Recherche et Diffusion de l'Information dans les Réseaux

Philippe Robert

Le 8 avril 2014

Présentation

Présentation

- **Directeur de recherche à l'INRIA**
- **Responsable de l'équipe de recherche "Réseaux, Algorithmes et Probabilités"**
- **Professeur à l'École Polytechnique**

Mathématicien Spécialité : Probabilités

Problématique générale

La raison d'être d'un réseau :

- **Diffuser**
- **Rechercher**

l'information

Quelques problèmes classiques

- Recherche de l'information
- Partage/Diffusion de l'information
- Contrôle d'accès
- Gérer les conflits d'accès
- ...

Plan de la conférence

Une brève histoire

Transmission de données dans les réseaux

Recherche d'information

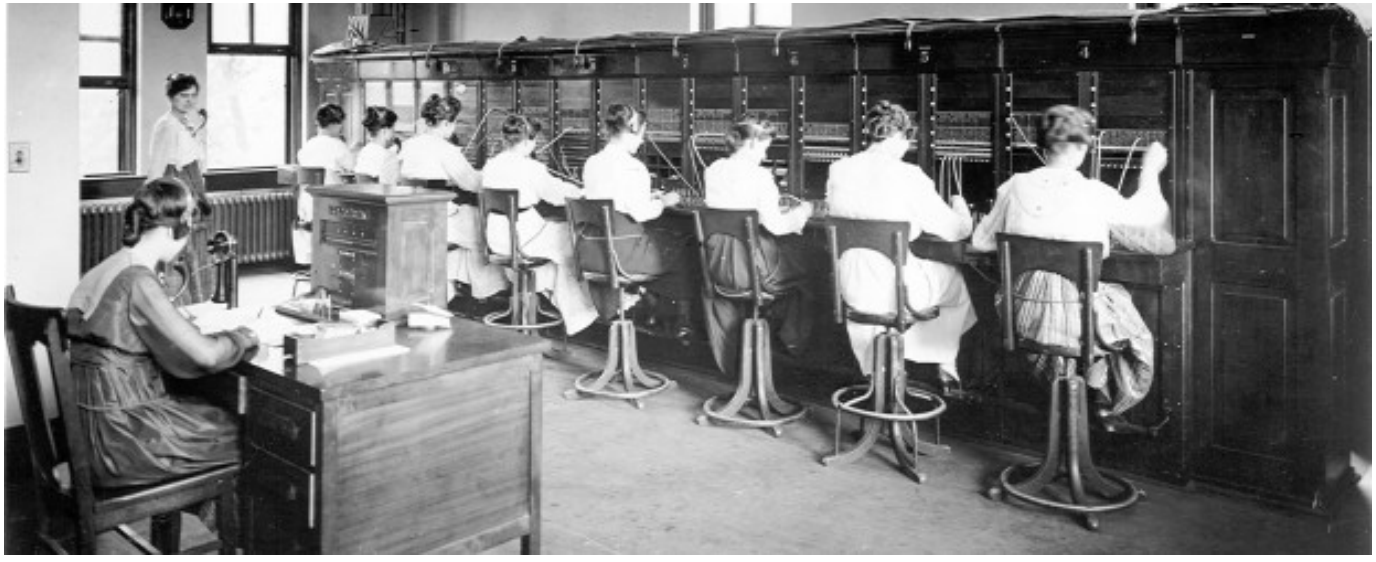
Internet maintenant

Histoire

Un bref aperçu historique

- 1900 : Réseaux téléphoniques

Réseau téléphonique



Un bref aperçu historique

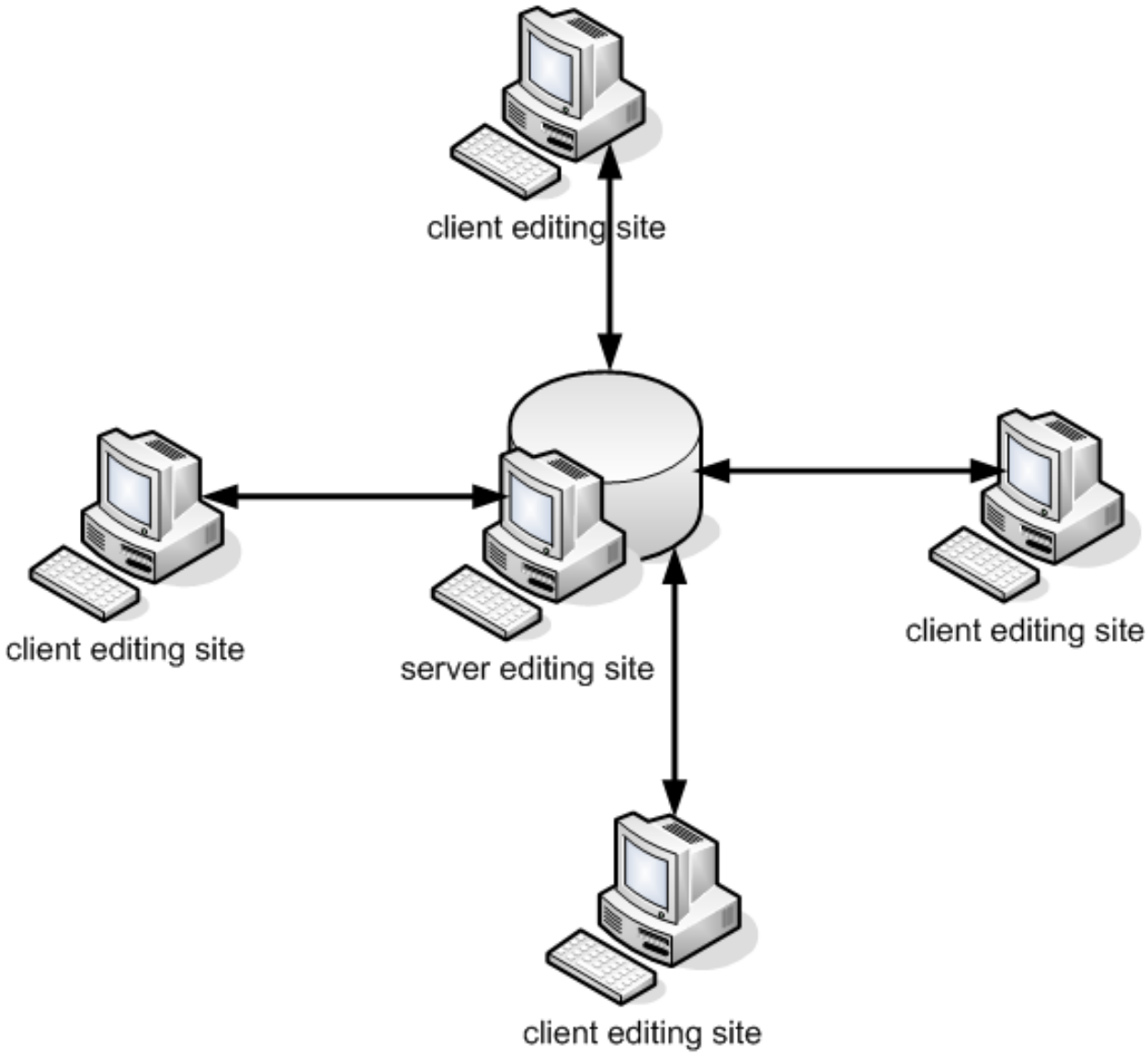
- 1900 : Réseaux téléphoniques

Un bref aperçu historique

- 1900 : Réseaux téléphoniques
- 1960 : Réseaux informatiques
  - Serveur central

IBM System/360

Le Modèle du Serveur Central



Un bref aperçu

- 1909 : Réseaux téléphoniques
- 1960 : Réseaux informatiques
  - Serveur central

Un bref aperçu

- 1909 : Réseaux téléphoniques
- 1960 : Réseaux informatiques
  - Serveur central
- 1980 : Systèmes distribués
  - Réseaux Locaux
  - Internet
  - Réseaux Mobiles

Système Distribué

- **Ensemble de Machines communicantes**
- **Pas de Contrôle Central**
- **Chacune agit de façon autonome**

Exemples de Systèmes Distribués

Exemples

- **Internet : Transmission des données**

Exemples de Systèmes Distribués

Exemples

- **Internet : Transmission des données**
- **Réseaux Pair à Pair**

  Recherche

  et Stockage des données : BitTorrent

  Téléphone : Skype

  ...

Les Systèmes Distribués dans la Nature

- **Bancs de Poisson**
- **Fourmilière, Essaims**

Les Systèmes Distribués dans la Nature

- **Bancs de Poisson**
- **Fourmilière, Essaims**
- **Cerveau**
- **...**

Innovation dans les réseaux

Progrès technologiques :

- **Vitesse des processeurs, des composants**
- **Nouveaux matériaux, . . .**

Innovation dans les réseaux

Progrès technologiques :

- **Vitesse des processeurs, des composants**
- **Nouveaux matériaux, . . .**
  Important mais n'est plus la source dominante d'innovation

Innovation dans les réseaux

Progrès technologiques :

- Vitesse des processeurs, des composants
- Nouveaux matériaux, ...

  Important mais n'est plus la source dominante d'innovation

Conception d'Algorithmes

- Langages/Programmes Informatiques
- Modélisation mathématique
- ...

  Importance croissante

Internet

L'unité d'information de l'Internet : LE PAQUET

Un paquet : un entête + les données

- **L'entête**
  contient entre autres l'adresse de la machine qui doit recevoir le paquet
- **Les données**
  une partie du contenu du fichier transféré

L'unité d'information de l'Internet : LE PAQUET

Un paquet : un entête + les données

- **L'entête**
  contient entre autres l'adresse de la machine qui doit recevoir le paquet
- **Les données**
  une partie du contenu du fichier transféré

Exemples

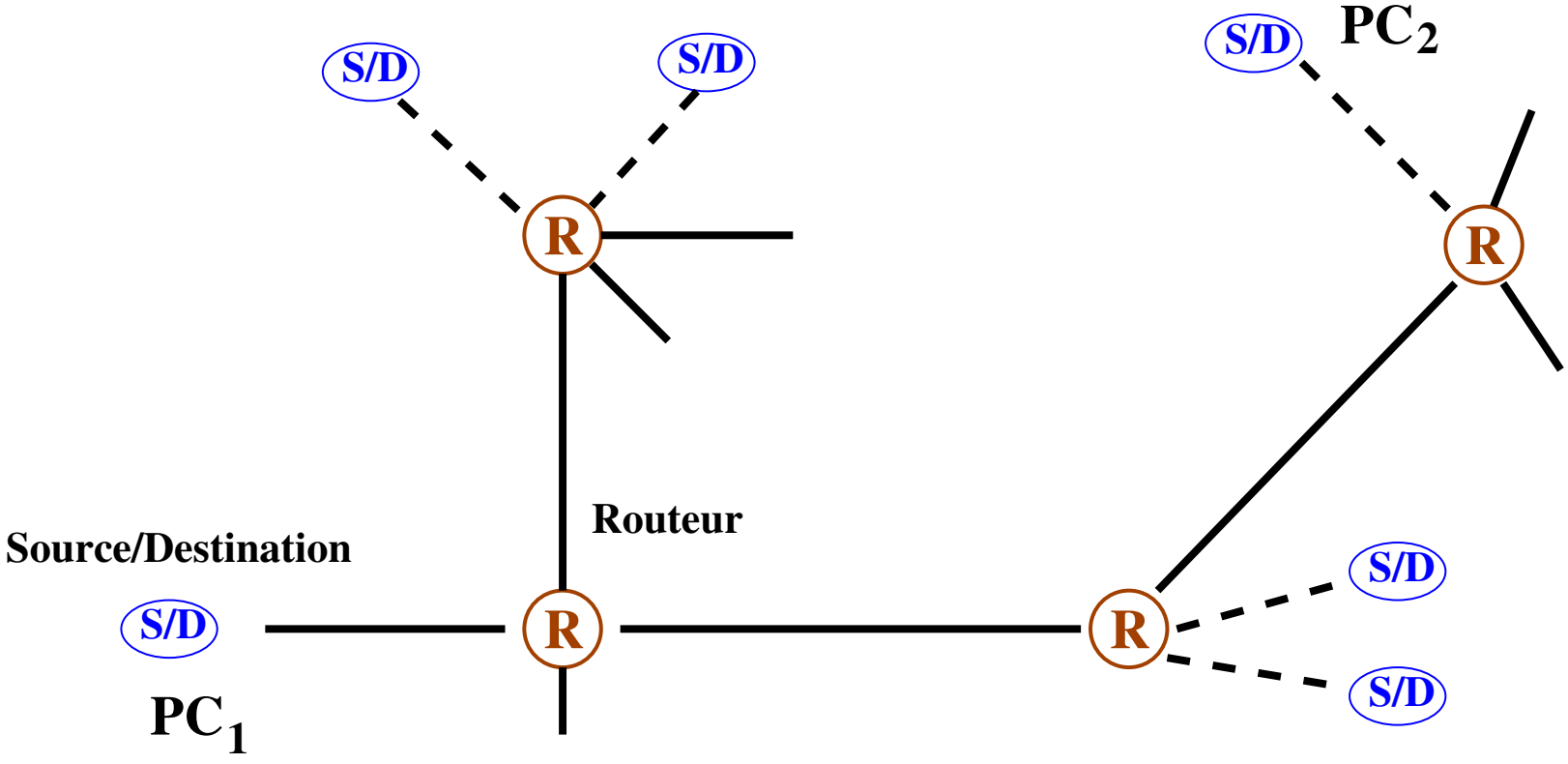
- Un CD mp3 : 400 000 paquets
- Un film : 4 000 000 paquets

Un réseau à commutation de paquets

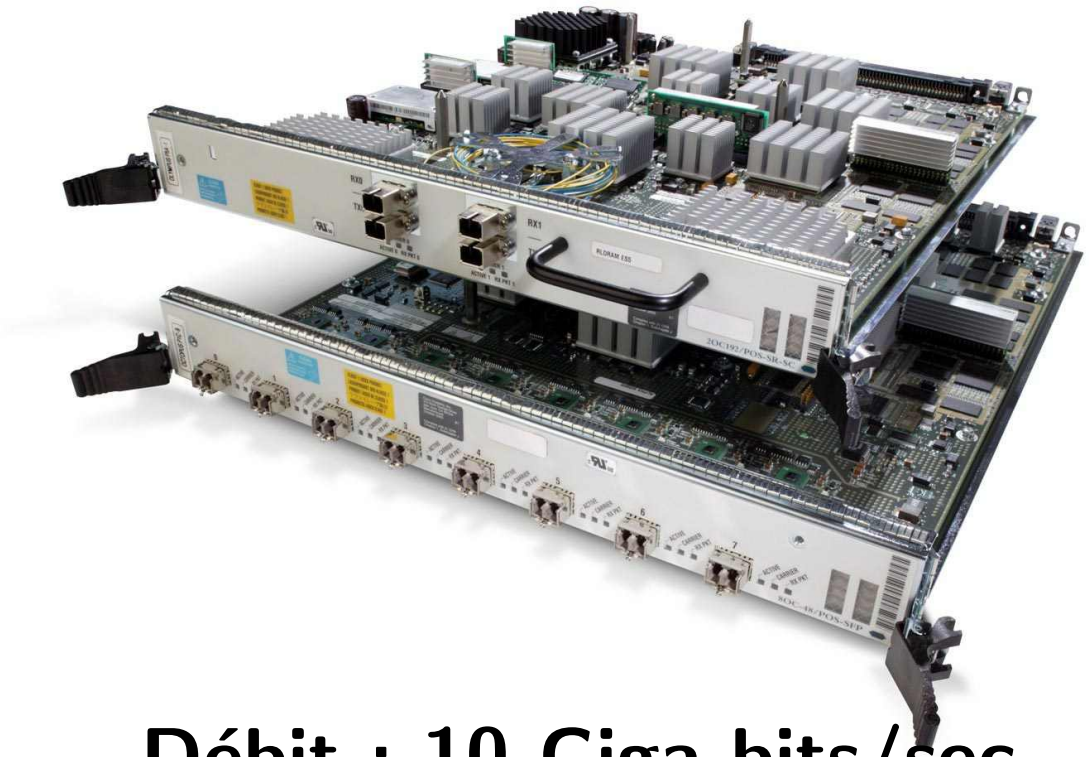
- Messages divisés en paquets
- Paquets acheminés individuellement

- Avantages
  - Système distribué
  - Flexibilité : Évolution facile

Internet : Une vue simplifiée



Un Routeur Cisco OC192

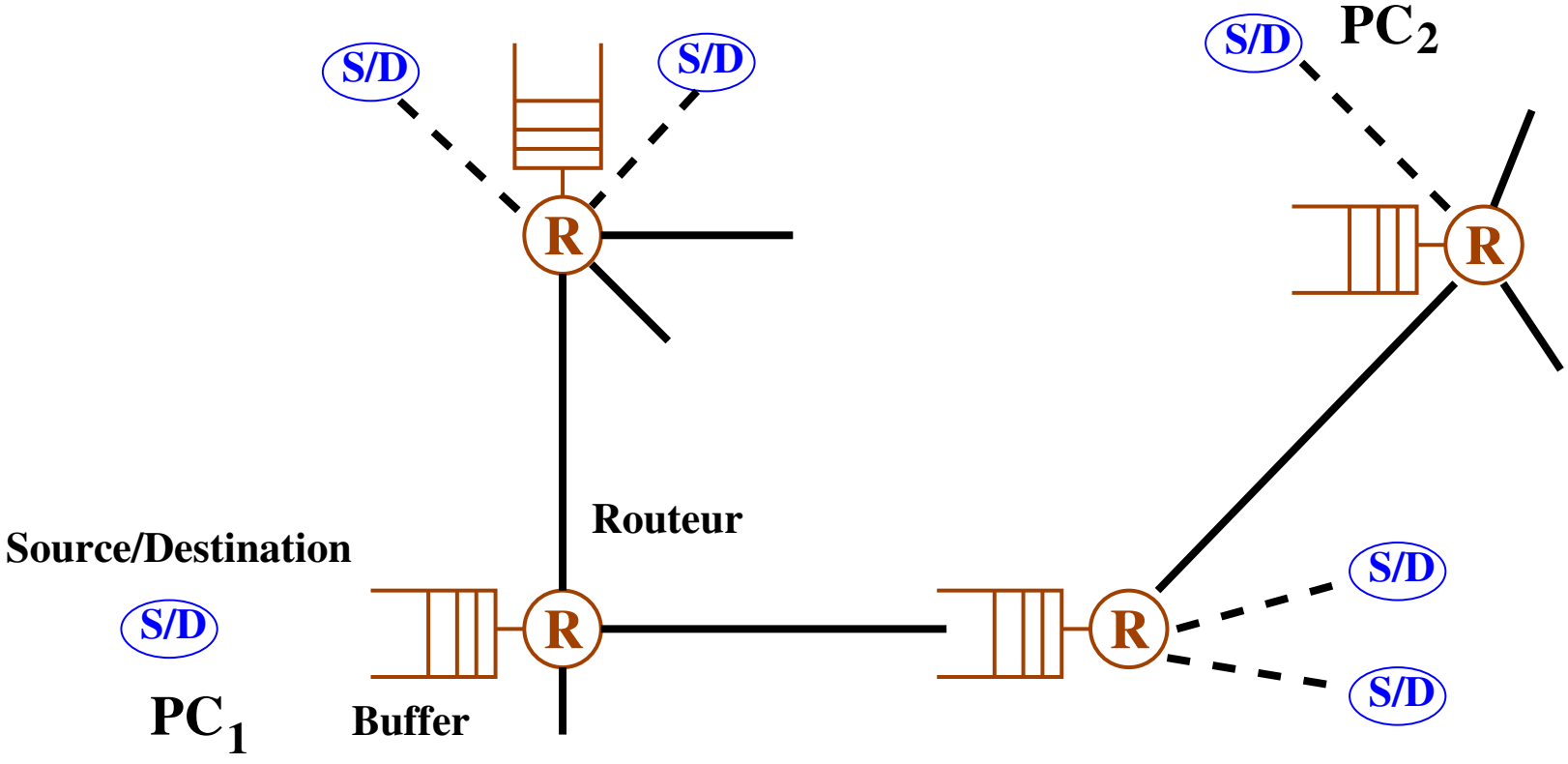


Débit : 10 Giga bits/sec

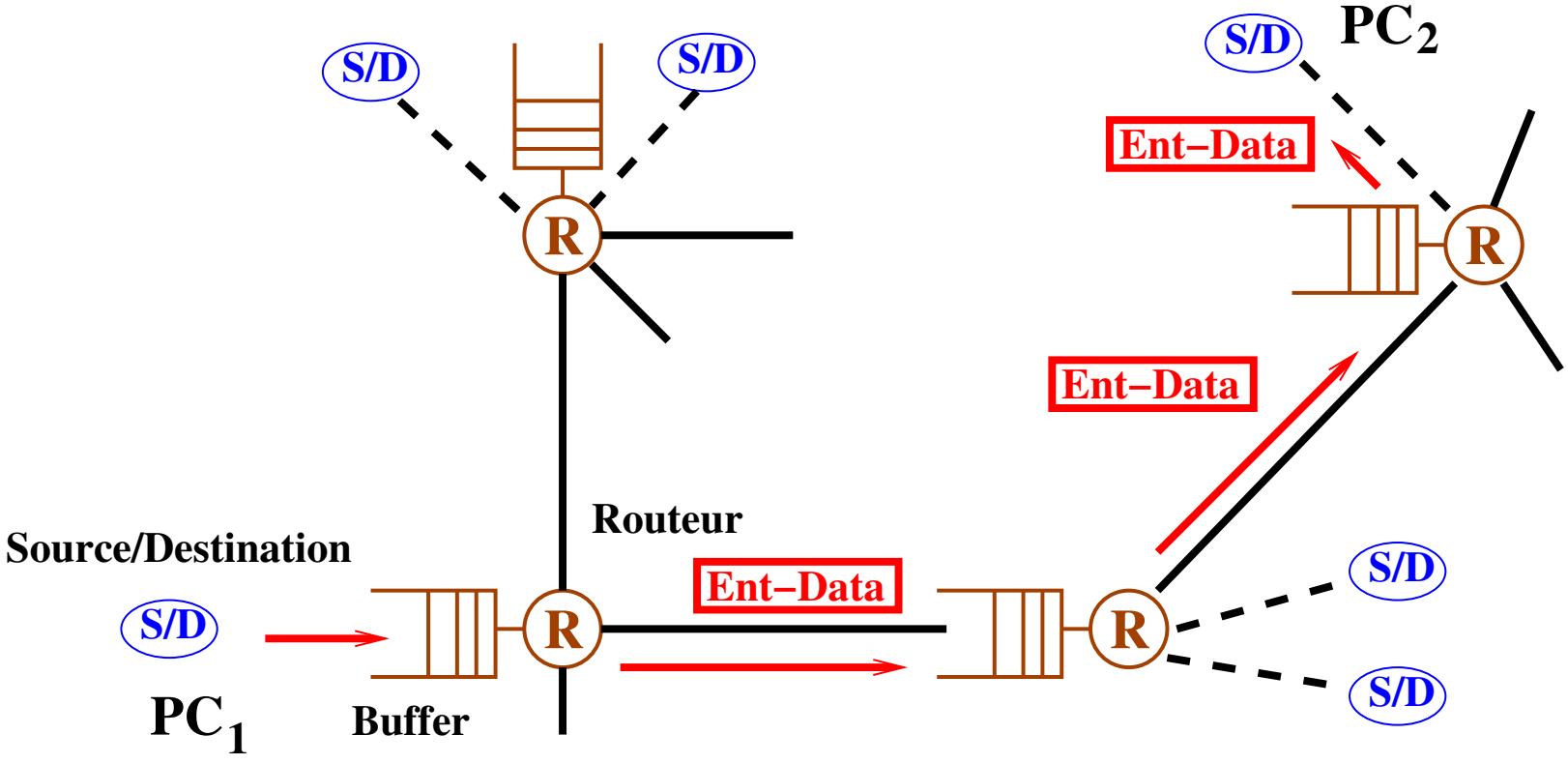
Trajet d'un paquet de Paris à Stanford (Californie)

0	@work	France
1	rocq-gw-ipv6.inria.fr	
2	gi4-1-inria-rtr-021.noc.renater.fr	
3	te2-5-paris1-rtr-021.noc.renater.fr	
4	te0-1-0-3-paris1-rtr-001.noc.renater.fr	
5	te0-1-0-4-paris2-rtr-001.noc.renater.fr	
6	hurricane-electric.franceix.net	
7	10gigabitethernet1-1.core1.par2.he.net	Paris
8	10gigabitethernet6-2.core1.lon1.he.net	Londres
9	10gigabitethernet7-4.core1.nyc4.he.net	New York
10	10gigabitethernet8-3.core1.chi1.he.net	Chicago
11	10gigabitethernet3-2.core1.den1.he.net	Denver
12	10gigabitethernet11-4.core1.sjc2.he.net	San Jose
13	10gigabitethernet3-2.core1.pao1.he.net	
14	stanford-university.he.net	

Internet : Une vue simplifiée



Internet : Une vue simplifiée



Transfert du fichier F de la machine PC_1 vers la machine PC_2

- **Sur chaque machine : un programme contrôle l'échange**

Transfert du fichier F de la machine PC_1 vers la machine PC_2

- Sur chaque machine : un programme contrôle l'échange

- PC_1 segmente en n paquets une copie de F et envoie le numéro 1, puis 2, ...n à PC_2

Transfert du fichier F de la machine PC_1 vers la machine PC_2

- Sur chaque machine : un programme contrôle l'échange

- PC_1 segmente en n paquets une copie de F et envoie le numéro 1, puis 2, ...n à PC_2

Mémoire des routeurs finie : en cas de congestion

$\Rightarrow$ Le réseau perd des paquets

Transmission de Données sur Internet

Problème : Comment transmettre de façon fiable dans un réseau qui ne l'est pas ?

Transmission de Données sur Internet

TCP : Transmission Control Protocol

- **Algorithme de transmission de données**
- **$>$ 95% du traffic Internet contrôlé par TCP**

Les principes de base TCP

Cerf and Kahn (1973)

- **Accusé de réception des messages**
- **Régulation des envois : à un instant une source a au plus W paquets en circulation dans le réseau**

W : Taille de la fenêtre de congestion

Les principes de base TCP (II)

Contrôle de la congestion Jacobson (1987)

- **Transmission de W paquets OK :**
  $W \rightarrow W + 1$

- **Un paquet est perdu :**
  $W \rightarrow W/2$

Conclusion sur TCP

+++ Adaptation aux conditions de trafic

-- Pas de garantie de débit, d'accès, ...

Conclusion sur TCP

+++ Adaptation aux conditions de trafic

-- Pas de garantie de débit, d'accès, ...

Remarquables propriétés d'auto-stabilisation

Google

Un peu d'histoire

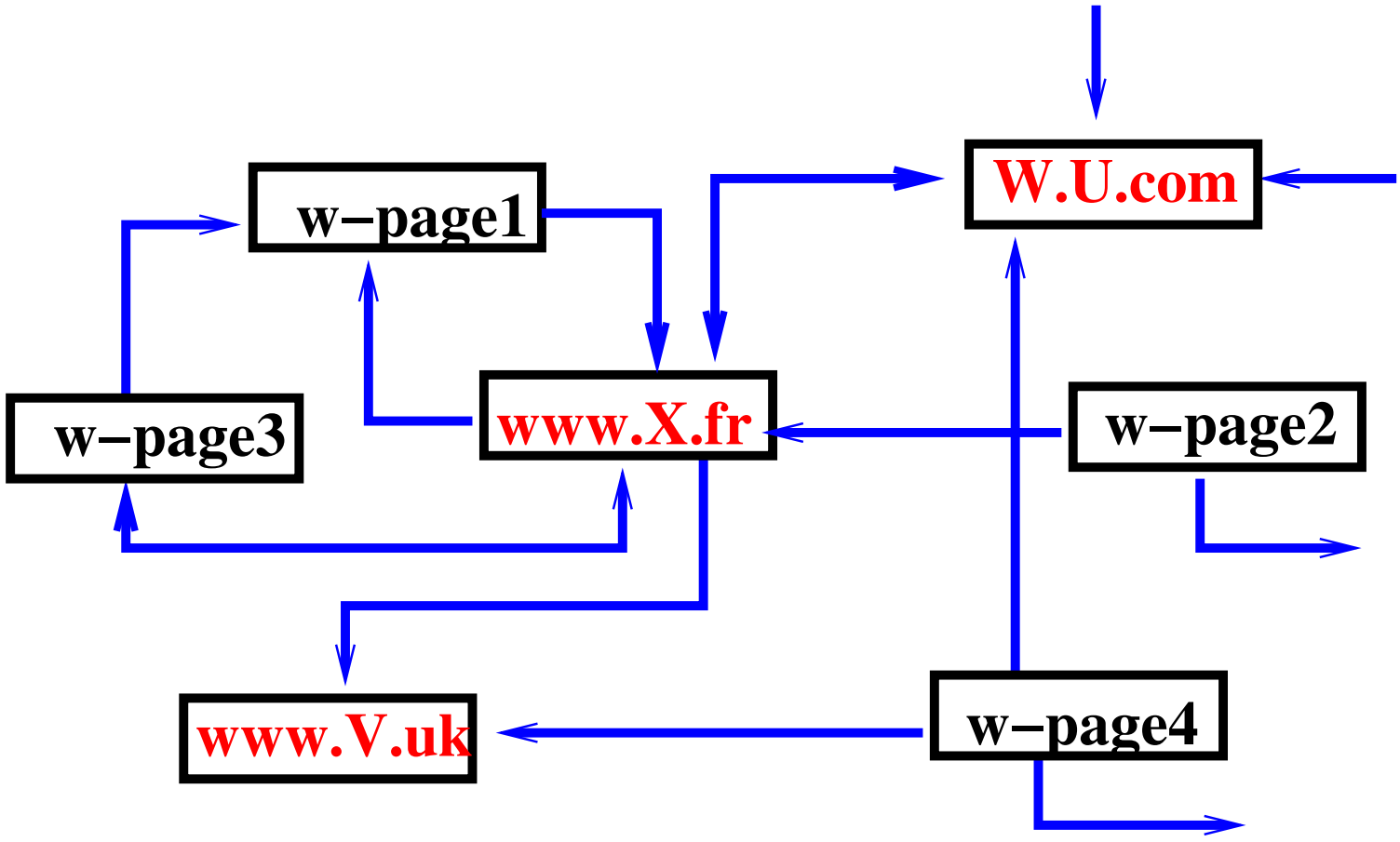
- **1995 : Premiers moteurs de recherche**
  AltaVista

- **1998 : Article Brin et Page**
  Fondateurs de Google
  "The anatomy of a largescale hypertextual web search engine"

Un peu d'histoire

- **1995 : Premiers moteurs de recherche**
  AltaVista

- **1998 : Article Brin et Page**
  Fondateurs de Google
  "The anatomy of a largescale hypertextual web search engine"
  - **Introduction de la notion de "Page rank"**
  - **Algorithme pour estimer celui-ci**



Le web : un graphe orienté

45 milliards de pages web indexées par Google

Comment Marche un Moteur de Recherche ?

Problème : Recherche d'un site web ayant une information sur le sujet "XYZ"

Comment Marche un Moteur de Recherche ?

Problème : Recherche d'un site web ayant une information sur le sujet "XYZ"

- **Première étape (facile)**
  Recherche de l'ensemble des sites web ayant ce mot "XYZ"

Comment Marche un Moteur de Recherche ?

Problème : Recherche d'un site web ayant une information sur le sujet "XYZ"

- **Première étape (facile)**
  Recherche de l'ensemble des sites web ayant ce mot "XYZ"

- **Deuxième étape**
  Quel est le site web le plus pertinent ?

Google snowden

Web News Images Videos More Search tools

About 22,300,000 results (0.27 seconds)

Edward **Snowden** - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Edward_**Snowden** Wikipedia

Edward Joseph **Snowden** (born June 21, 1983) is an American computer specialist, a former Central Intelligence Agency (CIA) employee, and former National ...

Prism - Global surveillance disclosures - Booz Allen Hamilton - Tempora

News for **snowden**

'The **Snowden** Files', by Luke Harding

Financial Times - 3 hours ago

The **Snowden** Files: The Inside Story of the World's Most Wanted Man, by Luke Harding. Guardian Faber Publishing RRP£12.99/Vintage ...

Books|The Needles in the Monumental NSA Haystack

New York Times - 1 day ago

Ex-NSA Chief Details **Snowden's** Hiring at Agency, Booz Allen

Wall Street Journal - 1 day ago

Edward **Snowden** | World news | The Guardian

www.theguardian.com › News › World news The Guardian

NSA whistleblower Edward **Snowden**: 'They're going to say I aided our enemies Video (7min 07sec): The second part of an exclusive first interview with former ...

Edward Snowden

System Administrator

Edward Joseph Snowden is an American computer speci
Central Intelligence Agency employee, and former Nation
Agency contractor who disclosed to several media outlet
number of top secret NSA documents. Wikipedia

Born: June 21, 1983 (age 30), Elizabeth City, North Caro
States

Nationality: American

Parents: Lonnie Snowden

Awards: Sam Adams Award

People also search for

Comment Marche un Moteur de Recherche ?

Principe : **Trouver une fonction π telle que :**
À une page web p on associe $\pi(p) \in [0, 1]$

p plus pertinent que q si $\pi(p) > \pi(q)$

Comment Marche un Moteur de Recherche ?

Principe : Trouver une fonction π telle que :
À une page web p on associe $\pi(p) \in [0, 1]$

p plus pertinent que q si $\pi(p) > \pi(q)$

$\mathcal{E}_{XYZ} = \{p : p \text{ page web contenant “XYZ”}\}$

Comment Marche un Moteur de Recherche ?

Principe : Trouver une fonction π telle que :
À une page web p on associe $\pi(p) \in [0, 1]$

p plus pertinent que q si $\pi(p) > \pi(q)$

$\mathcal{E}_{XYZ} = \{p : p \text{ page web contenant “XYZ”}\}$

Action : Afficher les pages $p_1, \ldots p_{10}$ ayant les 10 plus grandes valeurs pour π sur $\mathcal{E}_{XYZ}$

La fonction π pour Google

Brin et Page (1997)

π(q) : Importance de la page q

$$\mathcal{L}_{\mathbf{q}} = \{\mathbf{p} : \mathbf{p} \text{ a un lien vers la page web } \mathbf{q}\}$$

La fonction π pour Google

Brin et Page (1997)

$\pi(q)$: Importance de la page q

$$\mathcal{L}_q = \{p : p \text{ a un lien vers la page web } q\}$$

Principe :

Importance de p "transmise/héritée" de q :

$$\pi(q)M(q,p)$$

avec

$$M(q,p) = \frac{1}{|\mathcal{L}_q|}$$

La fonction π pour Google

Importance de q transmise à p

$$\pi(q)M(q,p)$$

La fonction π pour Google

Équation linéaire pour π

$$\pi(\mathbf{p}) = \sum_{\mathbf{q}:\mathbf{p}\in\mathcal{L}_\mathbf{q}} \pi(\mathbf{q})\mathbf{M}(\mathbf{q},\mathbf{p})$$

La fonction π pour Google

Équation linéaire pour π

$$\pi(p) = \sum_{q:p\in\mathcal{L}_q} \pi(q)M(q,p)$$

Le système est singulier de rang $|\mathcal{S}| - 1$

$$\sum_{p:p\in\mathcal{L}_q} M(q,p) = \sum_{p:p\in\mathcal{L}_q} \frac{1}{|\mathcal{L}_q|} = 1$$

$\mathcal{S}$: ensemble de toutes les pages web

La fonction π pour Google

Si $\mathcal{M} = (\mathrm{M}(\mathrm{p},\mathrm{q}), \mathrm{p},\mathrm{q} \in \mathcal{S})$, il existe un unique vecteur $(\pi(\mathrm{p}), \mathrm{p} \in \mathcal{S})$ tel que

$$\pi = \pi\mathcal{M}$$

et

$$\sum_{\mathrm{p}\in\mathcal{S}} \pi(\mathrm{p}) = 1$$

La fonction π pour Google

Si $\mathcal{M} = (\mathrm{M}(\mathrm{p},\mathrm{q}), \mathrm{p},\mathrm{q} \in \mathcal{S})$, il existe un unique vecteur $(\pi(\mathrm{p}), \mathrm{p} \in \mathcal{S})$ tel que

$$\pi = \pi\mathcal{M}$$

et

$$\sum_{\mathrm{p}\in\mathcal{S}} \pi(\mathrm{p}) = 1$$

Propriété : $\pi(\mathrm{p}) \in (0,1), \forall \mathrm{p} \in \mathcal{S}$

La fonction π pour Google

Si $\mathcal{M} = (\mathrm{M}(\mathrm{p},\mathrm{q}), \mathrm{p},\mathrm{q} \in \mathcal{S})$, il existe un unique vecteur $(\pi(\mathrm{p}), \mathrm{p} \in \mathcal{S})$ tel que

$$\pi = \pi\mathcal{M}$$

et

$$\sum_{\mathrm{p}\in\mathcal{S}} \pi(\mathrm{p}) = 1$$

Propriété : $\pi(\mathrm{p}) \in (0,1), \forall \mathrm{p} \in \mathcal{S}$

Un détail :

le nombre de pages dans $\mathcal{S}$ est 50 milliards !

La fonction π pour Google en pratique

Analyse numérique :

- **Produits matrice/vecteurs**
- **Techniques d'uniformisation**

La fonction π pour Google

La fonction π pour Google

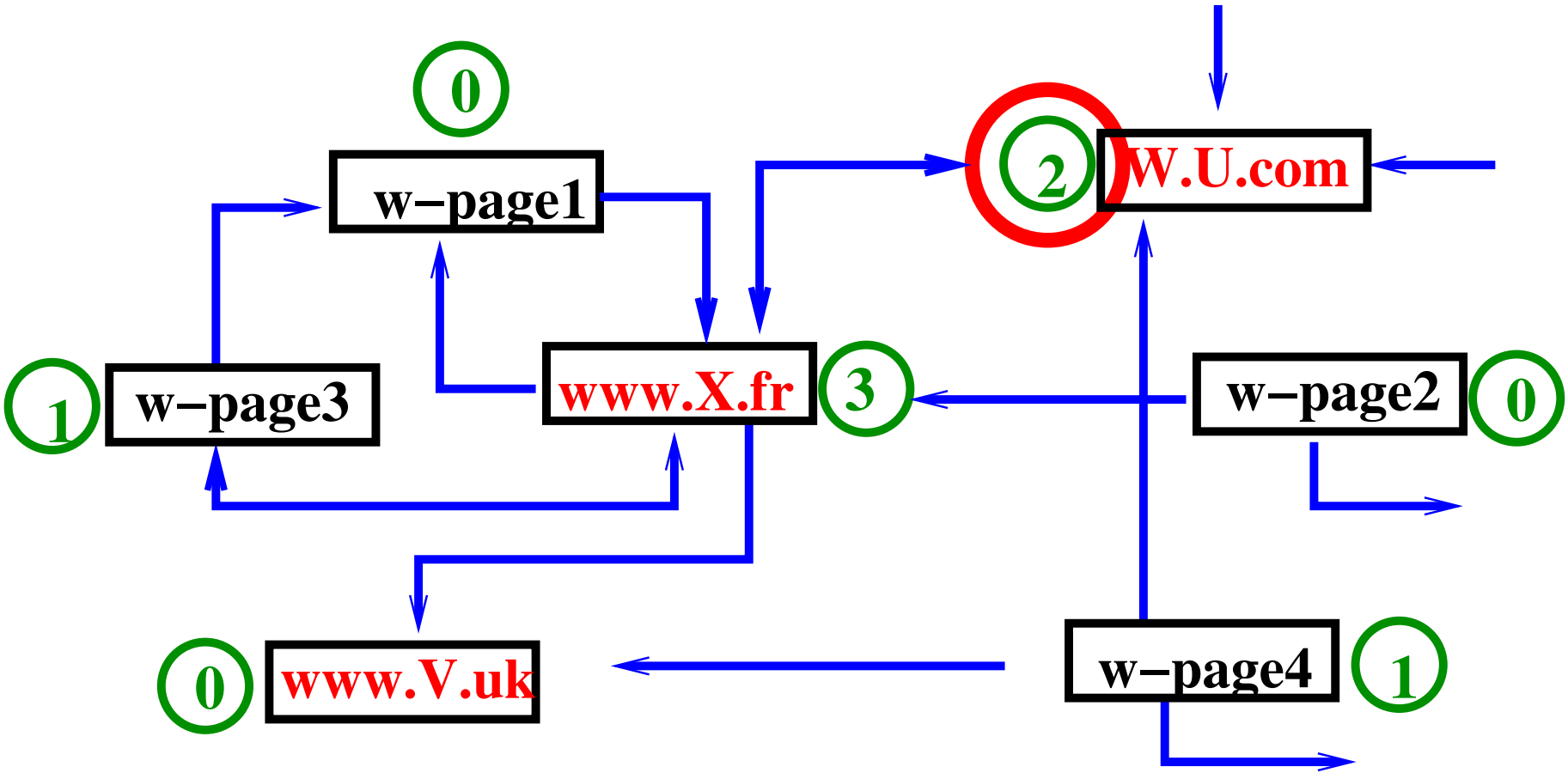
Modèle mathématique : surfeur aléatoire S

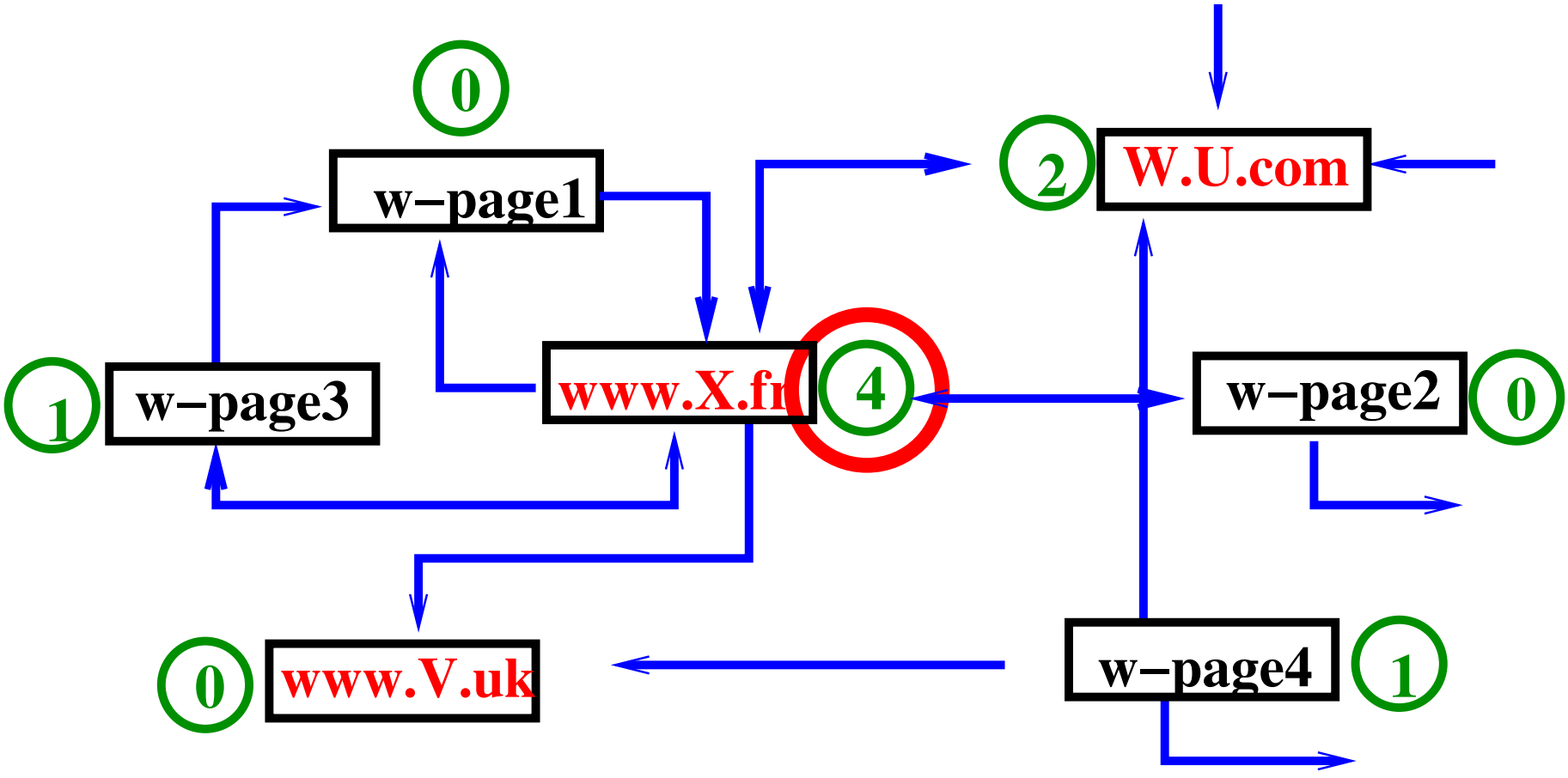
La fonction π pour Google

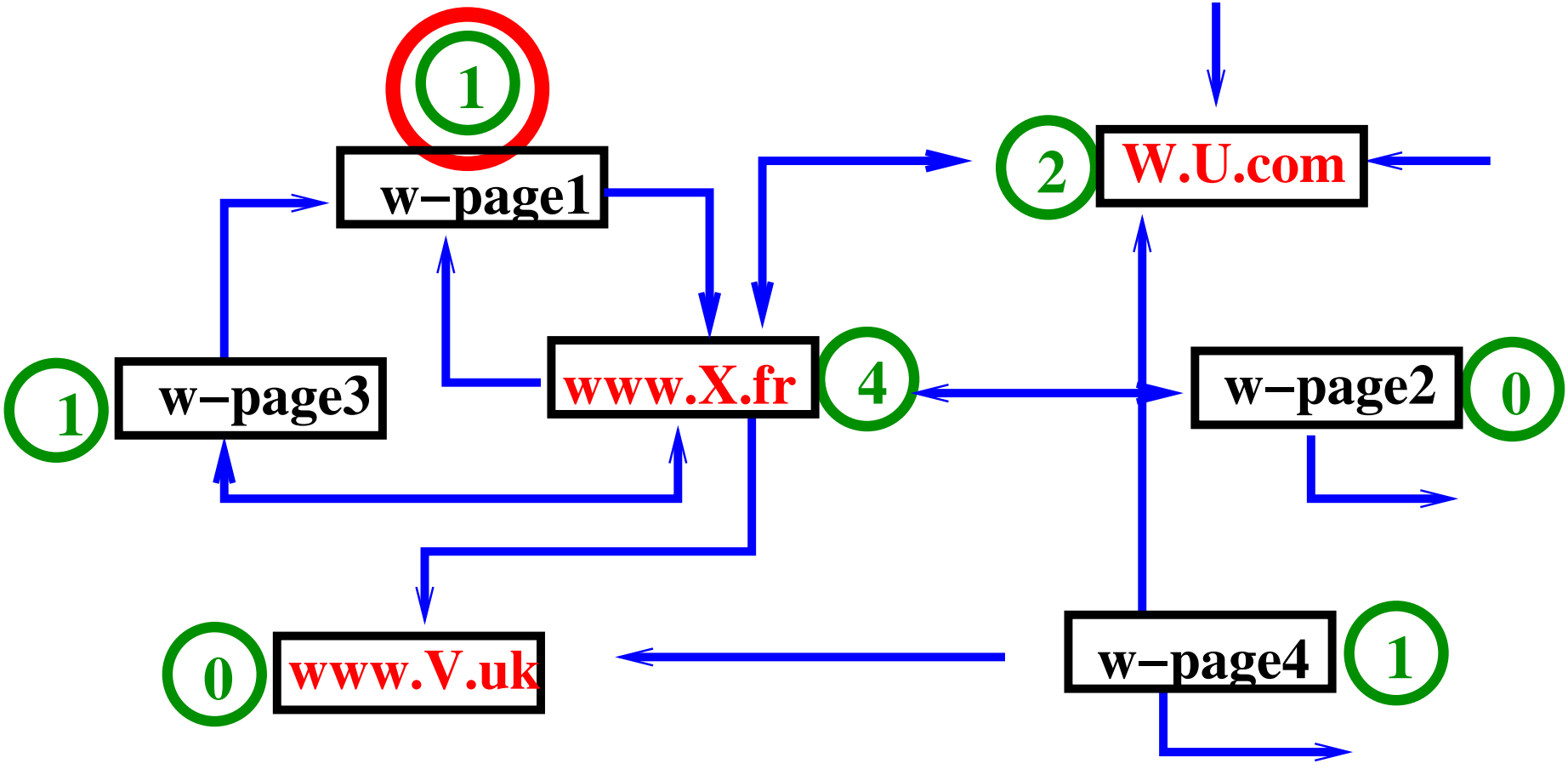
Modèle mathématique : surfeur aléatoire S

- **S navigue au hasard sur le web :**

  Si S sur une page web à l'instant t à $t+1$, S choisit au hasard un lien de cette page et va sur la page web correspondante, etc...


0
w–page1
2
W.U.com
1
w–page3
www.X.fr
3
w–page2
0
0
www.V.uk
w–page4
1


0
w–page1
2
W.U.com
1
w–page3
www.X.fr
4
w–page2
0
0
www.V.uk
w–page4
1


1
w–page1
2
W.U.com
1
w–page3
www.X.fr
4
w–page2
0
0
www.V.uk
w–page4
1

La fonction π pour Google

Durée du surf $T = 10^{12}$ (par exemple)

Si p est une page web,

$$f_T(p) = \frac{1}{T} N_T(p),$$

où $N_T(p)$: nb de passages à p entre 0 et T

La fonction π pour Google

Durée du surf $T = 10^{12}$ (par exemple)

Si p est une page web,

$$f_T(p) = \frac{1}{T} N_T(p),$$

où $N_T(p)$: nb de passages à p entre 0 et T

p_1 plus pertinent que p_2 si $f_T(p_1) > f_T(p_2)$

La fonction π pour Google

Durée du surf $T = 10^{12}$ (par exemple)

Si p est une page web,

$$f_T(p) = \frac{1}{T} N_T(p),$$

où $N_T(p)$: nb de passages à p entre 0 et T

p_1 plus pertinent que p_2 si $f_T(p_1) > f_T(p_2)$

Problèmes

- **Dépend de T ?**

La fonction π pour Google

Durée du surf $T = 10^{12}$ (par exemple)

Si p est une page web,

$$f_T(p) = \frac{1}{T} N_T(p),$$

où $N_T(p)$: nb de passages à p entre 0 et T

p_1 plus pertinent que p_2 si $f_T(p_1) > f_T(p_2)$

Problèmes

- **Dépend de T ?**
- **Dépend du point de départ du surfeur ?**

La fonction π pour Google

Durée du surf $T = 10^{12}$ (par exemple)

Si p est une page web,

$$f_T(p) = \frac{1}{T} N_T(p),$$

où $N_T(p)$: nb de passages à p entre 0 et T

p_1 plus pertinent que p_2 si $f_T(p_1) > f_T(p_2)$

Problèmes

- **Dépend de T ?**
- **Dépend du point de départ du surfeur ?**
- **Dépend des choix aléatoires du surfeur ?**

Résultats de Maths

La limite existe :

$$\nu(p) = \lim_{T \to +\infty} \frac{N_T(p)}{T}$$

Résultats de Maths

La limite existe :

$$\nu(p) = \lim_{T \to +\infty} \frac{N_T(p)}{T}$$

- **ne dépend donc pas de T (T assez grand)**

Résultats de Maths

La limite existe :

$$\nu(p) = \lim_{T \to +\infty} \frac{N_T(p)}{T}$$

- **ne dépend donc pas de T (T assez grand)**
- **ne dépend pas du point de départ**

ν est en fait l'unique solution de

$$\nu = \nu\mathcal{M} \text{ et } \sum_{x \in \mathcal{S}} \nu(x) = 1$$

Résultats de Maths

La limite existe :

$$\nu(p) = \lim_{T \to +\infty} \frac{N_T(p)}{T}$$

- ne dépend donc pas de T (T assez grand)
- ne dépend pas du point de départ

ν est en fait l'unique solution de

$$\nu = \nu \mathcal{M} \text{ et } \sum_{x \in \mathcal{S}} \nu(x) = 1$$

$\nu = \pi$

Conclusions

Idées brillantes de Brin et Page :

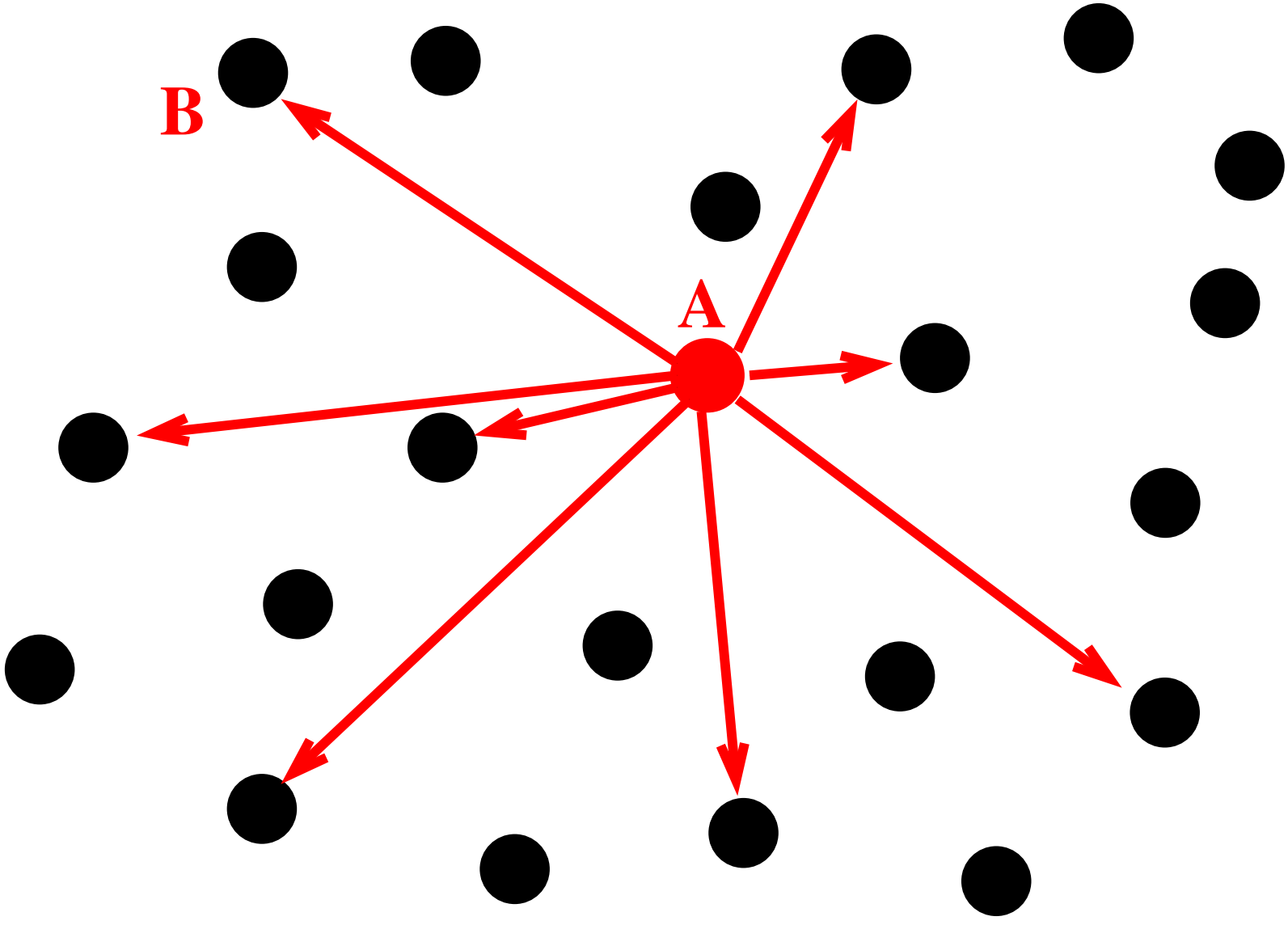
- **Modélisation :**
  Représentation mathématique du “page rank”

- **Algorithme de calcul de π**

Les réseaux sociaux

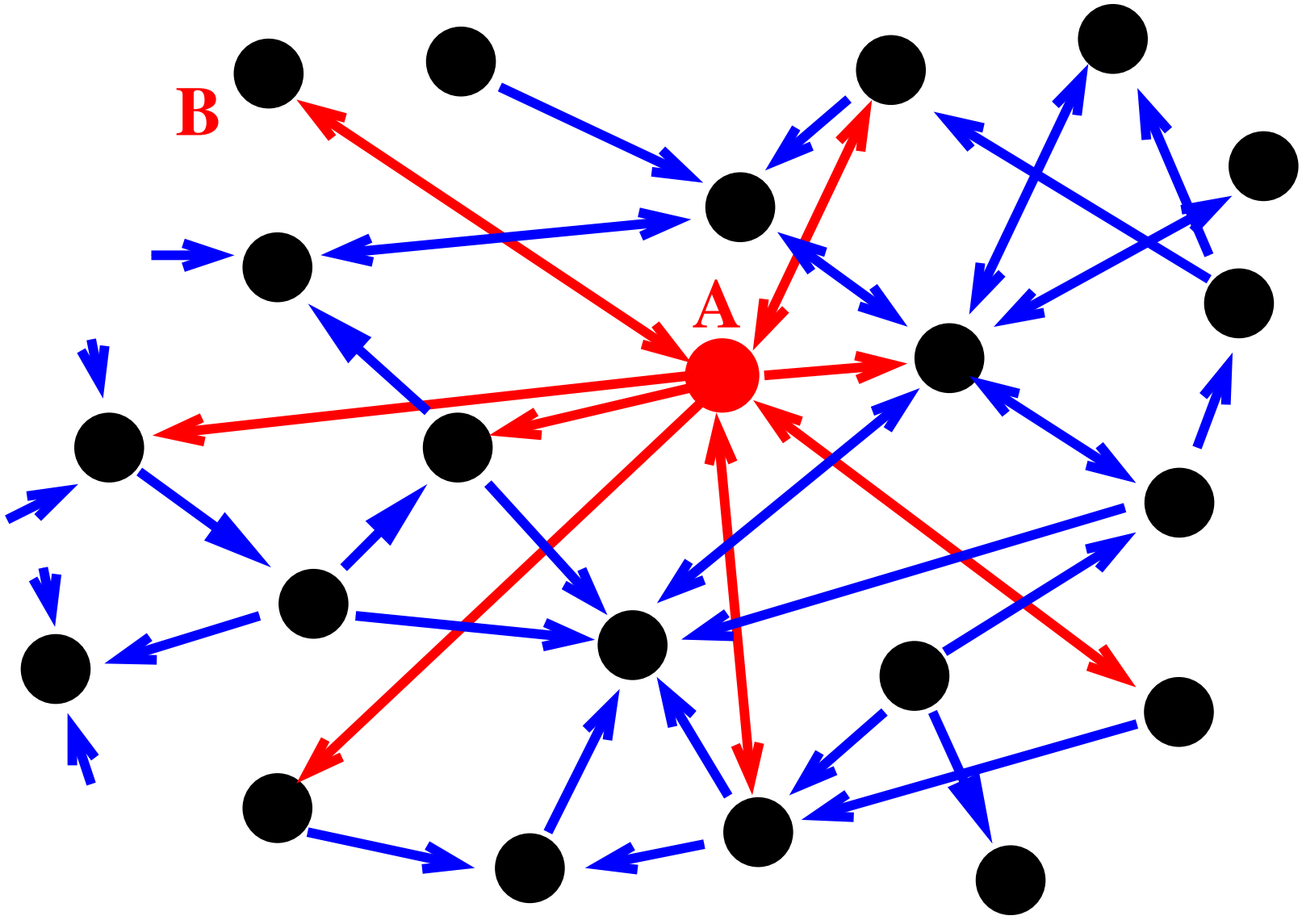
Réseaux sociaux

- **Réseaux par affinité : A ami de B : un lien de A vers B.**


B
A

Réseaux sociaux

- Réseaux par affinité : A ami de B : un lien de A vers B.
- Très grand nombre de nœuds
  Facebook : 1 milliard


B
A

Réseaux sociaux

Problème :
Comment extraire de l'information de ces réseaux ?

Réseaux sociaux

Problème :

Comment extraire de l'information de ces réseaux ?

Un domaine actif de recherche

- **Data Mining (Fouille de données)**
  Algorithmes pour structurer les données.
- **Typologie des Réseaux Sociaux**
  Caractérisation/Estimation des graphes
- **Navigation**
  Algorithmes pour se déplacer.

Réseaux sociaux

Problème :

Comment extraire de l'information de ces réseaux ?

Un domaine actif de recherche

- **Data Mining (Fouille de données)**
  Algorithmes pour structurer les données.
- **Typologie des Réseaux Sociaux**
  Caractérisation/Estimation des graphes
- **Navigation**
  Algorithmes pour se déplacer.

Enjeux économiques

The New York Times

Business Day
Technology

Facebook Reasserts Posts Can Be Used to Advertise

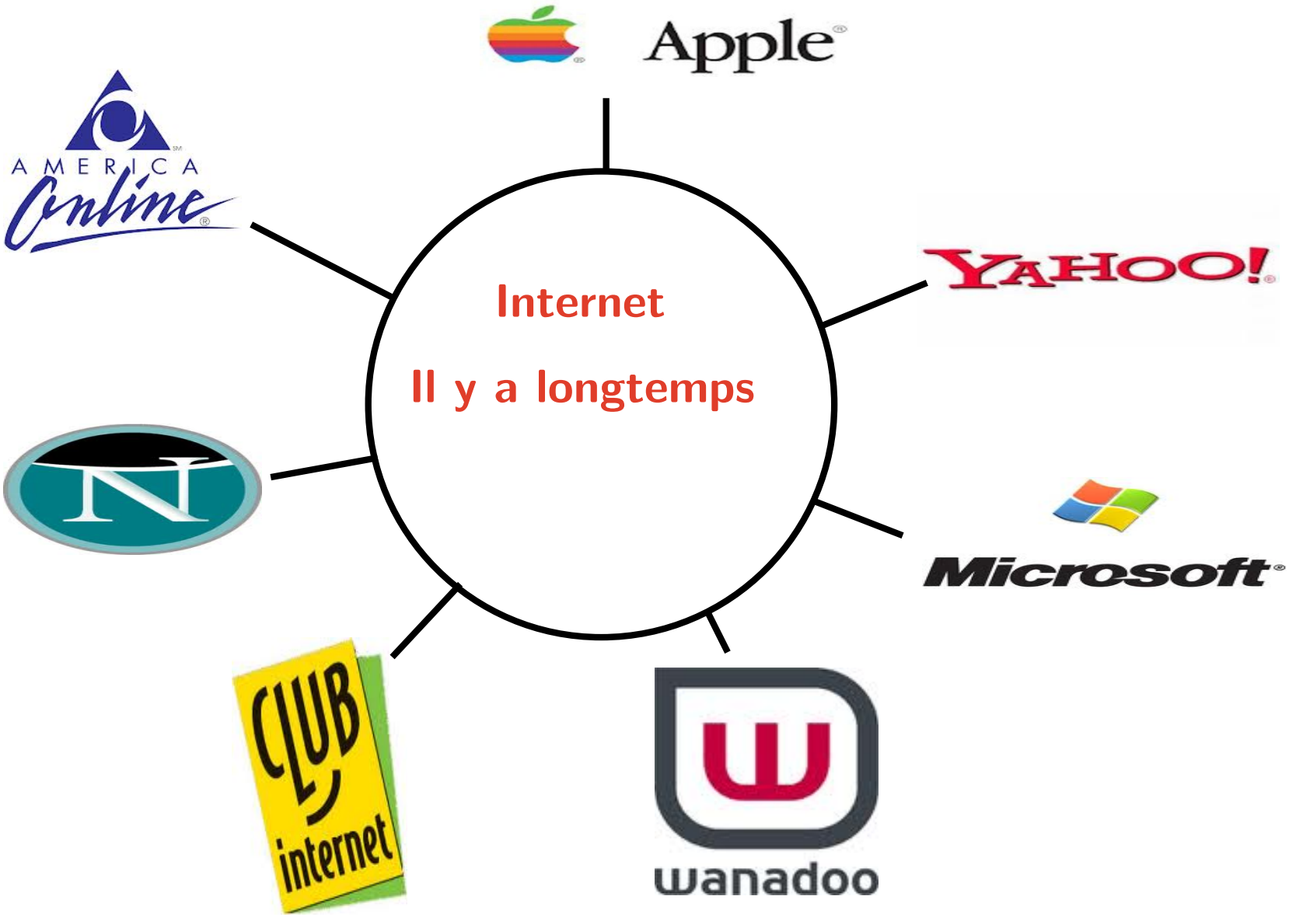
Nam Y. Huh/Associated Press

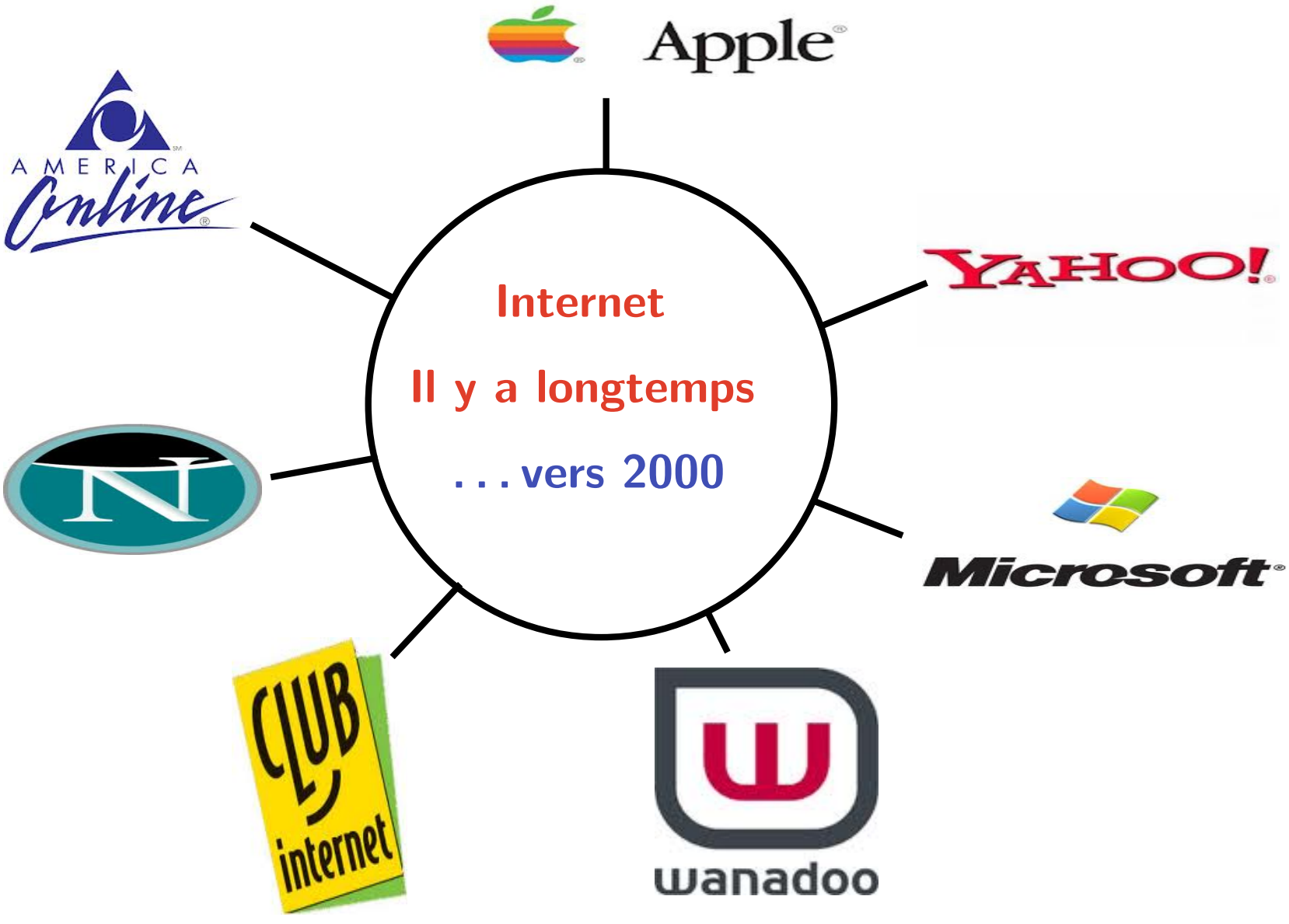
Mark Risinger, 16, checking his Facebook page in October. The social network's privacy practices always draw a great deal of attention.

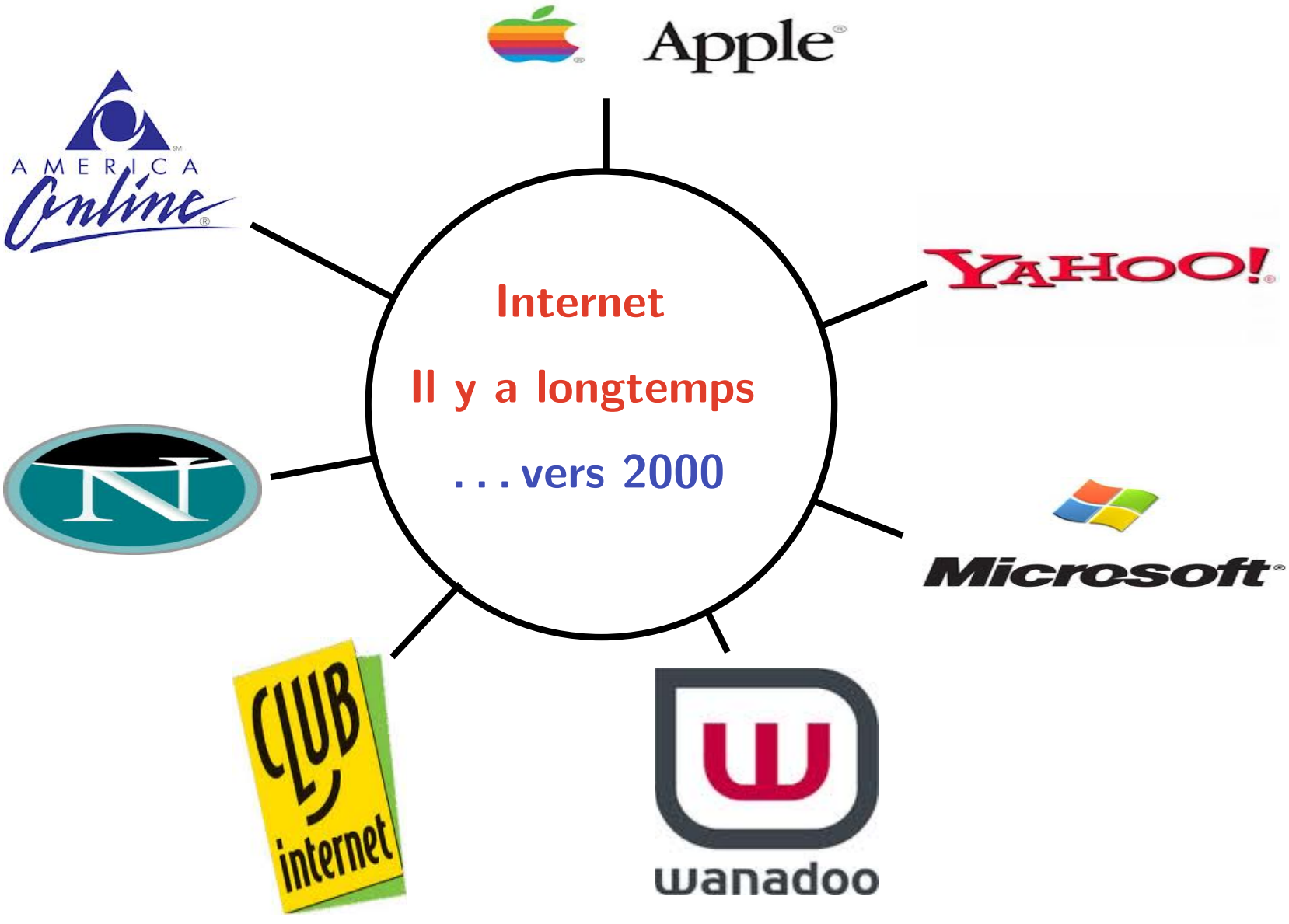
By VINDU GOEL
Published: November 15, 2013

SAN FRANCISCO — If you post something on Facebook, let there be no doubt that it can end up as an ad shown to your friends and

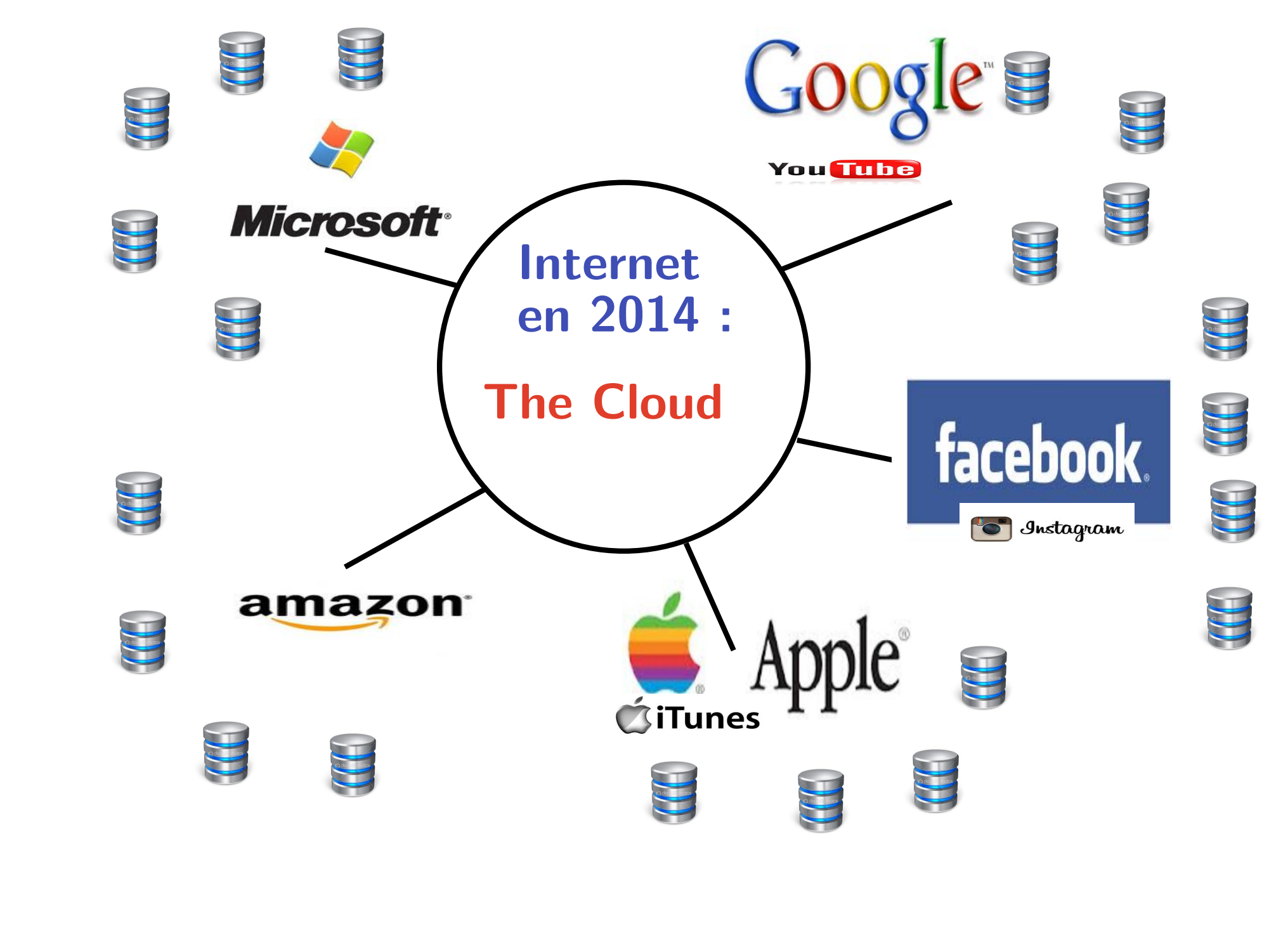
Internet maintenant

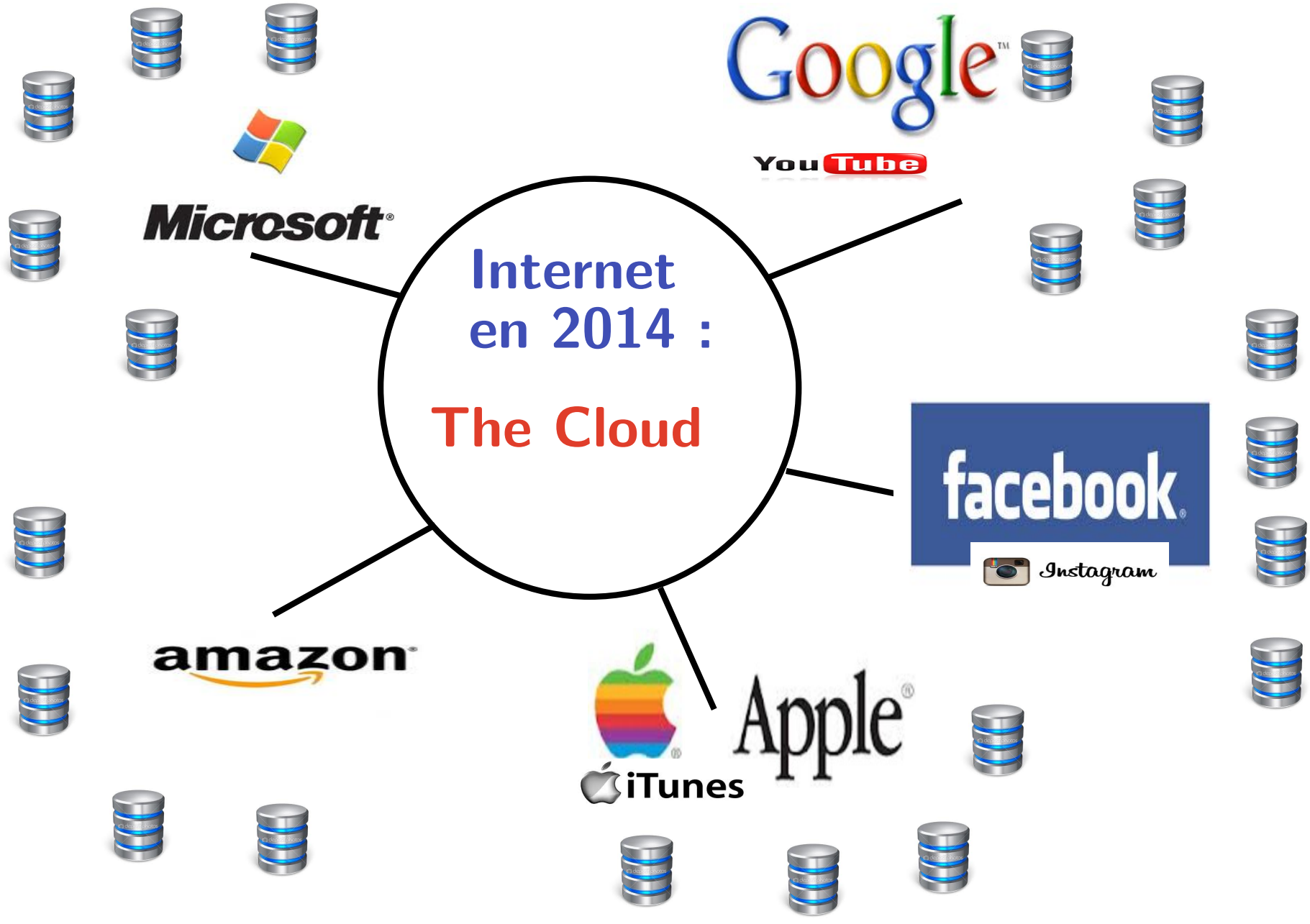

Apple
AMERICA Online
Internet
Il y a longtemps
YAHOO!
Microsoft
CLUB internet
wanadoo


Apple
AMERICA Online
YAHOO!
Internet
Il y a longtemps
. . . vers 2000
Microsoft
CLUB internet
wanadoo



Accès à certaines informations


Internet
en 2014 :
The Cloud
Google
YouTube
Microsoft
facebook
Instagram
amazon
Apple
iTunes



The Cloud : Calcul et Stockage des données

Les Nombres du Cloud

Nombre de serveurs (Estimation) :

- **Google : 900 000**
- **Microsoft : 518 000**
- **Amazon : 445 000**
- **HP/EDS : 380 000**
- **OVH : 120 000**
- **Facebook : 60 000**
- **...**

Les Modèles Économiques des Réseaux

1960 — 1985 IBM

Machine+programme Informatique spécifique

Les Modèles Économiques des Réseaux

1960 — 1985 IBM

Machine+programme Informatique spécifique

1985 — 201 ? Microsoft

Programme Informatique généraliste

Les Modèles Économiques des Réseaux

1960 — 1985 IBM

Machine+programme Informatique spécifique

1985 — 201 ? Microsoft

Programme Informatique généraliste

2000 — ? Google

Recherche sur Internet

Les Modèles Économiques des Réseaux

1960 — 1985 IBM

Machine+programme Informatique spécifique

1985 — 201 ? Microsoft

Programme Informatique généraliste

2000 — ? Google

Recherche sur Internet

2010 — Fermes de données :

Google, Amazon, Apple, Facebook, . . .

Les Modèles Économiques des Réseaux

1960 — 1985 IBM

Machine+programme Informatique spécifique

1985 — 201 ? Microsoft

Programme Informatique généraliste

2000 — ? Google

Recherche sur Internet

2010 — Fermes de données :

Google, Amazon, Apple, Facebook, . . .

+ Réseaux Sociaux

Facebook, Twitter, Snapchat, . . .

La Fin