Web Science

Master 1 IFI

Andrea G. B. Tettamanzi
Université de Nice Sophia Antipolis
Département Informatique
andrea.tettamanzi@unice.fr

CM - Séance 3

PageRank et comment Google transforme des mots en argent

Introduction

- Quelques chiffres sur Google (mise à jour : 25 mai 2014)
  - Capitalisation de marché : ~ 375 milliards de dollars
  - Chiffre d'affaires : 62 milliards de dollars
  - Bénéfices avant intérêts, impôts, etc. : 18.6 milliards de $
  - Environs 54 000 employés
- Par comparaison
  - PIB du Luxembourg : 57 milliards de dollars
  - Si Google était un pays, il serait le 70ème par PIB sur 193
  - En 2013, Google était 15ème au monde par capitalisation parmi les entreprises cotées en bourse
- Pas mal pour un « simple » moteur de recherche…

$590/s !!!

La clé du succès

- Le succès de Google se base sur deux algorithmes :
  - **PageRank**
  - **AdWords + AdSense**
- Le premier permet de trier les résultats des recherches :
  - valeur d'usage
  - il a imposé Google comme leader du marché
- Le deuxième génère l'impression de messages publicitaires ciblés aux intérêts du public d'une page Web :
  - valeur d'échange
  - AdWords permet d'acheter du trafic, AdSense de le vendre

Plan

- PageRank
- AdWords + AdSense
- TD

1ère Partie

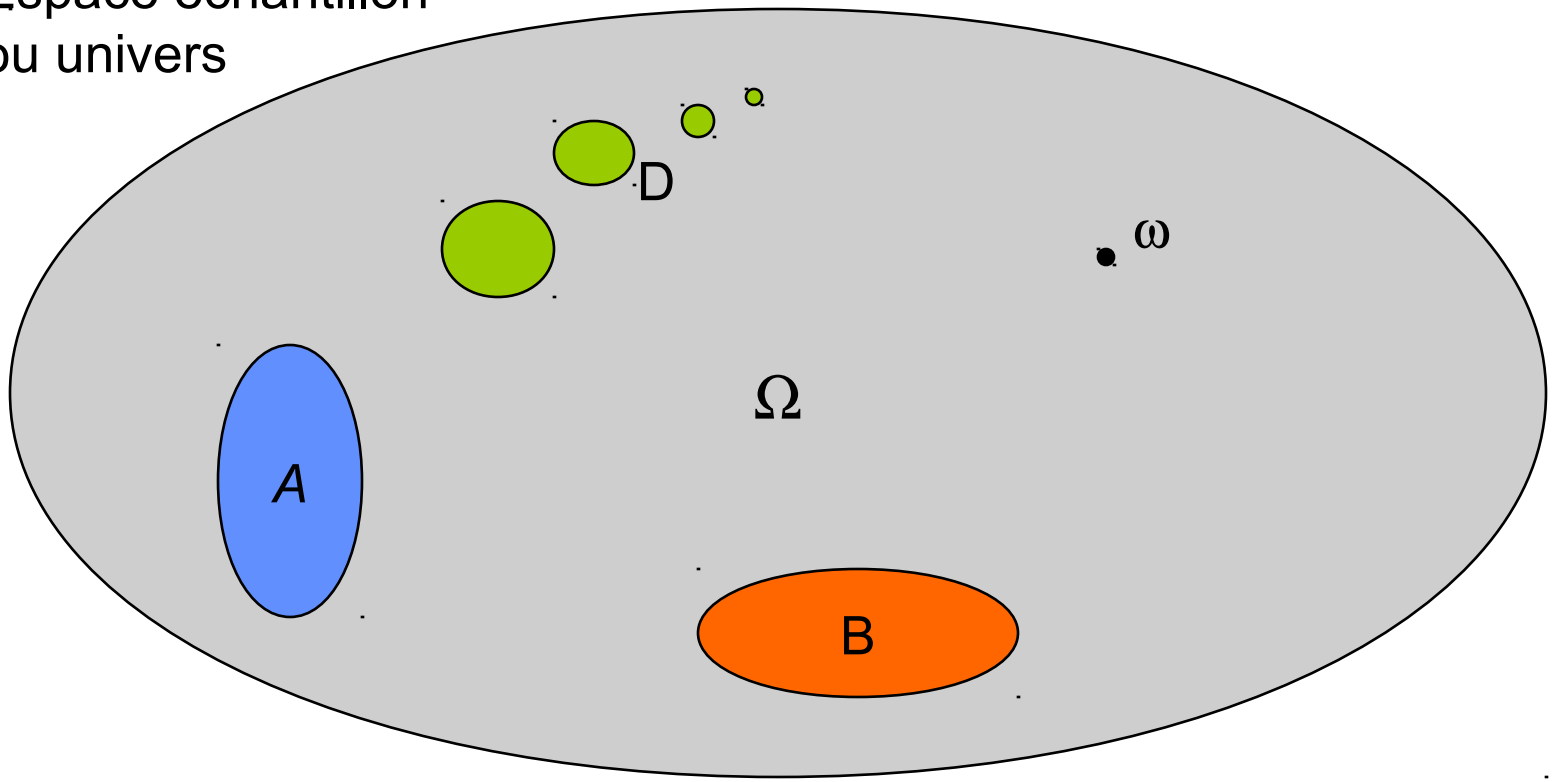
PageRank

Intuition de fond

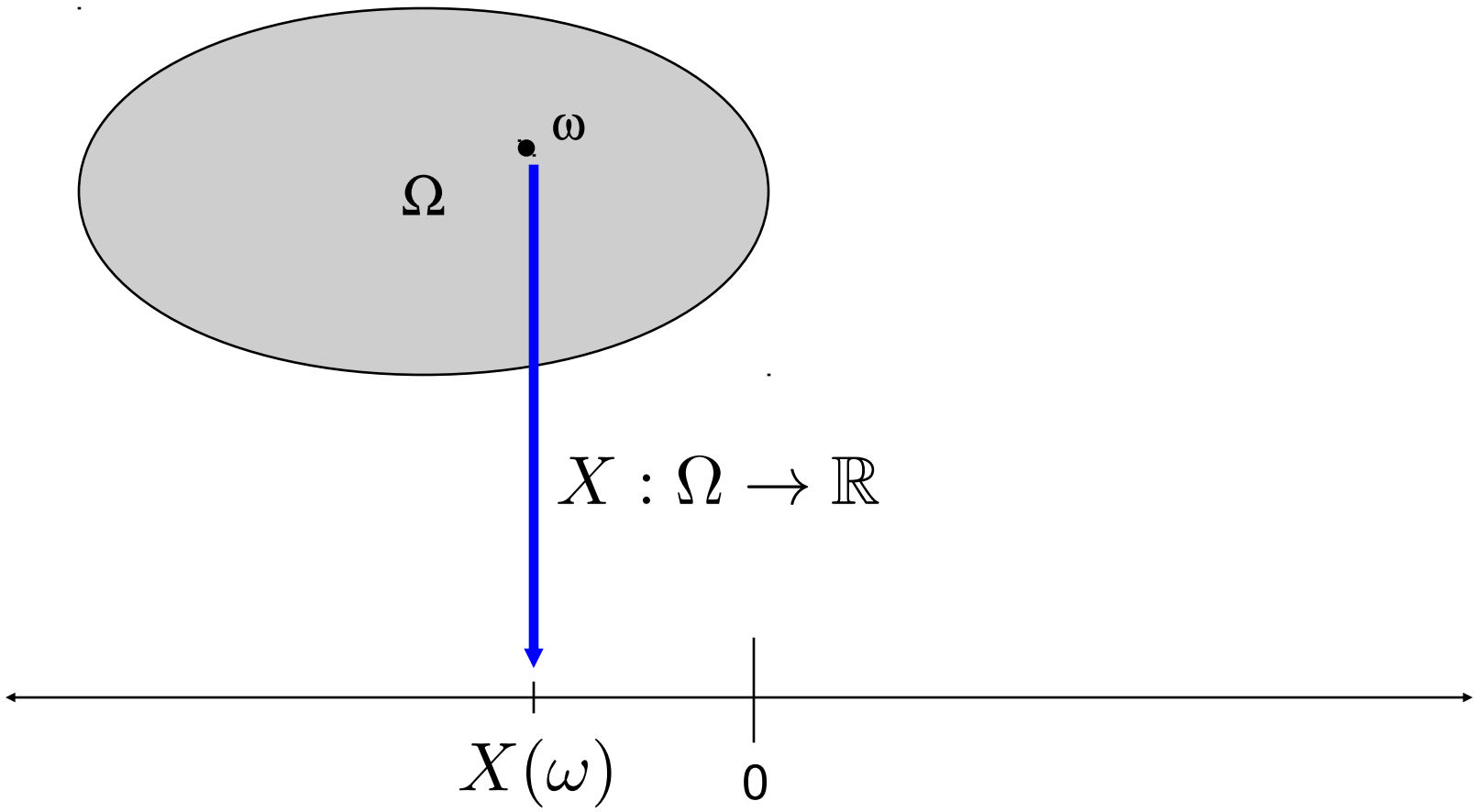
- Le Web comme graphe orienté
  - Ses nœuds sont les pages HTML
  - Ses arcs sont les hyperliens . . .
- Quelles pages visiterait un **surfeur aléatoire** ?
  - Le surfeur aléatoire commencerait par une page arbitraire
  - Il sauterait d'une page à la suivante en cliquant sur un hyperlien choisi au hasard
- Idée : mesurer l'importance d'un page par la probabilité qu'elle soit visitée à l'instant *t* par un surfeur aléatoire !
- Cette probabilité est la fréquence de visite de la page

Événements

Espace échantillon
ou univers



Variables aléatoires



Processus Stochastiques

Une suite de variables aléatoires

$$X_1, X_2, \ldots, X_t, \ldots$$

Chacune dotée de sa propre distribution de probabilité.

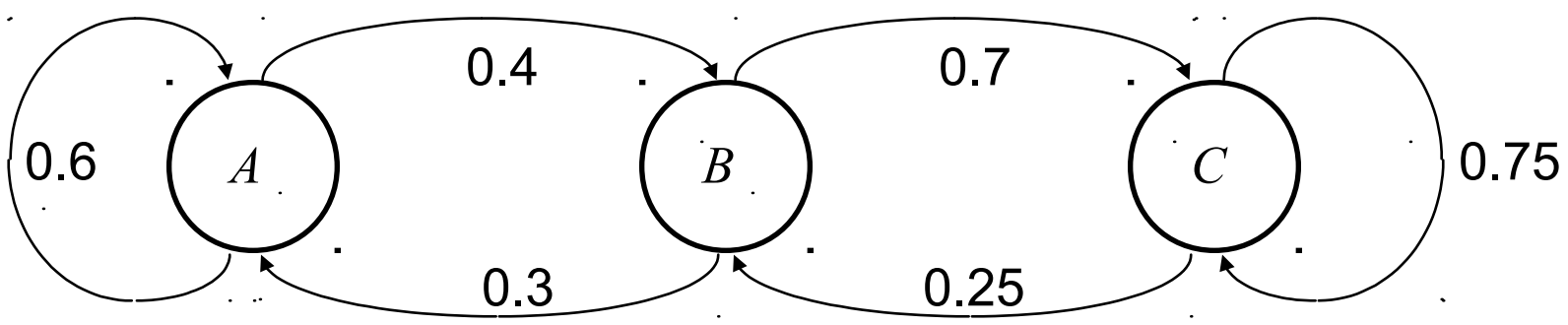
Notation: $\{X_t(\omega)\}_{t=0,1,\ldots}$

Chaînes de Markov

Un processus stochastique $\{X_t(\omega)\}_{t=0,1,\ldots}$

est une chaîne de Markov si et seulement si, pour tout t,

$$\Pr[X_t = x \mid X_0, X_1, \ldots, X_{t-1}] = \Pr[X_t = x \mid X_{t-1}]$$



Matrice de transition

$$\mathbf{T} = \begin{bmatrix} \Pr(X_t = x_1 \mid X_{t-1} = x_1) & \dots & \Pr(X_t = x_n \mid X_{t-1} = x_1) \\ \Pr(X_t = x_1 \mid X_{t-1} = x_2) & \dots & \Pr(X_t = x_n \mid X_{t-1} = x_2) \\ \vdots & & \vdots \\ \Pr(X_t = x_1 \mid X_{t-1} = x_n) & \dots & \Pr(X_t = x_n \mid X_{t-1} = x_n) \end{bmatrix}$$

T est une matrice stochastique :

$$\forall i, \quad \sum_{j=1}^{n} \Pr(X_t = x_j \mid X_{t-1} = x_i) = 1$$

Définition « idéalisée » de PageRank

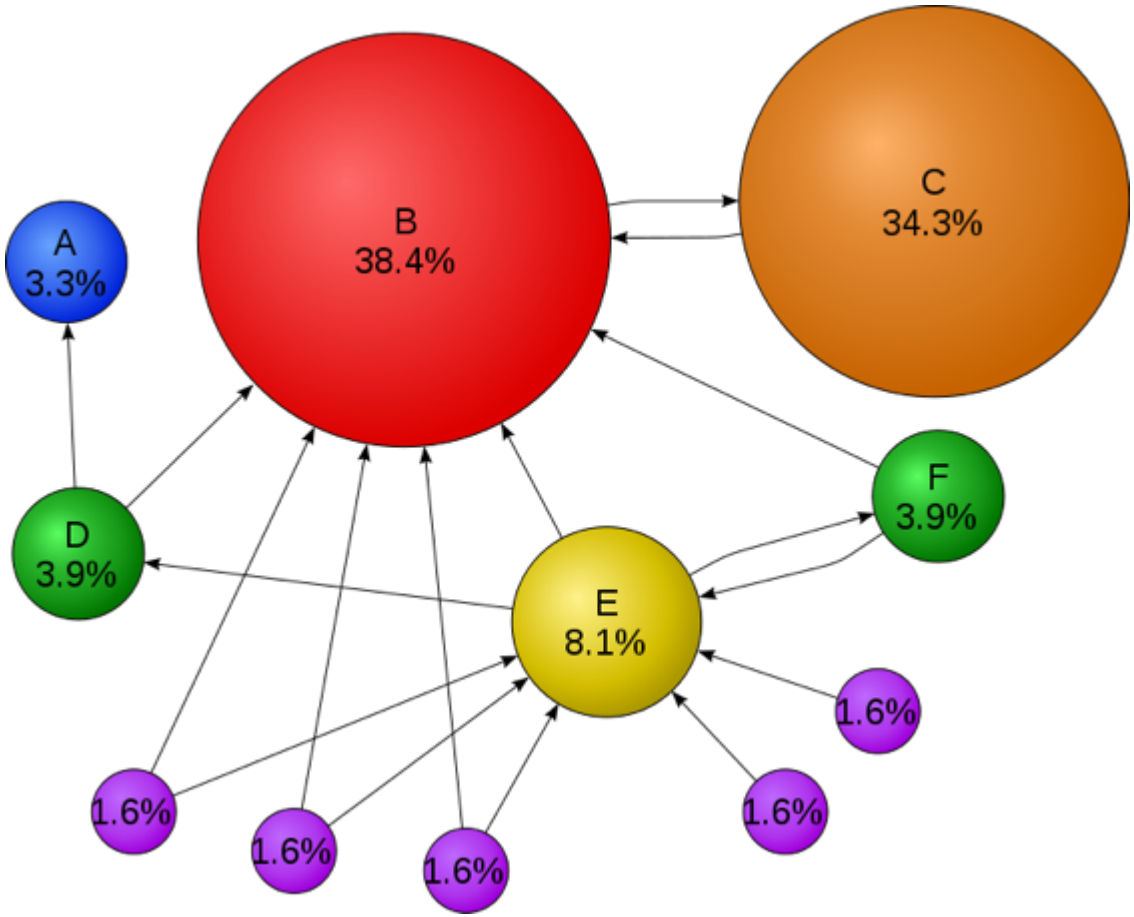
$q_i = \text{N}^o$ liens sortant de la page i

$\mathbf{H} = (h_{ij})$

$$h_{ij} = \begin{cases} 1/q_i & \text{s'il existe un lien de } i \text{ à } j; \\ 0 & \text{sinon.} \end{cases}$$

$$\pi_j = \sum_i \pi_i h_{ij} \quad \Longleftrightarrow \quad \pi = \pi\mathbf{H}$$

Exemple



Hypothèse de fond

Une page Web est importante si elle est référencée par d'autres pages importantes

Analyse de la définition

- Il y a trois facteurs qui déterminent le PageRank d'une page :
  - Le nombre de liens qui pointent vers elle ;
  - La propension des pages d'origine de ces liens à diriger le surfeur vers elle, c'est-à-dire, le nombre total de leurs liens sortants ;
  - Le PageRank des pages d'origine des liens
- Le modèle idéalisé a deux problèmes :
  - Les pages sans liens sortants (*dangling*), qui captureraient le surfeur
  - Le surfeur peut aussi rester piégé dans un *bucket* (= sceau), une composante joignable et fortement connectée, sans arcs sortants vers le reste du graphe

Modèle réel : matrice Google

- Les lignes de la matrice **H** avec tous les éléments égales à zéro, qui correspondent aux pages sans liens sortants, sont remplacées par une distribution uniforme ou arbitraire.
- Soit **S** la matrice ainsi modifiée.
- Pour résoudre le problème des *buckets*, Brin et Page proposent de remplacer la matrice **S** par la matrice Google :

$$\mathbf{G} = \delta\mathbf{S} + (1-\delta)\mathbf{E}$$

matrice de téléportation

damping factor

$$\mathbf{E} = \begin{bmatrix} 1/n & 1/n & \cdots & 1/n \\ \vdots & \vdots & & \vdots \\ 1/n & 1/n & \cdots & 1/n \end{bmatrix}$$

Interprétation de la matrice Google

- La définition de la matrice Google peut être expliquée comme suit
  - Avec probabilité δ, le surfeur aléatoire suit le lien suivant
  - Avec probabilité 1 – δ, le surfeur se lasse de suivre des liens et dirige le navigateur vers un nouveau URL, qui pourrait n'avoir rien à voir avec la page courante.
  - Dans ce cas, le surfeur est « téléporté » à cette nouvelle page
- Les inventeurs de PageRank proposent un facteur d'amortissement δ = 0.85 :
  - En moyenne, après avoir suivi 5 liens, le surfeur choisit une nouvelle page au hasard.
- Le vecteur PageRank est donc π tel que

$$\pi = \pi \mathbf{G}$$

Existence et unicité du vecteur PageRank

- Le vecteur π est un vecteur propre de **G** de valeur propre 1.
- La matrice **S** est stochastique, ainsi que la matrice **E**.
- La matrice **G** est donc stochastique.
- Si **G** est stochastique, l'équation $\pi = \pi\mathbf{G}$ a au moins une solution.
- D'après le théorème de Perron-Frobenius, si **A** est une matrice carrée non-négative irréductible, alors il existe un vecteur **x** tel que $\mathbf{x}\,\mathbf{A} = r\,\mathbf{x}$, où r est le rayon spectral de **A**.
- La matrice **S** est vraisemblablement réductible mais, grâce à la matrice de téléportation, **G** ne l'est sûrement pas.
- En plus, puisque **G** est stochastique, son rayon spectral est 1.
- Par conséquent, un vecteur PageRank > 0 existe et est unique.

PageRank et Théorie de Markov

- Le modèle de marche aléatoire sur le graphe du Web, modifié avec la téléportation, induit naturellement une chaîne de Markov avec un nombre fini (même si énorme) *n* d'états (= pages)
- **G** est la matrice de transition de cette chaîne de Markov
- Puisque **G** est irréductible, la chaîne est ergodique et a une distribution stationnaire unique, qui correspond au vecteur PageRank π.

Calcul du vecteur PageRank

- La **méthode de la puissance** est une méthode numérique qui permet de déterminer la valeur propre de module maximal d'une matrice à coefficients réels.
- On prend un vecteur **x** au hasard et on calcule la suite récurrente:

$$\mathbf{x}^{(0)} = \mathbf{x}, \quad \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)}\mathbf{A}/\|\mathbf{A}\|$$

- Cette suite converge vers la valeur propre de module maximal de la matrice **A**
- Pour calculer π, on part du vecteur **u** = (1/n, …, 1/n) et on s'arrête lorsque

$$\|\pi^{(t+1)} - \pi^{(t)}\| < \epsilon$$

2ème Partie

AdWords et AdSense

… ou de Comment Google transforme des mots en argent

Enjeu

- Mars 2000 : Éclatement de la « Bulle Internet »
  - Beaucoup de *start-ups* proposant une valeur d'usage, mais pas de valeur d'échange ne résistèrent pas
  - Meilleure idée que de simplement vendre de la publicité
  - Accumulation de capital linguistique grâce à ses services
  - Exploiter ce capital
- Un algorithme qui organise automatiquement la spéculation autour des mots a permis de créer le premier marché linguistique mondial
- Trademarks : il était déjà possible d'acheter certains mots
- Google a élargi et libéralisé ce marché

Annonces - Pourquoi ces annonces ?

1 [X-Helvetic Tours | helvetictours.ch](#)
www.helvetictours.ch/xHelveticTours
Des **vacances** à la mer aussi peu chères, ca fait vraiment du bien!

2 **Vacances** tout compris | clubmed.ch
www.clubmed.ch/
-15% sur vos **vacances** d'hiver 12/13 ou jusqu'à 480 CHF offerts now !

3 Voyages Jusqu'à -70% - Offres Imbattables: Vos Voyages
www.groupon.ch/Voyages
Jusqu'à -70% avec Groupon. Ici !

Vacances 2012. L'été à la mer, en France, Corse, Var, Bretagne ...
www.**vacances**.com/
Nombreuses annonces, de professionnels et de particuliers, pour les **vacances** d'hiver. Chalets, studios, appartements au ski. Week-ends, voyages, séjours en ...
↳ Location - Espagne - Location Aquitaine - Provence et Côte d'Azur

Le calendrier scolaire - Ministère de l'Éducation nationale
www.education.gouv.fr › ... › Le ministère › Repères, histoire et patrimoine
Vacances, Zone A, Zone B, Zone C ... **Vacances** de la Toussaint ... Le départ en **vacances** a lieu après la classe, la reprise des cours le matin des jours indiqués.
↳ Vacances Scolaires 2011-2012 - Le calendrier scolaire ... - MENE0914826A

Vacances Look Voyages : séjour pas cher en famille, club tout ...
www.look-voyages.fr/
Votre séjour en club de **vacances** Lookea, en hôtel ou en formule circuit au meilleur prix avec Look Voyages. Départ dernière minute, pas cher ou en promo !

Villages et Clubs de **vacances** en tout inclus Thomas Cook
tt.thomascook.fr/village-club-**vacances**/
Ambiance. Une ambiance animée en journée comme en soirée. CHAQUE SEMAINE, NOS ANIMATEURS CONCOCTENT LE PROGRAMME DU VILLAGE : ...

Annonces - Pourquoi ces annonces ?

4 Tech Travel
www.techtravel.ch/
les meilleures offres de voyages Culturels, Escapades, Plages

5 Site Officiel Air Transat
www.airtransat.ch/
Les plus bas prix pour les vols vers le Canada. Réservez en ligne!

6 Voyager Moins Cher
www.voyagermoinscher.com/
Voyage et billet d'avion Promotions et dernière minute.

7 **Vacances** en France
www.declicfrance.com/**Vacance**_France
Trouvez vos **vacances** en France parmi plus de 5000 offres !

8 **Vacances** Go Voyages
www.govoyages.com/**Vacances**
Choisissez vos **vacances** au prix le plus bas chez Go Voyages!

9 **Vacances** à l'étranger
www.sejour-express.com/
Trouvez les moins chères sur le comparateur de **vacances**

10 **Vacances** en Club
www.club-**vacances**-express.fr/
Trouvez les moins chères sur le comparateur de location en Club

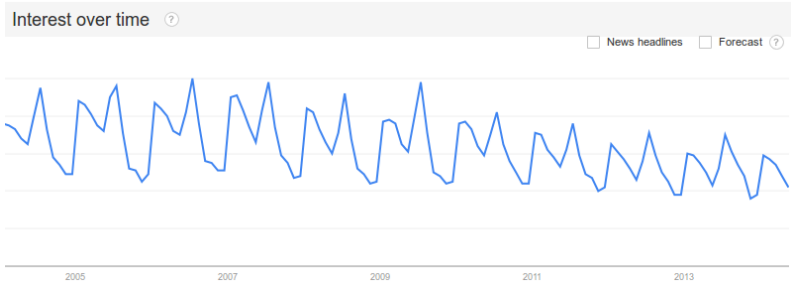
AdWords

- Mécanisme d'enchère sur des mots pour placer des annonces
- Tous les mots(-clefs) peuvent donner lieu à des enchères
- L'algorithme classe automatiquement les annonces selon un calcul en quatre étapes :
  - Enchère sur un mot (*E*) : l'annonceur fixe un prix max qu'il est prêt à payer en cas de clic
  - Calcul du score qualité *Q* de l'annonce (pertinence) : secret !
  - Calcul du rang de l'annonce : $R = E\,Q$, et classement i
  - Calcul du prix à payer en cas de clic :

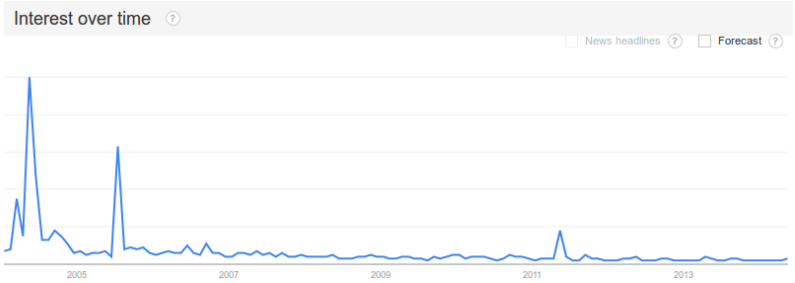
$$P_i = E_{i+1}\frac{Q_i}{Q_{i+1}}$$

GoogleTrends

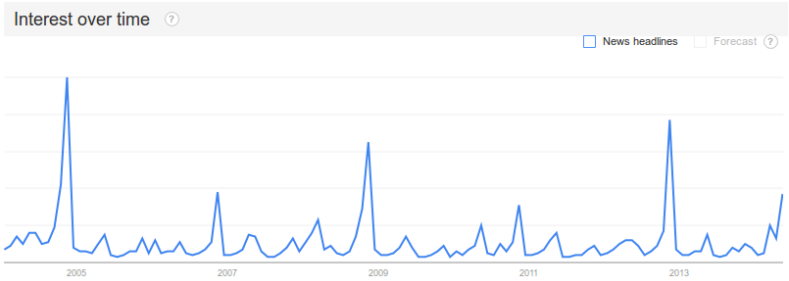
Vacances



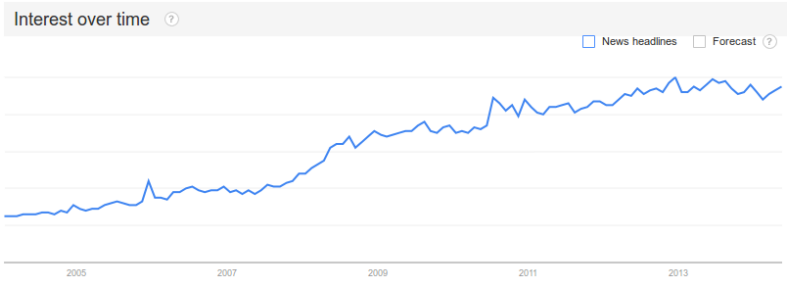
al-Qaeda



Elections



Porn



Achat et vente de trafic

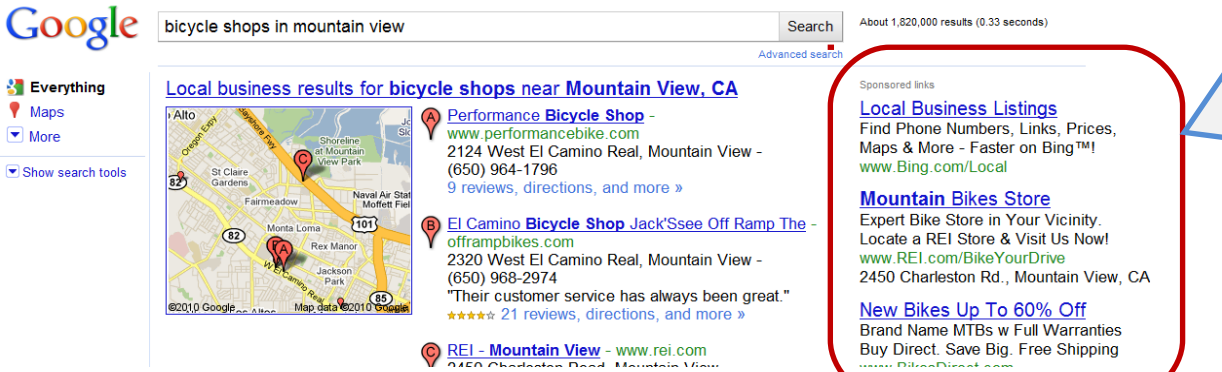
Les annonceurs font des enchères sur les mots pour en acheter le trafic

Les sites Web vendent leur trafic à Google pour afficher les annonces

Avantages pour les utilisateurs

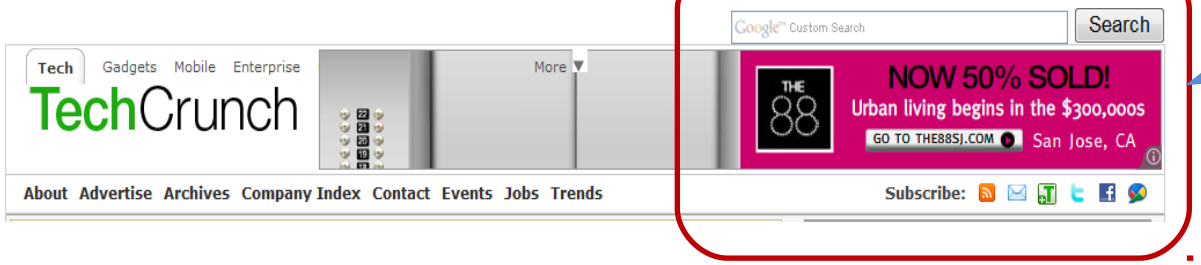
- Des services « gratuits » (recherche, info, email, maps, etc.)
- Une publicité utile, pertinente, non-invasive
- Une expérience agréable des contenus en ligne

Deux sources de bénéfice



Publicité sur les sites de Google, tels que

- google.com
- gmail.com
- orkut.com
- Youtube.com



Publicité sur les sites des clients adSense

Merci de votre attention