement : $x = uA + e \in \mathbb{Z}_{/q}^{m}$ et $c = b\lfloor \frac{q}{2} \rfloor + u \cdot p + e_0 \in \mathbb{Z}_{/q}$ où $u \in \mathbb{Z}_{/q}^{n}$ uniforme et $(e, e_0) \in \mathbb{Z}_{/q}^{m+1}$ « erreur ».

On a donc $(x|p) = u(A|p) + (e|e_0) + (0|b\lfloor \frac{q}{2} \rfloor) \in \mathbb{Z}_{/q}^{m+1}$ qui est soit proche de L, soit de $L + (0|\lfloor \frac{q}{2} \rfloor)$.

- La distinction entre ces deux cas est rendue possible par la connaissance du petit vecteur $(-s|1) \in \Lambda^{\perp}(A|p)$ (car on a $(A|p)(-s|1)^{\mathrm{tr}} = -As^{\mathrm{tr}} + p = 0$).

Moralité : Connaître un petit vecteur dans le réseau dual L^* permet de séparer nettement L en hyperplans.

Réseaux euclidiens et cryptographie
David Madore
Plan
Généralités sur les réseaux euclidiens
L'algorithme LLL
Réseaux et cryptographie

Preuve de sécurité (idée)

Preuve en deux points :

- Savoir distinguer une clé publique $p \in \mathbb{Z}_{/q}^n$ (avec $p = As^{\mathrm{tr}}$ où $s \in \mathbb{Z}_{/q}^m$ petit vecteur) d'une clé aléatoire uniforme $\in \mathbb{Z}_{/q}^{n\times(m+1)}$ revient à savoir résoudre LWE.

En effet, se donner $p = As^{\mathrm{tr}}$ revient à se donner s modulo $\Lambda^\perp(A)$, c'est-à-dire un tirage $vB + s$ avec v uniforme, où $B \in \mathbb{Z}_{/q}^{(m-n)\times n}$ définit $\Lambda(B) = \Lambda^\perp(A)$. C'est bien un problème LWE.

- Savoir déchiffrer pour une clé $A' \in \mathbb{Z}_{/q}^{n\times(m+1)}$ aléatoire uniforme revient à savoir résoudre LWE.

En effet, il s'agit de distinguer $uA' + e'$ (avec u uniforme).

Slim Essid

Enseignant-chercheur au département Traitement du Signal et de l'Image de Télécom ParisTech

Factorisations en matrices positives pour l'analyse de signaux audio

Les techniques de factorisation en matrices positives (NMF, Nonnegative Matrix Factorisation) permettent d'expliquer une matrice d'observations à entrées positives comme la combinaison linéaire positive de formes élémentaires elles-mêmes à coefficients positifs. Ces techniques sont devenues incontournables ces quinze dernières années pour diverses tâches d'analyse de données. Elles s'avèrent performantes dans différents domaines d'applications, allant de l'analyse du langage à l'analyse de signaux électroencéphalographiques (EEG) en passant par les images et les signaux audiovisuels.

Dans cette présentation seront expliqués les modèles NMF et les méthodes permettant de les estimer, avant de présenter les applications que l'on peut en faire. On donnera en particulier des exemples d'application à l'extraction de thèmes à partir de textes, à l'analyse de musique, et à la réjection d'artéfacts dans des enregistrements de signaux EEG.

An introduction to Nonnegative Matrix Factorisation

Slim ESSID

Telecom ParisTech

June 2015

Credits

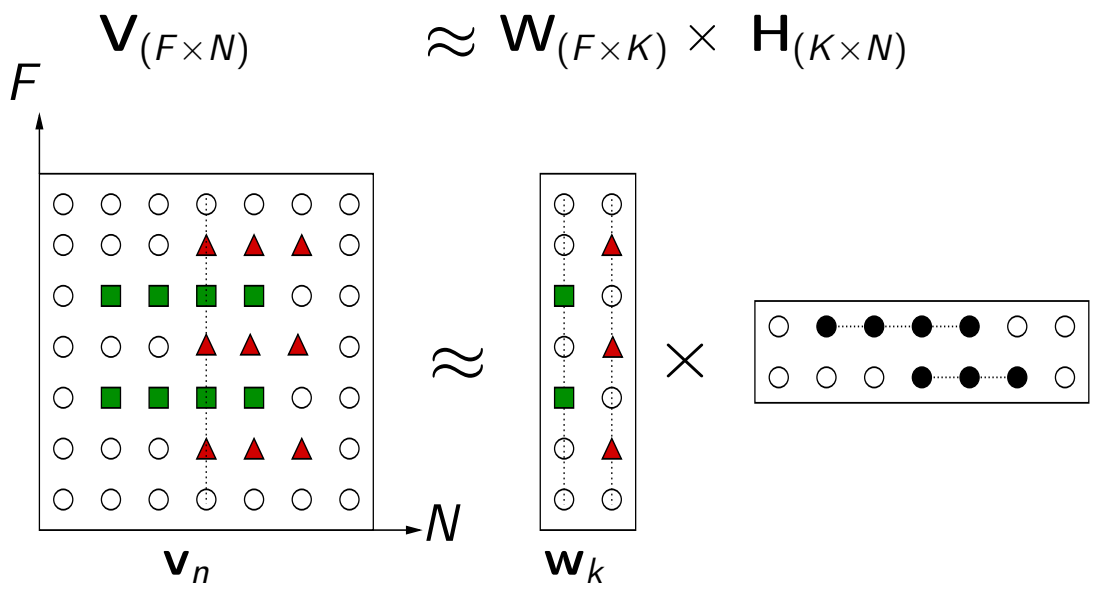
Some illustrations, slides and demos are reproduced courtesy of:

- A. Ozerov,
- C. Févotte,
- N. Seichepine,
- R. Hennequin,
- F. Vallet,
- A. Liutkus.

- Introduction
- NMF models
- Algorithms for solving NMF
- Applications
- Conclusion

Explaining data by factorisation

General formulation

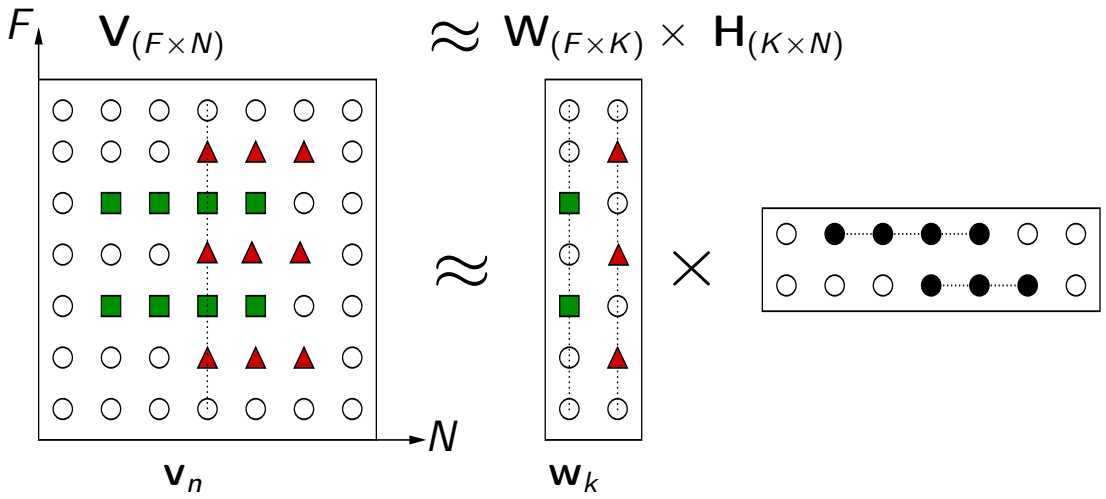


$$\mathbf{v}_n \approx \sum_{k=1}^{K} h_{kn}\mathbf{w}_k$$

Illustration by C. Févotte

Explaining data by factorisation

General formulation



data matrix | "explanatory variables" "basis", "dictionary", "patterns", "topics" | "regressors", "activation coefficients", "expansion coefficients"

Illustration by C. Févotte

Data is often nonnegative by nature[1]

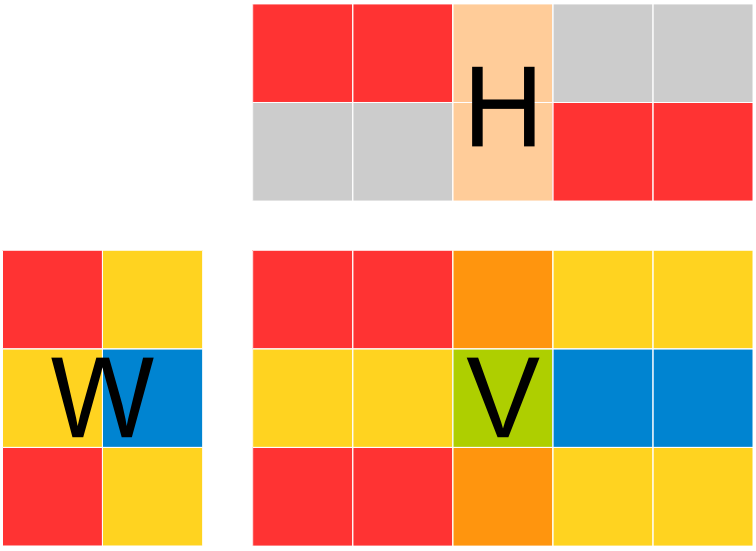
- pixel intensities;
- amplitude spectra;
- occurrence counts;
- food or energy consumption;
- user scores;
- stock market values;
- ...

For the sake of **interpretability** of the results, optimal processing of **nonnegative data** may call for processing under **nonnegativity constraints**.

[1] slide adapted from (Févotte, 2012).

The Nonnegative Matrix Factorisation model

NMF provides an unsupervised linear representation of the data:



$$\mathbf{V} \approx \mathbf{WH};$$

- $\mathbf{W} = [w_{fk}]$ s.t. $w_{fk} \geq 0$ and
- $\mathbf{H} = [h_{kn}]$ s.t. $h_{kn} \geq 0$.

Illustration by N. Seichepine

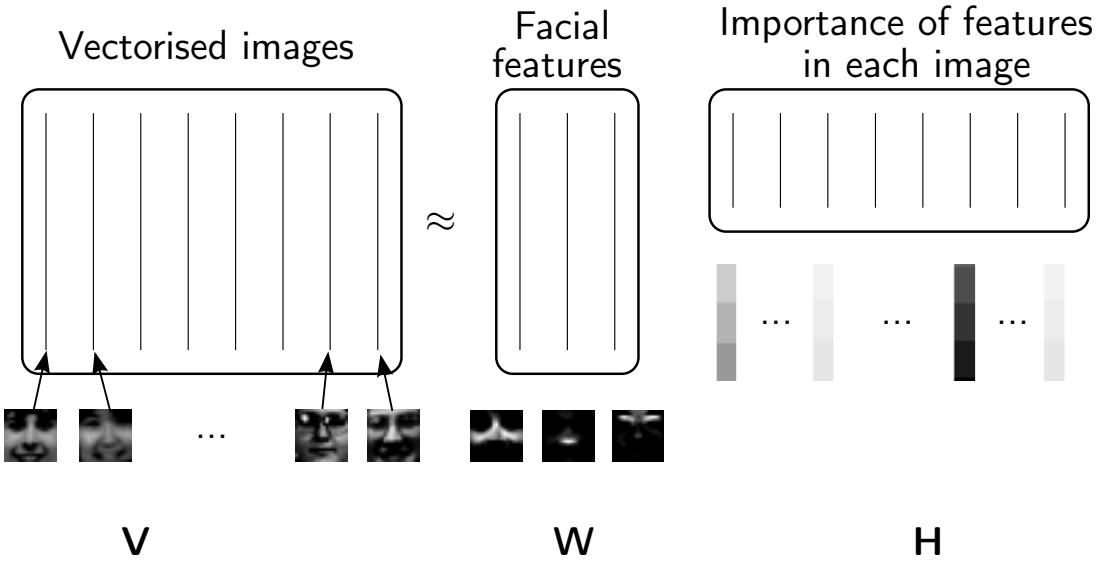
Explaining face images by NMF[2]

Image example: 49 images among 2429 from MIT's CBCL face dataset

[2] slide adapted from (Févotte, 2012).

Explaining face images by NMF

Method



NMF outputs

Image example

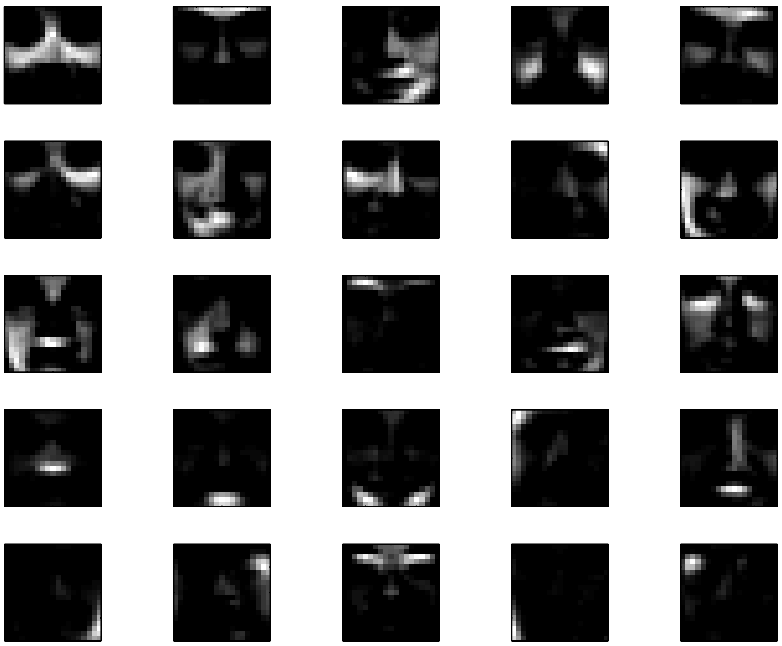


Illustration by C. Févotte

Notations I

- $\mathbf{V}$: the $F \times N$ **data matrix**:
  - F features (rows),
  - N observations/examples/feature vectors (columns);
- $\mathbf{v}_n = (v_{1n}, \cdots, v_{Fn})^T$: the n-th **feature vector** observation among a collection of N observations $\mathbf{v}_1, \cdots, \mathbf{v}_N$;
- $\mathbf{v}_n$ is a column vector in $\mathbb{R}_+^F$; $\mathbf{v}_n$ is a row vector;

- $\mathbf{W}$: the $F \times K$ **dictionary matrix**:
  - w_{fk} is one of its coefficients,
  - $\mathbf{w}_k$ a dictionary/basis vector among K elements;

Notations II

- $\mathbf{H}$: the $K \times N$ **activation/expansion matrix:**
  - $\mathbf{h}_n$: the **column vector** of activation coefficients for observation $\mathbf{v}_n$:

$$\mathbf{v}_n \approx \sum_{k=1}^{K} h_{kn} \mathbf{w}_k \ ;$$

  - $\mathbf{h}_{k:}$: the **row vector** of activation coefficients relating to basis vector $\mathbf{w}_k$.

NMF optimization criteria

NMF approximation $\mathbf{V} \approx \mathbf{WH}$ is usually obtained through:

$$\min_{\mathbf{W},\mathbf{H} \geq 0} D(\mathbf{V}|\mathbf{WH}),$$

where $D(\mathbf{V}|\widehat{\mathbf{V}})$ is a *separable matrix divergence*:

$$D(\mathbf{V}|\widehat{\mathbf{V}}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d(v_{fn}|\hat{v}_{fn}),$$

and $d(x|y)$ defined for all $x, y \geq 0$ is a *scalar divergence* such that:

- $d(x|y)$ is continuous over x and y;
- $d(x|y) \geq 0$ for all $x, y \geq 0$;
- $d(x|y) = 0$ if and only if $x = y$.

Popular (scalar) divergences

Euclidean (EUC) distance (Lee and Seung, 1999)

$$d_{EUC}(x|y) = (x-y)^2$$

Kullback-Leibler (KL) divergence (Lee and Seung, 1999)

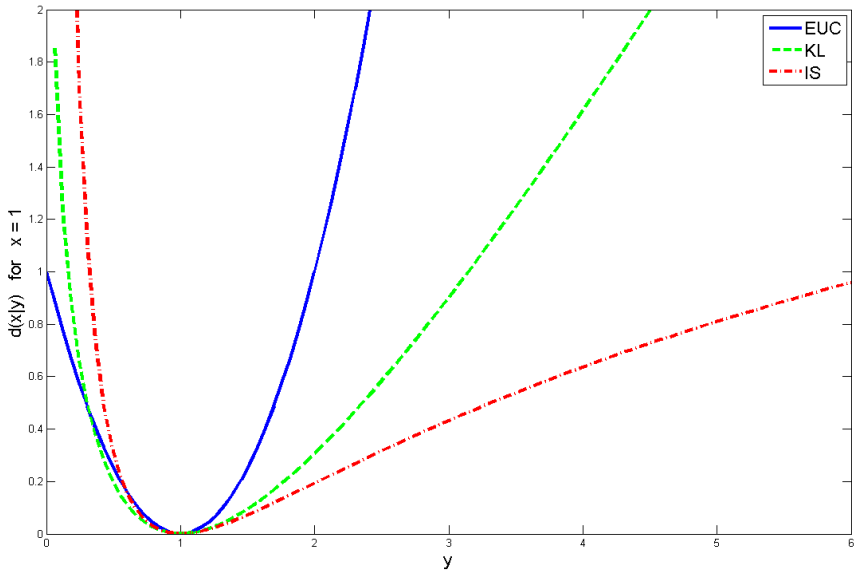
$$d_{KL}(x|y) = x \log \frac{x}{y} - x + y$$

Itakura-Saito (IS) divergence (Févotte et al., 2009)

$$d_{IS}(x|y) = \frac{x}{y} - \log \frac{x}{y} - 1$$

Convexity properties

Divergence $d(x\|y)$	EUC	KL	IS
Convex on x	yes	yes	yes
Convex on y	yes	yes	**no**



Scale invariance properties[3]

$$
\begin{aligned}
d_{EUC}(\lambda x|\lambda y) &= \lambda^2\, d_{EUC}(x|y) \\
d_{KL}(\lambda x|\lambda y) &= \lambda\, d_{KL}(x|y) \\
d_{IS}(\lambda x|\lambda y) &= d_{IS}(x|y)
\end{aligned}
$$

The IS divergence is **scale-invariant** $\rightarrow$ it provides higher accuracy in the representation of data with large dynamic range (*e.g.* audio spectra).

[3] slide adapted from (Févotte, 2012).

Weighted NMF

Conventional NMF optimization criterion:

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \sum_{f=1}^{F}\sum_{n=1}^{N} d(v_{fn}|\hat{v}_{fn})\,.$$

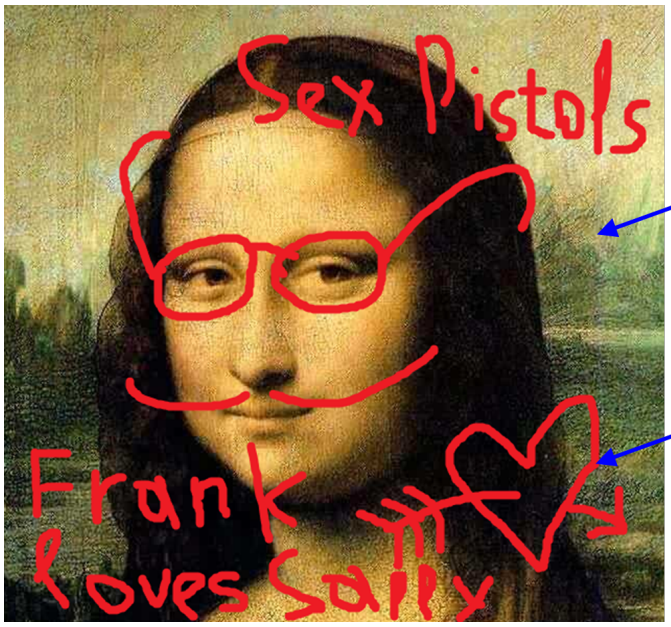
Weighted NMF optimization criterion:

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \sum_{f=1}^{F}\sum_{n=1}^{N} b_{fn}\, d(v_{fn}|\hat{v}_{fn})\,,$$

where b_{fn} $(f = 1, \ldots, F,\ n = 1, \ldots, N)$ are some nonnegative weights representing the contribution of data point v_{fn} to NMF learning.

Weighted NMF application example I

Learning from partial observations (e.g., for **image inpainting** as in (Mairal et al., 2010)):

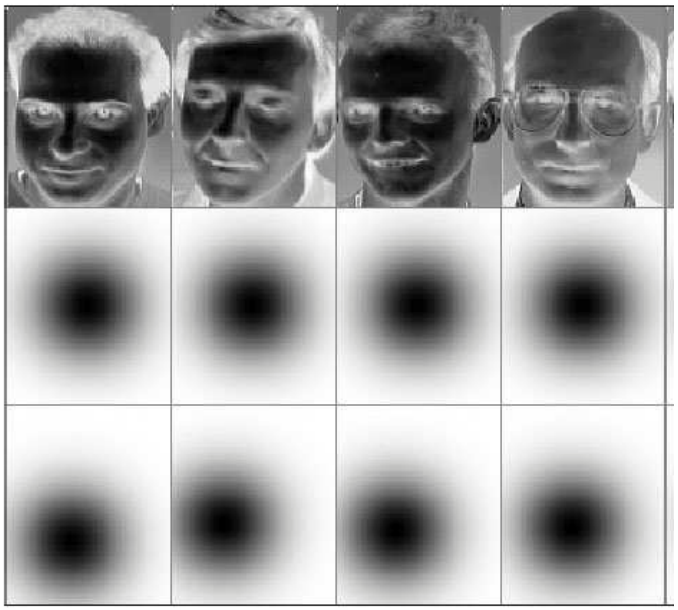


Observed value
$b_{fn} = 1$

Missing value
$b_{fn} = 0$

Weighted NMF application example II

Face feature extraction (example and figure from (Blondel et al., 2008)):



Data $\mathbf{V}$

Weights $\mathbf{B} = \{b_{fn}\}_{f,n}$

Image-centered weights

Face-centered weights

Optimization problem

An efficient solution of the NMF optimization problem

$$\min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH}) \Leftrightarrow \min_{\boldsymbol{\theta}} C(\boldsymbol{\theta})\,;\;\; C(\boldsymbol{\theta}) \stackrel{\text{def}}{=} D(\mathbf{V}|\mathbf{WH})$$

where $\boldsymbol{\theta} \stackrel{\text{def}}{=} \{\mathbf{W}, \mathbf{H}\}$ denotes the NMF parameters, must cope with the following difficulties:

- the **nonnegativity constraints** must be taken into account;
- the solution is **not unique**...

NMF is ill-posed

The solution is not unique

Given $\mathbf{V} = \mathbf{WH}$; $\mathbf{W} \geq 0$, $\mathbf{H} \geq 0$; any matrix $\mathbf{Q}$ such that:

- $\mathbf{WQ} \geq 0$
- $\mathbf{Q}^{-1}\mathbf{H} \geq 0$

provides an alternative factorisation $\mathbf{V} = \tilde{\mathbf{W}}\tilde{\mathbf{H}} = (\mathbf{WQ})(\mathbf{Q}^{-1}\mathbf{H})$.

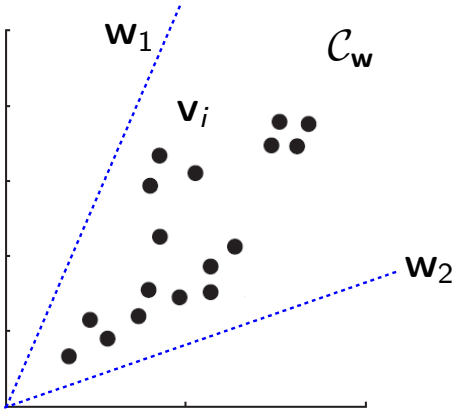
In particular, $\mathbf{Q}$ can be any **nonnegative generalised permutation matrix**; *e.g.*, in $\mathbb{R}^3$:

$$\mathbf{Q} = \begin{bmatrix} 0 & 0 & 2 \\ 0 & 3 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

This case is not so problematic: merely accounts for **scaling** and **permutation** of basis vectors $\mathbf{w}_k$.

Geometric interpretation and ill-posedness

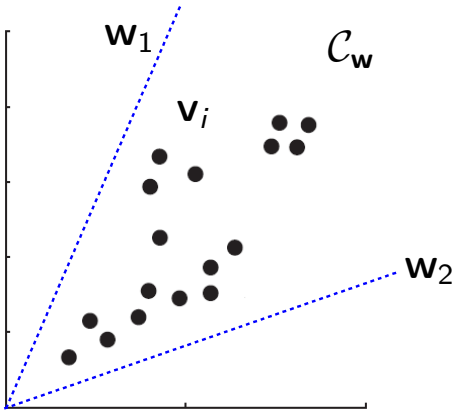
NMF assumes the data is well described by a **simplicial convex cone** $\mathcal{C}_{\mathbf{w}}$ generated by the columns of $\mathbf{W}$:

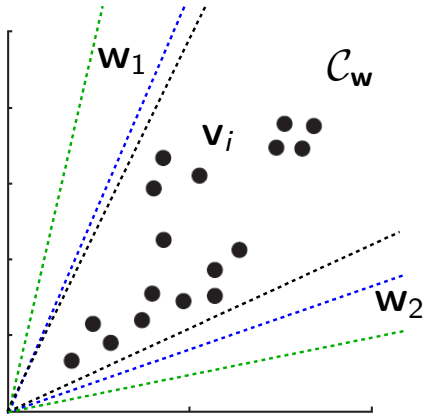


$$\mathcal{C}_{\mathbf{w}} = \left\{ \sum_{k=1}^{K} \lambda_k \mathbf{w}_k ;\ \lambda_k \geq 0 \right\}$$

Geometric interpretation and ill-posedness

NMF assumes the data is well described by a **simplicial convex cone** $\mathcal{C}_{\mathbf{W}}$ generated by the columns of $\mathbf{W}$:



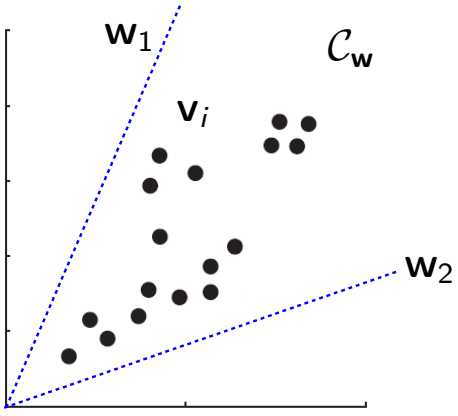


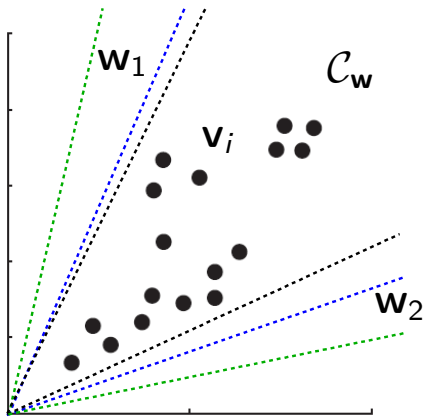
$$\mathcal{C}_{\mathbf{W}} = \left\{ \sum_{k=1}^{K} \lambda_k \mathbf{w}_k ;\ \lambda_k \geq 0 \right\}$$

Problem: which $\mathcal{C}_{\mathbf{W}}$?

Geometric interpretation and ill-posedness

NMF assumes the data is well described by a **simplicial convex cone** $\mathcal{C}_{\mathbf{W}}$ generated by the columns of $\mathbf{W}$:





$$\mathcal{C}_{\mathbf{W}} = \left\{ \sum_{k=1}^{K} \lambda_k \mathbf{w}_k ;\ \lambda_k \geq 0 \right\}$$

Problem: which $\mathcal{C}_{\mathbf{W}}$?

→ Need to impose **constraints** on the set of possible solutions to select the most "useful" ones.

Alternating optimization strategy

The problem is usually easier to optimize over one matrix (say $\mathbf{H}$) given the other matrix (say $\mathbf{W}$) is known and fixed.

Indeed, for several divergences $D(\mathbf{V}|\mathbf{WH})$ is even convex separately w.r.t. $\mathbf{H}$ and w.r.t. $\mathbf{W}$, but not w.r.t. $\{\mathbf{W}, \mathbf{H}\}$.

For this reason many state-of-the-art NMF optimization algorithms rely on the following iterative alternating optimization strategy.

Alternating optimization a.k.a block-coordinate descent (one iteration):

- update $\mathbf{W}$, given $\mathbf{H}$ fixed,
- update $\mathbf{H}$, given $\mathbf{W}$ fixed.

Multiplicative update rules

A heuristic approach introduced by (Lee and Seung, 2001) to solve $\min_{\boldsymbol{\theta}} C(\boldsymbol{\theta})$

Multiplicative update (MU) rule for $\mathbf{H}$ (similarly for $\mathbf{W}$) is defined as:

$$h_{kn} \leftarrow h_{kn} \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_- / \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_+ ,$$

where

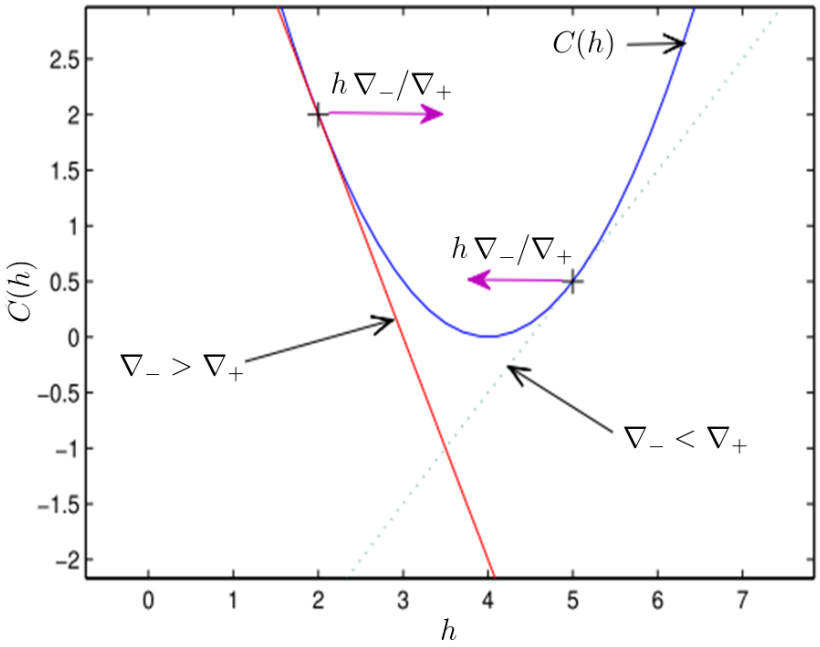
$$\nabla_{h_{kn}} C(\boldsymbol{\theta}) = \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_+ - \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_- ,$$

and the summands are both nonnegative.

NOTE: The nonnegativity of $\mathbf{W}$ and $\mathbf{H}$ is guaranteed by construction.

Intuitive explanation

We consider for simplicity $\nabla_h C(h) = \nabla_+ - \nabla_-$



Discussion

The only two things guaranteed by this approach:

- the newly updated value lies in the **direction of partial derivative decrease**;
- the newly updated value is **always nonnegative**.

Nothing more can be guaranteed in general, and all the other algorithm properties depend on the **"positive-negative" decomposition chosen:**

$$\nabla_{h_{kn}} C(\boldsymbol{\theta}) = [\nabla_{h_{kn}} C(\boldsymbol{\theta})]_+ - [\nabla_{h_{kn}} C(\boldsymbol{\theta})]_- \ .$$

Majorisation-minimisation viewpoint

For **many divergences** and **certain "positive-negative" decompositions** each MU rule can be interpreted as a **Majorisation-Minimisation (MM)** procedure (Hunter and Lange, 2004):

To minimise $C(s)$, *e.g.*, $s = w_{fk}$ or $s = h_{kn}$:

- build $G(s|\tilde{s})$ such that $G(s|\tilde{s}) \geq C(s)$ and $G(\tilde{s}|\tilde{s}) = C(\tilde{s})$;
- optimize iteratively $G(s|\tilde{s})$ instead of $C(s)$.

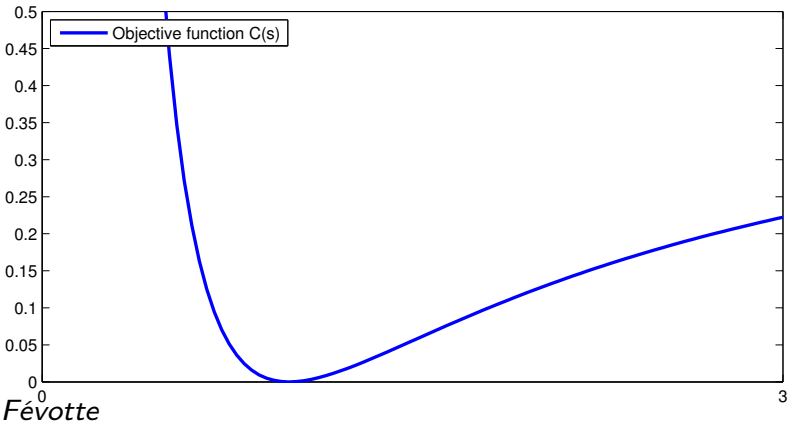


Illustration by C. Févotte

Majorisation-minimisation viewpoint

For **many divergences** and **certain "positive-negative" decompositions** each MU rule can be interpreted as a **Majorisation-Minimisation (MM)** procedure (Hunter and Lange, 2004):

To minimise $C(s)$, *e.g.*, $s = w_{fk}$ or $s = h_{kn}$:

- build $G(s|\tilde{s})$ such that $G(s|\tilde{s}) \geq C(s)$ and $G(\tilde{s}|\tilde{s}) = C(\tilde{s})$;
- optimize iteratively $G(s|\tilde{s})$ instead of $C(s)$.

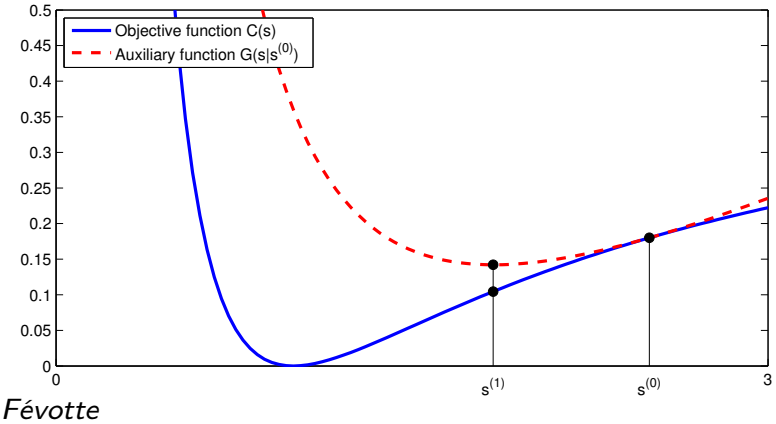


Illustration by C. Févotte

Majorisation-minimisation viewpoint

For **many divergences** and **certain "positive-negative" decompositions** each MU rule can be interpreted as a **Majorisation-Minimisation (MM)** procedure (Hunter and Lange, 2004):

To minimise $C(s)$, *e.g.*, $s = w_{fk}$ or $s = h_{kn}$:

- build $G(s|\tilde{s})$ such that $G(s|\tilde{s}) \geq C(s)$ and $G(\tilde{s}|\tilde{s}) = C(\tilde{s})$;
- optimize iteratively $G(s|\tilde{s})$ instead of $C(s)$.

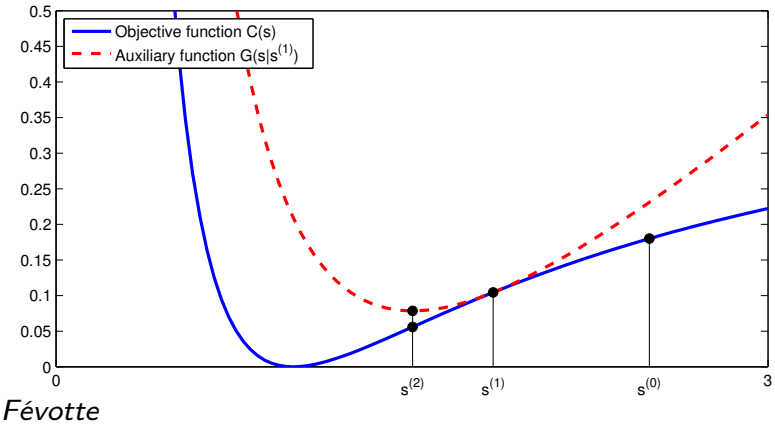


Illustration by C. Févotte

Algorithms for solving NMF | Multiplicative update rules

Majorisation-minimisation viewpoint

For **many divergences** and **certain "positive-negative" decompositions** each MU rule can be interpreted as a **Majorisation-Minimisation (MM)** procedure (Hunter and Lange, 2004):

To minimise $C(s)$, *e.g.*, $s = w_{fk}$ or $s = h_{kn}$:

- build $G(s|\tilde{s})$ such that $G(s|\tilde{s}) \geq C(s)$ and $G(\tilde{s}|\tilde{s}) = C(\tilde{s})$;
- optimize iteratively $G(s|\tilde{s})$ instead of $C(s)$.

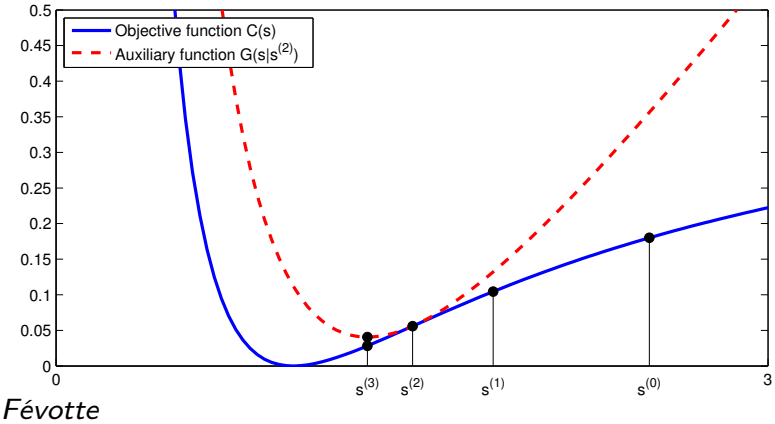


Illustration by C. Févotte

Majorisation-minimisation viewpoint

For **many divergences** and **certain "positive-negative" decompositions** each MU rule can be interpreted as a **Majorisation-Minimisation (MM)** procedure (Hunter and Lange, 2004):

To minimise $C(s)$, *e.g.*, $s = w_{fk}$ or $s = h_{kn}$:

- build $G(s|\tilde{s})$ such that $G(s|\tilde{s}) \geq C(s)$ and $G(\tilde{s}|\tilde{s}) = C(\tilde{s})$;
- optimize iteratively $G(s|\tilde{s})$ instead of $C(s)$.

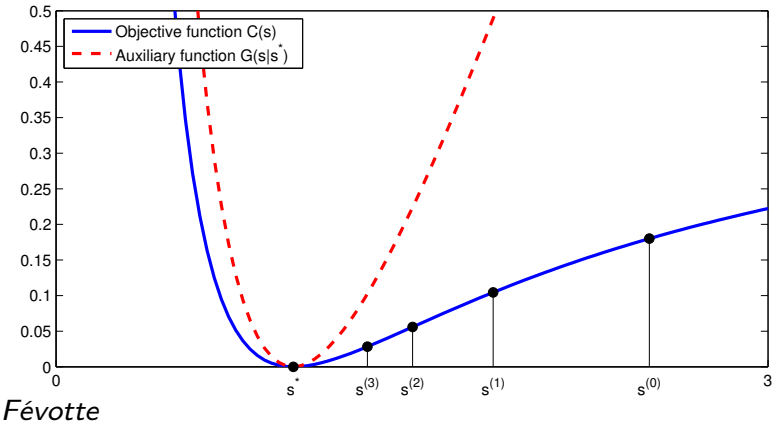


Illustration by C. Févotte

Algorithms for solving NMF | Multiplicative update rules

Majorisation-minimisation viewpoint

For **many divergences** and **certain "positive-negative" decompositions** each MU rule can be interpreted as a **Majorisation-Minimisation (MM)** procedure (Hunter and Lange, 2004):

To minimise $C(s)$, *e.g.*, $s = w_{fk}$ or $s = h_{kn}$:

- build $G(s|\tilde{s})$ such that $G(s|\tilde{s}) \geq C(s)$ and $G(\tilde{s}|\tilde{s}) = C(\tilde{s})$;
- optimize iteratively $G(s|\tilde{s})$ instead of $C(s)$.

- **NOTE:** The MM procedure guarantees the cost is non-increasing at each iteration:

$$C(s^{(t+1)}) \leq G(s^{(t+1)}|s^{(t)}) \leq G(s^{(t)}|s^{(t)}) = C(s^{(t)}).$$

Slim ESSID (Telecom ParisTech) | Introduction to NMF | TPT - UPS – June 2015 | 28 / 53

Summary

Multiplicative Update rules:

Advantages:

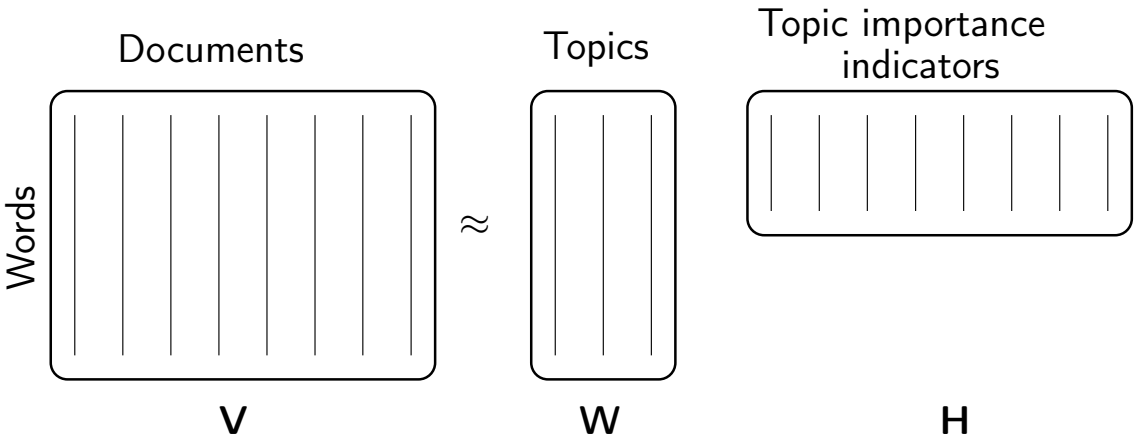
- easy to implement;
- non-negativity of **W** and **H** is guaranteed.

Drawbacks:

- monotonicity is not always guaranteed;
- among other algorithms the convergence rate is not the highest one.

Topics recovery

Assume $\mathbf{V} = [v_{fn}]$ is a **term-document** co-occurrence matrix:
v_{fn} is the frequency of occurrences of word m_f in document d_n;



Text document analysis example

After sklearn topics extraction demo (Pedregosa et al., 2011)

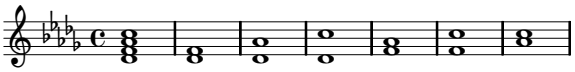
Analysing the 20 newsgroups dataset with NMF, the following topics are automatically determined:

- **Topic #0:** god people bible israel jesus christian true moral think christians believe don say human israeli church life children jewish
- **Topic #1:** drive windows card drivers video scsi software pc thanks vga graphics help disk uni dos file ide controller work
- **Topic #2:** game team nhl games ca hockey players buffalo edu cc year play university teams baseball columbia league player toronto
- **Topic #3:** window manager application mit motif size display widget program xlib windows user color event information use events values
- **Topic #4:** pitt gordon banks cs science pittsburgh univ computer soon disease edu reply pain health david article medical medicine

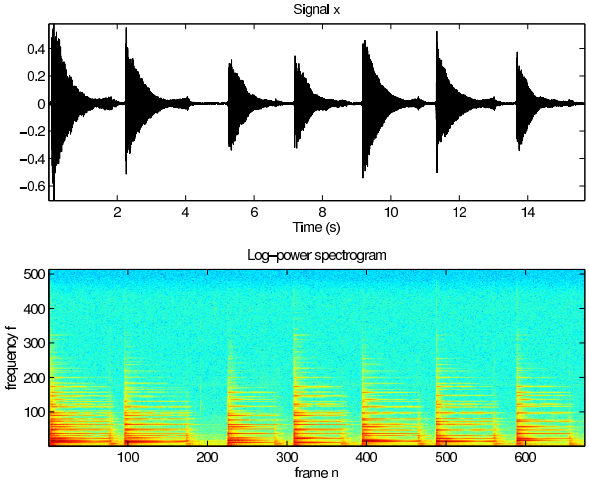
Topics described by most frequent words in each dictionary element $\mathbf{W}_k$.

NMF-based music transcription

Demo slide courtesy of C. Févotte (Fevotte et al., 2009)



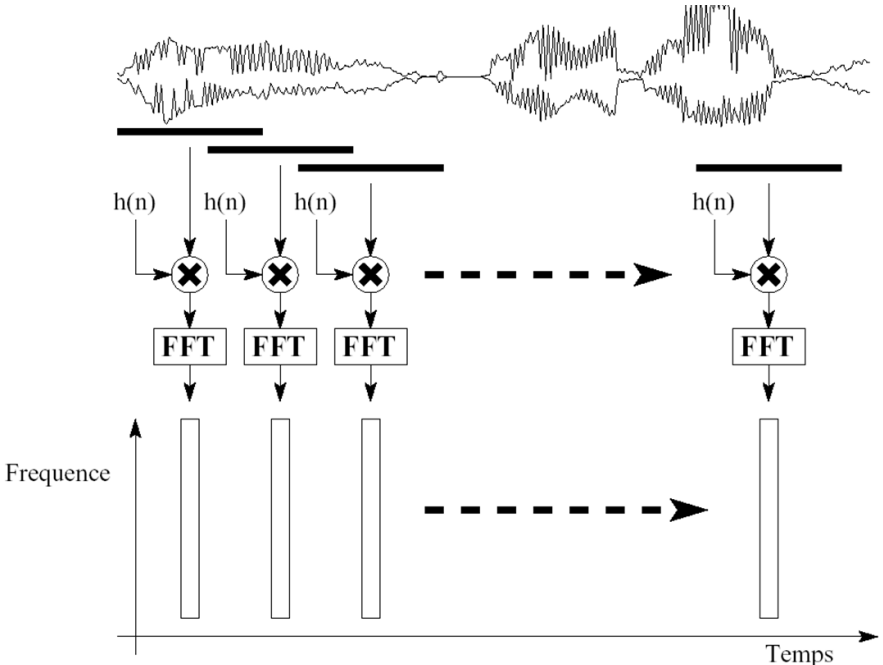
(MIDI numbers: 61, 65, 68, 72)



Three representations of the data.

Spectral analysis

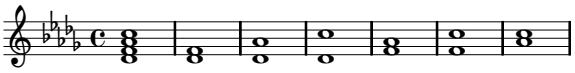
Short-Term Fourier Transform (STFT)



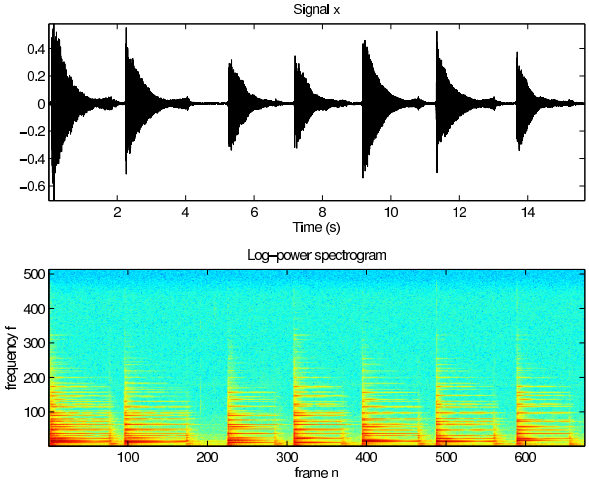
Drawing by J. Laroche

NMF-based music transcription demo

Demo slide courtesy of C. Févotte (Fevotte et al., 2009)



(MIDI numbers: 61, 65, 68, 72)

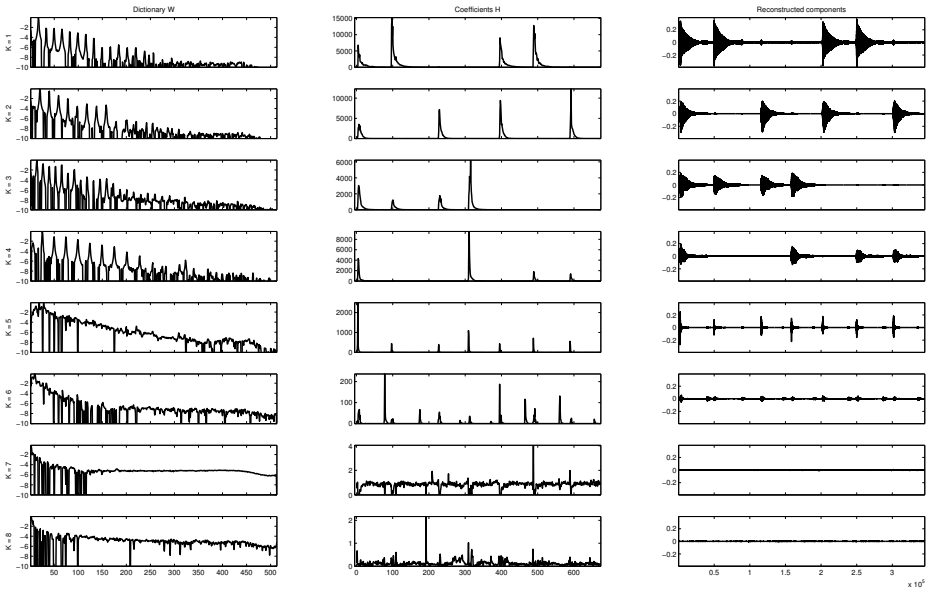


Three representations of the data.

Music transcription demo

Demo slide courtesy of C. Févotte (Fevotte et al., 2009)

NMF decomposition with $K = 8$

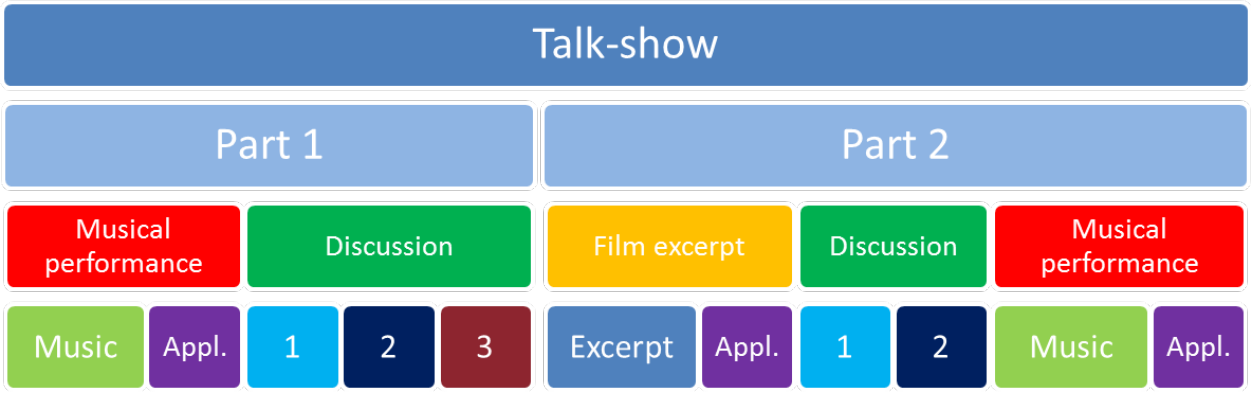


Pitch estimates: 65.0 68.0 61.0 72.0 0 0 0 0

(True values: 61, 65, 68, 72)

The video structuring problem

Goal: automatically extract a **temporal organization** of a document into units conveying a homogeneous type of (audio/video) content.



Video Structuring

Using NMF for temporal segmentation and soft-clustering (Essid and Fevotte, 2013)

Discovering the video editing structure (Essid and Fevotte, 2012)

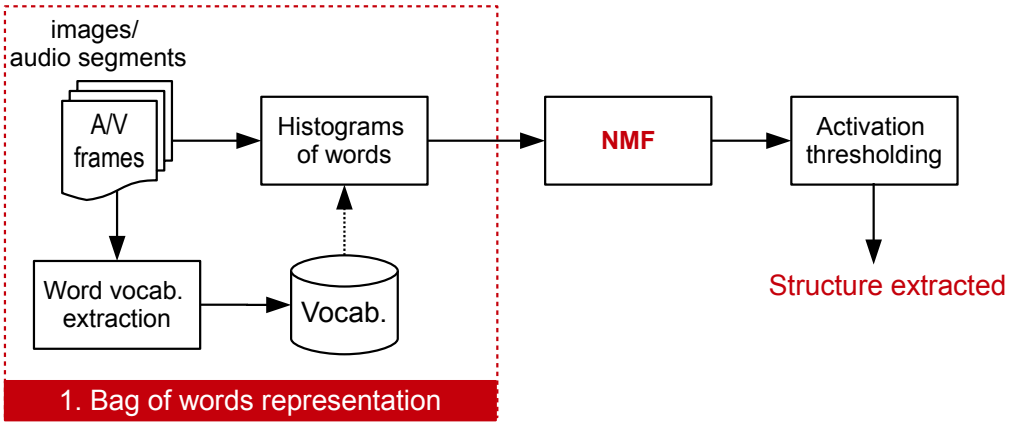
Performing **speaker diarization** (Seichepine et al., 2013)

"Who spoke when?"

illustration by N. Seichepine

A generic video structuring system using NMF

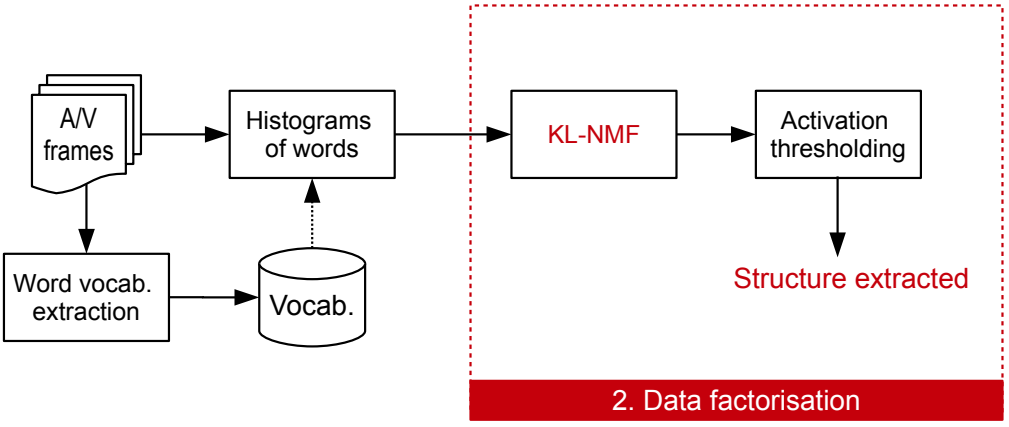
Challenge: perform the task in a **non-supervised fashion**.
Proposed approach: a **generic** structuring scheme using **NMF** (Essid and Fevotte, 2013):



1. create a low-level (visual/audio) vocabulary and use it to extract **histogram of (visual/audio) words** from the sequence of observation frames;

A generic video structuring system using NMF

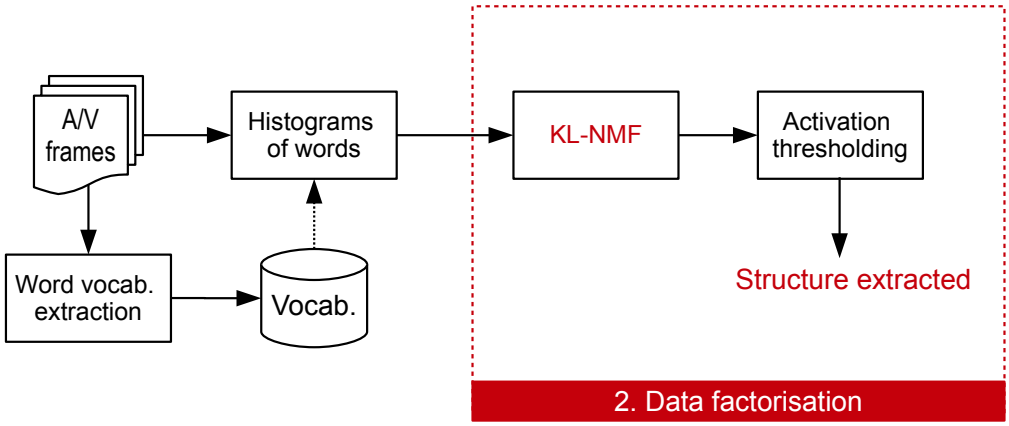
Challenge: perform the task in a **non-supervised fashion**.
Proposed approach: a **generic** structuring scheme using **NMF** (Essid and Fevotte, 2013):

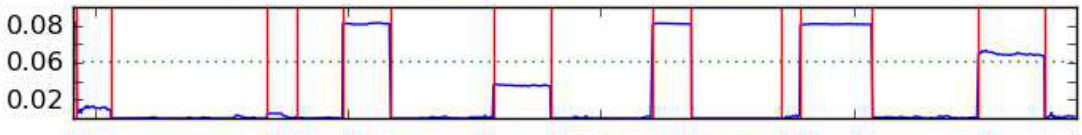


2. apply a variant of **smooth NMF** using the **Kullback-Leibler** divergence to extract **latent structuring events** and their **activations** across the duration of the document.

A generic video structuring system using NMF

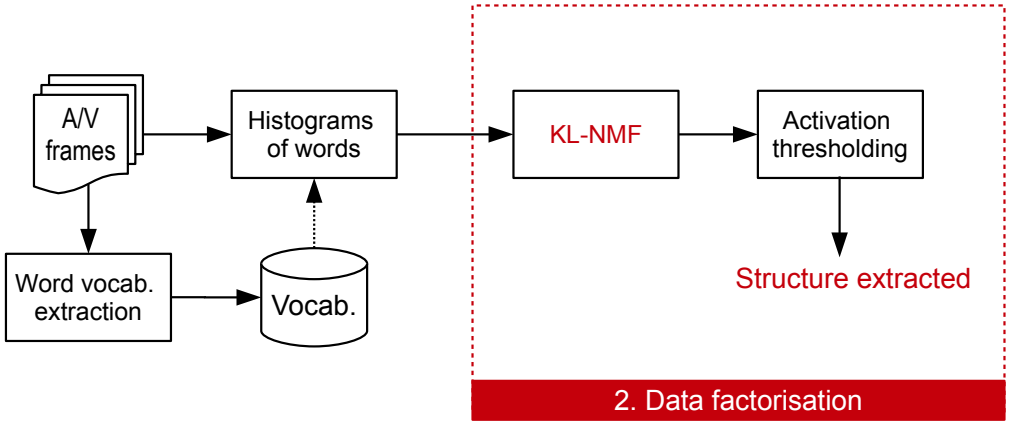
Challenge: perform the task in a **non-supervised fashion**.
Proposed approach: a **generic** structuring scheme using **NMF** (Essid and Fevotte, 2013):





A generic video structuring system using NMF

Challenge: perform the task in a **non-supervised fashion**.
Proposed approach: a **generic** structuring scheme using **NMF** (Essid and Fevotte, 2013):



Activations should be **temporally smooth**: structuring events naturally exhibit a "certain" temporal continuity.

Smooth KL-NMF

Using the Kullback-Leibler (KL) divergence as a measure of fit

Given histogram data (whose columns are frame-wise descriptors), we seek a factorization $\mathbf{V} \approx \mathbf{WH}$; $w_{fk} \geq 0$; $h_{kn} \geq 0$ that minimises

$$C(\mathbf{W}, \mathbf{H}) = D(\mathbf{V}|\mathbf{WH}) + \beta S(\mathbf{H})\,;$$

- $D(\mathbf{V}|\mathbf{WH}) = \sum_{fn} d_{KL}(v_{fn}|\sum_k w_{fk} h_{kn})$: **fit-to-data term** such that $d_{KL}(x|y) = x \log \frac{x}{y} - x + y$;
- $S(H)$ is a **regularisation** term that controls the **temporal smoothness** of the activation coefficients:

$$S(H) = \frac{1}{2} \sum_{k=1}^{K} \sum_{n=2}^{N} (h_{kn} - h_{k(n-1)})^2.$$

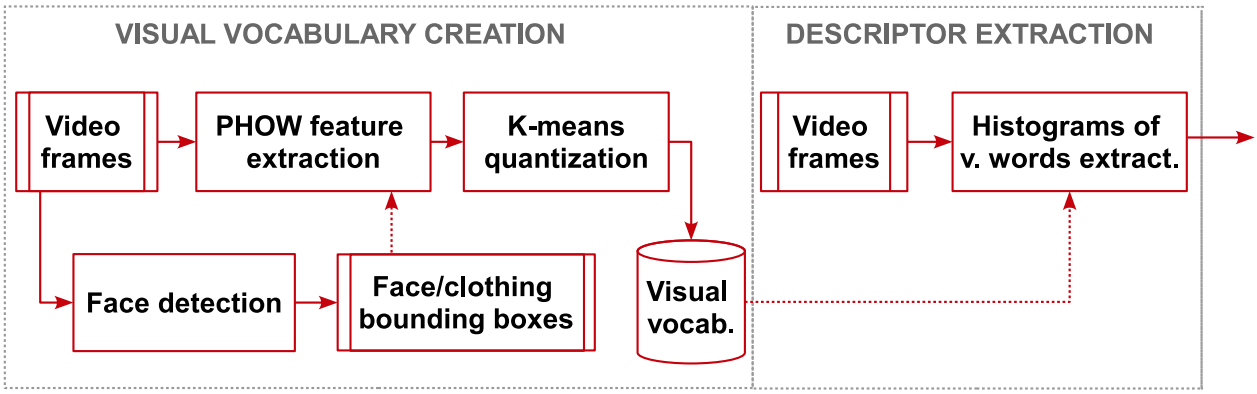
Applications

Onscreen person-oriented structuring

Discover the video editing structure: label the video frames as follows in a **non-supervised** fashion:

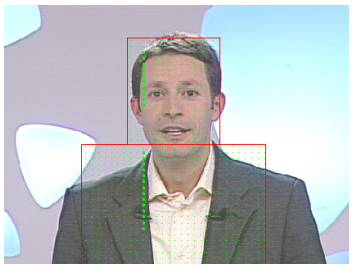
Using the **Canal9 political debates** database (Vinciarelli et al., 2009).

Visual features



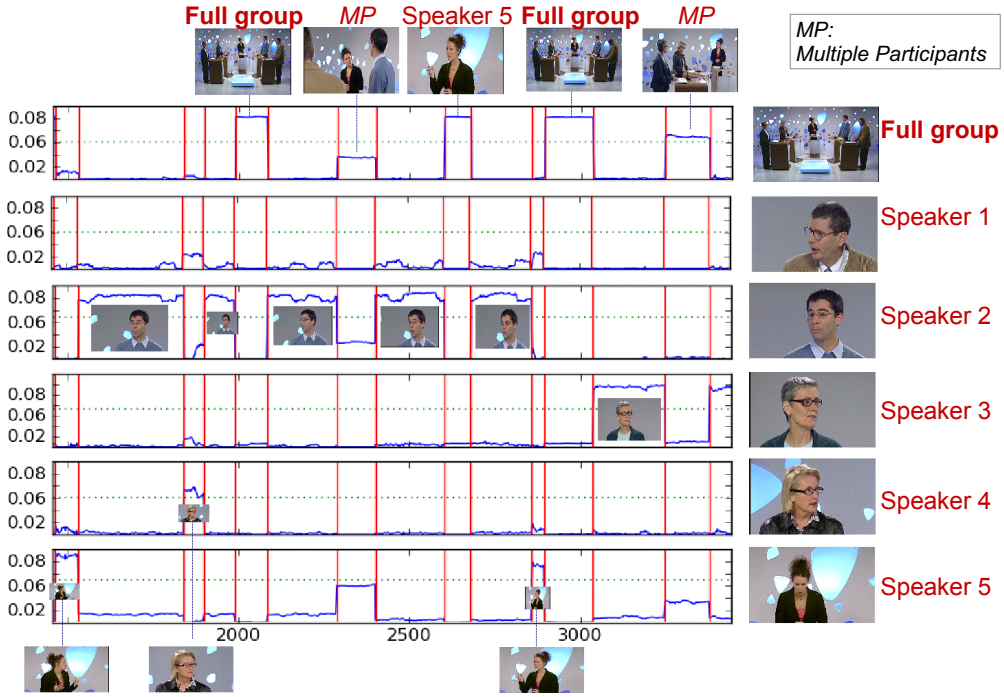
Visual vocabulary creation

- **PHOW** features (Bosch et al., 2007): histograms of orientation gradients over 3 scales, on 8-pixel step grid; extracted from **faces** and **clothing** regions, determined automatically for current video;
- quantization over 128 bins using K-means.



Results

Visualising the activations



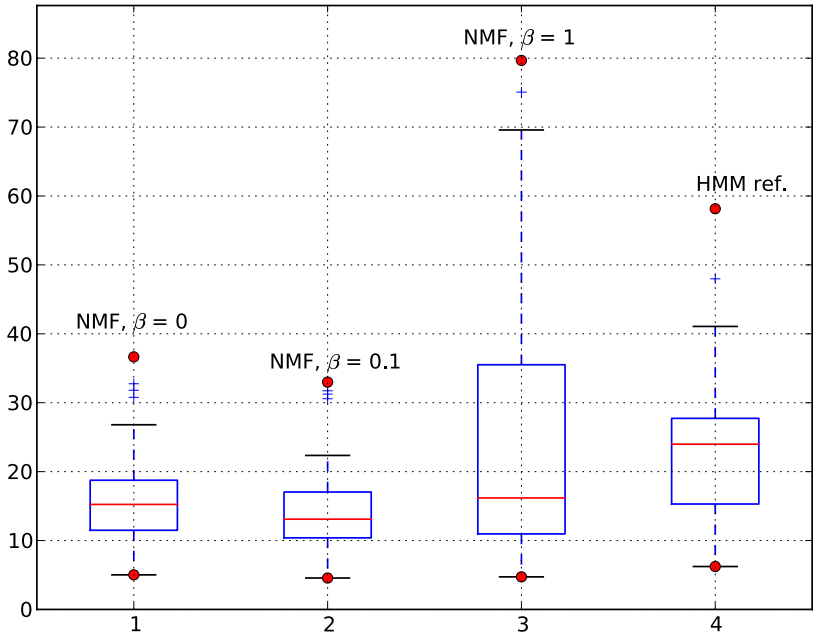
Experimental validation

Canal9 political debates database (Vinciarelli et al., 2009)

- broadcasts featuring a moderator and 2 to 4 guests;
- moderators, guest and background vary;
- 7 hours of video content: 10 minutes from each of the first 41 shows;
- 189 distinct persons; 28521 video shots.

Results

Shot-type classification error rates



Take-home messages I

- NMF is a **versatile** data decomposition technique that has proven effective for **diverse applications** across **numerous disciplines**,
  - it tends to provide "meaningful" and "natural" **part-based** data representations,
  - it can be used both for feature learning, topic extraction, clustering, segmentation, source separation, coding...
- For NMF to be successful, it has to be estimated using **appropriate cost-functions** reflecting prior knowledge about the data.

Take-home messages II

- Many algorithms are available to estimate NMF, mostly alternating updates of **W** and **H**; variants include:
  - **multiplicative updates:** heuristic, simple and easy to implement, but slow and instable,
  - **majorisation-minimisation:** well-founded for a variety of cost functions, stable, still slow,
  - **gradient-descent** and **Newton**: fast but unstable.
- NMF is a state-of-the-art technique for a number of audio-processing tasks (transcription, source separation...),
- it has a great potential for video analysis tasks, especially temporal structure analysis.

Ongoing and future research

- How to properly estimate the **model-order** K?
- How to achieve **better** and **faster** "convergence"?
- How to perform **non-linear** data decompositions?
- How to handle **big data**?

A selection of NMF software

Software	Language	Main features
beta_ntf	Python	Weighted tensor decomposition, all β-divergences, MM
sklearn.decomposition.NMF	Python	ℓ_2-norm, gradient-descent, sparsity
IMM DTU NMF toolbox	Matlab	ℓ_2-norm, MM, gradient-descent, ALS
Févotte's matlab scripts	Matlab	ℓ_2-norm, KL and IS-div, MM, probabilistic
Seichepine's matlab scripts	Matlab	Soft **co-factorisation**, ℓ_2-norm, KL and IS-div, ℓ_1/ℓ_2-norm **temporal smoothing**, MM
svmnmf	Matlab	Geometric SVM-based NMF, **kernel**-based non-linear decompositions, fast
libNMF	C	ℓ_2-norm, MM, gradient-descent, ALS, multi-core, fast

Bibliography I

V. D. Blondel, N.-D. Ho, and P. V. Dooren. Weighted non-negative matrix factorization and face feature extraction. In *Image and Vision Computing*, 2008.

A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *IEEE 11th International Conference on Computer Vision*. IEEE, 2007. URL http://www.computer.org/portal/web/csdl/doi/10.1109/ICCV.2007.4409066.

S. Essid and C. Fevotte. Decomposing the Video Editing Structure of a Talk-show using Nonnegative Matrix Factorization. In *International Conference on Image Processing (ICIP)*, Orlando, FL, USA, 2012.

S. Essid and C. Fevotte. Smooth Nonnegative Matrix Factorization for Unsupervised Audiovisual Document Structuring. *IEEE Transactions on Multimedia*, 15(2):415–425, 2013. ISSN 1520-9210. doi: 10.1109/TMM.2012.2228474.

C. Fevotte, N. Bertin, and J.-L. Durrieu. Nonnegative Matrix Factorization with the Itakura-Saito Divergence. With Application to Music Analysis. *Neural Computation*, 21(3), Mar. 2009.

C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

D. R. Hunter and K. Lange. A tutorial on MM algorithms. *Amer. Stat.*, 58(1):30–37, Feb. 2004.

D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401: 788–791, 1999.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural and Information Processing Systems 13*, pages 556–562, 2001.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11(10-60), 2010.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Bibliography II

N. Seichepine, S. Essid, C. Fevotte, and O. Cappe. Soft nonnegative matrix co-factorization with application to multimodal speaker diarization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, 2013.

A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. Canal9: A database of political debates for analysis of social interactions. In *IEEE International Workshop on Social Signal Processing*, Amsterdam, 2009. Ieee. ISBN 978-1-4244-4800-5. doi: 10.1109/ACII.2009.5349466. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5349466.

Pierre Senellart

Professeur en informatique au département Informatique et Réseaux de Télécom ParisTech

Bases de données probabilistes : modèles et applications aux données du Web

Le Web est une vaste source d'informations, mais ces informations sont souvent incertaines, imprécises, partielles, contradictoires. De plus, les outils d'extraction d'information depuis le Web peuvent ajouter un degré d'incertitude supplémentaire.

Nous présentons des techniques permettant de modéliser cette incertitude, sous la forme de probabilités associés aux tuples d'une base de données ou d'annotations logiques complexes, avec une distribution de probabilité sous-jacente. Nous expliquons comment des requêtes de l'algèbre relationnelle peuvent être exécutées sur ces modèles de données probabilistes, afin d'obtenir les probabilités associées aux résultats de requête. Nous montrons que l'évaluation de ces probabilités est la plupart du temps intractable, mais que certains cas de tractabilité peuvent être isolés.

Cet exposé couvre des recherches récentes en théorie des bases de données, les relie à des notions élémentaires abordées dans le programme de mathématiques, d'informatique et d'option informatique des classes préparatoires (théorie des probabilités, bases de données relationnelles, logique propositionnelle, complexité algorithmique) et en présente des applications dans le domaine des moteurs de recherche Web.

Probabilistic Databases: Models and Applications to Web Data

Pierre Senellart

Journées Télécom–UPS, 29 May 2015

Part I: Uncertainty in the Real World

Uncertain data

Numerous sources of uncertain data:

- Measurement errors
- Data integration from contradicting sources
- Imprecise mappings between heterogeneous schemata
- Imprecise automatic process (information extraction, natural language processing, etc.)
- Imperfect human judgment
- Lies, opinions, rumors

Uncertain data

Numerous sources of uncertain data:

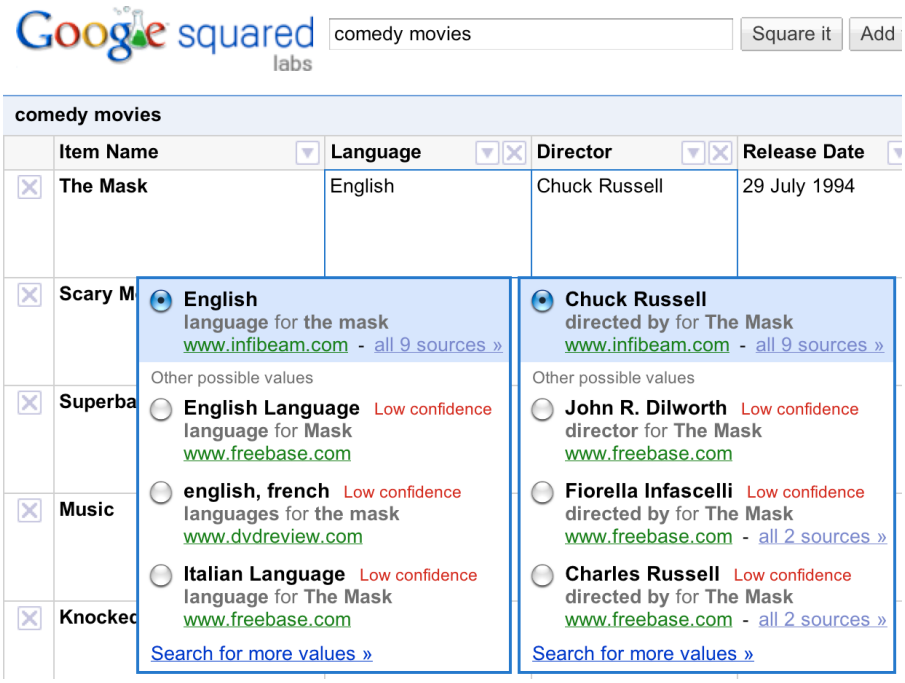
- Measurement errors
- Data integration from contradicting sources
- Imprecise mappings between heterogeneous schemata
- Imprecise automatic process (information extraction, natural language processing, etc.)
- Imperfect human judgment
- Lies, opinions, rumors

Use case: Web information extraction

instance	iteration	date learned	confidence
arabic, egypt	406	08-sep-2011	(Seed) 100.0
chinese, republic_of_china	439	24-oct-2011	100.0
chinese, singapore	421	21-sep-2011	(Seed) 100.0
english, britain	439	24-oct-2011	100.0
english, canada	439	24-oct-2011	(Seed) 100.0
english, england001	439	24-oct-2011	100.0
arabic, morocco	422	23-sep-2011	100.0
cantonese, hong_kong	406	08-sep-2011	100.0
english, uk	436	19-oct-2011	100.0
english, south_vietnam	427	27-sep-2011	99.9
french, morocco	422	23-sep-2011	99.9
greek, turkey	430	07-oct-2011	99.9

Never-ending Language Learning (NELL, CMU),
`http://rtw.ml.cmu.edu/rtw/kbbrowser/`

Use case: Web information extraction



Google Squared (terminated), screenshot from [Fink et al., 2011]

Use case: Web information extraction

Subject	Predicate	Object	Confidence
Elvis Presley	diedOnDate	1977-08-16	97.91%
Elvis Presley	isMarriedTo	Priscilla Presley	97.29%
Elvis Presley	influences	Carlo Wolff	96.25%

YAGO, `http://www.mpi-inf.mpg.de/yago-naga/yago`

Uncertainty in Web information extraction

- The information extraction system is imprecise
- The system has some confidence in the information extracted, which can be:
  - a probability of the information being true (e.g., conditional random fields)
  - an ad-hoc numeric confidence score
  - a discrete level of confidence (low, medium, high)
- What if this uncertain information is not seen as something final, but is used as a source of, e.g., a query answering system?

Different types of uncertainty

Two dimensions:

- Different types:
  - Unknown value: NULL in an RDBMS
  - Alternative between several possibilities: either A or B or C
  - Imprecision on a numeric value: a sensor gives a value that is an approximation of the actual value
  - Confidence in a fact as a whole: cf. information extraction
  - Structural uncertainty: the schema of the data itself is uncertain
- Qualitative (NULL) or Quantitative (95%, low-confidence, etc.) uncertainty

Managing uncertainty

Objective

Not to pretend this imprecision does not exist, and manage it as rigorously as possible throughout a long, automatic and human, potentially complex, process.

Managing uncertainty

Objective

Not to pretend this imprecision does not exist, and manage it as rigorously as possible throughout a long, automatic and human, potentially complex, process.

Especially:

- Represent all different forms of uncertainty
- Use probabilities to represent quantitative information on the confidence in the data
- Query data and retrieve uncertain results
- Allow adding, deleting, modifying data in an uncertain way
- Bonus (if possible): Keep as well lineage/provenance information, so as to ensure traceability

Why probabilities?

- Not the only option: fuzzy set theory [Galindo et al., 2005], Dempster-Shafer theory [Zadeh, 1986]
- Mathematically rich theory, nice semantics with respect to traditional database operations (e.g., joins)
- Some applications already generate probabilities (e.g., statistical information extraction or natural language probabilities)
- In other cases, we "cheat" and pretend that (normalized) confidence scores are probabilities: see this as a first-order approximation

Objective of this talk

- Present data models for uncertain data management in general, and probabilistic data management in particular:
  - relational
  - XML
- Briefly discuss querying of probabilistic data

Part II: Probabilistic Models of Uncertainty

- Probabilistic Relational Models
- Probabilistic XML

Possible worlds semantics

Possible world: A regular (deterministic) relational or XML database

Incomplete database: (Compact) representation of a set of possible worlds

Probabilistic database: (Compact) representation of a probability distribution over possible worlds, either:

- finite: a set of possible worlds, each with their probability
- continuous: more complicated, requires defining a σ-algebra, and a measure for the sets of this σ-algebra

Part II: Probabilistic Models of Uncertainty

- Probabilistic Relational Models
- Probabilistic XML

The relational model

- Data stored into tables
- Every table has a precise schema (type of columns)
- Adapted when the information is very structured

Patient	Examin. 1	Examin. 2	Diagnosis
A	23	12	α
B	10	23	β
C	2	4	γ
D	15	15	α
E	15	17	β

Codd tables, a.k.a. SQL NULLs

Patient	Examin. 1	Examin. 2	Diagnosis
A	23	12	α
B	10	23	$\bot_1$
C	2	4	γ
D	15	15	$\bot_2$
E	$\bot_3$	17	β

- Most simple form of incomplete database
- Widely used in practice, in DBMS since the mid-1970s!
- All NULLs ($\bot$) are considered distinct
- Possible world semantics: all (infinitely many under the open world assumption) possible completions of the table
- In SQL, three-valued logic, weird semantics:

```
SELECT * FROM Tel WHERE tel_nr = '333' OR tel_nr <> '333'
```

C-tables [Imielinski and Lipski, 1984]

Patient	Examin. 1	Examin. 2	Diagnosis	Condition
A	23	12	α	
B	10	23	$\perp_1$	
C	2	4	γ	
D	$\perp_2$	15	$\perp_1$	
E	$\perp_3$	17	β	$18 < \perp_3 < \perp_2$

- NULLs are labeled, and can be reused inside and across tuples
- Arbitrary correlations across tuples
- Closed under the relational algebra (Codd tables only closed under projection and union)
- Every set of possible worlds can be represented as a database with c-tables

Tuple-independent databases (TIDs)

[Lakshmanan et al., 1997, Dalvi and Suciu, 2007]

Patient	Examin. 1	Examin. 2	Diagnosis	Probability
A	23	12	α	0.9
B	10	23	β	0.8
C	2	4	γ	0.2
C	2	14	γ	0.4
D	15	15	α	0.6
D	15	15	β	0.4
E	15	17	β	0.7
E	15	17	α	0.3

- Allow representation of the confidence in each row of the table
- Impossible to express dependencies across rows
- Very simple model, well understood

Block-independent databases (BIDs)

[Barbará et al., 1992, Ré and Suciu, 2007]

Patient	Examin. 1	Examin. 2	Diagnosis	Probability	
A	23	12	α	0.9	
B	10	23	β	0.8	
C	2	4	γ	0.2	$\oplus$
C	2	14	γ	0.4	
D	15	15	β	0.6	$\oplus$
D	15	15	α	0.4	
E	15	17	β	0.7	$\oplus$
E	15	17	α	0.3	

- The table has a primary key: tuples sharing a primary key are mutually exclusive (probabilities must sum up to ≤ 1)
- Simple dependencies (exclusion) can be expressed, but not more complex ones

Probabilistic c-tables [Green and Tannen, 2006]

Patient	Examin. 1	Examin. 2	Diagnosis	Condition
A	23	12	α	w_1
B	10	23	β	w_2
C	2	4	γ	w_3
C	2	14	γ	$\neg w_3 \wedge w_4$
D	15	15	β	w_5
D	15	15	α	$\neg w_5 \wedge w_6$
E	15	17	β	w_7
E	15	17	α	$\neg w_7$

- The w_i's are Boolean random variables
- Each w_i has a probability of being true (e.g., $\Pr(w_1) = 0.9$)
- The w_i's are independent
- Any finite probability distribution of tables can be represented using probabilistic c-tables

Two actual PRDBMS: Trio and MayBMS

Two main probabilistic relational DBMS:

Trio [Widom, 2005] Various uncertainty operators: unknown value, uncertain tuple, choice between different possible values, with probabilistic annotations. See example later on.

MayBMS [Koch, 2009] Implementation of the probabilistic c-tables model. In addition, uncertain tables can be constucted using a `REPAIR-KEY` operator, similar to BIDs.

Two actual PRDBMS: Trio and MayBMS

Two m

```
test=# select * from R;
 dummy | weather | ground |  p
-------+---------+--------+------
 dummy | rain    | wet    | 0.35
 dummy | rain    | dry    | 0.05
 dummy | no rain | wet    |  0.1
 dummy | no rain | dry    |  0.5
(4 rows)
```

own

ible

ter on.

Ma

bles

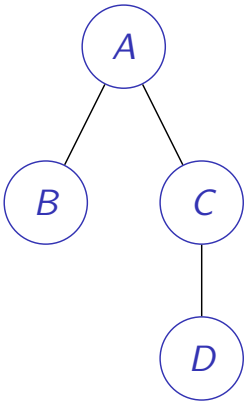
d using

```
test=# create table S as
repair key Dummy in R weight by P;
SELECT
test=# select Ground, conf() from S group by Ground;
 ground | conf
--------+------
 dry    | 0.55
 wet    | 0.45
(2 rows)
```

Part II: Probabilistic Models of Uncertainty

- Probabilistic Relational Models
- Probabilistic XML

The semistructured model and XML




```
<a>
  <b>...</b>
  <c>
    <d>...</d>
  </c>
</a>
```

- Tree-like structuring of data
- No (or less) schema constraints
- Allow mixing tags (structured data) and text (unstructured content)
- Particularly adapted to tagged or heterogeneous content

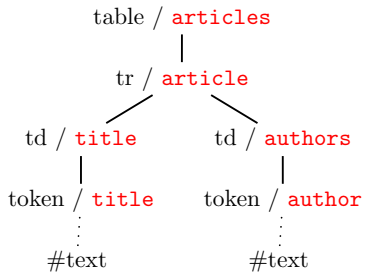
Why Probabilistic XML?

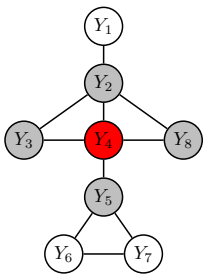
- Extensive literature about probabilistic relational databases [Dalvi et al., 2009, Widom, 2005, Koch, 2009]
- Different typical querying languages: conjunctive queries vs XPath and tree-pattern queries (possibly with joins)
- Cases where a tree-like model might be appropriate:
  - No schema or few constraints on the schema
  - Independent modules annotating freely a content warehouse
  - Inherently tree-like data (e.g., mailing lists, parse trees) with naturally occurring queries involving the descendant axis

Remark

Some results can be transferred from one model to the other. In other cases, connection much trickier! [Amarilli and Senellart, 2013]

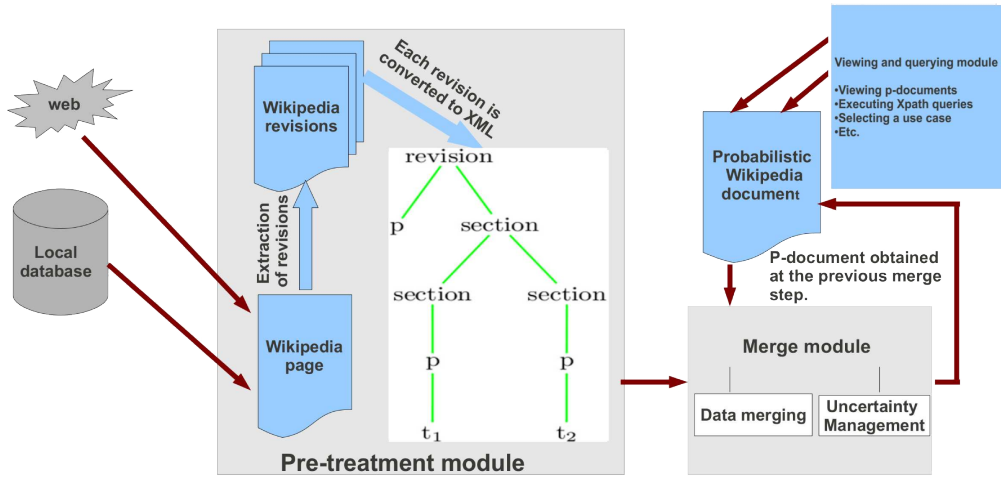
Web information extraction [Senellart et al., 2008]





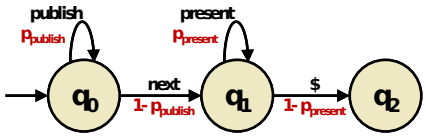
- Annotate HTML Web pages with possible labels
- Labels can be learned from a corpus of annotated documents
- Conditional random fields for XML: estimate probabilities of annotations given annotations of neighboring nodes
- Provides probabilistic labeling of Web pages

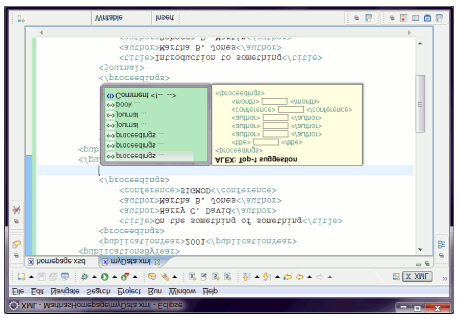
Uncertain version control
[Abdessalem et al., 2011, Ba et al., 2013]



Use trees with probabilistic annotations to represent the uncertainty in the correctness of a document under open version control (e.g., Wikipedia articles)

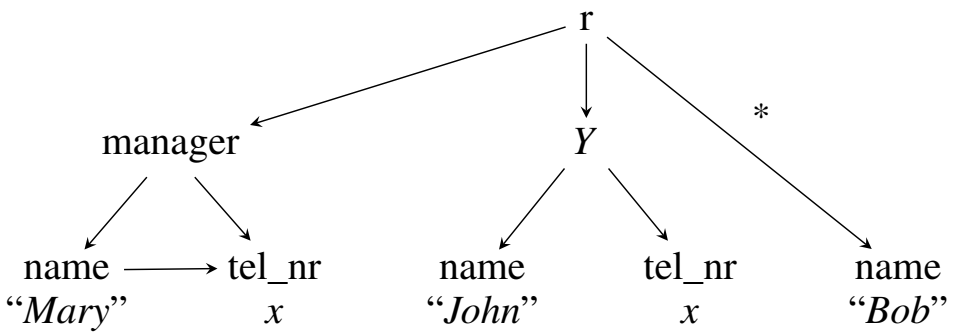
Probabilistic summaries of XML corpora
[Abiteboul et al., 2012a,b]





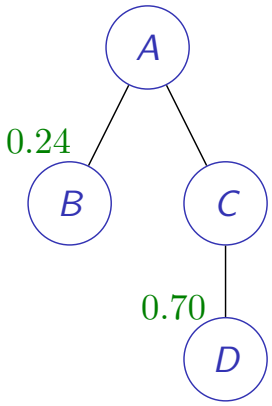
- Transform an XML schema (deterministic top-down tree automaton) into a probabilistic generator (probabilistic tree automaton) of XML documents
- Probability distribution optimal with respect to a given corpus
- Application: Optimal auto-completions in an XML editor

Incomplete XML [Barceló et al., 2009]



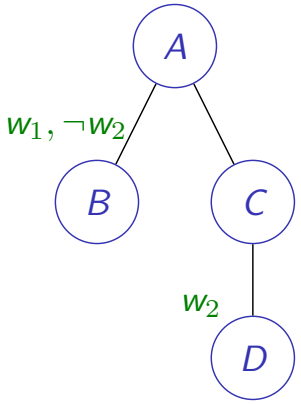
- Models all XML documents where these patterns exist (i.e., this subtree can be matched)
- Can be used for query answering, etc.

Simple probabilistic annotations



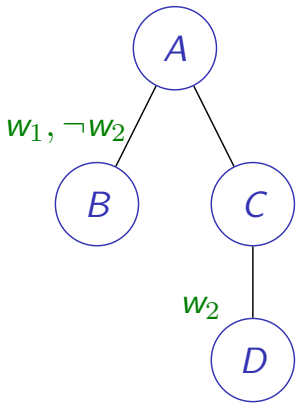
- Probabilities associated to tree nodes
- Express parent/child dependencies
- Impossible to express more complex dependencies
- $\Rightarrow$ some sets of possible worlds are not expressible this way!

Annotations with event variables

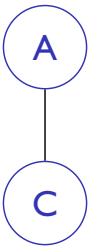


Event	Prob.
w_1	0.8
w_2	0.7

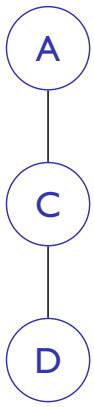
Annotations with event variables



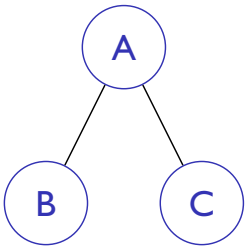
semantics →



$p_1 = 0.06$



$p_2 = 0.70$



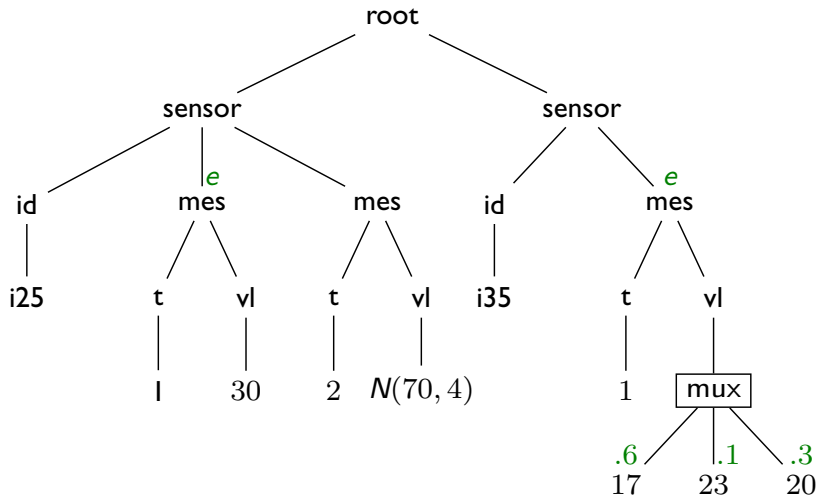
$p_3 = 0.24$

Event	Prob.
w_1	0.8
w_2	0.7

- Expresses arbitrarily complex dependencies
- Obviously, analogous to probabilstic c-tables

A general probabilistic XML model

[Abiteboul et al., 2009]



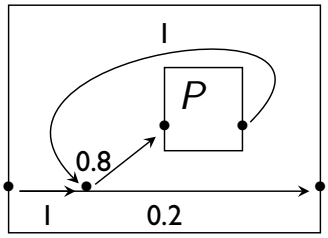
- *e*: event "it did not rain" at time 1
- mux: mutually exclusive options
- $N(70, 4)$: normal distribution

- Compact representation of a set of possible worlds
- Two kinds of dependencies: global (*e*) and local (mux)
- Generalizes all previously proposed models of the literature

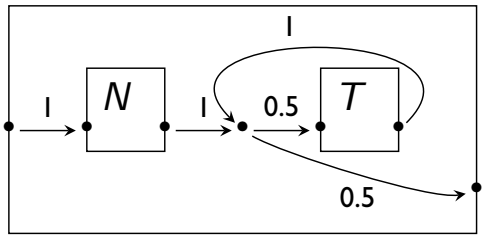
Recursive Markov chains [Benedikt et al., 2010]

```
<!ELEMENT directory (person*)>
<!ELEMENT person (name,phone*)>
```

D: `directory`



P: `person`



- Probabilistic model that extends PXML with local dependencies
- Allows generating documents of unbounded width or depth

Part III: Querying Probabilistic Databases

- Semantics
- Lineage computation and #P-Hardness
- Special tractable case within Probabilistic XML

Part III: Querying Probabilistic Databases

- Semantics
- Lineage computation and #P-Hardness
- Special tractable case within Probabilistic XML

Semantics Of Query Answering: Example

Person

name	city	probability
Ivan	Moscow	0.3
Jean	Paris	0.8
Pedro	Madrid	0.4

Query:
SELECT name FROM Person

Semantics Of Query Answering: Example

Person

name	city	probability
Ivan	Moscow	0.3
Jean	Paris	0.8
Pedro	Madrid	0.4

Query:
SELECT name FROM Person

Pr = 0.3*0.8*0.4

name	city
Ivan	Moscow
Jean	Paris
Pedro	Madrid

Pr = 0.3*0.2*0.4

name	city
Ivan	Moscow
Pedro	Madrid

...

Semantics Of Query Answering: Example

Person

name	city	probability
Ivan	Moscow	0.3
Jean	Paris	0.8
Pedro	Madrid	0.4

Query:
SELECT name FROM Person

Pr = 0.3*0.8*0.4

name	city
Ivan	Moscow
Jean	Paris
Pedro	Madrid

Pr = 0.3*0.2*0.4

name	city
Ivan	Moscow
Pedro	Madrid

...

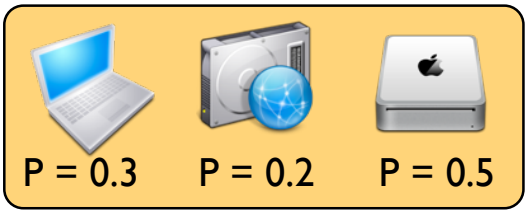
Possible answers: ({Ivan, Juan, Pedro}, 0.3*0.8*0.4), ({Ivan, Pedro}, 0.3*0.2*0.4), ...

Possible tuples: (Ivan, 0.3), (Jean, 0.8), (Pedro, 0.4)

Semantics Of Query Answering

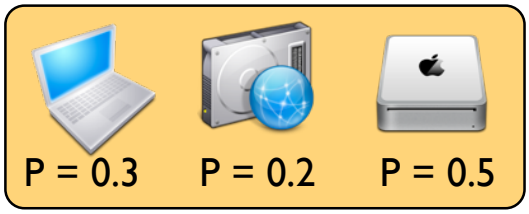
Possible Answers Semantics

Probabilistic DB:



Possible Tuples Semantics

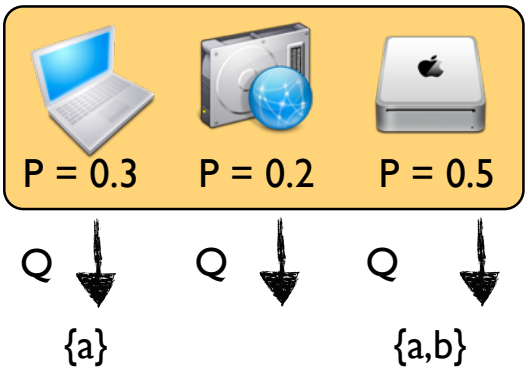
Probabilistic DB:



Semantics Of Query Answering

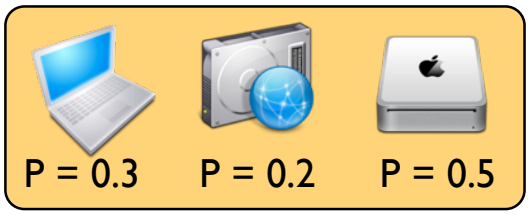
Possible Answers Semantics

Probabilistic DB:



Possible Tuples Semantics

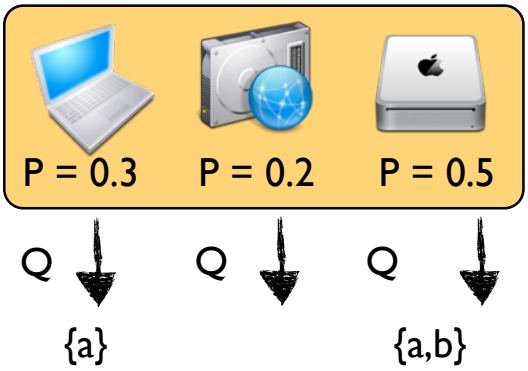
Probabilistic DB:



Semantics Of Query Answering

Possible Answers Semantics

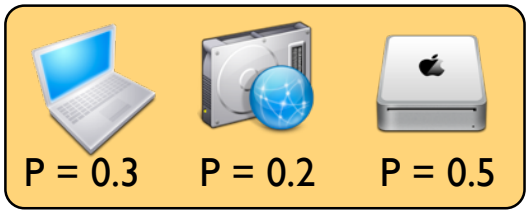
Probabilistic DB:



Answer: ({a}, 0.3); ({a,b}, 0.5)

Possible Tuples Semantics

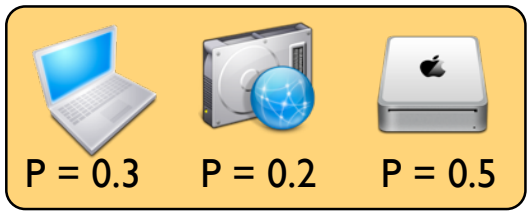
Probabilistic DB:



Semantics Of Query Answering

Possible Answers Semantics

Probabilistic DB:



Q Q Q

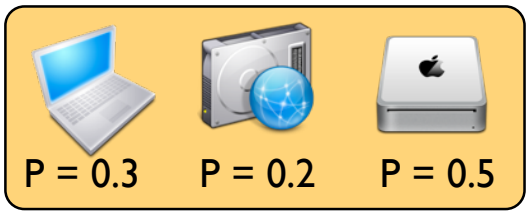
{a} {a,b}

Answer: ({a}, 0.3); ({a,b}, 0.5)

Probability distribution on sets of tuples

Possible Tuples Semantics

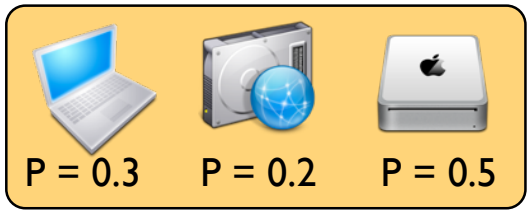
Probabilistic DB:



Semantics Of Query Answering

Possible Answers Semantics

Probabilistic DB:



Q Q Q

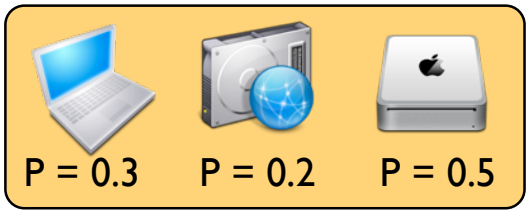
{a} {a,b}

Answer: ({a}, 0.3); ({a,b}, 0.5)

Probability distribution on sets of tuples

Possible Tuples Semantics

Probabilistic DB:



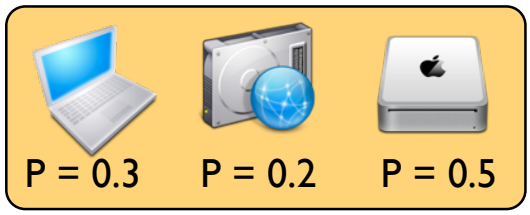
Q Q Q

{a} {a,b}

Semantics Of Query Answering

Possible Answers Semantics

Probabilistic DB:



Q Q Q

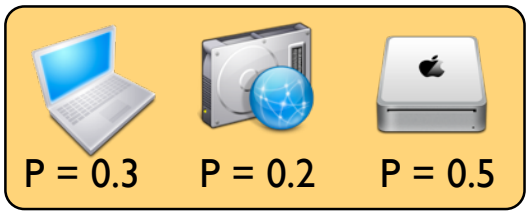
{a} {a,b}

Answer: ({a}, 0.3); ({a,b}, 0.5)

Probability distribution on sets of tuples

Possible Tuples Semantics

Probabilistic DB:



Q Q Q

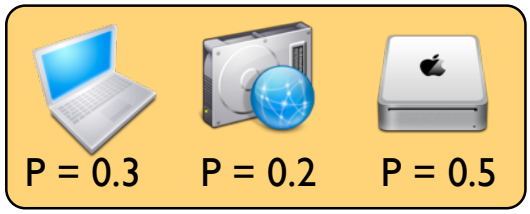
{a} {a,b}

Answer: (a, 0.8), (b, 0.5)

Semantics Of Query Answering

Possible Answers Semantics

Probabilistic DB:



Q Q Q

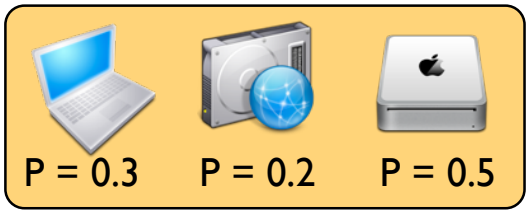
{a} {a,b}

Answer: ({a}, 0.3); ({a,b}, 0.5)

Probability distribution on sets of tuples

Possible Tuples Semantics

Probabilistic DB:



Q Q Q

{a} {a,b}

Answer: (a, 0.8), (b, 0.5)

Probability distribution on tuples

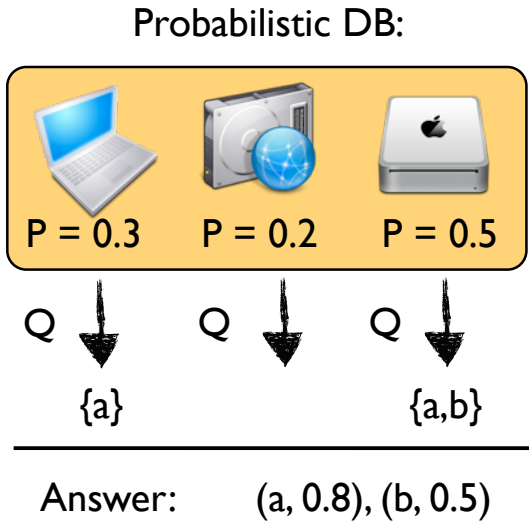
Wednesday, October 26, 2011

Possible Answer vs Possible Tuple Semantics

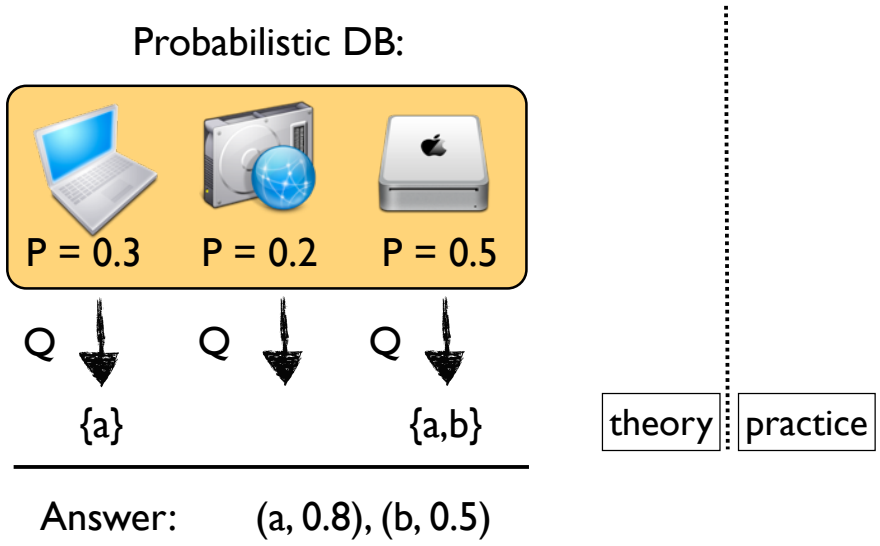
[Dalvi,Suciu'09]

- Possible answers semantics:
  - Precise
  - Can be used to compose queries
  - Difficult user interface
- Possible tuples semantics:
  - Less precise, but simple; sufficient for most apps
  - Cannot be used to compose queries
  - Simple user interface

Goals of Query Answering

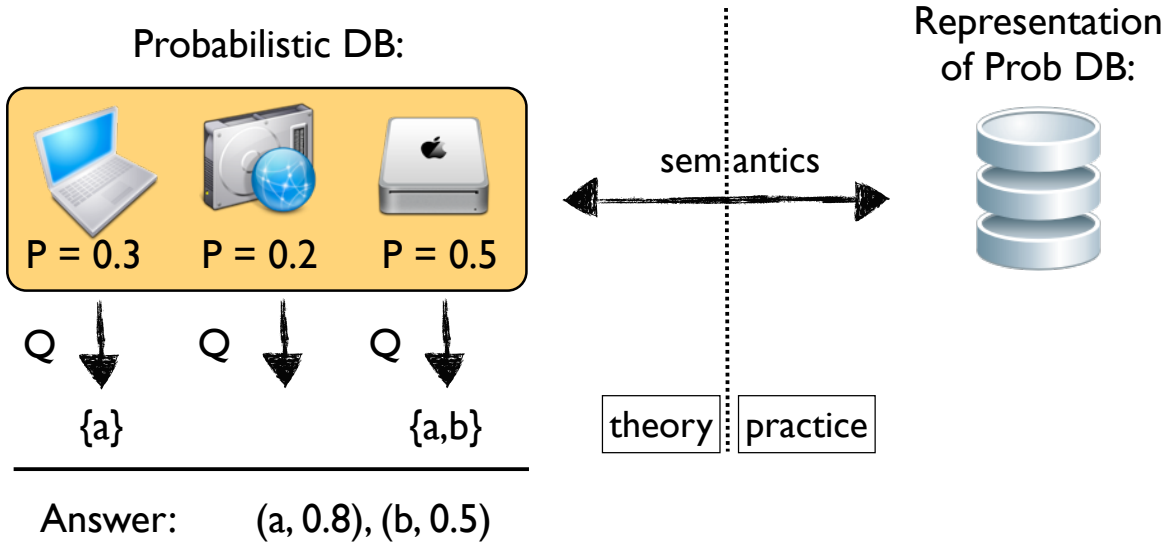


Goals of Query Answering



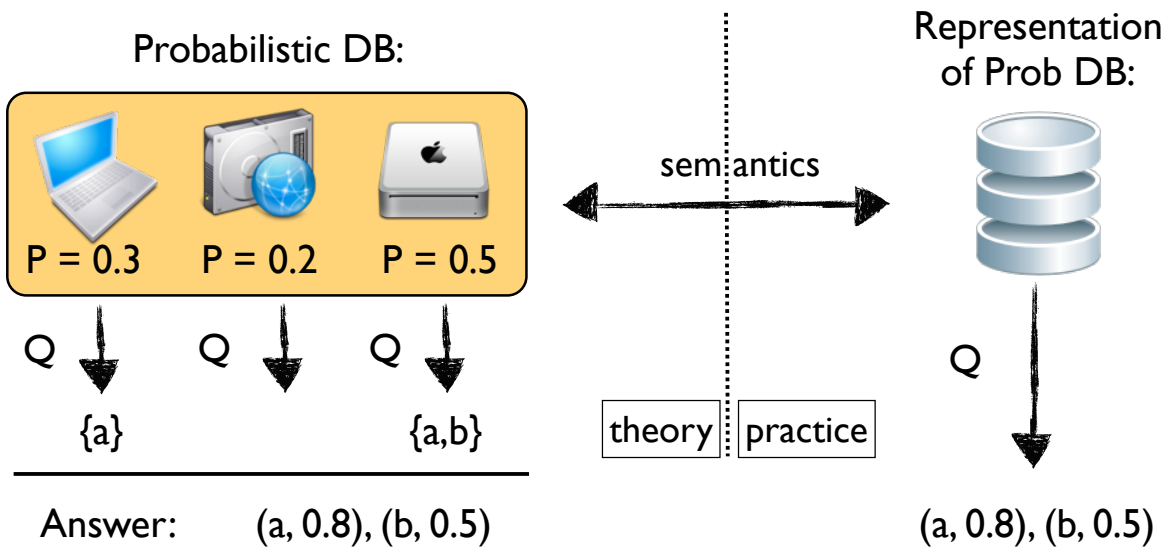
- There may be EXP many worlds ➜ naive evaluation is exponential
- Can we do better?

Goals of Query Answering



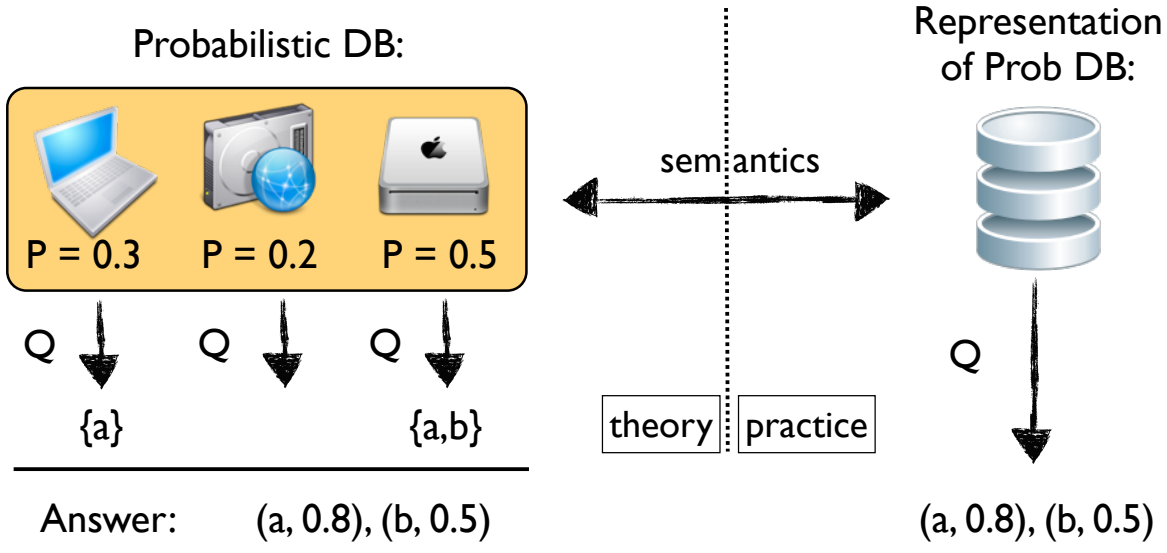
- There may be EXP many worlds ➜ naive evaluation is exponential
- Can we do better?

Goals of Query Answering



- There may be EXP many worlds ➜ naive evaluation is exponential
- Can we do better?

Goals of Query Answering



- There may be EXP many worlds ➜ naive evaluation is exponential
- Can we do better?
- Goal: to find out how to query representation system directly

Part III: Querying Probabilistic Databases

- Semantics
- Lineage computation and #P-Hardness
- Special tractable case within Probabilistic XML

General Lineage: Examples of Operators (1)

Drivers

ID	person	car	Lineage
31	Jimmy	Toyota	$x \wedge y$
32	Jimmy	Honda	y
33	Hank	Honda	$x \vee z$

Saw

ID	witness	car	Lineage
21	Cathy	Honda	w

Pr(x is true) = 0.2 Pr(z is true) = 0.8
Pr(y is true) = 0.4 Pr(w is true) = 0.5

Project = π_{person} (Drives)

Project

person	Lineage
Jimmy	$(x \wedge y) \vee y$
Hank	$x \vee z$

Select = $\sigma_{car="honda"}$ (Drives)

Select

person	car	Lineage
Jimmy	Honda	y
Hank	Honda	$x \vee z$

General Lineage: Examples of Operators (1)

Drivers

ID	person	car	Lineage
31	Jimmy	Toyota	x ∧ y
32	Jimmy	Honda	y
33	Hank	Honda	x ∨ z

Saw

ID	witness	car	Lineage
21	Cathy	Honda	w

Pr(x is true) = 0.2 Pr(z is true) = 0.8
Pr(y is true) = 0.4 Pr(w is true) = 0.5

Project = π_{person} (Drives)

Project

person	Lineage
Jimmy	(x ∧ y) ∨ y
Hank	x ∨ z

Select = $\sigma_{car="honda"}$ (Drives)

Select

person	car	Lineage
Jimmy	Honda	y
Hank	Honda	x ∨ z

General Lineage: Examples of Operators (2)

Drivers

ID	person	car	Lineage
31	Jimmy	Toyota	x ∧ y
32	Jimmy	Honda	y
33	Hank	Honda	x ∨ z

Saw

ID	witness	car	Lineage
21	Cathy	Honda	w

Pr(x is true) = 0.2 Pr(z is true) = 0.8
Pr(y is true) = 0.4 Pr(w is true) = 0.5

Join = Saw $\bowtie_{car}$ Drives

Several = $\pi_{person}(\sigma_{person="Hank"}$ (Saw $\bowtie_{car}$ Drives)

Join

person	car	witness	Lineage
Jimmy	Honda	Cathy	y ∧ w
Hank	Honda	Cathy	(x ∨ z) ∧ w

Several

person	Lineage
Hank	(x ∨ z) ∧ w

General Lineage: Examples of Operators (2)

Drivers

ID	person	car	Lineage
31	Jimmy	Toyota	x ∧ y
32	Jimmy	Honda	y
33	Hank	Honda	x ∨ z

Saw

ID	witness	car	Lineage
21	Cathy	Honda	w

Pr(x is true) = 0.2 Pr(z is true) = 0.8
Pr(y is true) = 0.4 Pr(w is true) = 0.5

Join = Saw $\bowtie_{car}$ Drives

Several = $\pi_{person}(\sigma_{person="Hank"}$ (Saw $\bowtie_{car}$ Drives)

Join

person	car	witness	Lineage
Jimmy	Honda	Cathy	y ∧ w
Hank	Honda	Cathy	(x ∨ z) ∧ w

Several

person	Lineage
Hank	(x ∨ z) ∧ w

General Lineage: Examples of Operators (3)

Saw-day

ID	witness	car	Lineage
31	Cathy	Honda	z
32	Bob	BMW	y ∧ w

Saw-night

ID	witness	car	Lineage
21	Cathy	Honda	w

Pr(x is true) = 0.2 Pr(z is true) = 0.8
Pr(y is true) = 0.4 Pr(w is true) = 0.5

Union = Saw-day ∪ Saw-night

Difference = Saw-day \ Saw-night

Union

witness	car	Lineage
Cathy	Honda	z ∨ w
Bob	BMW	y ∧ w

Difference

witness	car	Lineage
Cathy	Honda	z ∧ (¬w)
Bob	BMW	y ∧ w

General Lineage: Examples of Operators (3)

Saw-day

ID	witness	car	Lineage
31	Cathy	Honda	z
32	Bob	BMW	y ∧ w

Saw-night

ID	witness	car	Lineage
21	Cathy	Honda	w

Pr(x is true) = 0.2 Pr(z is true) = 0.8
Pr(y is true) = 0.4 Pr(w is true) = 0.5

Union = Saw-day ∪ Saw-night

Difference = Saw-day \ Saw-night

Union

witness	car	Lineage
Cathy	Honda	z ∨ w
Bob	BMW	y ∧ w

Difference

witness	car	Lineage
Cathy	Honda	z ∧ (¬w)
Bob	BMW	y ∧ w

Query Probabilities from Lineage

Join = Saw $\bowtie_{car}$ Drives

Pr(x is true) = 0.2 Pr(z is true) = 0.8
Pr(y is true) = 0.4 Pr(w is true) = 0.5

Join

person	car	witness	Lineage
Jimmy	Honda	Cathy	y ∧ w
Hank	Honda	Cathy	(x ∨ z) ∧ w

Theorem: SPJUD-query evaluation over PrRBDs with boolean-formulas lineage is #P-hard, i.e. intractable

Query Probabilities from Lineage

Join = Saw $\bowtie_{car}$ Drives

Pr(x is true) = 0.2 Pr(z is true) = 0.8
Pr(y is true) = 0.4 Pr(w is true) = 0.5

Join

person	car	witness	Lineage
Jimmy	Honda	Cathy	y ∧ w
Hank	Honda	Cathy	(x ∨ z) ∧ w

- Pr(Jimmy ∈ (Saw $\bowtie_{car}$ Drives)) = Pr(y∧w) = Pr(y) x Pr(w) = 0.4 x 0.5 =0.2

Theorem: SPJUD-query evaluation over PrRBDs with boolean-formulas lineage is #P-hard, i.e. intractable

Query Probabilities from Lineage

Join = Saw $\bowtie_{car}$ Drives

Pr(x is true) = 0.2 Pr(z is true) = 0.8
Pr(y is true) = 0.4 Pr(w is true) = 0.5

Join

person	car	witness	Lineage
Jimmy	Honda	Cathy	y ∧ w
Hank	Honda	Cathy	(x ∨ z) ∧ w

- Pr(Jimmy ∈ (Saw $\bowtie_{car}$ Drives)) = Pr(y∧w) = Pr(y) x Pr(w) = 0.4 x 0.5 =0.2
- Pr(Hank ∈ (Saw $\bowtie_{car}$ Drives)) = Pr((x∨z) ∧ w))

  = Pr(x∨z) x Pr (w)
  = [Pr(x) + Pr(z) - Pr(x∧z)] x 0.5
  = [Pr(x) + Pr(z) - Pr(x) x Pr(z)] x 0.5
  = [0.2 + 0.8 - 0.2 x 0.8] x 0.5 = 0.42

Theorem: SPJUD-query evaluation over PrRBDs with boolean-formulas lineage is #P-hard, i.e. intractable

Query Probabilities from Lineage

Join = Saw $\bowtie_{car}$ Drives

Pr(x is true) = 0.2 Pr(z is true) = 0.8
Pr(y is true) = 0.4 Pr(w is true) = 0.5

Join

person	car	witness	Lineage
Jimmy	Honda	Cathy	$y \wedge w$
Hank	Honda	Cathy	$(x \vee z) \wedge w$

- Pr(Jimmy $\in$ (Saw $\bowtie_{car}$ Drives)) = Pr($y \wedge w$) = Pr(y) $\times$ Pr(w) = 0.4 $\times$ 0.5 =0.2
- Pr(Hank $\in$ (Saw $\bowtie_{car}$ Drives)) = Pr($(x \vee z) \wedge w$))

= Pr($x \vee z$) $\times$ Pr (w)
= [Pr(x) + Pr(z) - Pr($x \wedge z$)] $\times$ 0.5
= [Pr(x) + Pr(z) - Pr(x) $\times$ Pr(z)] $\times$ 0.5
= [0.2 + 0.8 - 0.2 $\times$ 0.8] $\times$ 0.5 = 0.42

In general:
Pr(lineage) = Pr(φ)
where φ is a prop. formula

Theorem: SPJUD-query evaluation over PrRBDs with boolean-formulas lineage is #P-hard, i.e. intractable

#P Functions

- Probability computation is a function and not a decision problem
- Usually complexity is studied for decision problems: P(x) = yes/no
- Complexity classes for probability computation are for classes of functions
- #P functions: f(x) = n
  - there is a PTIME non-deterministic Turing machine M_f
  - n = the number of accepting runs of M_f on x, i.e., of $M_f(x)$
- #P functions are counting counterparts of NP decision problems
- Example of #P-complete function:
  #2DNF: count number of evaluations for 2DNF propositional formulas
- #P-comp. functions are counter counterparts of NP-comp. problems

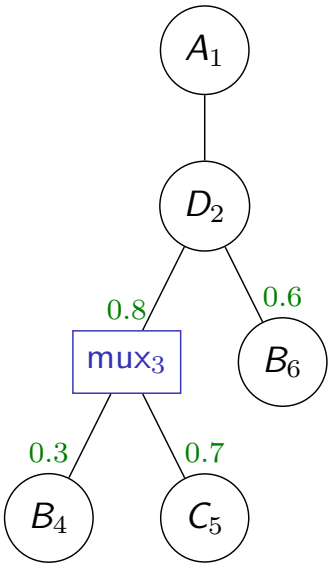
Part III: Querying Probabilistic Databases

- Semantics
- Lineage computation and #P-Hardness
- Special tractable case within Probabilistic XML

Algorithm for TP over local dependencies

[Kimelfeld and Sagiv, 2007]

Bottom-up dynamic programming algorithm. Query: /A//B



	A_1	D_2	mux_3	B_4	C_5	B_6
/B				1	0	1
//B				1	0	1
/A//B				0	0	0

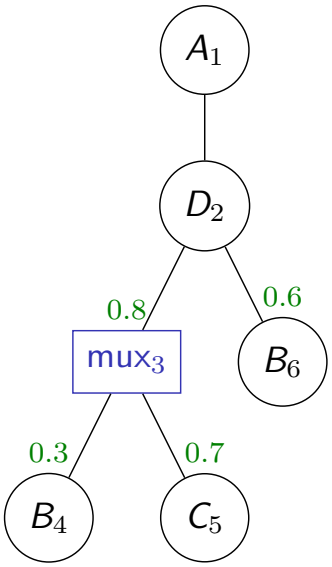
mux convex sum

ordinary inclusion-exclusion

Algorithm for TP over local dependencies

[Kimelfeld and Sagiv, 2007]

Bottom-up dynamic programming algorithm. Query: /A//B



	A_1	D_2	mux$_3$	B_4	C_5	B_6
/B			0.3	I	0	I
//B			0.3	I	0	I
/A//B			0	0	0	0

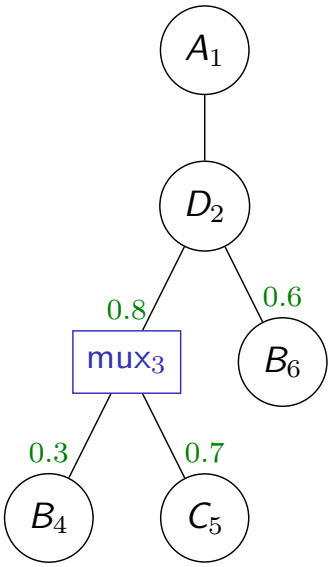
mux convex sum

ordinary inclusion-exclusion

Algorithm for TP over local dependencies

[Kimelfeld and Sagiv, 2007]

Bottom-up dynamic programming algorithm. Query: /A//B



	A_1	D_2	mux$_3$	B_4	C_5	B_6
/B		0	0.3	I	0	I
//B		0.696	0.3	I	0	I
/A//B		0	0	0	0	0

mux convex sum

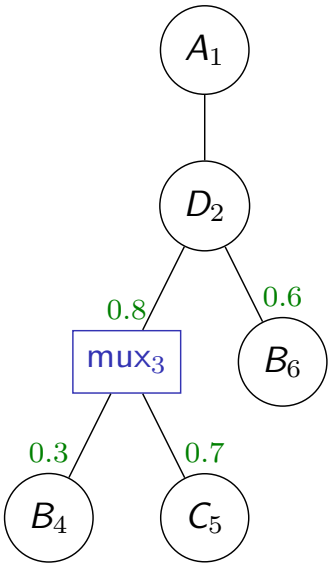
ordinary inclusion-exclusion

$$\Pr(D_2 \models //B) = 1 - (1 - 0.8 \times \Pr(\text{mux}_3 \models /B)) \times (1 - 0.6 \times \Pr(B_6 \models /B))$$
$$= 1 - (1 - 0.8 \times 0.3) \times (1 - 0.6) = 0.696$$

Algorithm for TP over local dependencies

[Kimelfeld and Sagiv, 2007]

Bottom-up dynamic programming algorithm. Query: /A//B



	A_1	D_2	mux_3	B_4	C_5	B_6
/B	0	0	0.3	1	0	1
//B	0.696	0.696	0.3	1	0	1
/A//B	0.696	0	0	0	0	0

mux convex sum

ordinary inclusion-exclusion

Part IV: To go further

Systems

Trio	`http://infolab.stanford.edu/trio/`, useful to see lineage computation
MayBMS	`http://maybms.sourceforge.net/`, full-fledged probabilistic relational DBMS, on top of PostgreSQL, usable for actual applications.
ProApproX	`http://www.infres.enst.fr/~souihli/Publications.html` to play with various approximation and exact query evaluation methods for probabilistic XML.

Reading material

- An influential paper on incomplete databases [Imielinski and Lipski, 1984]
- A book on probabilistic relational databases, focused around TIDs/BIDs and MayBMS [Suciu et al., 2011]
- An in-depth presentation of MayBMS [Koch, 2009]
- A gentle presentation of relational and XML probabilistic models [Kharlamov and Senellart, 2011]
- A survey of probabilistic XML [Kimelfeld and Senellart, 2013]

Merci.

Talel Abdessalem, M. Lamine Ba, and Pierre Senellart. A probabilistic XML merging tool. In *Proc. EDBT*, pages 538–541, Uppsala, Sweden, March 2011. Demonstration.

Serge Abiteboul, Benny Kimelfeld, Yehoshua Sagiv, and Pierre Senellart. On the expressiveness of probabilistic XML models. *VLDB Journal*, 18(5):1041–1064, October 2009.

Serge Abiteboul, Yael Amsterdamer, Daniel Deutch, Tova Milo, and Pierre Senellart. Finding optimal probabilistic generators for XML collections. In *Proc. ICDT*, pages 127–139, Berlin, Germany, March 2012a.

Serge Abiteboul, Yael Amsterdamer, Tova Milo, and Pierre Senellart. Auto-completion learning for XML. In *Proc. SIGMOD*, pages 669–672, Scottsdale, USA, May 2012b. Demonstration.

Antoine Amarilli and Pierre Senellart. On the connections between relational and XML probabilistic data models. In *Proc. BNCOD*, pages 121–134, Oxford, United Kingdom, July 2013.

M. Lamine Ba, Talel Abdessalem, and Pierre Senellart. Uncertain version control in open collaborative editing of tree-structured documents. In *Proc. DocEng*, Florence, Italy, September 2013.

Daniel Barbará, Hector Garcia-Molina, and Daryl Porter. The management of probabilistic data. *IEEE Transactions on Knowledge and Data Engineering*, 4(5):487–502, 1992.

Pablo Barceló, Leonid Libkin, Antonella Poggi, and Cristina Sirangelo. XML with incomplete information: models, properties, and query answering. In *Proc. PODS*, pages 237–246, New York, NY, 2009. ACM.

Michael Benedikt, Evgeny Kharlamov, Dan Olteanu, and Pierre Senellart. Probabilistic XML via Markov chains. *Proceedings of the VLDB Endowment*, 3(1):770–781, September 2010. Presented at the VLDB 2010 conference, Singapore.

Omar Benjelloun, Anish Das Sarma, Alon Y. Halevy, and Jennifer Widom. Uldbs: Databases with uncertainty and lineage. In *VLDB*, pages 953–964, 2006.

Nilesh Dalvi, Chrisopher Ré, and Dan Suciu. Probabilistic databases: Diamonds in the dirt. *Communications of the ACM*, 52(7), 2009.

Nilesh N. Dalvi and Dan Suciu. Efficient query evaluation on probabilistic databases. In *VLDB*, pages 864–875, 2004.

Nilesh N. Dalvi and Dan Suciu. Efficient query evaluation on probabilistic databases. *VLDB Journal*, 16(4), 2007.

Robert Fink, Andrew Hogue, Dan Olteanu, and Swaroop Rath. SPROUT^2: a squared query engine for uncertain web data. In *SIGMOD*, 2011.

José Galindo, Angelica Urrutia, and Mario Piattini. *Fuzzy Databases: Modeling, Design And Implementation*. IGI Global, 2005.

Todd J. Green and Val Tannen. Models for incomplete and probabilistic information. In *Proc. EDBT Workshops, IIDB*, Munich, Germany, March 2006.

Tomasz Imielinski and Witold Lipski. Incomplete information in relational databases. *Journal of the ACM*, 31(4):761–791, 1984.

Evgeny Kharlamov and Pierre Senellart. Modeling, querying, and mining uncertain XML data. In Andrea Tagarelli, editor, *XML Data Mining: Models, Methods, and Applications*. IGI Global, 2011.

Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Updating probabilistic XML. In *Proc. Updates in XML*, Lausanne, Switzerland, March 2010.

B. Kimelfeld and Y. Sagiv. Matching twigs in probabilistic XML. In *Proc. VLDB*, Vienna, Austria, September 2007.

Benny Kimelfeld and Pierre Senellart. Probabilistic XML: Models and complexity. In Zongmin Ma and Li Yan, editors, *Advances in Probabilistic Databases for Uncertain Information Management*, pages 39–66. Springer-Verlag, May 2013.

Benny Kimelfeld, Yuri Kosharovsky, and Yehoshua Sagiv. Query evaluation over probabilistic XML. *VLDB J.*, 2009.

Christoph Koch. MayBMS: A system for managing large uncertain and probabilistic databases. In Charu Aggarwal, editor, *Managing and Mining Uncertain Data*. Springer-Verlag, 2009.

Laks V. S. Lakshmanan, Nicola Leone, Robert B. Ross, and V. S. Subrahmanian. ProbView: A flexible probabilistic database system. *ACM Transactions on Database Systems*, 22(3), 1997.

Dan Olteanu, Jiewen Huang, and Christoph Koch. Approximate confidence computation in probabilistic databases. In *Proc. ICDE*, 2010.

Christopher Ré and Dan Suciu. Materialized views in probabilistic databases: for information exchange and query optimization. In *Proc. VLDB*, 2007.

Pierre Senellart, Avin Mittal, Daniel Muschick, Rémi Gilleron, and Marc Tommasi. Automatic wrapper induction from hidden-Web sources with domain knowledge. In *Proc. WIDM*, pages 9–16, Napa, USA, October 2008.

Asma Souihli and Pierre Senellart. Optimizing approximations of DNF query lineage in probabilistic XML. In *Proc. ICDE*, pages 721–732, Brisbane, Australia, April 2013.

Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch. *Probabilistic Databases*. Morgan & Claypool, 2011.

Jennifer Widom. Trio: A system for integrated management of data, accuracy, and lineage. In *Proc. CIDR*, Asilomar, CA, USA, January 2005.

Lotfi A. Zadeh. A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination. *AI Magazine*, 7(2), 1986.

François Roueff

Professeur au département Traitement du Signal et de l'Image de Télécom ParisTech

Analyse de séries temporelles : modélisation, inférence statistique et application au problème de la prédiction

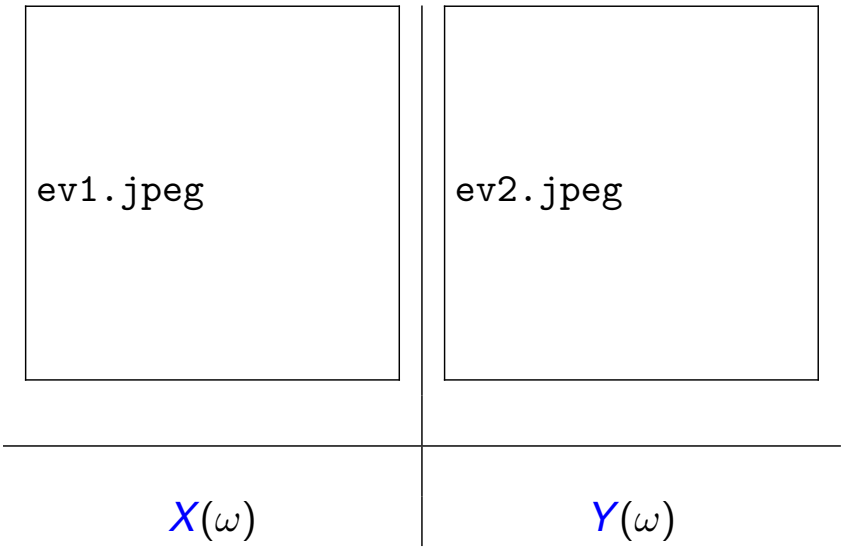
L'analyse des séries temporelles s'est beaucoup développée au cours des dernières décennies avec des applications dans des domaines variés tels que le traitement au signal, l'économétrie, ou la climatologie. Dans ce contexte, le problème de la prédiction est assez facile à poser : ayant observé une suite de nombre jusqu'à un instant donné, à quelles valeurs futures peut-on s'attendre ? La modélisation aléatoire des données fournit un cadre mathématique à la fois rigoureux, intuitif et pratique pour répondre à cette question. Nous donnerons les idées de bases de cette approche et illustrerons leur application pratique.

La multiplication des données stockées disponibles peuvent laisser croire à la possibilité d'une modélisation de plus en plus complexe (et donc plus fine) des séries temporelles. Cette complexification se heurte néanmoins à deux obstacles essentiels pour ce qui est de la prédiction : 1) l'obstacle algorithmique : un algorithme de prédiction se doit d'être exécutable en "temps réel" ; 2) l'obstacle de la non-stationnarité : comment prendre en compte le fait qu'un modèle statistique doive lui-même évoluer au cours du temps par l'effet de facteurs externes ? Nous donnerons quelques approches récentes qui permettent de répondre à ces questions.

Analyse des séries temporelles

Plan

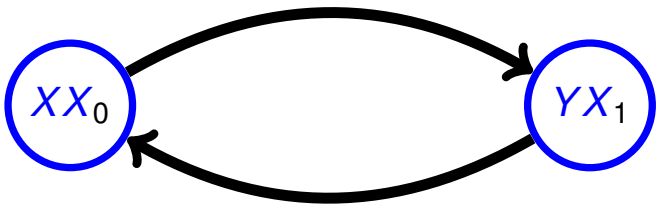
Qu'est-ce que l'information ?



$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)\,?$$

Qu'est-ce que la causalité ?

La notion de causalité n'est pas clairement définie uniquement à partir des probabilités :



$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B \,|\, X \in A)$$

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(Y \in B)\mathbb{P}(X \in A \,|\, Y \in B)$$

$$\mathbb{P}(X_0 \in A, X_1 \in B) = \mathbb{P}(X_0 \in A)\mathbb{P}(X_1 \in B \,|\, X_0 \in A)$$

D'où l'importance du temps $t = 0, 1$!

Applications

L'analyse des séries temporelles reposant sur la modélisation aléatoire trouve de nombreuses applications :

- Santé : analyse de signaux physiologiques (imagerie médicale).
- Ingénierie : surveillance, détection d'anomalies, localisation/poursuite.
- Données audio : analyse de la parole, synthèse, codage.
- Écologie : données climatiques, hydrologie.
- Économétrie : données économiques/financières.
- Assurance : analyse de risques.

Rythme cardiaque

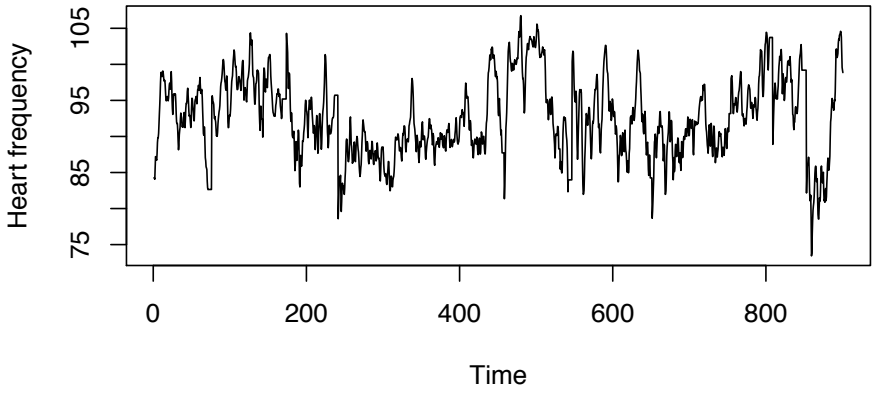


FIGURE: Rythme cardiaque d'une personne au repos au cours de 900 seconds. (nombre de battements par minute évalué toutes les 1/2 sec.)

Trafic Internet

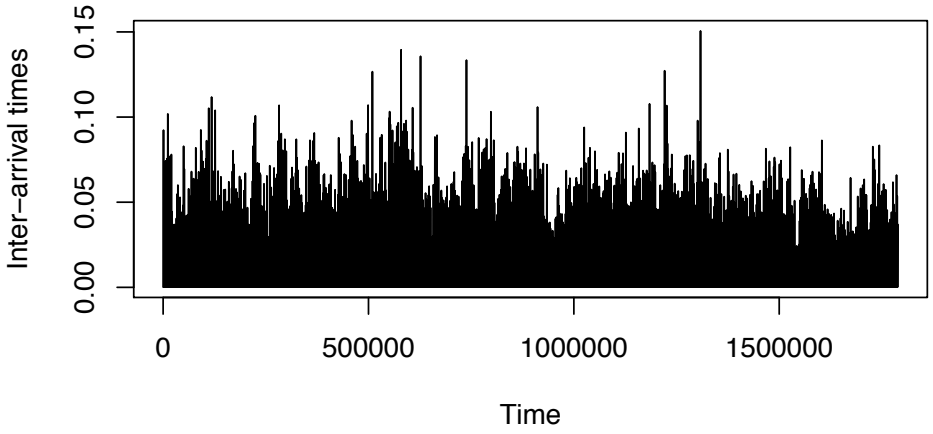


FIGURE: Temps d'inter–arrivées de paquets TCP (en seconde) durant 2 heures de trafic Internet sur un lien `http://ita.ee.lbl.gov/`.

Signal de parole

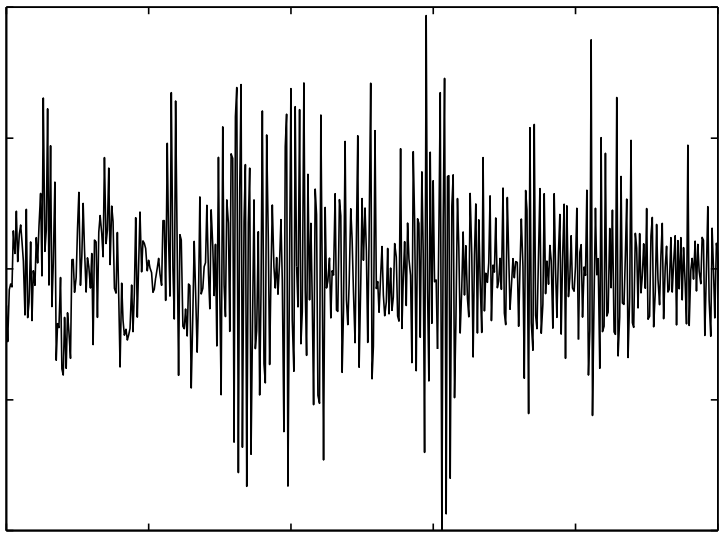


FIGURE: Un signal de parole échantillonné à 8000 Hz. Enregistrement du phonème sh (de sharp).

Données climatiques : vitesse du vent

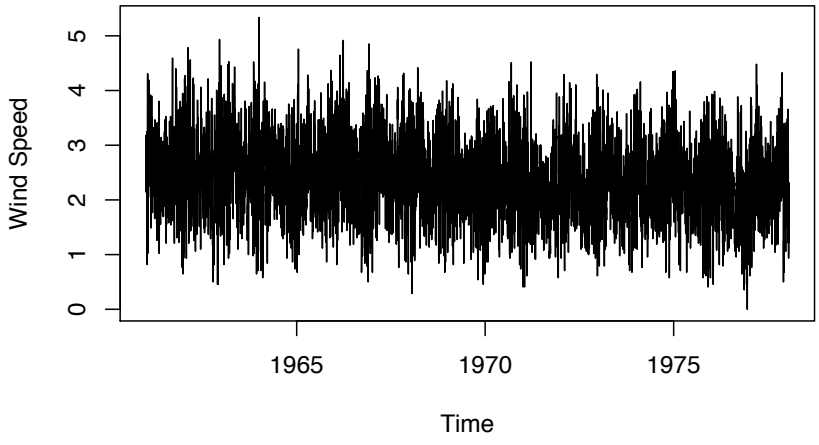


FIGURE: Enregistrement quotidien de la vitesse du vent à Kilkenny (Irlande) en noeuds (1 noeud = 0.5148 metres/second).

Données climatiques : indices de température

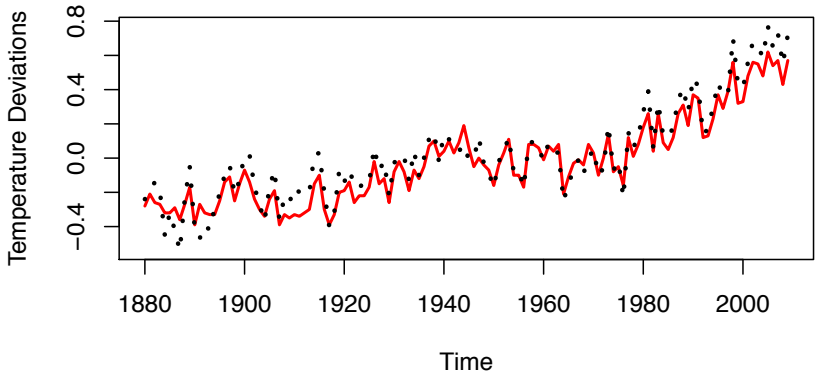


FIGURE: Indices moyens de température terre–océan (ligne rouge pleine) et surface–air (ligne noire pointillée).
`http://data.giss.nasa.gov/gistemp/graphs/`.

Produit national brut

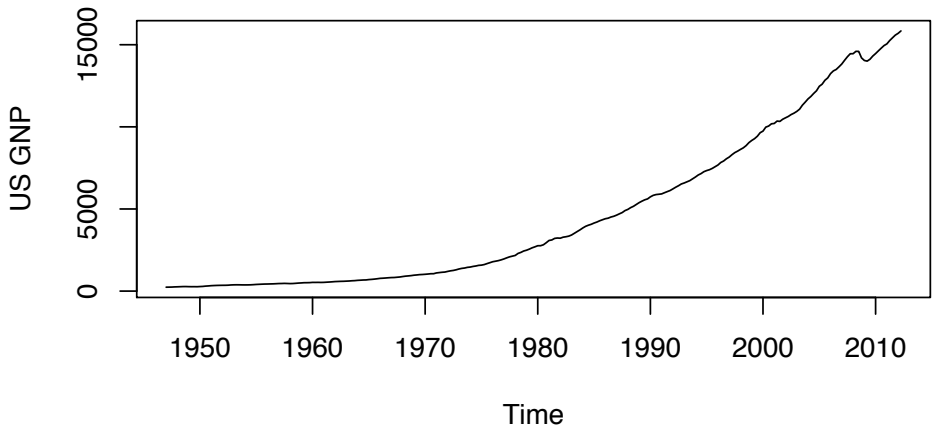


FIGURE: Produit national brut (PNB) des États-Unis en milliards de $s.
`http://research.stlouisfed.org/fred2/series/GNP`.

Taux trimestriel

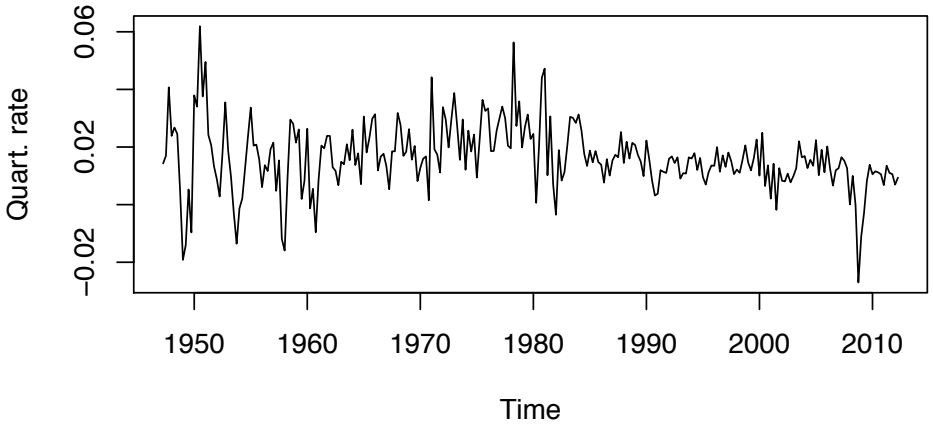


FIGURE: Taux trimestriel du PNB des États-Unis.

Indice financier

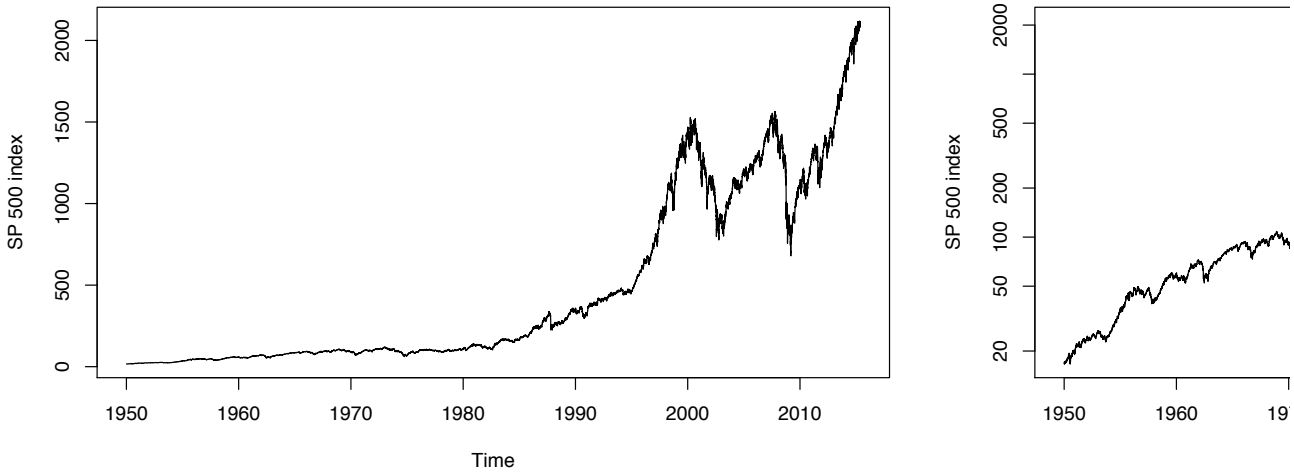


FIGURE: Valeur quotidienne à l'ouverture de l'indice Standard and Poor 500 (New York Stock Exchange (NYSE) et NASDAQ).

Rendements financiers

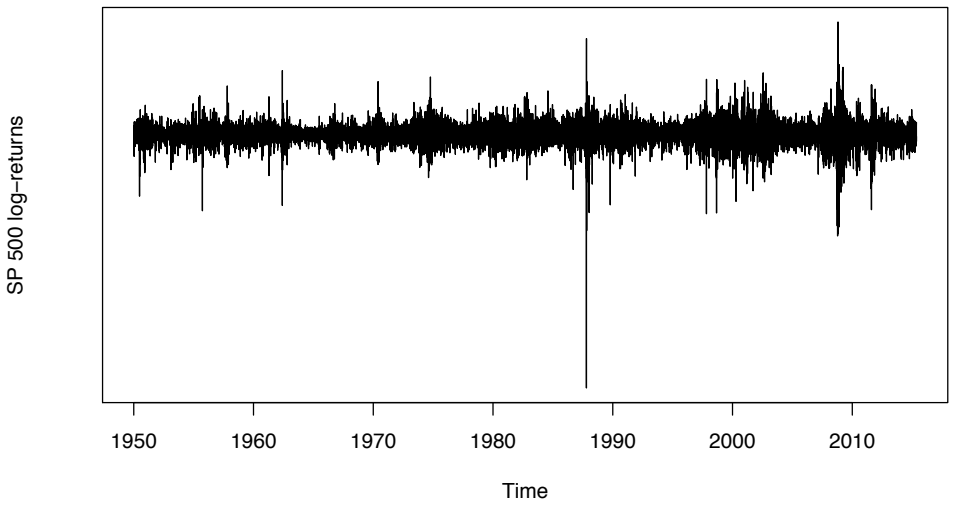


FIGURE: Log–rendements de l'indice SP500.

Buts principaux de l'analyse des séries temporelles

- Modélisation aléatoire : tendance (saisonnière, linéaire, . . .) + bruit (doté de "propriétés structurelles").
- Inférence Statistique : estimer les paramètres du modèle, tester des hypothèses (détecter la présence d'une tendance, d'un signal, classifier des signaux).
- Prédiction : pour un modèle aléatoire donné, utiliser les données du passé pour "deviner" les valeurs futures.
- Filtrage et poursuite : estimer une quantité cachée (observée indirectement) et les suivre au cours du temps.
- Détection d'un changement : découvrir aussi rapidement que possible si la suite de valeurs observées a un comportement statistique qui s'est modifié au cours du temps (détection d'anomalies).

Stationnarité et covariance

- Une série temporelle est modélisé comme la réalisation d'un processus stochastique *i.e.* une suite $X = (X_t)_{t\in\mathbb{Z}}$ de v.a. définies sur le même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. [1]
- L'hypothèse de base en série temporelle est que $(X_t)_{t\in\mathbb{Z}}$ et $(X_{t+1})_{t\in\mathbb{Z}}$ ont la même loi : on dit que X est stationnaire.
- Alors si $X_t \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ (*i.e.* $\mathbb{E}[|X_t|^2] < \infty$), la moyenne $\mu = \mathbb{E}[X_t]$ ne dépend pas de t et la covariance $\gamma(s-t) = \mathrm{Cov}\,(X_s, X_t)$ dépend uniquement de l'écart $s-t$.
- On dit que X est stationnaire au second ordre, si les propriétés du point précédent sont vérifiées (sans nécessairement supposer que X est stationnaire).

1. On peut alors munir $\mathbb{R}^{\mathbb{Z}}$ d'une tribu qui rend $\omega \mapsto X(\omega) = (X_t(\omega))_{t\in\mathbb{Z}}$ mesurable.

Exemples

- Si $X = (X_t)_{t\in\mathbb{Z}}$ est une suite de v.a. i.i.d. centrées et L^2, on dit que X est un bruit blanc fort. Alors $\mu = 0$ et $\gamma(\tau) = 0$ pour $\tau \neq 0$.
- Si X est stationnaire au second ordre avec $\mu = 0$ et $\gamma(\tau) = 0$ pour $\tau \neq 0$, on dit que c'est un bruit blanc faible.
- Si X est stationnaire (au second ordre) alors Y défini par $Y_t = X_t + \sum_{k=1}^{q} \theta_k X_{t-k}\,, \quad t \in \mathbb{Z}$, est stationnaire (au second ordre).
- Si X est stationnaire au second ordre et $|\phi| \neq 1$, alors il existe un unique processus Y stationnaire au second ordre vérifiant $Y_t = \phi Y_{t-1} + X_t\,, t \in \mathbb{Z}$
- En particulier si $|\phi| < 1$, on peut écrire $Y_t = \sum_{k\geq 0} \phi^k X_{t-k}$.

Méthode des moments

- La plupart des modèles repose sur un nombre fini de paramètres $\theta_1, \ldots, \theta_p$ qu'il s'agit d'estimer à partir d'un historique de données $X_1, \ldots, X_n$.
- La méthode des moments consiste à déduire des estimations de $\theta_1, \ldots, \theta_p$ à partir d'estimations de μ et γ, en général :

$$\widehat{\mu}_n = \frac{1}{n} \sum_{k=1}^{n} X_k \ ,$$

$$\widehat{\gamma}_n(t) = \frac{1}{n} \sum_{k=1}^{n-|t|} (X_k - \widehat{\mu}_n) \left(X_{k+|t|} - \widehat{\mu}_n\right) \ .$$

Démo en R

Théorie asymptotique

- Comme X n'est en général pas i.i.d. les théorèmes habituels (loi des grands nombres, théorème de la limite centrale (TLC)) ne s'appliquent pas.
- On peut néanmoins avoir des résultats assez généraux, par exemple :
  - Si $\gamma(t) \to 0$ quand $t \to \infty$, alors $\widehat{\mu}_n \overset{L^2}{\to} \mu$ quand $n \to \infty$.
  - Si $\sum_t |\gamma(t)| < \infty$, alors $\widehat{\mu}_n \overset{\text{p.s.}}{\to} \mu$ quand $n \to \infty$.
  - TLC (si hypothèses plus fortes)...

Du TLC au théorème de Donsker

Soit (X_k) suite de v.a. i.i.d. de moyenne μ et de variance σ^2. On note

$$S_n(t) = \frac{1}{n} \sum_{k=1}^{[nt]} X_k, \quad 0 \le t \le 1 \, .$$

Alors le théorème de la limite centrale théorème de Donsker donne, quand $n \to \infty$,

$$W_n(t) := \frac{\sqrt{n}}{\sigma} (S_n \quad (t) - \ t\mu) \Longrightarrow \mathcal{N}(0,1) W(t)$$

où $W(t)$, $0 \le t \le 1$ est un mouvement brownien.

Comment se débarrasser de la moyenne

Pour se débarrasser de la moyenne μ, on écrit

$$W_n(t) - tW_n(1) \quad \Longrightarrow W(t) - tW(1)$$

$$\| $$

$$\frac{\sqrt{n}}{\sigma}\left(S_n(t) - t\mu\cancel{t\mu}\right) - \cancel{t}t\frac{\sqrt{n}}{\sigma}\left(S_n(1) - \mu\cancel{\mu}\right) \Longrightarrow W(t) - tW(1)$$

$$\| $$

$$\frac{\sqrt{n}}{\sigma}\left(S_n(t) - tS_n(1)\right) \Longrightarrow W(t) - tW(1)$$

$W(t) - tW(1)$, $0 \leq t \leq 1$, s'appelle un pont brownien.

Détection de rupture

Soit $\hat{\sigma}_n$ un estimateur consistant de σ, on en conclut :

$$\sup_{t\in[0,1]}\left|\frac{\sqrt{n}}{\hat{\sigma}_n}\left(S_n(t) - tS_n(1)\right)\right| \Longrightarrow \sup_{t\in[0,1]}\left|W(t) - tW(1)\right| .$$

Si en revanche il y a une rupture dans la moyenne, *i.e.* il existe $\mu_1 \neq \mu_2$ et $0 < r < 1$ tels que

$$\mathbb{E}[X_k] = \begin{cases} \mu_1 & \text{si } 1 \leq k \leq [r\,n] \\ \mu_2 & \text{si } [r\,n] < k \leq n \end{cases},$$

alors on a

$$\sup_{t\in[0,1]}\left|\frac{\sqrt{n}}{\hat{\sigma}_n}\left(S_n(t) - tS_n(1)\right)\right| \sim \frac{\sqrt{n}}{\sigma}(1-r)r(\mu_1 - \mu_2) .$$

Détection de rupture

En effet, on peut écrire

$$\begin{aligned} S_n(r) - rS_n(1) &= (1-r-r)S_n(r) - r\{S_n(1)-S_n(r) - S_n(r)\} \\ &= (1-r)\left(S_n(r)-r\mu_1 - r\mu_1\right) \text{ (d'ordre } 1/\sqrt{n}) \\ &\quad - r\left(\{S_n(1) - S_n(r)-(1-r)\mu_2\} - (1-r)\mu_2\right) \text{ (d'ordr} \\ &\quad + (1-r)r(\mu_1-\mu_2)(1-r)r(\mu_1-\mu_2) \end{aligned}$$

Démo en R

Espérance conditionnelle

La meilleure prédiction de Y sachant X est donné par

$$\widehat{Y} = \mathbb{E}[Y \,|\, X] = \psi(X) \;,$$

où (si $\mathbb{E}[|Y|^2] < \infty$) ψ est la fonction mesurable qui minimise

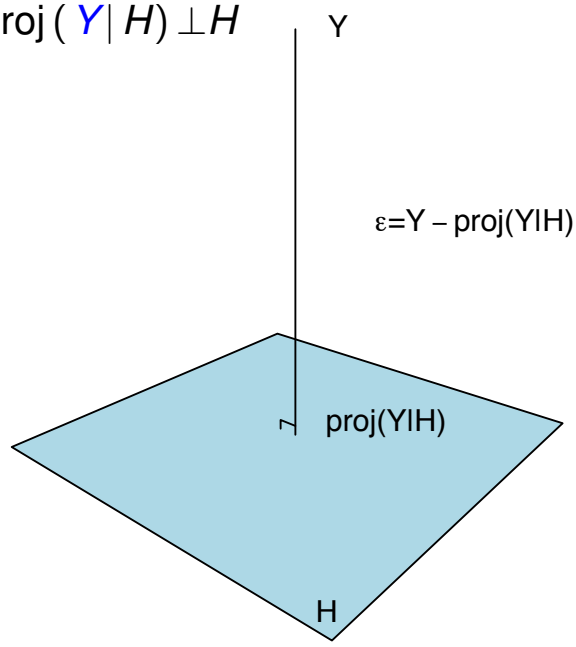
$$\phi \mapsto \mathbb{E}\left[(Y - \phi(X))^2\right] \;.$$

Il s'agit donc d'une projection $\widehat{Y} = \text{proj}\,(Y|\,H)$ sur un sous-espace fermé $H = L^2(\Omega, \sigma(X), \mathbb{P})$ dans l'espace de Hilbert $L^2(\Omega, \mathcal{F}, \mathbb{P})$ muni du produit scalaire

$$\langle X, Y \rangle = \mathbb{E}[XY] \;.$$

Prédiction = projection

(i) $\text{proj}\,(Y|\,H) \in H$
(ii) $\epsilon = Y - \text{proj}\,(Y|\,H) \perp H$



Prédiction linéaire

On peut simplifier la résolution du problème de la prédiction en se contraignant à une prédiction linéaire :

$$H = \text{Vect}\,(1, X) = \{a + bX\ ,\ a, b \in \mathbb{R}\}$$

(en supposant que $\mathbb{E}[|X|^2] < \infty$).

La solution est alors

$$\text{proj}\,(Y | \text{Vect}\,(1, X)) = \mathbb{E}[Y] + \frac{\text{Cov}\,(X, Y)}{\text{Var}\,(X)}(X - \mathbb{E}[X])$$

Définition

Soit $(X_t)_{t\in\mathbb{Z}}$ une série temporelle stationnaire au second ordre centrée de covariance γ. Son passé "lineaire" jusqu'au temps t est défini par

$$\mathcal{H}_t^X = \overline{\text{Vect}}\,(X_s, s \leq t) \ .$$

Prédiction linéaire et processus des innovations

- Le meilleur prédicteur linéaire est défini par

$$\text{proj}\left(X_t | \mathcal{H}_{t-1}^X\right) = \arg \min_{Y\in\mathcal{H}_{t-1}^X} \mathbb{E}\left[|X_t - Y|^2\right] \ .$$

- Le processus des innovations $(\epsilon_t)_{t\in\mathbb{Z}}$ de X est défini par

$$\epsilon_t = X_t - \text{proj}\left(X_t | \mathcal{H}_{t-1}^X\right), \qquad t \in \mathbb{Z} \ .$$

Propriétés

On montre que $(\epsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc faible.

Étape 1 On a par définition $\text{proj}\left(X_t | \mathcal{H}_{t-1}^X\right) \in \mathcal{H}_{t-1}^X$, donc $\epsilon_t \in \mathcal{H}_t^X$.

Étape 2 En particulier $\mathbb{E}[\epsilon_t] = 0$.

Étape 3 De plus, pour tout $s < t$, on a alors $\epsilon_s \in \mathcal{H}_s^X \subseteq \mathcal{H}_{t-1}^X \perp \epsilon_t$. Donc $\langle \epsilon_s, \epsilon_t \rangle = 0$.

Étape 4 Il reste à montrer que $\text{Var}\,(\epsilon_t)$ ne dépend pas de t. Cela vient du fait que $(X_t)_{t\in\mathbb{Z}}$ et $(X_{t+1})_{t\in\mathbb{Z}}$ ont les mêmes covariances.

Exemples

- $X = (X_t)_{t\in\mathbb{Z}}$ est un bruit blanc faible si et seulement si $X = \epsilon$.
- Soit Z un bruit blanc faible et $|\phi| \neq 1$, et soit X l'unique processus stationnaire au second ordre vérifiant

$$X_t = \phi X_{t-1} + Z_t \ , t \in \mathbb{Z}$$

  Alors, si $|\phi| < 1$, $\epsilon = Z$.
- En effet, $\phi X_{t-1} \in \mathcal{H}^X_{t-1}$ et comme on peut écrire $X_t = \sum_{k\geq 0} \phi^k Z_{t-k}$ on a $\mathcal{H}^X_{t-1} \subseteq \mathcal{H}^Z_{t-1} \perp Z_t$.
- Mais c'est faux si $|\phi| > 1$!

Prédiction en pratique

On dispose d'une suite $X_1, \ldots, X_T$.

Étape 1 Modélisation (comment les données sont générées ?)

Étape 2 Estimation des paramètres du modèle.

Étape 3 Calcul du meilleur prédicteur correspondant.

Un petit plus : agrégation de N prédicteurs

Soient N prédicteurs $(\hat{X}_t^{(i)})_{t=1\ldots,T}$, $i = 1, \ldots, N$. Soient pour un $\eta > 0$ et à tout instant $t = 1, \ldots, T$, les poids

$$\hat{\alpha}_t^{(i)} \propto \mathrm{e}^{-\eta \sum_{s=1}^{t-1} (\hat{X}_s^{(i)} - X_s)^2} \text{ qui somment à 1.}$$

Le prédicteur agrégé est défini par $\hat{X}_t = \sum_{i=1}^{N} \hat{\alpha}_t^{(i)} \hat{X}_t^{(i)}, \quad t = 1, \ldots, T$.

On fait l'hypothèse que $|\hat{X}_t^{(i)} - X_t| \leq C$ pour tout t et tout i. Alors, on peut trouver η suffisamment petit (en fonction de C) tel que

$$\frac{1}{T} \sum_{t=1}^{T} \left(\hat{X}_t - X_t \right)^2 \leq \inf_{1 \leq i \leq N} \frac{1}{T} \sum_{t=1}^{T} \left(\hat{X}_t^{(i)} - X_t \right)^2 + \frac{\ln N}{\eta T} .$$

Antonio Casilli

Maître de conférences en humanités numériques à Télécom ParisTech

Etudier les troubles psychiques de l'adolescence à travers l'analyse des réseaux sociaux

Pendant longtemps, il a été impossible d'obtenir des données de qualité sur la fréquentation des sites web et des communautés de personnes atteintes de troubles mentaux. A cause de leur caractère sensible, ces contenus (textes, photos, témoignages…) sont souvent cachés et leur analyse s'avère problématique sur un plan éthique et légal. Mais en passant pasr l'étude des réseaux de socialisation d'adolescents et jeunes adultes, des enquêtes permettent désormais une connaissance plus fine des structures de soutien, des pratiques et des usages numériques de ces individus.

En nous penchant sur le projet ANAMIA développé par des chercheurs de l'Institut Mines Télécom pour étudier les troubles de conduites alimentaires sur Internet, nous montrerons comment la recherche contemporaine se donne les moyens d'obtenir des résultats surprenants. Des méthodologies innovantes (simulations informatiques, visualisations de données, collecte dynamiques de réseaux personnels) permettent désormais de déjouer certaines de nos idées reçues à propos de la santé, de l'isolement social, de la pathologie, de la liberté de soin et des droits des patients.

Etudier les troubles psychiques de l'adolescence à travers l'analyse des réseaux sociaux

Antonio A. CASILLI

Telecom ParisTech

1 avril 2015

English OK

ASSEMBLÉE NATIONALE

Les députés | Dans l'Hémicycle | Commissions et autres instances | Documents parlementaires | Europe et international | Découvrir l'Assemblée | Informations pratiques

Accueil > Documents parlementaires > Amendements

Version PDF | Dossier législatif | Texte de référence | Compte rendu

APRÈS ART. 5 QUATER *N°1052*

ASSEMBLÉE NATIONALE

27 mars 2015

SANTÉ - (N° 2673)

Commission	
Gouvernement	

ADOPTÉ

AMENDEMENT N°1052

présenté par

Mme Olivier, Mme Coutelle, Mme Lemorton, Mme Hurel, Mme Untermaier, Mme Mazetier, M. Rouillard, Mme Clergeau, Mme Lacuey, M. Denaja, Mme Carrey-Conte, M. Ferrand, M. Premat, Mme Récalde, M. Cresta, Mme Zanetti, M. Bies, M. Ménard, M. Le Roux, Mme Fabre, Mme Povéda, M. Vignal, M. Roig, M. Assaf, Mme Pochon, M. Dufau, Mme Maquet, M. Delcourt, M. Bays, M. Marsac, Mme Martinel, M. Bardy, M. Jaltot, Mme Alaux, Mme Sandrine Doucet, Mme Descamps-Crosnier, M. Arif, M. Blazy, Mme Françoise Dubois, Mme Santais, M. Olive, Mme Troallic, Mme Gueugneau, Mme Imbert, Mme Tallard, M. Kalinowski et Mme Dessus

---

ARTICLE ADDITIONNEL

APRÈS L'ARTICLE 5 QUATER, insérer l'article suivant:

I. – La section 1 du chapitre III du titre II du livre II du code pénal est complétée par un article 223-2-1 ainsi rédigé :

« *Art. 223-2-1.* – Le fait de provoquer une personne à rechercher une maigreur excessive en encourageant des restrictions alimentaires prolongées ayant pour effet de l'exposer à un danger de mort ou de compromettre directement sa santé est puni d'un an d'emprisonnement et de 10 000 € d'amende. » ;

II – Le livre II *bis* de la troisième partie du code de la santé publique est complété par un titre II ainsi rédigé :

« Titre II

« Lutte contre la maigreur excessive

« *Art.* L. 3233. – Le fait de provoquer directement une personne à rechercher une maigreur excessive est réprimé par l'article 223-3 du code pénal. »

EXPOSÉ SOMMAIRE

Cet amendement vise à combattre les troubles alimentaires tels que l'anorexie ou la boulimie développant une disposition visant à réprimer l'incitation à la maigreur excessive.

Le phénomène pro-ana

Pro-ana ?

TIME
IN PARTNERSHIP WITH CNN
Edward Jones
SEARCH TIME.COM
Health & Science
Main • Ecocentric Blog • Going Green • Er

Anorexia Goes High Tech

By JESSICA REAVES Tuesday, Jul. 31, 2001

A pro-anorexia web site

PRINT EMAIL REPRINTS MORE

20 minutes.fr

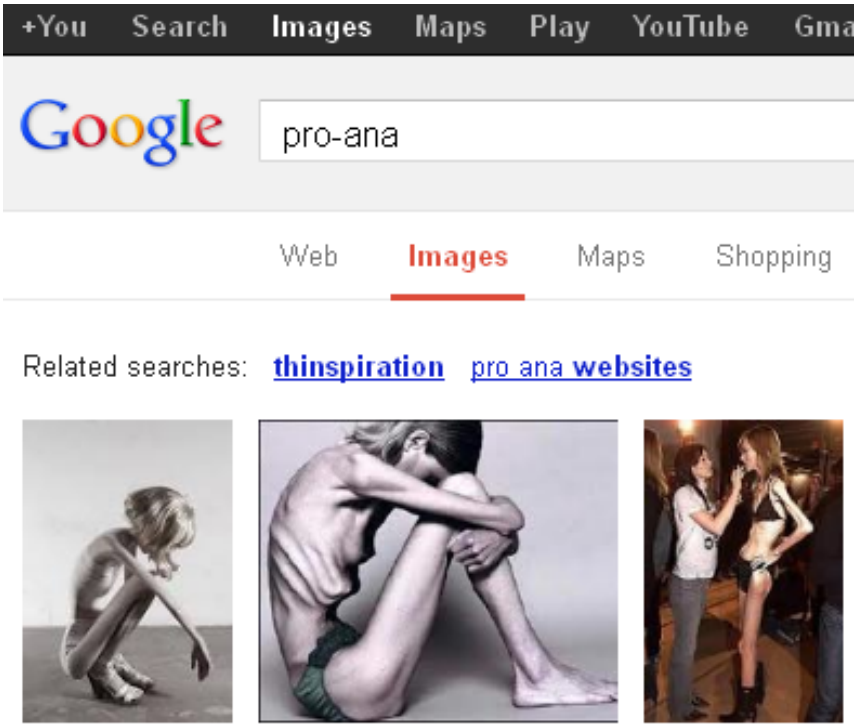
MARDI 15 AVRIL 2008 mise à jour 08h03

accueil | 20' à la seconde | sports | high-tech

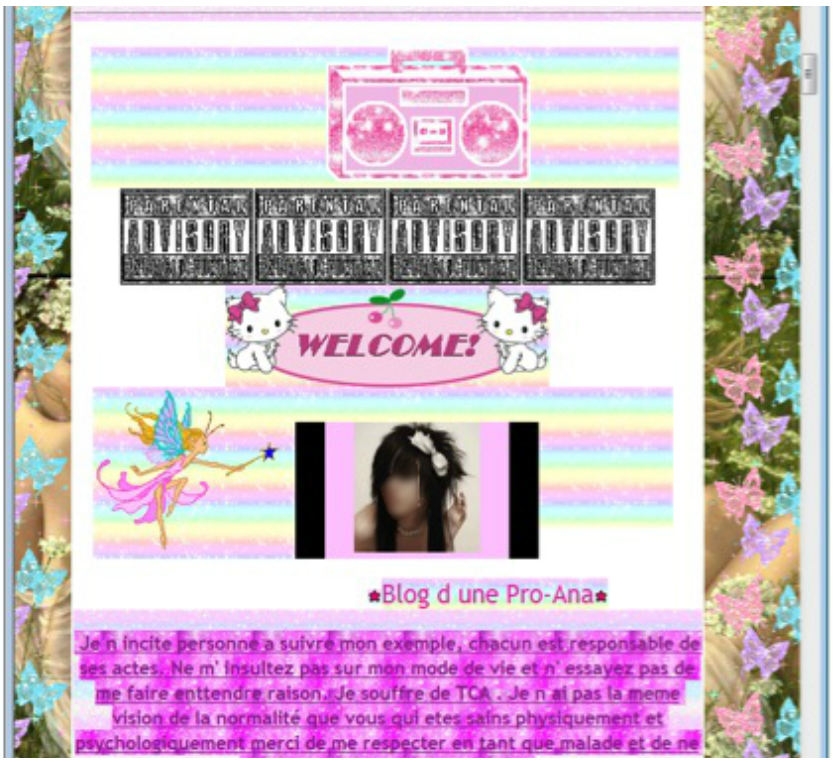
Les blogs pro-anorexie au pilori

Joel Saget AFP/Archives | *Une jeune femme anorexique allongée sur le lit de sa chambre à l'hôpital Sainte-Anne, le 15 février 2007 à Paris.*

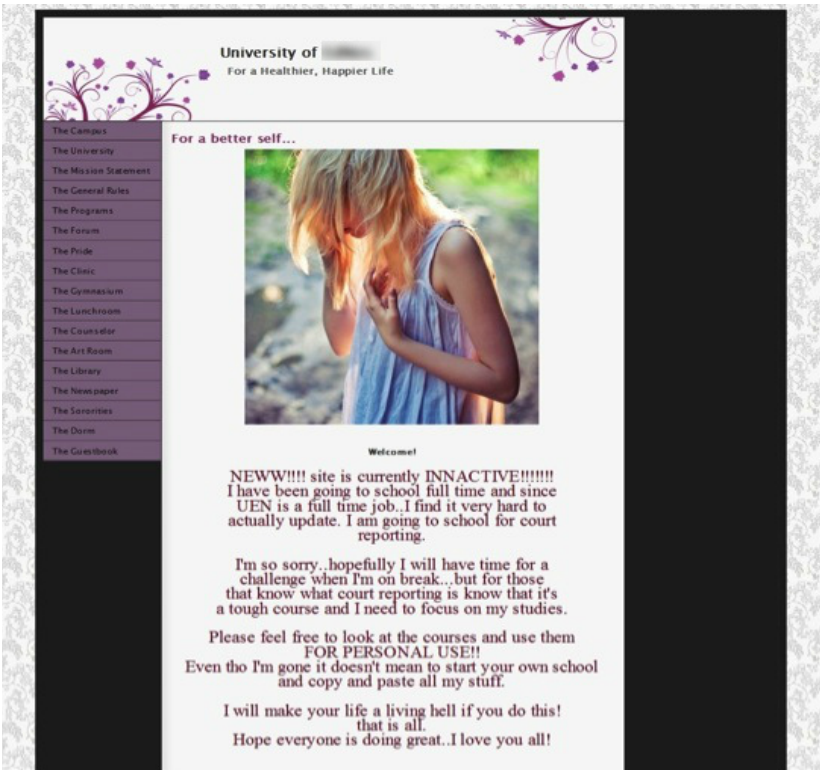
Pro-ana ?



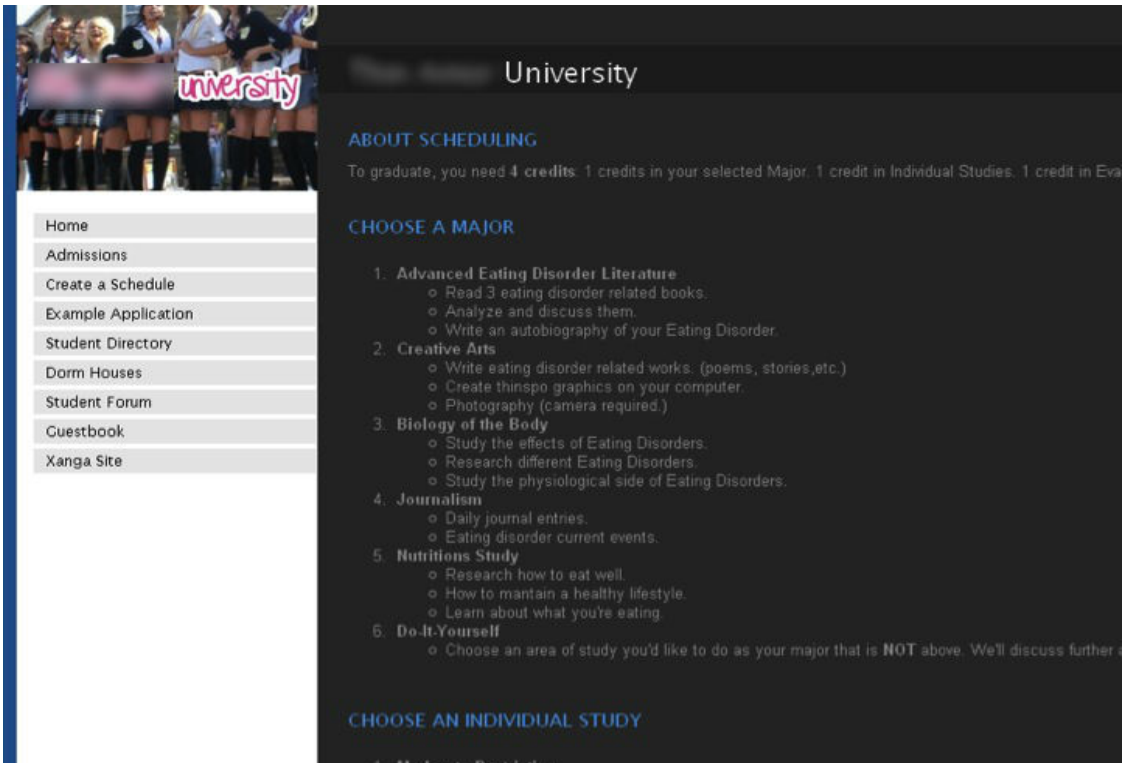
Pro-ana ?



Pro-ana ?

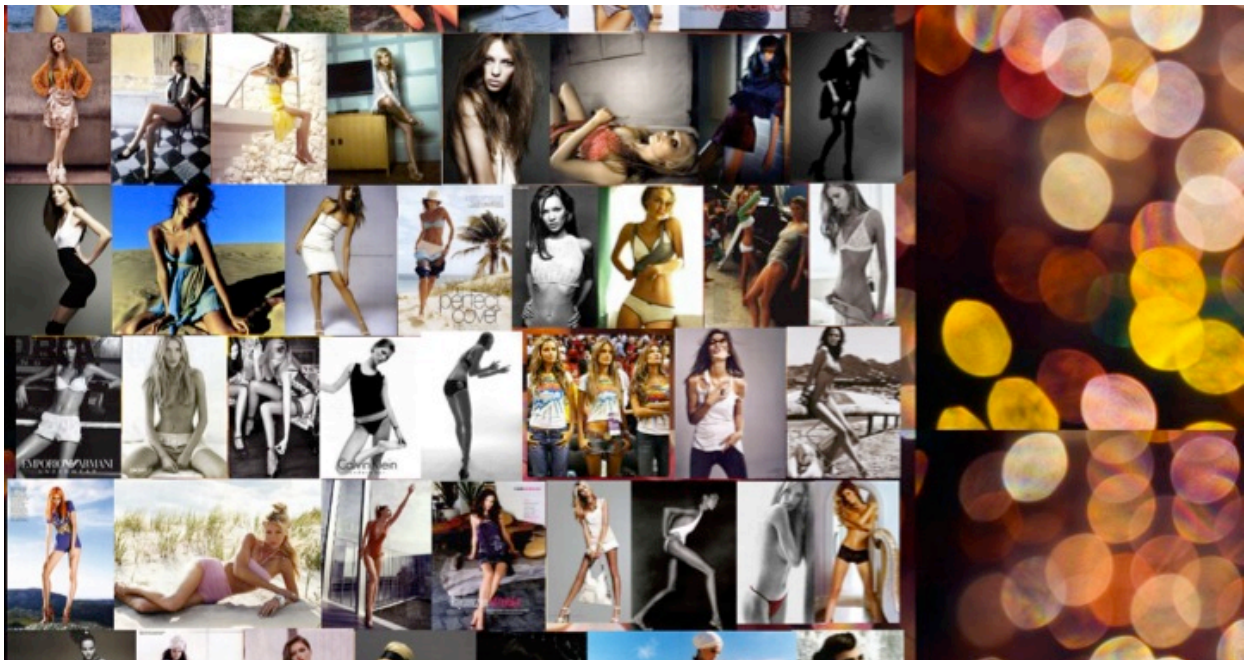


Pro-ana ?



Pro-ana ?

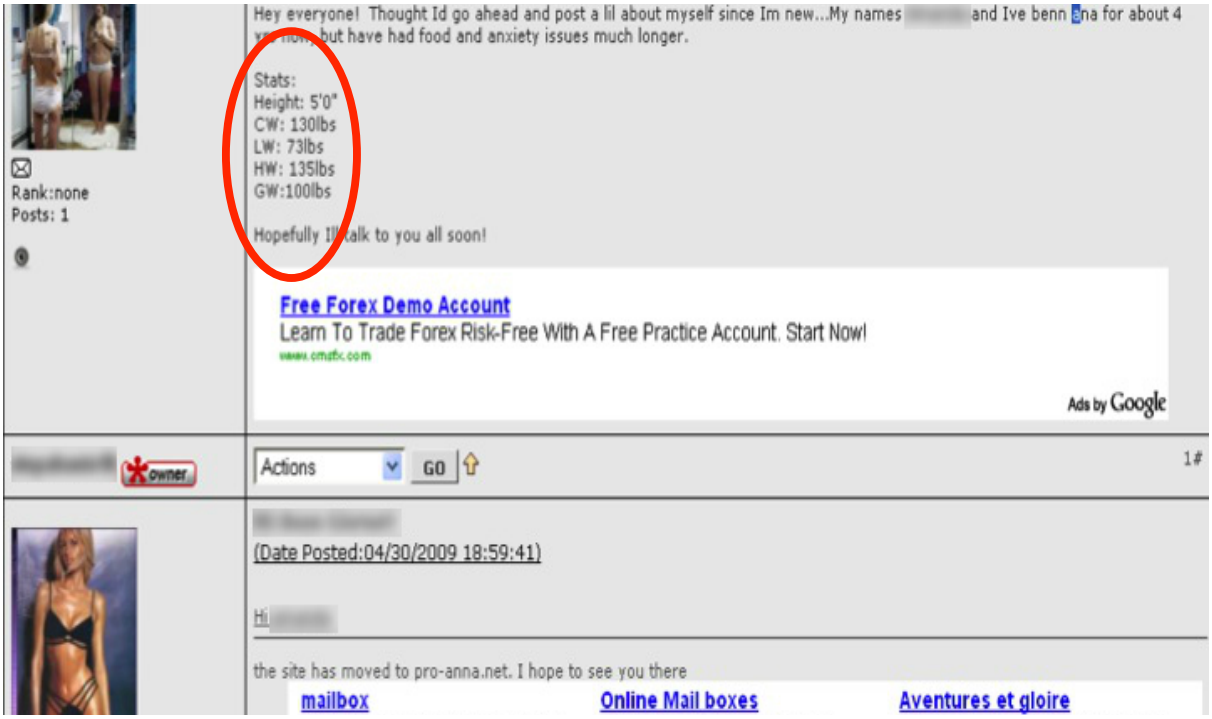
Pro-ana ?



Pro-ana ?

Pro-ana ?

Pro-ana ?

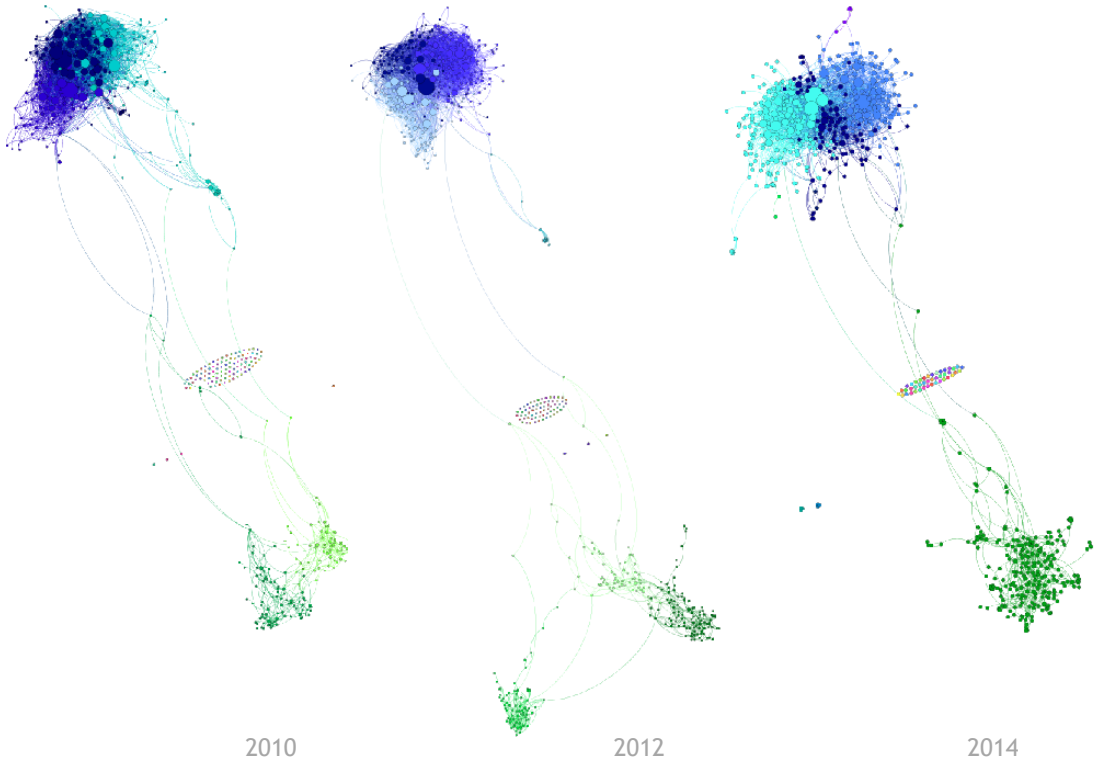


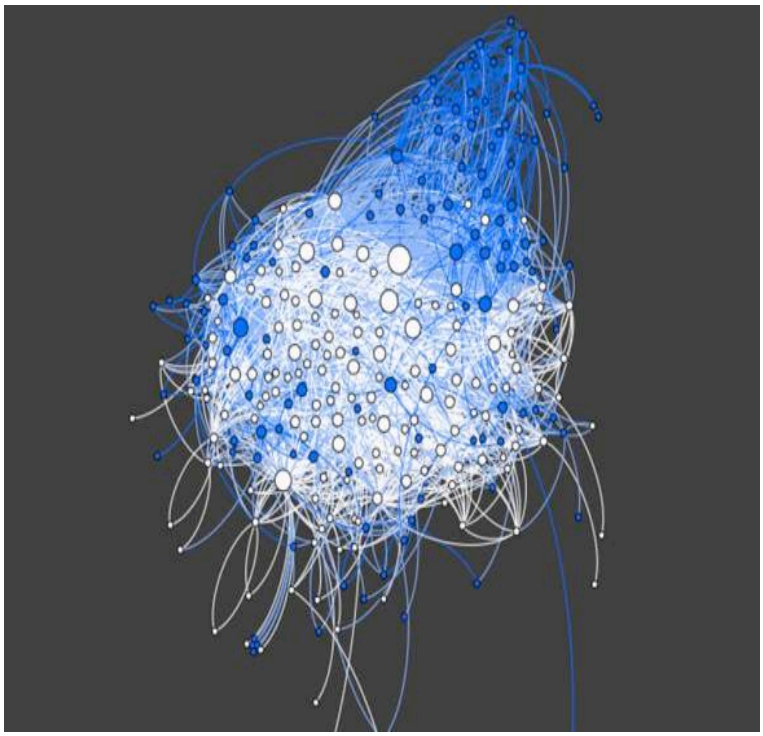
Le projet interdisciplinaire ANAMIA

Anamia

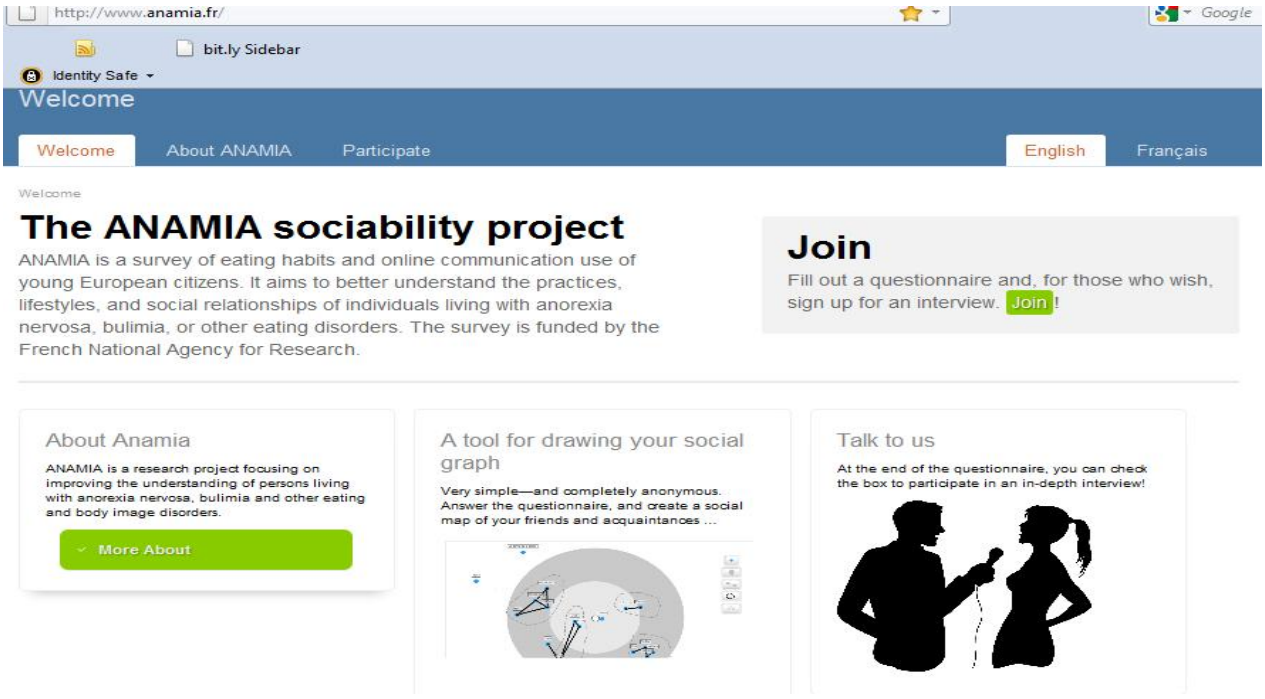
Les jeunes et le Web des troubles alimentaires – http://www.anamia.fr

Anamia

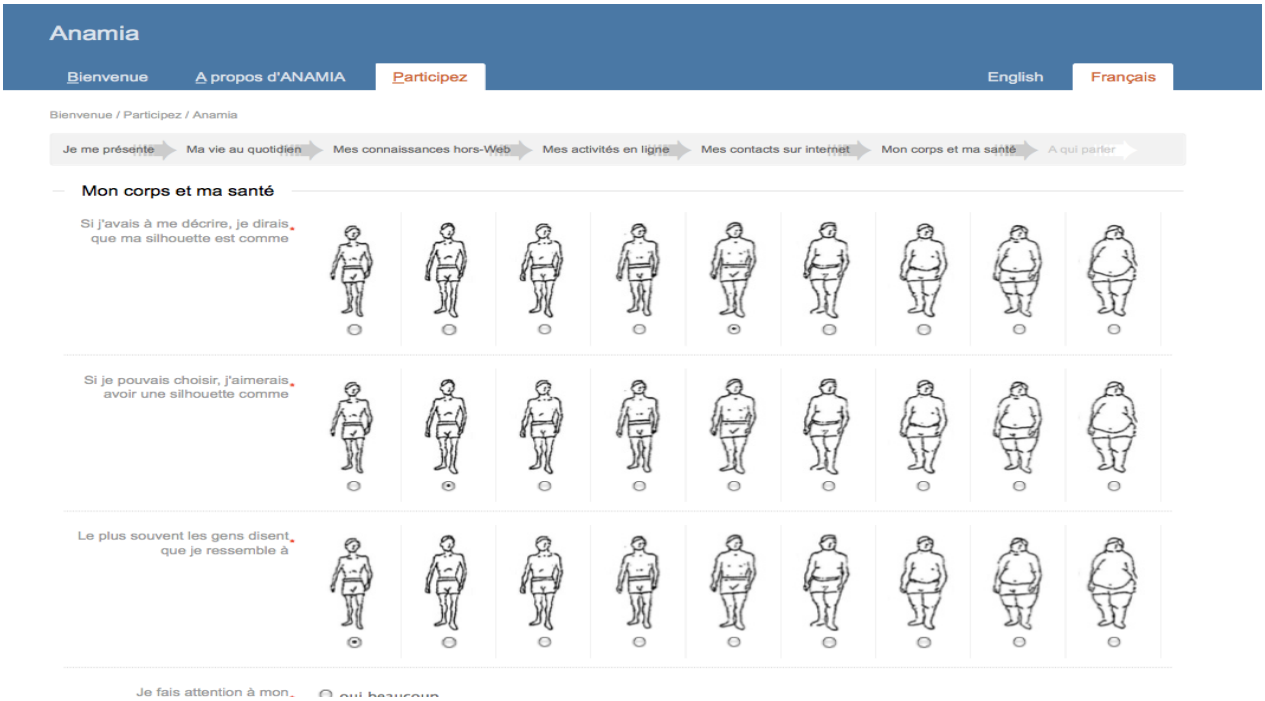




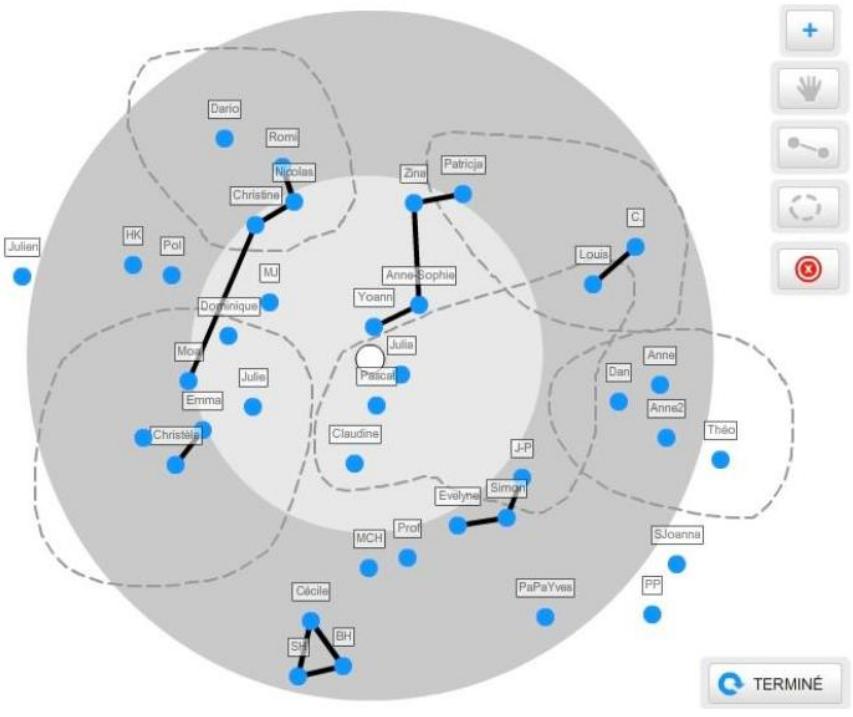
Anamia


http://www.anamia.fr/
Google
bit.ly Sidebar
Identity Safe
Welcome
Welcome
About ANAMIA
Participate
English
Français
Welcome
The ANAMIA sociability project
ANAMIA is a survey of eating habits and online communication use of young European citizens. It aims to better understand the practices, lifestyles, and social relationships of individuals living with anorexia nervosa, bulimia, or other eating disorders. The survey is funded by the French National Agency for Research.
Join
Fill out a questionnaire and, for those who wish, sign up for an interview. Join !
About Anamia
ANAMIA is a research project focusing on improving the understanding of persons living with anorexia nervosa, bulimia and other eating and body image disorders.
More About
A tool for drawing your social graph
Very simple—and completely anonymous. Answer the questionnaire, and create a social map of your friends and acquaintances ...
Talk to us
At the end of the questionnaire, you can check the box to participate in an in-depth interview!

Anamia



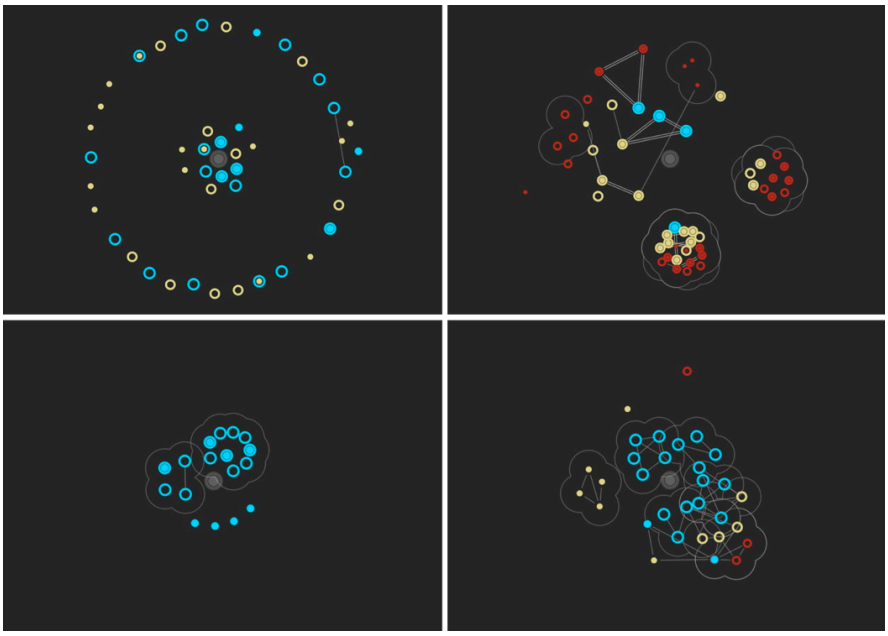
Anamia



Anamia

ED NOS

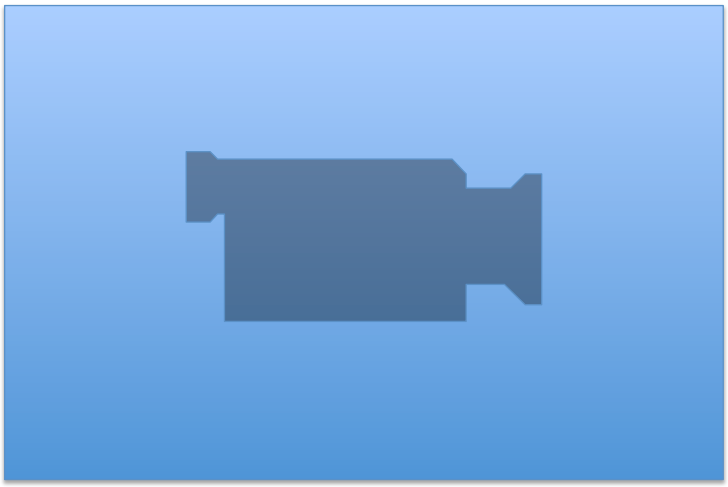
Anorexia



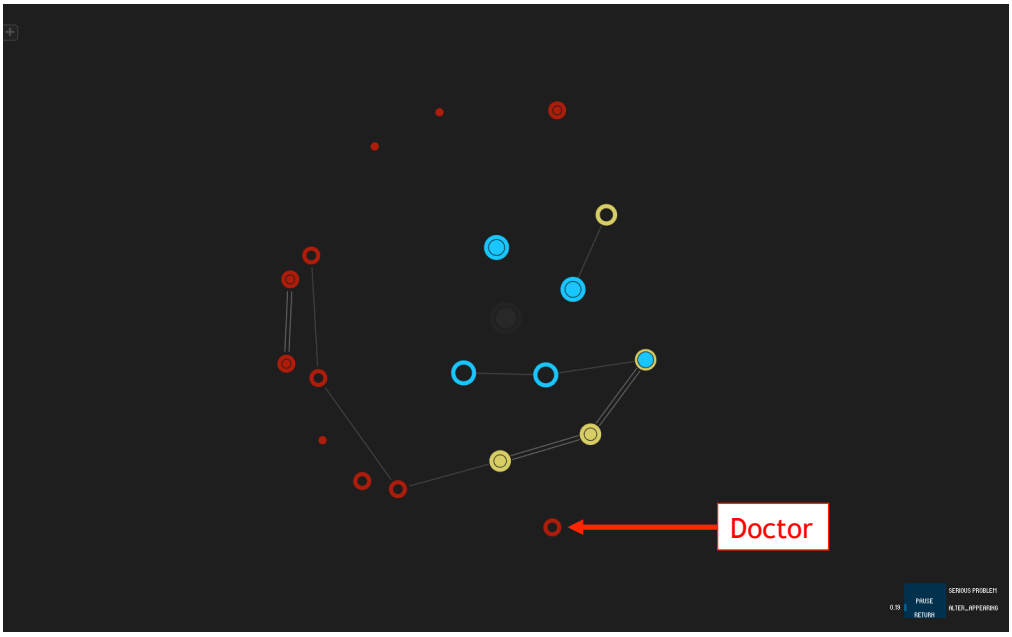
Bulimia

Binge Eating

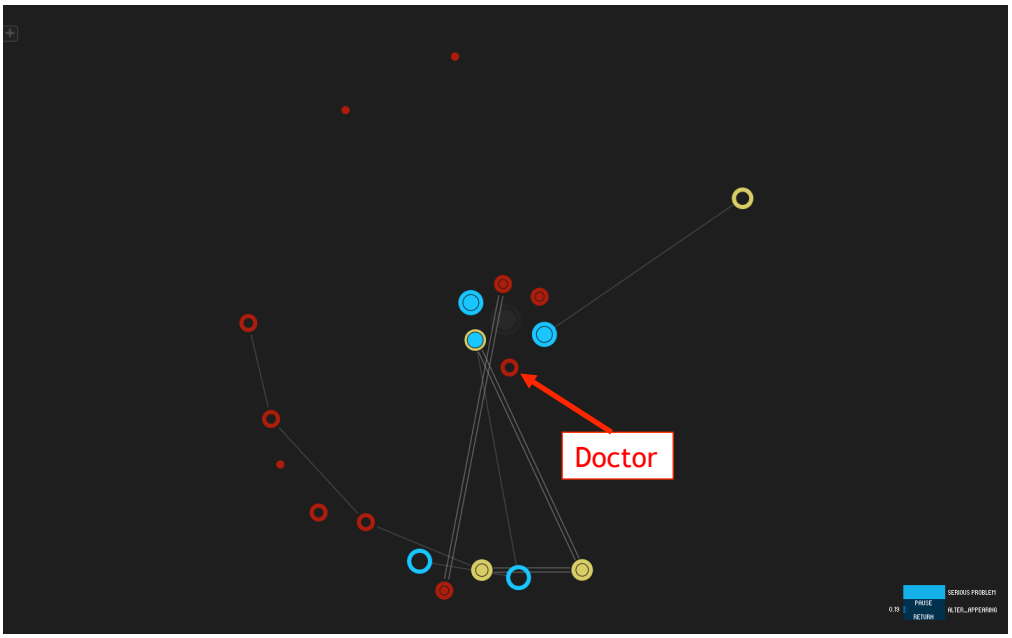
Anamia



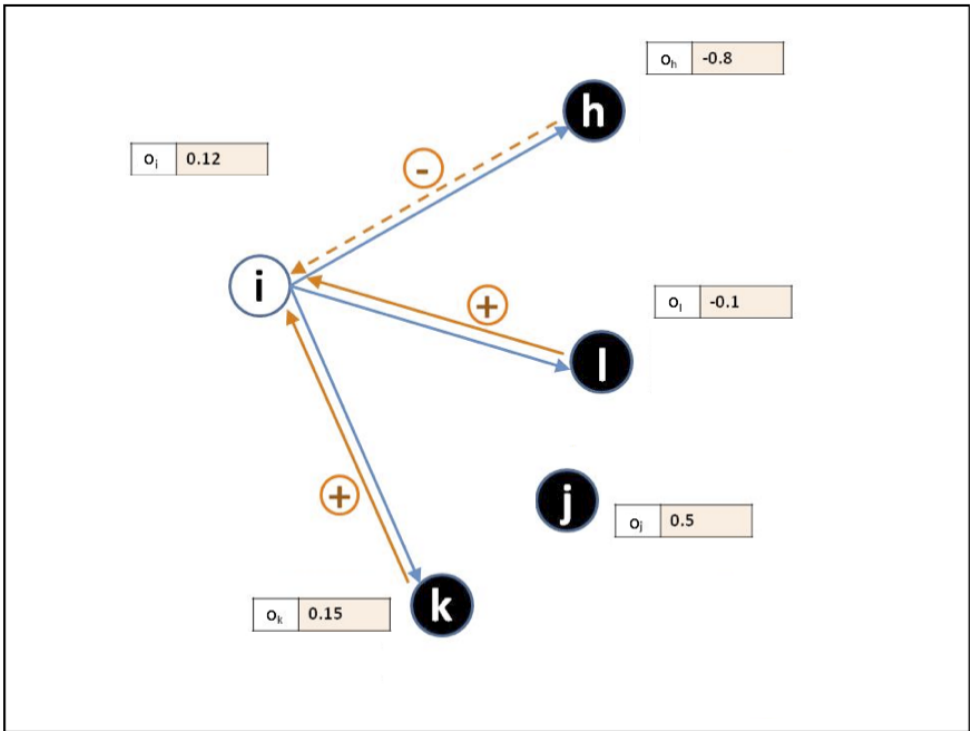
Anamia



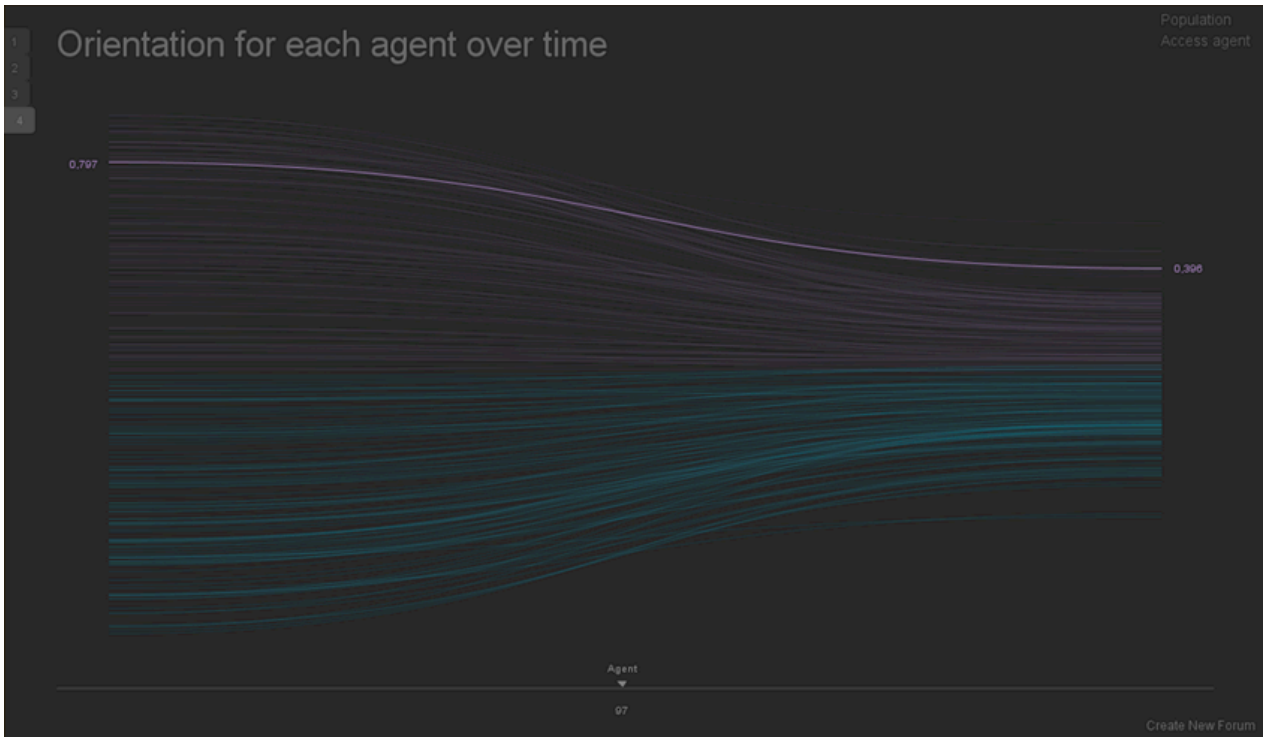
Anamia



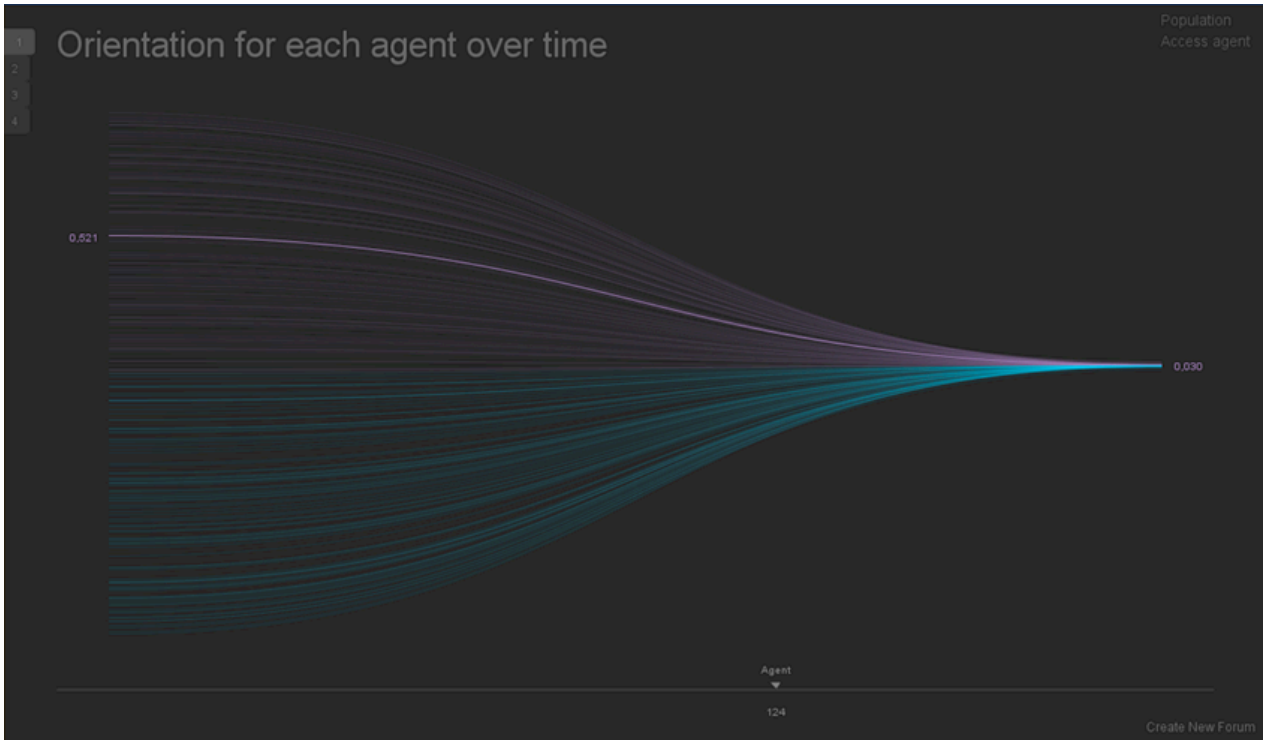
Radicalisation ?



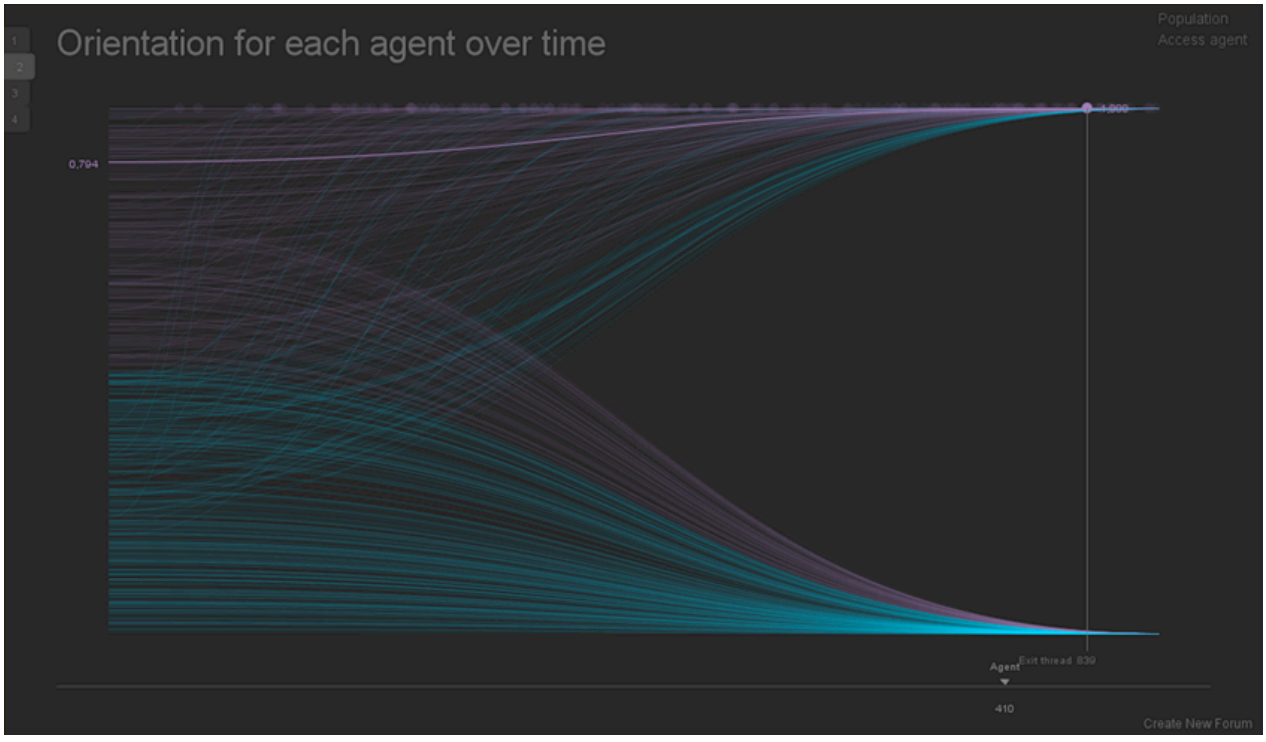
Radicalisation ?



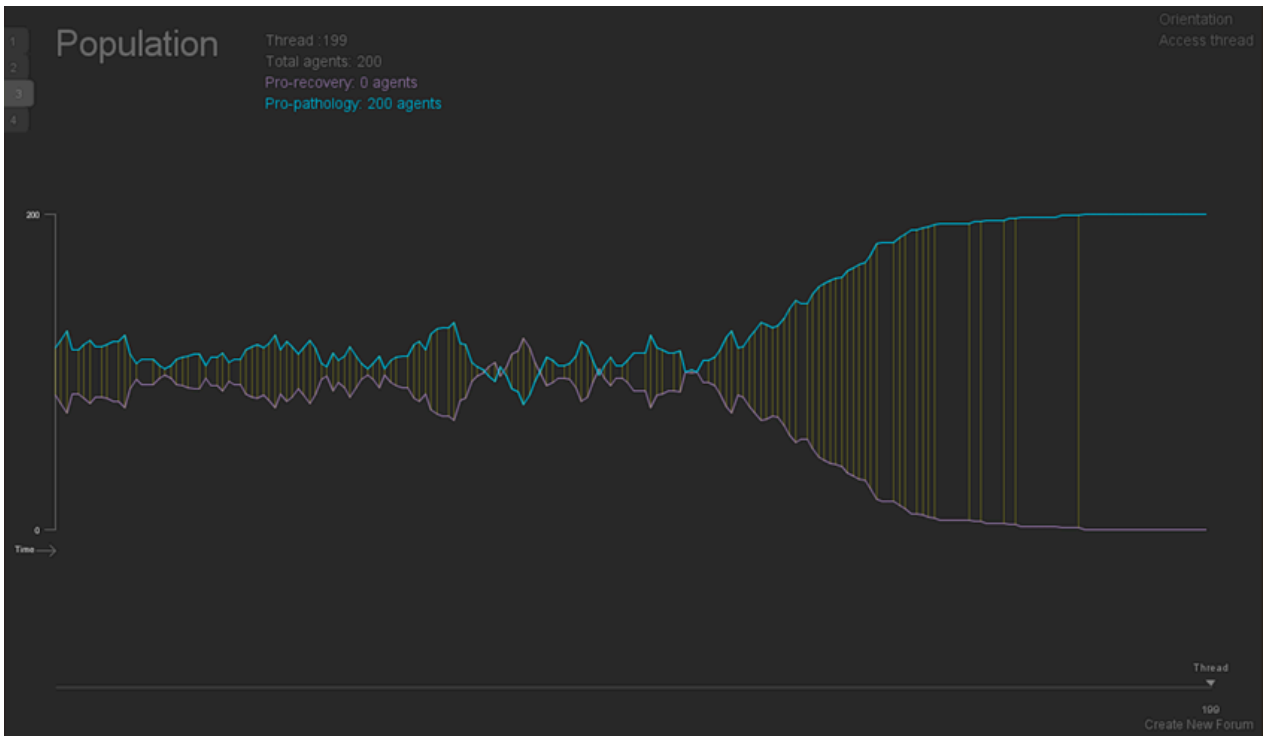
Radicalisation ?



Radicalisation ?

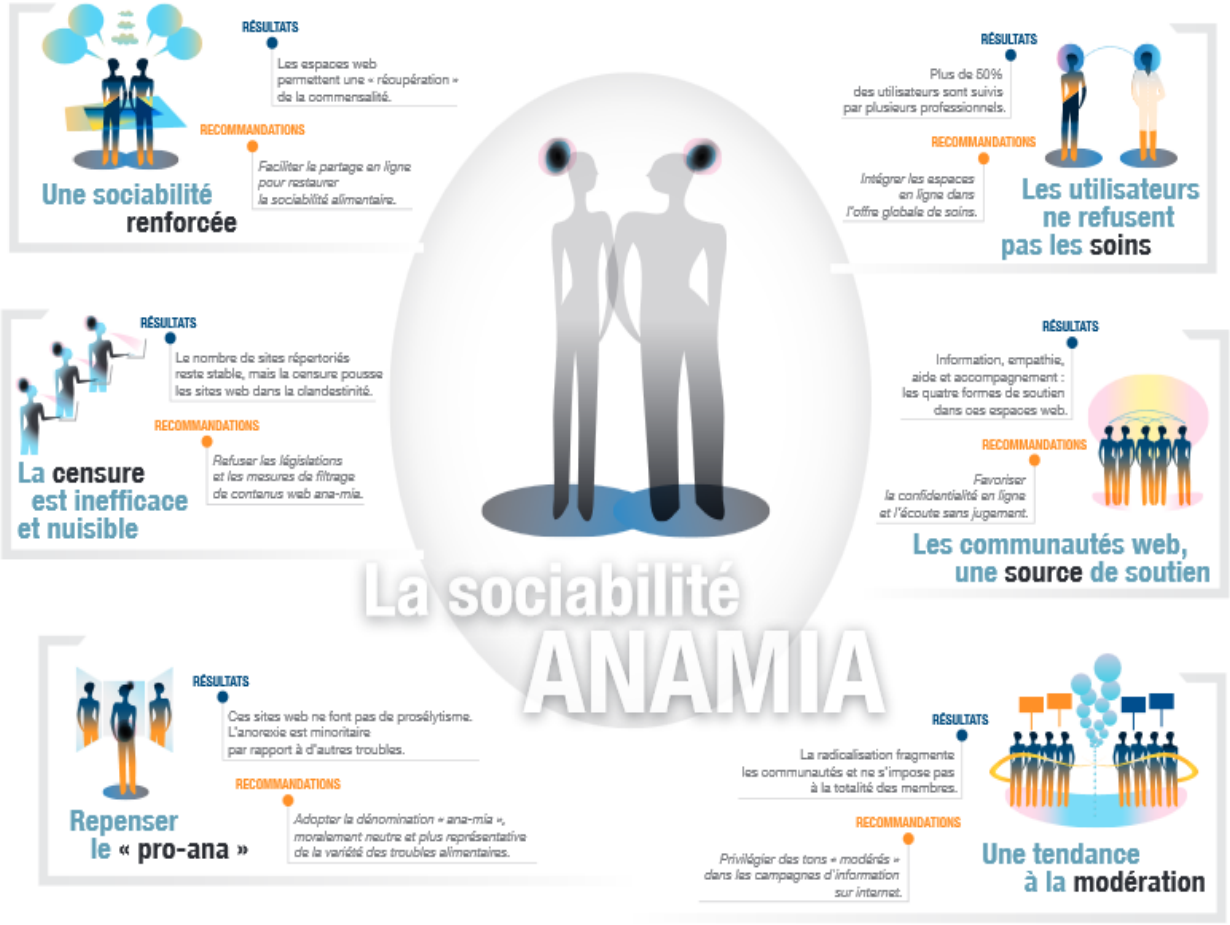


Radicalisation ?



Radicalisation ?

- Dans les communautés d'Internet, la radicalisation n'est pas une fatalité
- La radicalisation n'est que l'un des scénarios possibles (et pas majoritaire)
- Plus on est activement engagé, plus les membres changent, moins on se radicalise
- La censure impacte la radicalisation : mois de circulation, plus de logique d'entre-soi



Merci !

antonio.casilli@telecom-paristech.fr

Stages LIESSE 2015
à Télécom ParisTech

Journées de formation à destination des professeurs de Classes Préparatoires aux Grandes Écoles

- ***Initiation à la programmation Python : Application en traitement du signal et des images.***
lundi 13 et mardi 14 avril 2015 (session «vrais débutants»)

- ***Initiation à Scilab: Applications audiophoniques.***
lundi 4 mai 2015 (session «vrais débutants»)

- ***Martingales et analyse stochastique : Applications à la finance***
jeudi 7 mai 2015

- ***Informatique théorique : Théorie des langages, analyse lexicale, analyse syntaxique***
mardi 12 mai 2015
(sur le site de Sophia Antipolis)

- ***Bases de données relationnelles : Mise en pratique***
mercredi 13 mai 2015

- ***Algorithmes de tri et chemins dans les graphes***
lundi 18 et mardi 19 mai 2015

Journée Télécom-UPS
(en prélude de l'assemblée générale de l'UPS du 30 mai 2015)

Le Numérique Pour Tous
- vendredi 29 mai 2015 -

Inscription en ligne : www.telecom-paristech.fr/liesse/
Contact : liesse@telecom-paristech.fr

Télécom ParisTech
46 rue Barrault
75013 Paris
www. telecom-paristech.fr