Web

Master 1 IFI

Andrea G. B. Tettamanzi

Université Nice Sophia Antipolis

Département Informatique

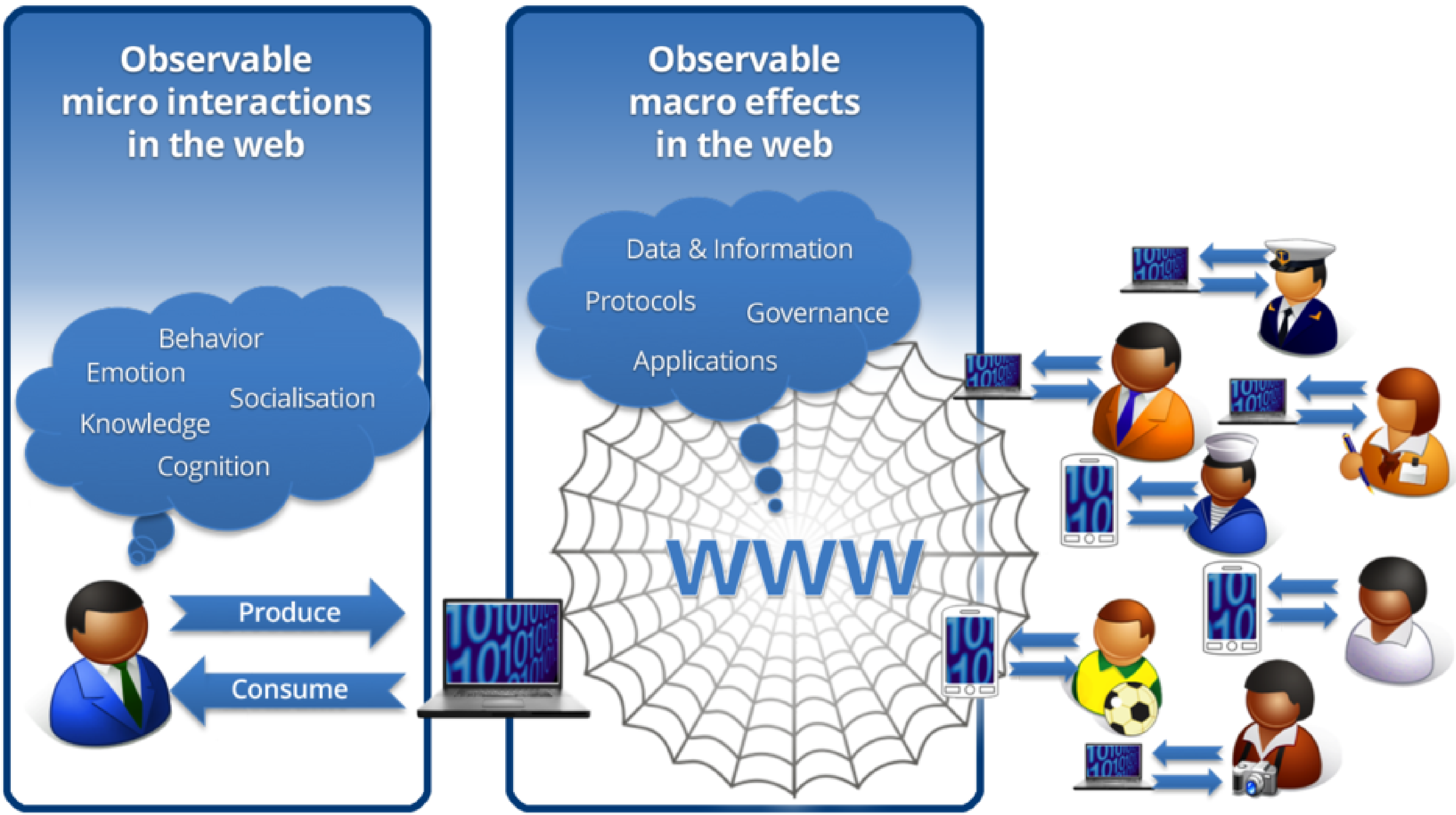
andrea.tettamanzi@unice.fr

Unit 8

Web Science
–
PageRank and how Google turns words into money

Web Science

- A deliberately ambiguous phrase.
  - Physical Science is an analytic discipline that aims to find laws that generate or explain observed phenomena
  - Computer Science is predominantly (though not exclusively) synthetic, in that formalisms and algorithms are created in order to support particular desired behavior
- Web science is a merging of these two paradigms
  - The Web needs to be studied and understood
  - ...and it needs to be engineered.
- At the micro scale, the Web is a piece of engineering
- At the macro scale, the Web is an emergent phenomenon
- Interdisciplinary: CS + Mathematics + Sociology + Economics



(cc-attrib Steffen Staab. http://west.uni-koblenz.de)

Web Science Topics

- Modeling Web-related structures, data, users and behaviors
- Analysis of online social and information networks, social media analysis
- Social machines, crowd computing, collective intelligence, and collaborative production (e.g., prediction markets)
- Web Economics
- Sentiment Analysis and Opinion Mining
- Legal Aspects of the Web
- Ethical Challenges
- Web access, literacy, divides, inclusion, exclusions, and development
- Humanities, arts, and culture on the Web

Introduction to PageRank

- Key statistics about Alphabet Inc. (= Google), as of April 9, 2019
  - Market capitalization: ~ $838 billion (2014: $375 billion)
  - Revenue : $136 billion (2014: $62 billion)
  - EBITDA : $34.9 billion (2014: $18.6 billion)

$1,107/s !!!

  - Full-time employees: 98,771 (2014: 54,000)
- As a comparison:
  - GDP of Ukraine: ~ $126.4 billion
  - If Google were a country, it would be 57th by GDP out of 186
  - In 2016, Alphabet was 94th among the world's corporations by capitalization and 2nd among publicly traded companies
- Not bad for "just" a search engine…

The Key of Success

- Google’s success is based on two algorithms :
  - **PageRank**
  - **AdWords + AdSense**
- The former allows Google to rank search results:
  - It gives Google its **use value**
  - It has imposed Google as a market leader
- The latter generates the impression of advertisements targeted on the interests of the audience of a Web page:
  - It gives Google its **exchange value**
  - AdWords allows buying traffic, AdSense allows selling traffic

Agenda

- PageRank
- AdWords + AdSense
- Lab work

Part I

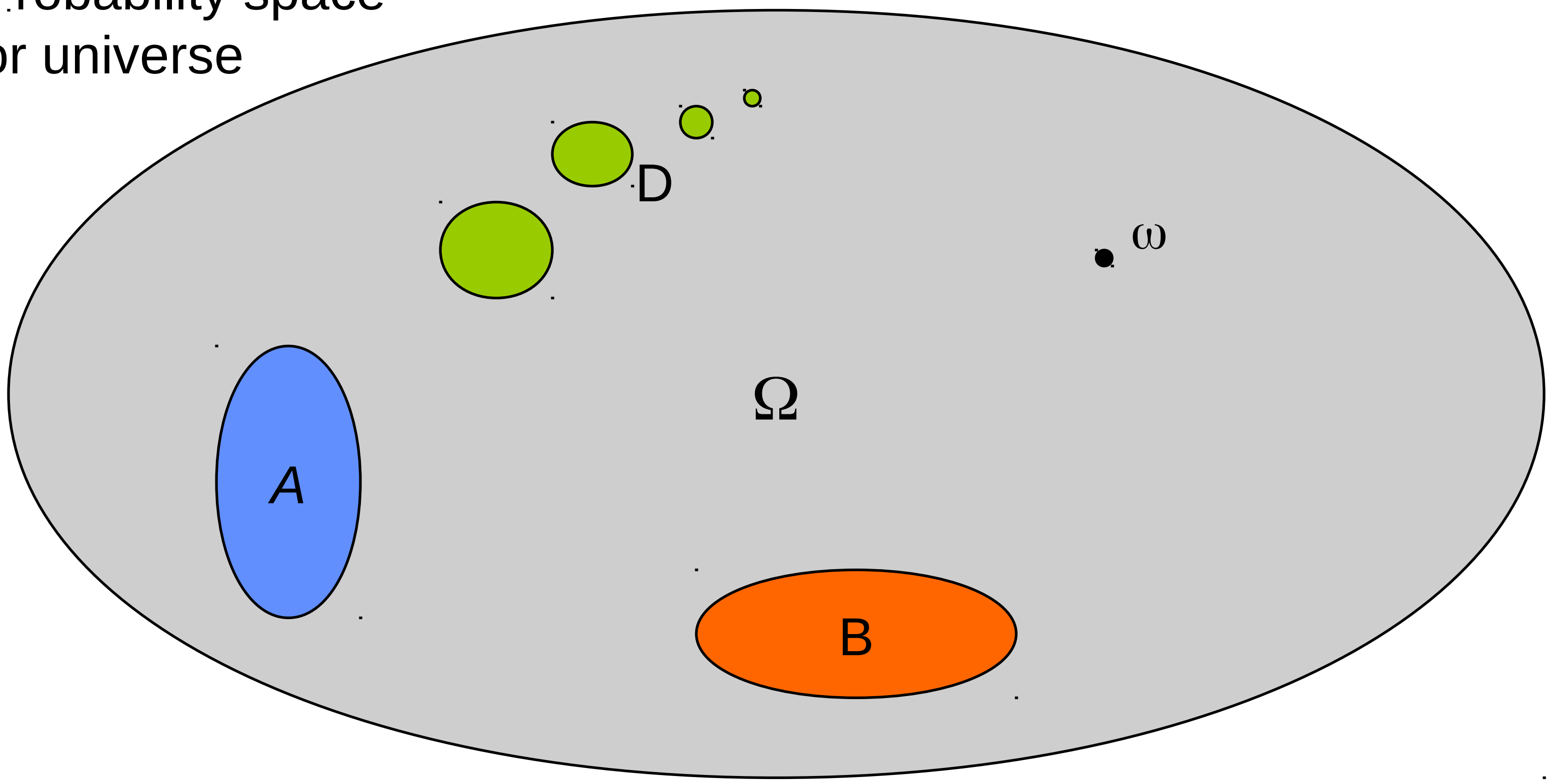
PageRank

Basic Intuition

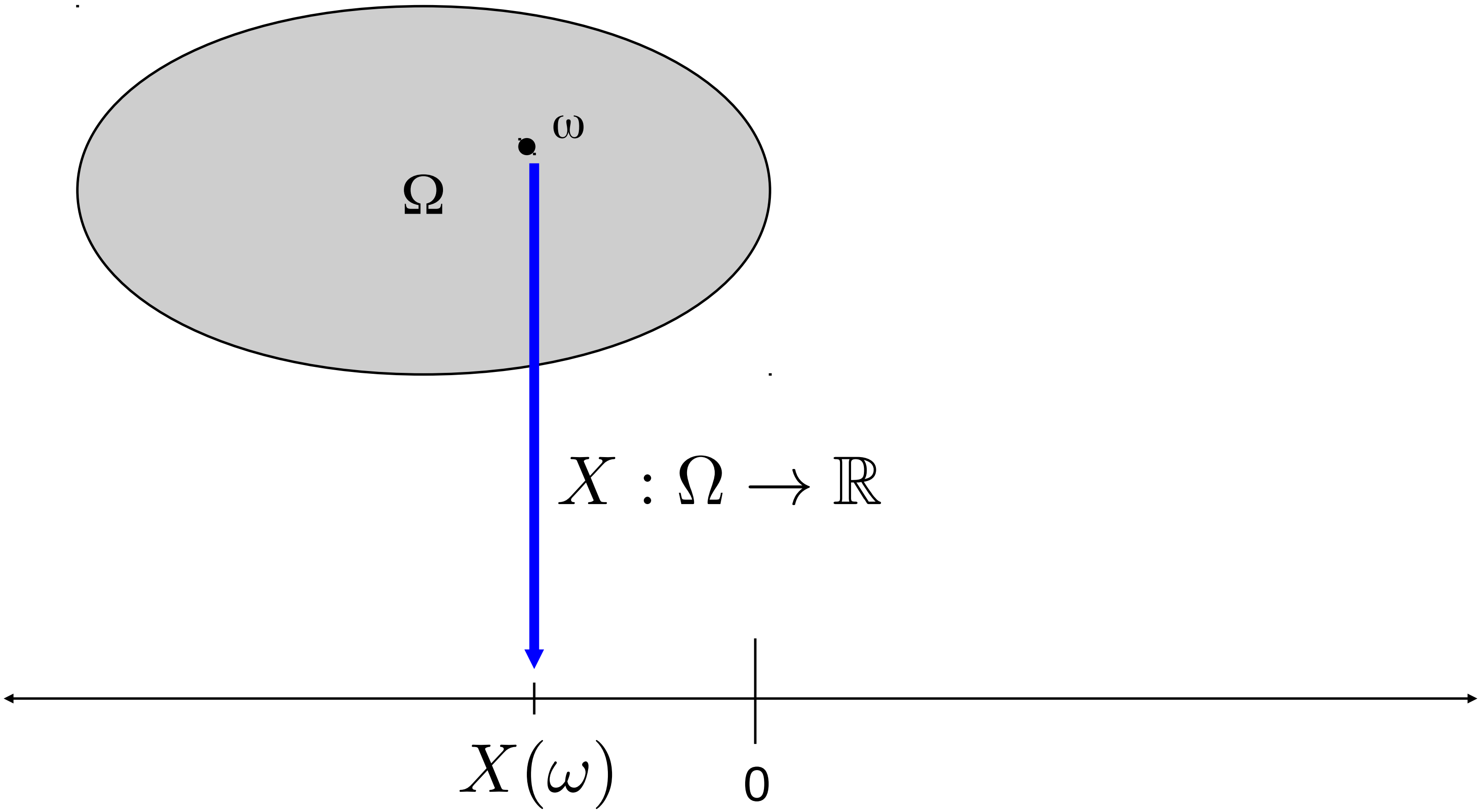
- The WWW as a directed graph
  - Its <u>nodes</u> are the HTML pages
  - Its <u>arcs</u> are the . . . hyperlinks
- Which pages would a **random surfer** visit?
  - The random surfers would start at a random page
  - They would jump from one page to the next by clicking a random hyperlink
  - Idea: measure the importance of a page by the probability that it is visited at time *t* by a random surfer!
- This probability is the visit frequency of the page

Events

Probability space
or universe



Random Variables



Random Processes

A sequence of random variables

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

Each equipped with its own probability distribution.

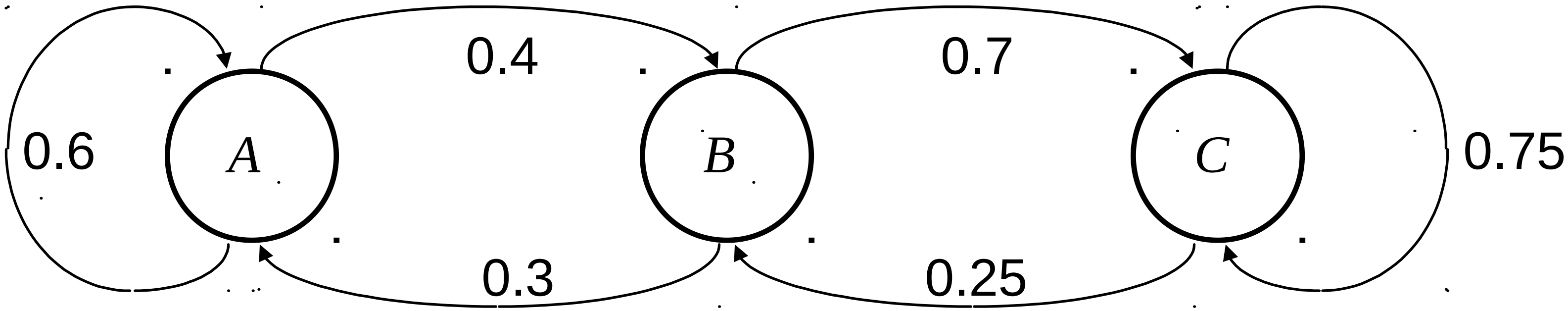
Notation: $\{X_t(\omega)\}_{t=0,1,\ldots}$

Markov Chains

A random process $\{X_t(\omega)\}_{t=0,1,\ldots}$

is a Markov chain if and only if, for all *t*,

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



Transition Matrix

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

T is a stochastic matrix:

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

“Idealized” Definition of PageRank

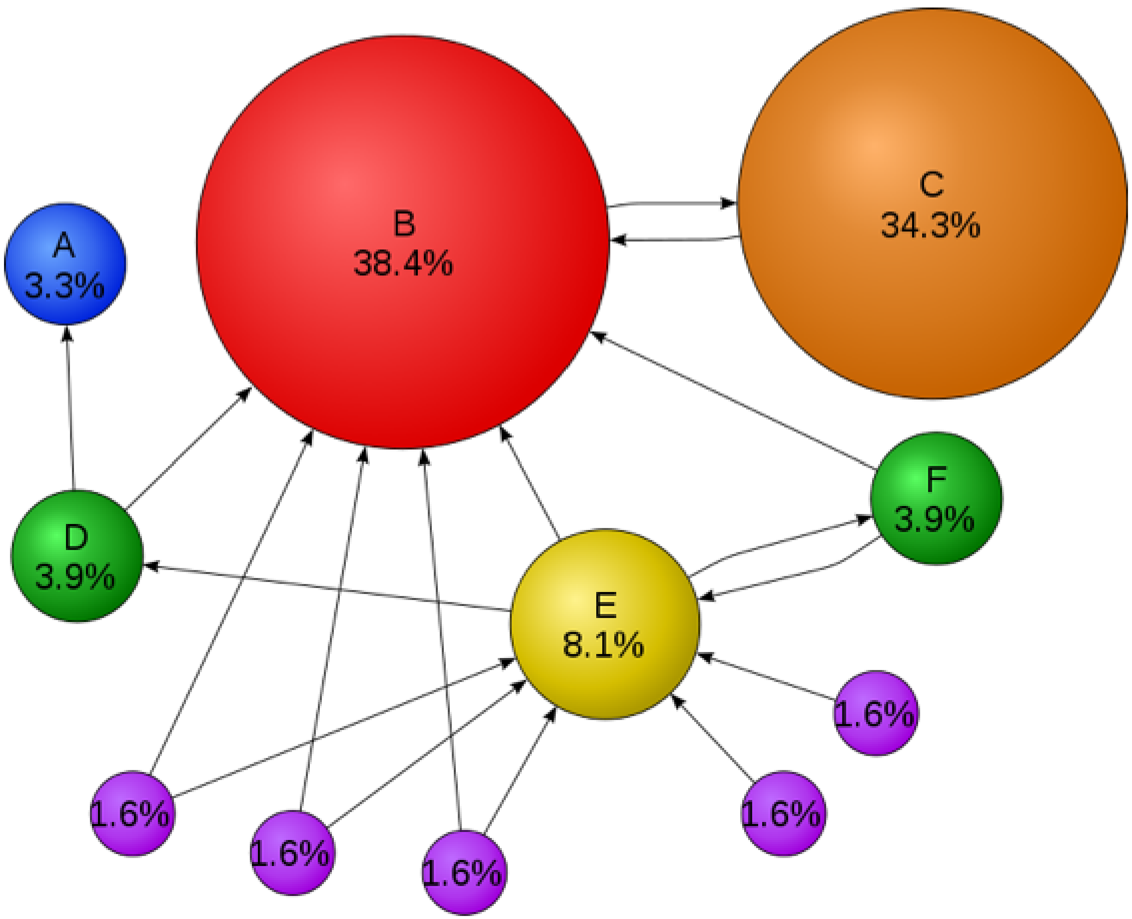
$$q_i = \#\text{ outgoing links from page } i$$

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

$$h_{ij} = \begin{cases} 1/q_i & \text{there exists a link from } i \text{ to } j; \\ 0 & \text{otherwise.} \end{cases}$$

$$\boxed{\pi_j = \sum_i \pi_i h_{ij}} \iff \boxed{\pi = \pi \mathbf{H}}$$

Example



Basic Hypothesis

A Web page is important insofar as it is referenced by other important pages

Analysis of the Definition

- There are three factors that determine the PageRank of a page:
  - The number of links pointing towards it;
  - The propensity of the pages containing those links to direct surfers towards it, i.e., the total number of outgoing links;
  - The PageRank of the pages containing those links
- The idealized model has two problems:
  - Pages without outgoing links (*dangling pages*)*,* which can capture surfers.
  - A surfer may also get trapped in a *bucket*, a reachable and strongly connected component, without outgoing arcs towards the rest of the graph.

Real Model: the Google Matrix

- The lines of matrix **H** having all zero elements, corresponding to pages without outgoing links, are replaced by a uniform or arbitrary distribution.
- Let **S** be the matrix thus modified.
- To solve the problem with *buckets*, Brin and Page propose to replace matrix **S** by the Google matrix:

$$\mathbf{G} = \underbrace{\delta}_{\text{damping factor}}\mathbf{S} + (1-\delta)\underbrace{\mathbf{E}}_{\text{Teleportation matrix}}$$

Teleportation matrix

damping factor

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

Interpreting the Google Matrix

- The definition of the Google matrix may be explained as follows
  - With probability δ, the random surfer follows the next link
  - With probability 1 – δ, the random surfer gets tired following links and directs the browser to a novel URL, which has nothing to do with the current page.
  - In this case, the surfer is “teleported” to this novel page
- The inventors of PageRank suggest a damping factor δ = 0.85 :
  - On average, after following 5 links, the surfer chooses a new random page.
- The PageRank vector is therefore π such that

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

Existence and Uniqueness of the PageRank vector

- The π vector is an eigenvector of **G** of eigenvalue 1.
- The **S** matrix is stochastic, as is matrix **E**.
- The **G** matrix is, therefore, stochastic as well.
- If **G** is stochastic, equation $\pi = \pi\mathbf{G}$ has at least one solution.
- According to Perron-Frobenius' Theorem, if **A** is an irreducible non-negative square matrix, then there exists a vector **x** such that $\mathbf{x}\,\mathbf{A} = r\,\mathbf{x}$, where r is the spectral radius of **A**.
- The **S** matrix is likely to be reducible; however, thanks to the teleportation matrix, **G** is certainly irreducible.
- Furthermore, since **G** is stochastic, its spectral radius is 1.
- As a consequence, a PageRank vector > 0 exists and is unique.

PageRank and Markov Theory

- The random walk model on the Web graph, modified with teleportation, naturally induces a Markov chain with a finite (albeit huge) number *n* of states (= pages)
- **G** is the transition matrix of such Markov chain
- Since **G** is irreducible, the chain is ergodic and it has a unique stationary distribution, corresponding to the PageRank vector π.

Computing the PageRank Vector (1)

- The **power method** is a numerical method which allows to determine the greatest (in absolute value) eigenvalue of a matrix with real coefficients.
- We take a random vector **x** and we compute the recurrence:

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

- This sequence converges to the greatest (in absolute value) eigenvalue of matrix **A**
- To compute π, we start from vector **u** = (1/n, …, 1/n) and we stop as soon as

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

Computing the PageRank Vector (2)

- The convergence speed of the power method applied to matrix **G** is of the same order as the rate by which δ^k goes to 0.
- For instance, for $\delta = 0.85$:
  - 43 iterations → precision of 3 decimal digits
  - 142 iteration → precision of 10 decimal digits
- We also observe that the power method applied to matrix **G** can be expressed in terms of matrix **H**
- **H** is an extremely sparse matrix, which can be stored in a memory space of size $O(n)$
- According to rumors, Google recomputes π once per month
- “Google dance”: oscillation of π during the computation

Part II

AdWords + AdSense

… or how Google turns words into money

What is it all about?

- March 2000 : the bursting of the “Internet” or “Dot-Com” Bubble
  - Many *start-ups* which offered a use value but no exchange value did not survive
  - Google had a better idea than simply selling advertising space
  - It accumulated “linguistic capital” thanks to its services
  - The idea was to exploit this capital
- An algorithm which automatically organizes speculation on words has allowed Google to create the first global linguistic market
- *Trademarks*: it was already possible to purchase certain words
- Google has boosted and liberalized that market

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 ?

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

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

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

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

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

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

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

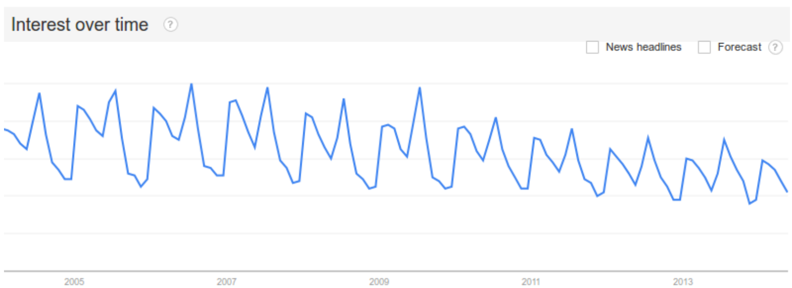
AdWords

- Auction mechanism on words to place advertisements
- All (key)words can bring about an auction
- The algorithm automatically ranks the advertisements according to a calculation in four steps:
  - Bid on a word (*E*): the advertiser fixes a maximum price she is willing to pay per click
  - Compute the quality score *Q* for the ad (relevance): secret !
  - Compute the rating of the ad, $R = E\,Q$, and its rank i
  - Compute the price to pay per click:

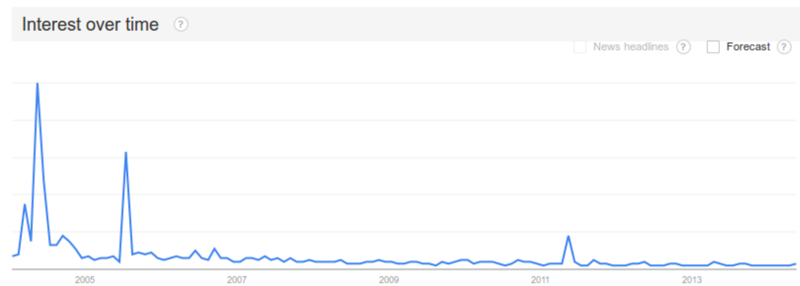
$$P_i = E_{i+1} \frac{Q_i}{Q_{i+1}}$$

GoogleTrends

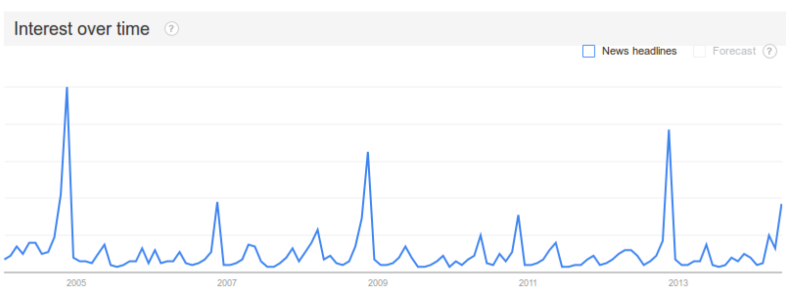
Holidays



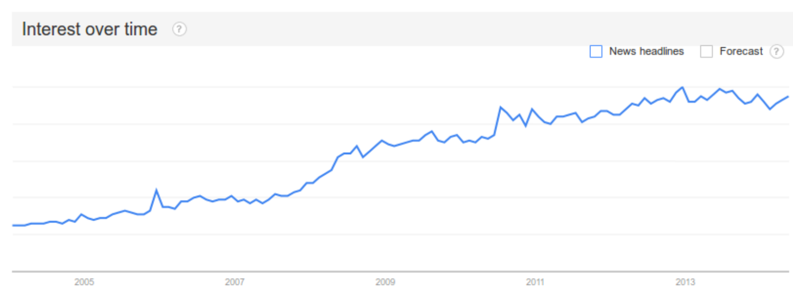
al-Qaeda



Elections



Porn



Buying and Selling Traffic

Ad Words

Ad Sense

Advertisers → $ → Google → $ → Web Page Owners/Creators

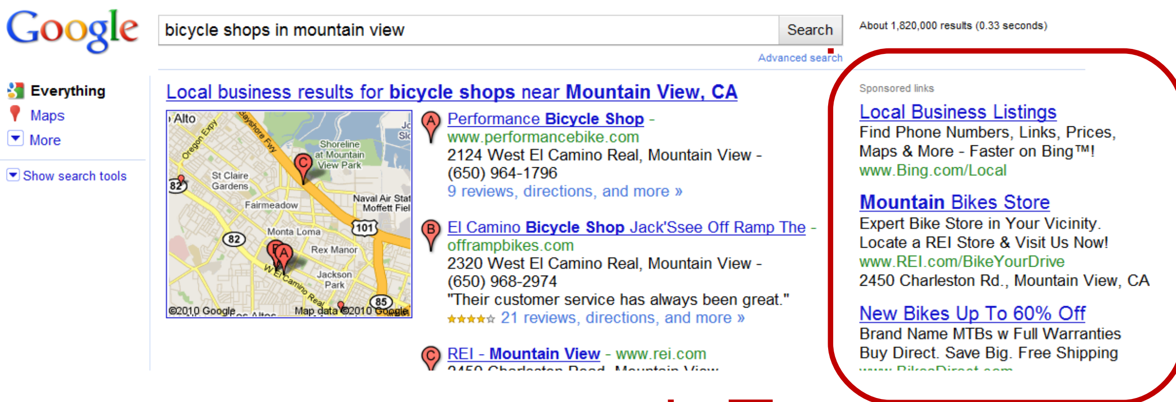
The advertisers bid on the words to buy their traffic

The Web sites sell their traffic to Google to show the ads

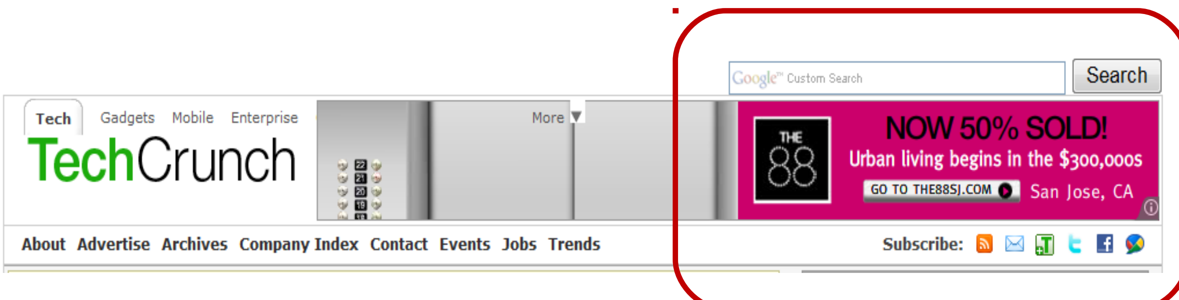
Advantages for the users

- “Free” services (search, docs, email, maps, translate, etc.)
- Useful, relevant, non-invasive advertisement
- Great user experience of on-line contents

Two Sources of Revenue



Advertisement on the Google sites, such as
- google.com
- gmail.com
- orkut.com
- Youtube.com



Advertisement on the adSense customer sites