Introduction à l'exploration du web par la théorie des graphes

Mathieu Jacomy
WebAtlas
Telecom ParisTech

Web Atlas

SciencesPo. | médialab

TELECOM ParisTech

Gephi

Dans cette présentation

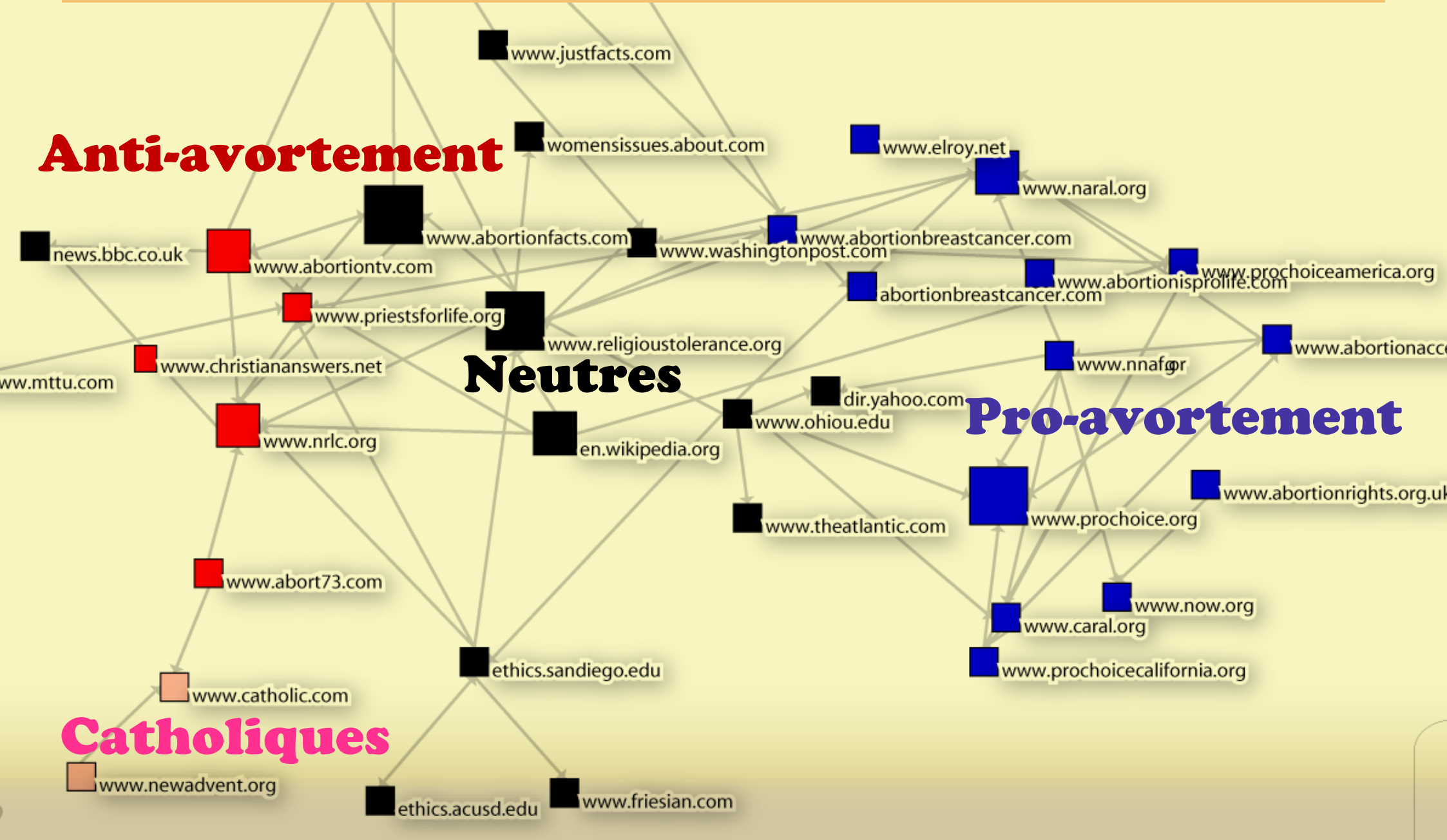
Un petit **exemple** introductif

Beaucoup de **théorie des graphes**
(sans trop de maths)

Le web et comment on l'explore

Un introduction à deux **outils** :
Le Navicrawler et Gephi

Un exemple basique : qui est proche de qui ?



Qu'est-ce que ça représente ? (1/3)

La position des éléments
dépend de leurs relations

Elle n'est pas une projection statistique
(Elle ne dépend pas de leur "couleur")

C'est pourquoi
on peut "lire" les positions

Qu'est-ce que ça représente ? (2/3)

Exemple simple :

Positions -> Formuler des hypothèses
Relations -> Valider les hypothèses

Cas complexe :

Les relations sont embrouillées
-> On valide avec des statistiques

Qu'est-ce que ça représente ? (3/3)

Lecture “en plein” :

Les neutres font le pont
entre pro- et anti-avortement

…

Lecture “en creux” :

Les pro- et anti- ne se connectent pas

…

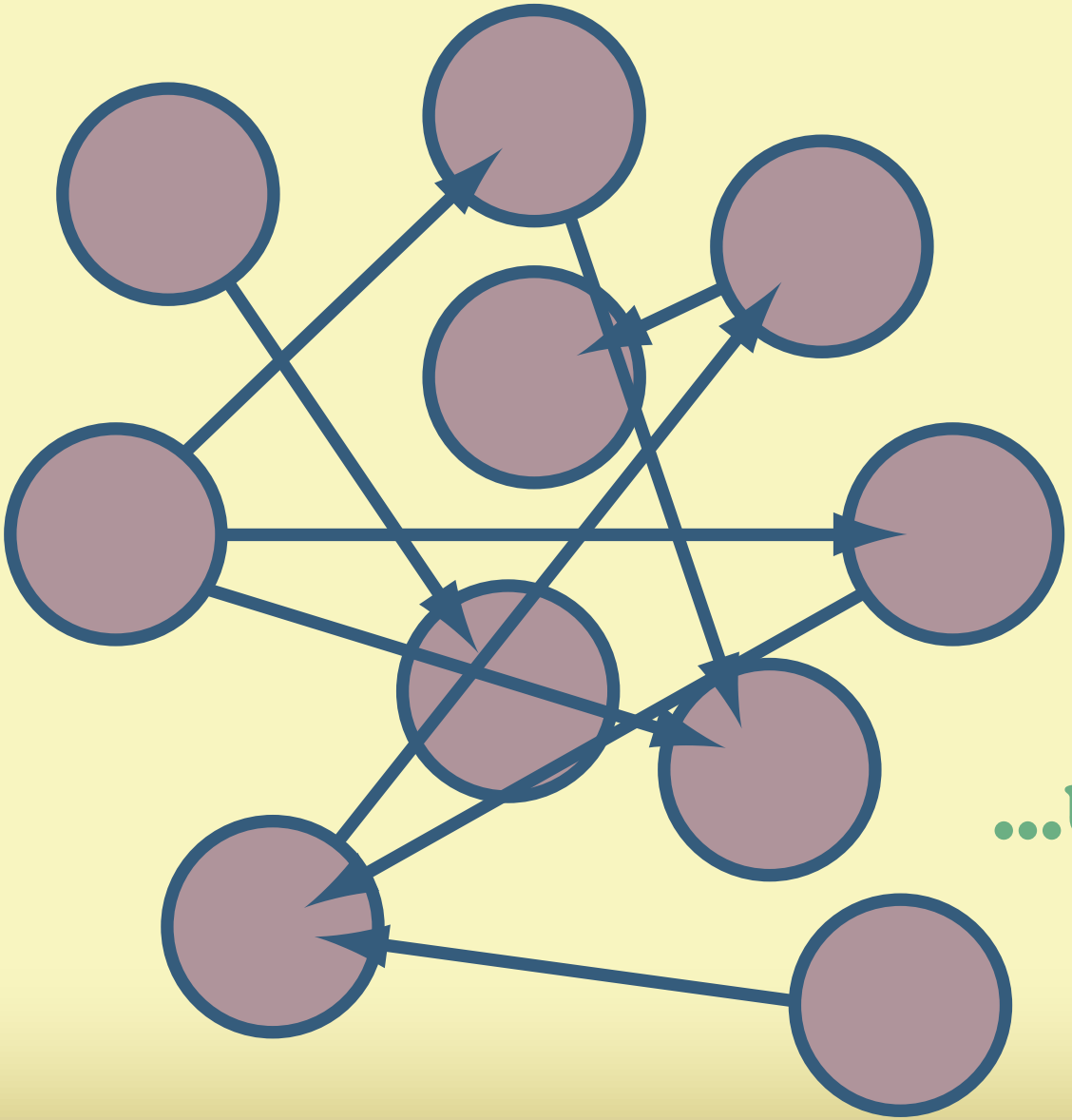
Quelle sorte de complexité ?

1) L'objet "graphe"

Le graphe (1/11)

Deux acteurs, l'un connaît l'autre
...ou Deux sites web, l'un a un lien vers l'autre
...ou Deux mots, l'un est associé à l'autre...

Le graphe (2/11)



...Un graphe

Le graphe (3/11)

Dans le cas de données **analogiques**,
le graphe est un outil de **modélisation**

Exemple : le réseau routier

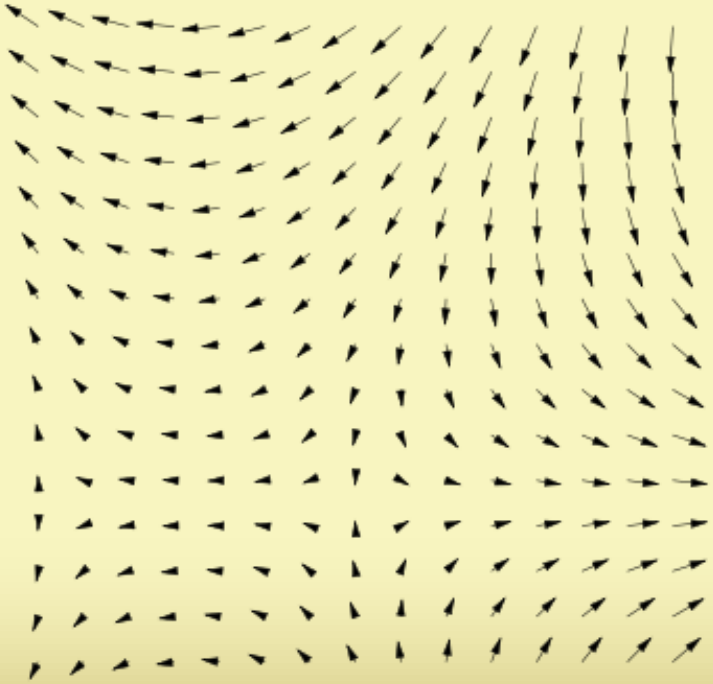
Dans le cas de données **formelles**,
le graphe est une **projection**

Exemple : graphe scientométrique d'une BDD
(c'est la base qui est le modèle)

Le graphe (4/11)

Le graphe est une structure mathématique

En ce sens, c'est un espace
(au sens par ex. de Bourbaki)



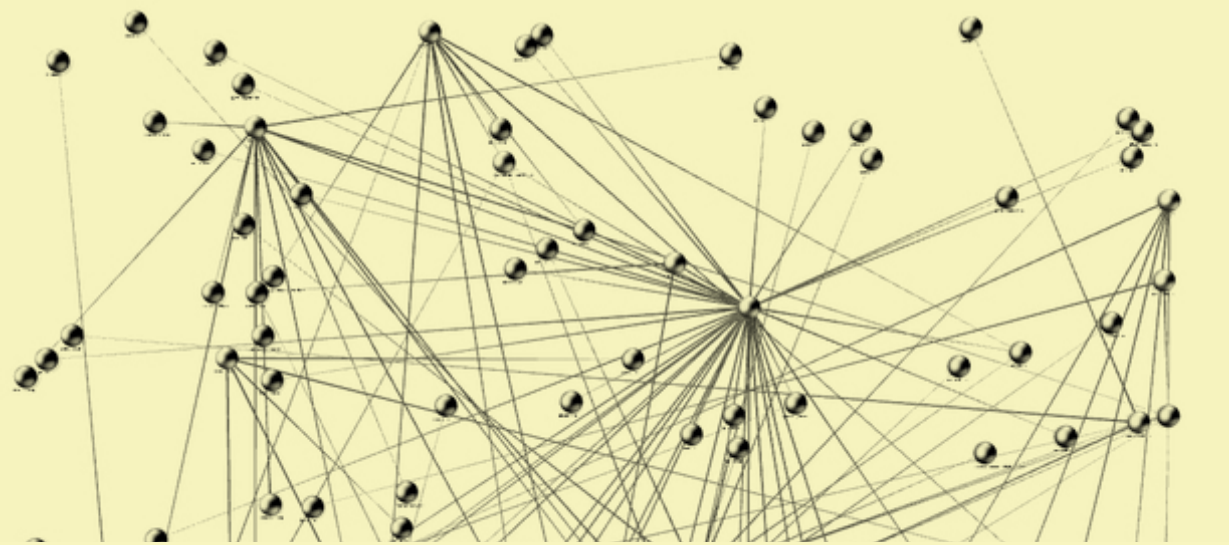
Comme un espace vectoriel :

Lieux

Qui renvoient
les uns aux autres

Traçant des chemins

Le graphe (5/11)



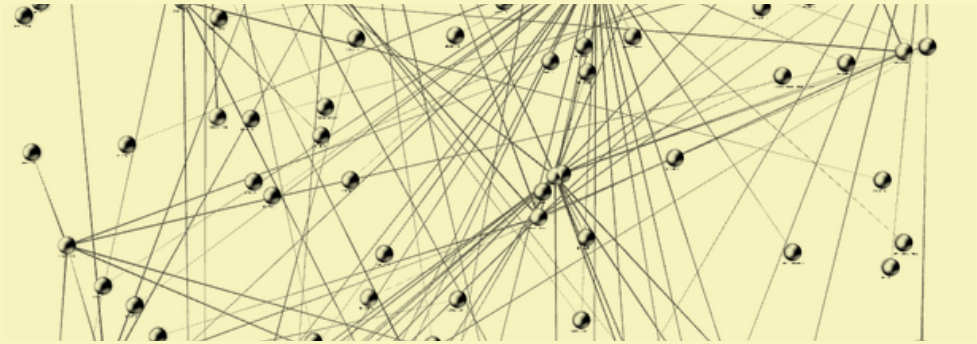
Les graphes sont moches

Le graphe (6/11)

Tout comme
cette robe

Le graphe (7/11)

Les graphes



N'ont **pas** de forme donnée

Ils sont cousus sur eux-mêmes : "**Non-planaires**"

La robe

Est informe

Et non-planaire : on ne peut pas la repasser

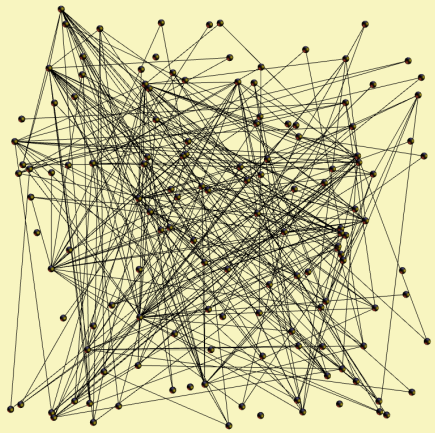
Le graphe (8/11)

On peut tout de même,
“artificiellement”
donner une forme à un graphe

Algorithmes de **spatialisation**
+
Travail **sémiologique**
(au sens de Bertin)

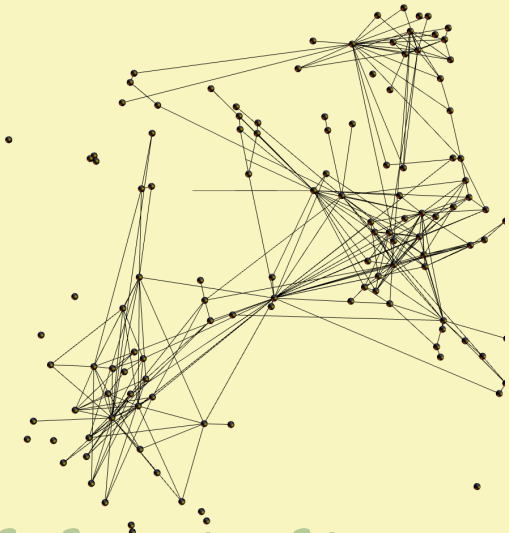
Le graphe (9/11)

1



2

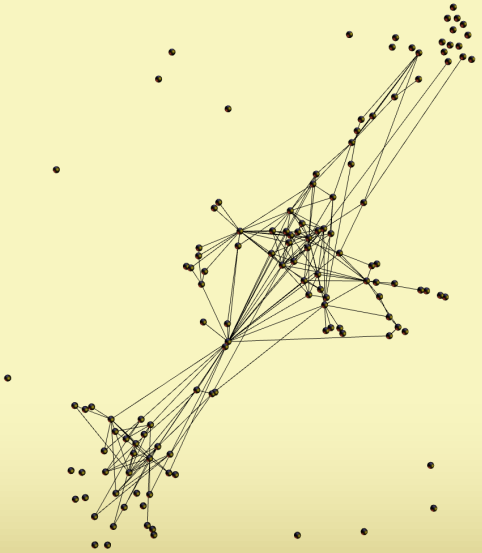
3



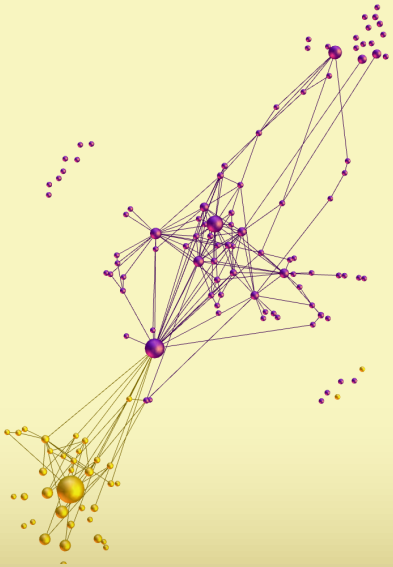
Toutes les “cartes” que nous verrons ont suivi ce traitement

4

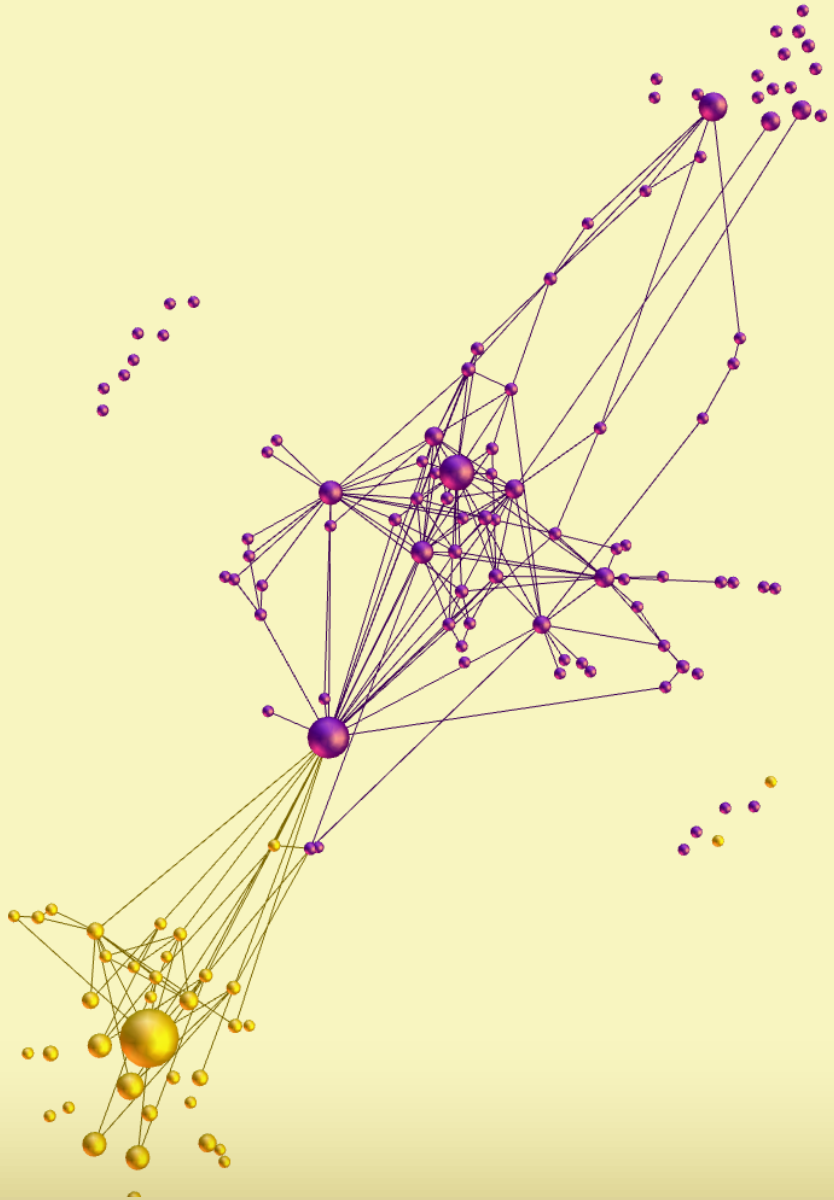
5



6



Le graphe (10/11)

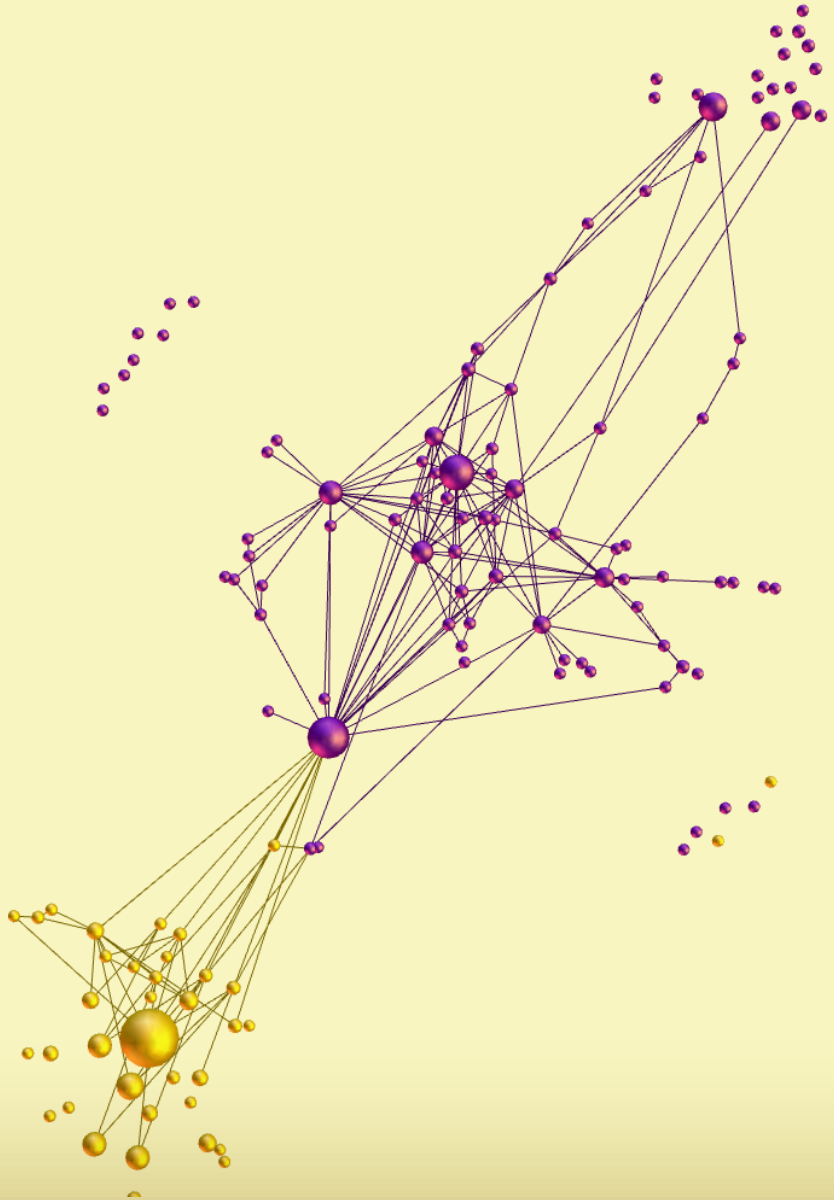


Graphe de sites web
En jaune : les blogs

On peut voir un effet de “blogosphère”

(la spatialisation ne dépend que des liens)

Le graphe (11/11)



Apparaissent les corrélations contenu-structure

Quelles sont les propriétés de ces graphes ?

Quelle sorte de complexité ?

2) Les propriétés des graphes “à invariance d’échelle”

Propriétés des graphes (1/25)

Nous nous confrontons à des graphes
d'un type particulier :
Ni aléatoires, ni strictement hiérarchiques

Nous allons d'abord voir leurs **propriétés**

Puis nous verrons **où on les trouve**

Propriétés des graphes (2/25)

6 degrés de séparation

(S. Milgram, D. Watts)

"Vous êtes à **6 poignées de mains**
de quiconque sur terre"

"Le diamètre du web est de **17 clics**"

C'est l'effet **petit-monde**
On parle de **réseaux "small-world"**

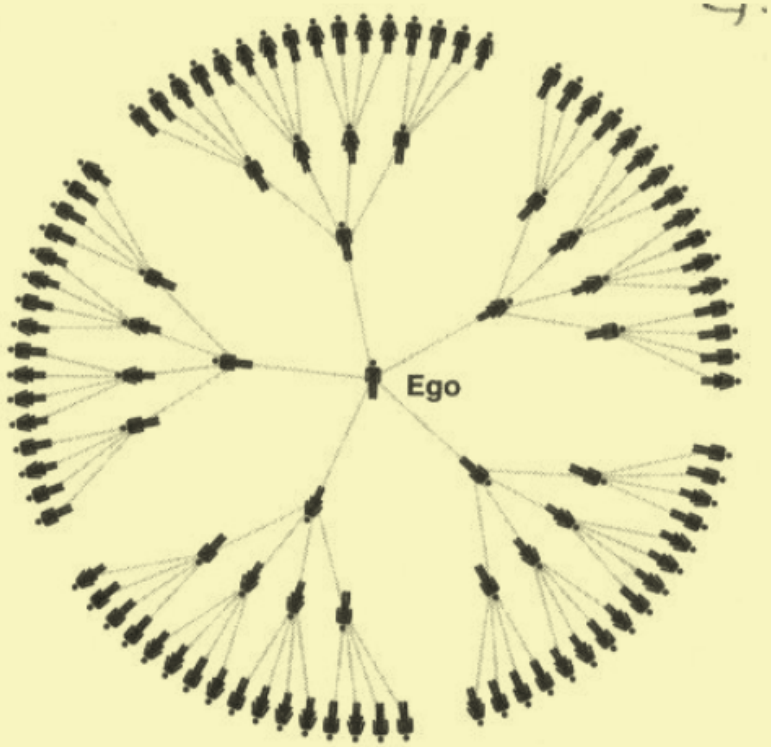
Propriétés des graphes (3/25)

Le diamètre d'un graphe

Distance entre deux noeuds :
Le plus court chemin

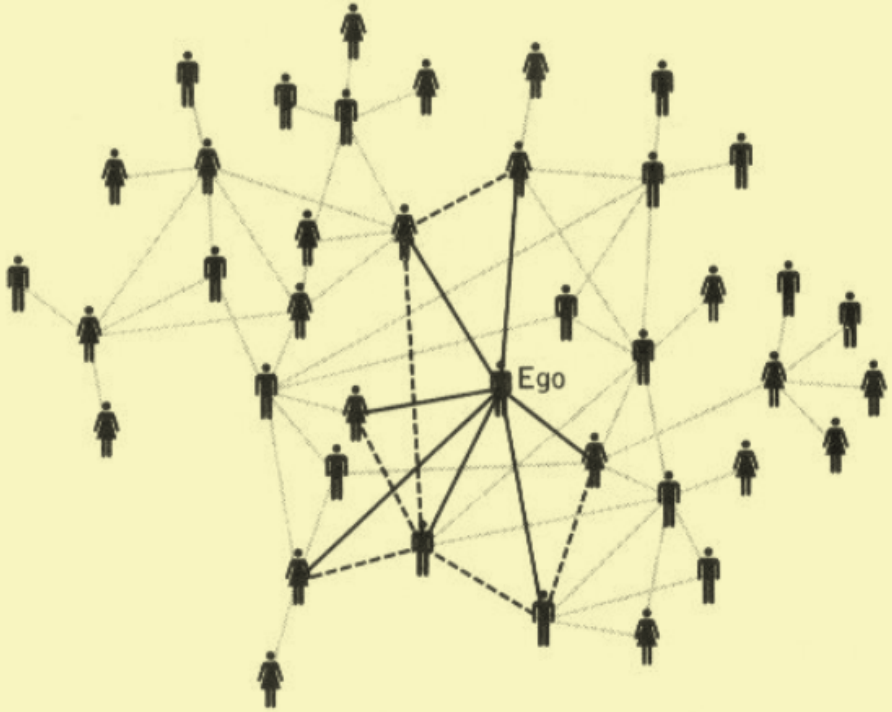
Diamètre d'un graphe:
La plus longue des distances entre deux noeuds

Propriétés des graphes (4/25)



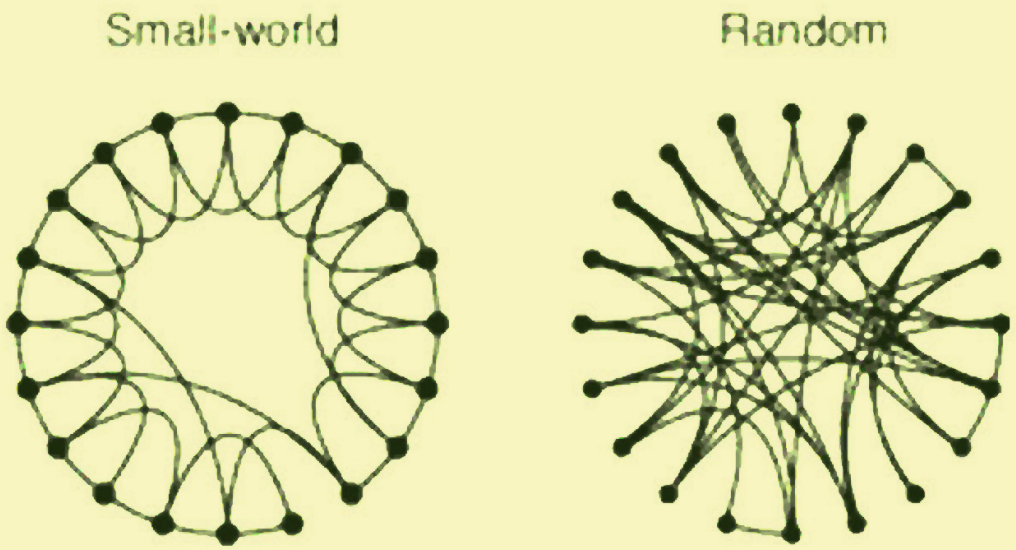
100 amis de 100 amis...
...à 6 degrés :

Plus que la population mondiale



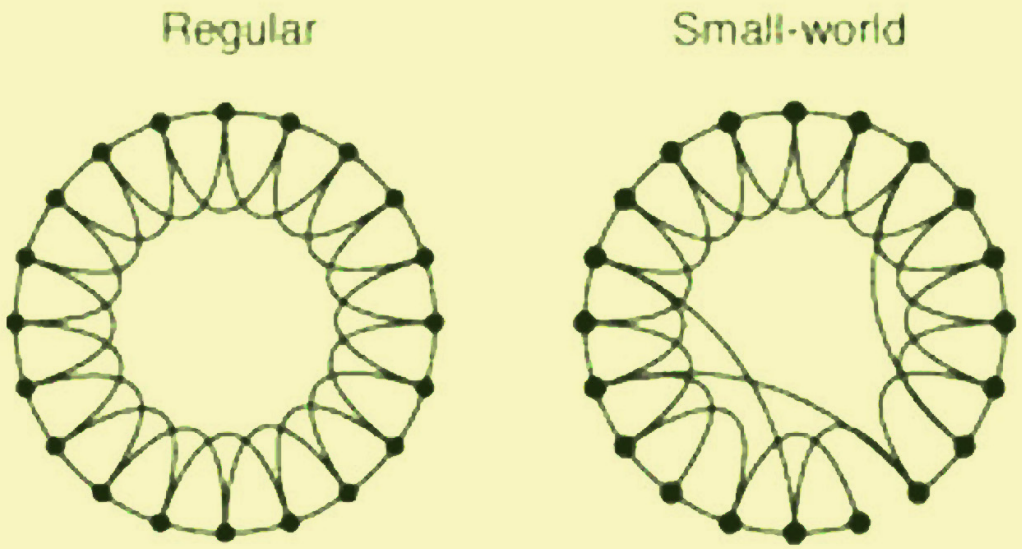
Mais sur les réseaux sociaux, il y a de la **redondance**
(les amis de mes amis...)

Propriétés des graphes (5/25)



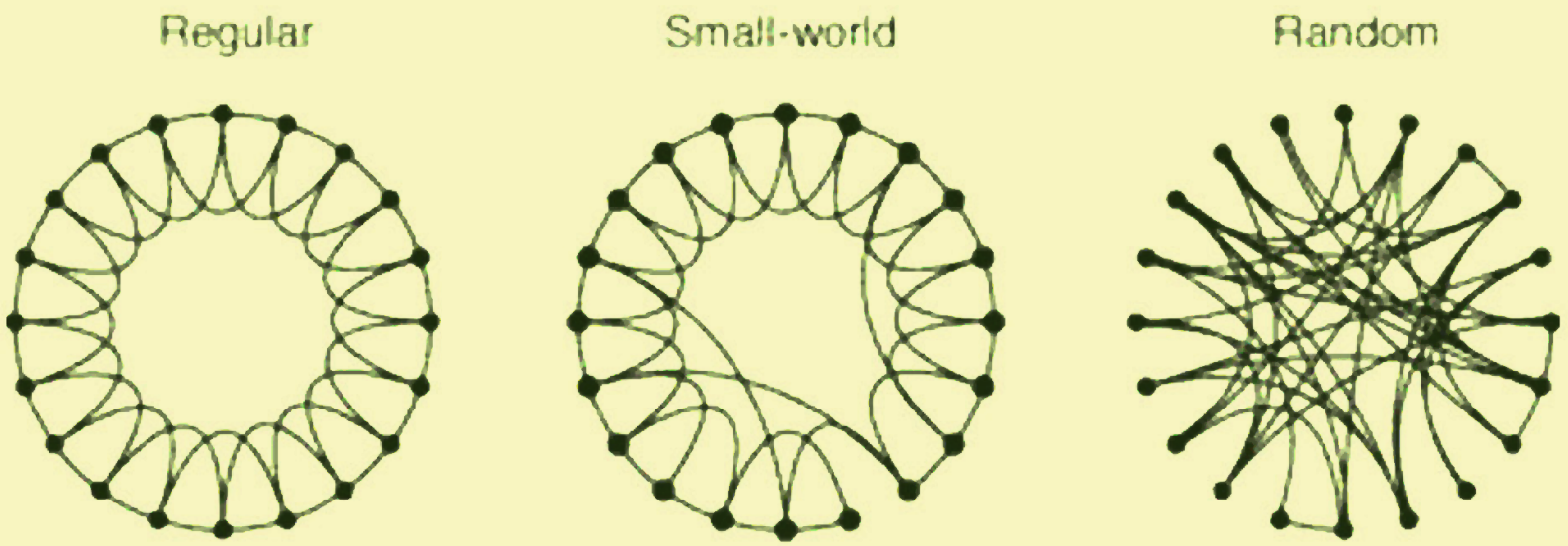
La redondance est caractéristique des graphes “small-world” (par rapport à un graphe aléatoire)

Propriétés des graphes (6/25)



La présence de raccourcis
est également caractéristique
(par rapport à un réseau régulier)
(réduction du diamètre)

Propriétés des graphes (7/25)



$p = 0$ → Increasing randomness → $p = 1$

Clustering Degree = High
Distance Degree = High

Clustering Degree = Low
Distance Degree = Low

Clustering Degree = High
Distance Degree = Low

Propriétés des graphes (8/25)

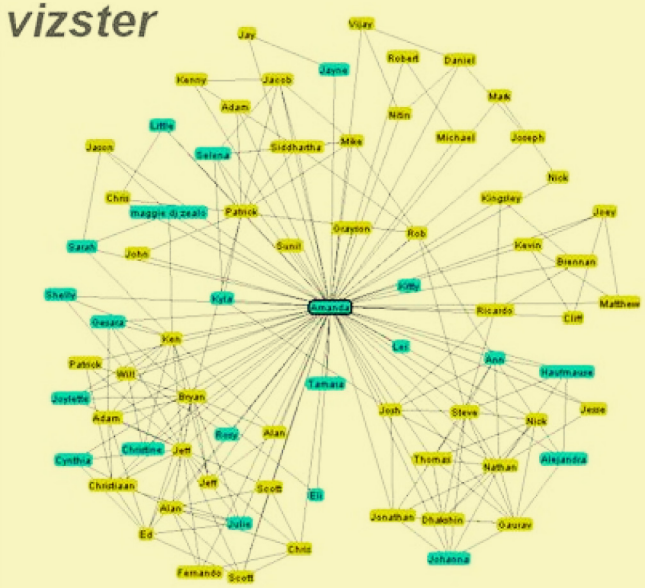
Il a fallu beaucoup de temps
avant que l'on envisage des réseaux :

1) d'un **diamètre court**

2) et **redondants**. C'est-à-dire :
Fortement **clusterisés**

(Duncan Watts)

Propriétés des graphes (9/25)



Graphe egocentré

“Les amis de mes amis sont mes amis”

On voit apparaître des **clusters** :
Collègues professionnels, famille…

Propriétés des graphes (10/25)

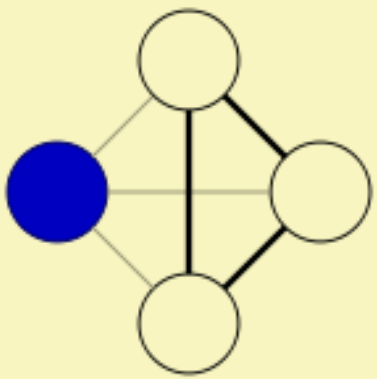
Coefficient de clustering

... d'un graphe :
Nombre de liens existants
sur nombre de liens possibles

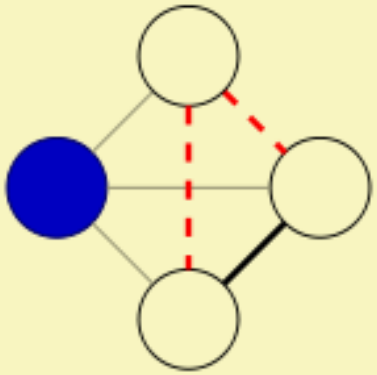
Coeff. de clustering d'un noeud :
Coefficient de clustering
du graphe de ses voisins

Propriétés des graphes (11/25)

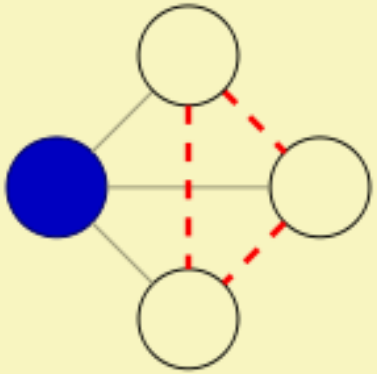
Coefficient de clustering



c = 1



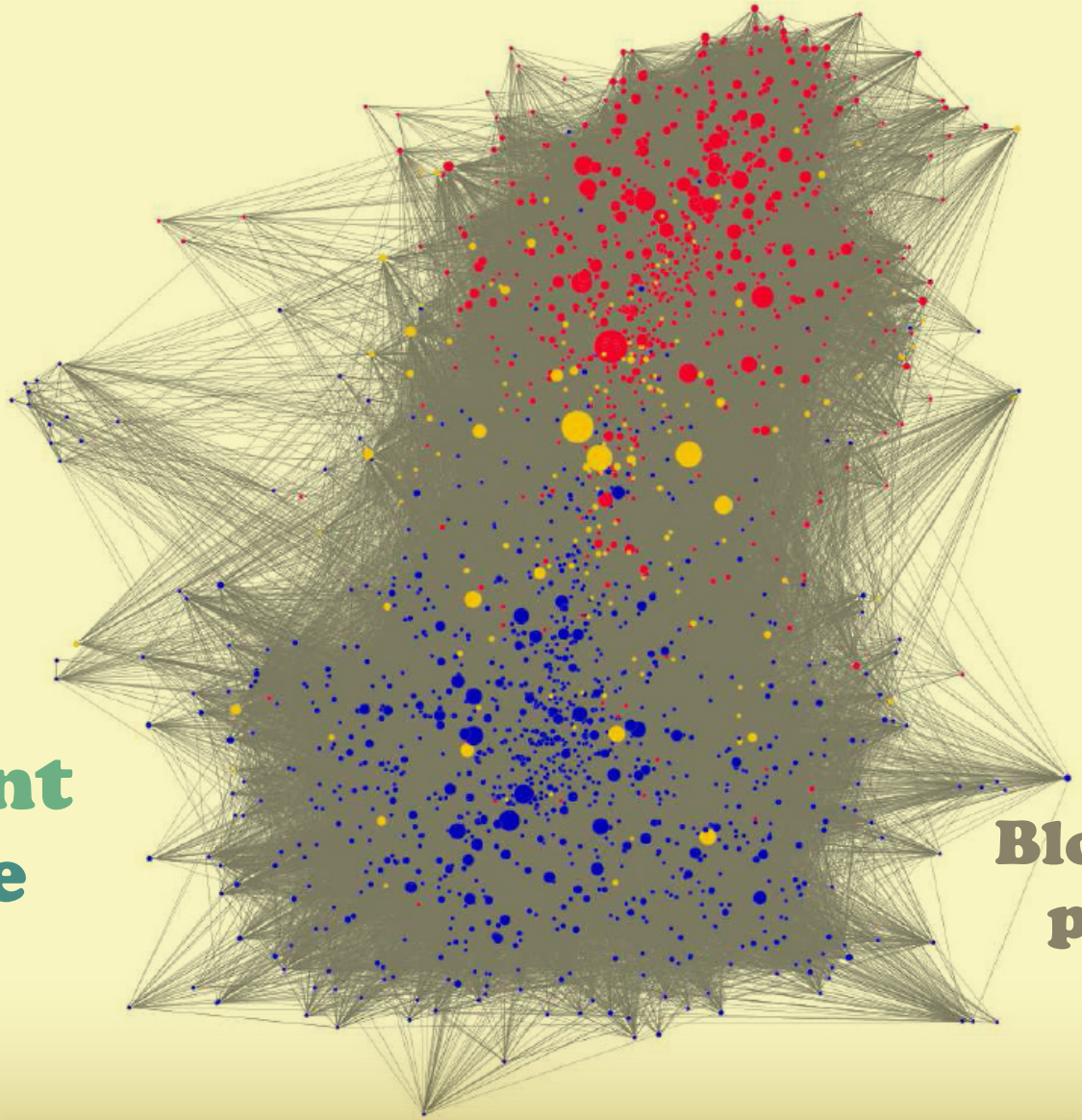
c = 1/3



c = 0

Propriétés des graphes (12/25)

On veut
clusteriser
les graphes
pour les
simplifier
et lire
plus facilement
leur **structure**



Blogosphère
politique
US
(RTGI)

Propriétés des graphes (13/25)

Cluster

Noeuds qui ont

plus de liens entre eux

qu'avec les autres noeuds

Classe (de connectivité)

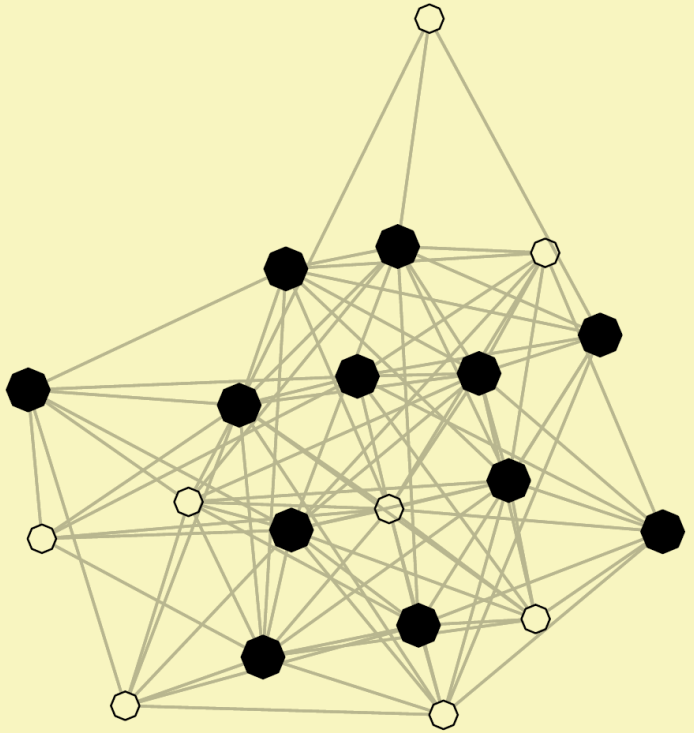
Noeuds qui partagent

une même façon de se lier

(c'est une définition plus large)

Propriétés des graphes (14/25)

Deux classes :
Plus de liens internes
Moins de liens externes
-> Deux clusters



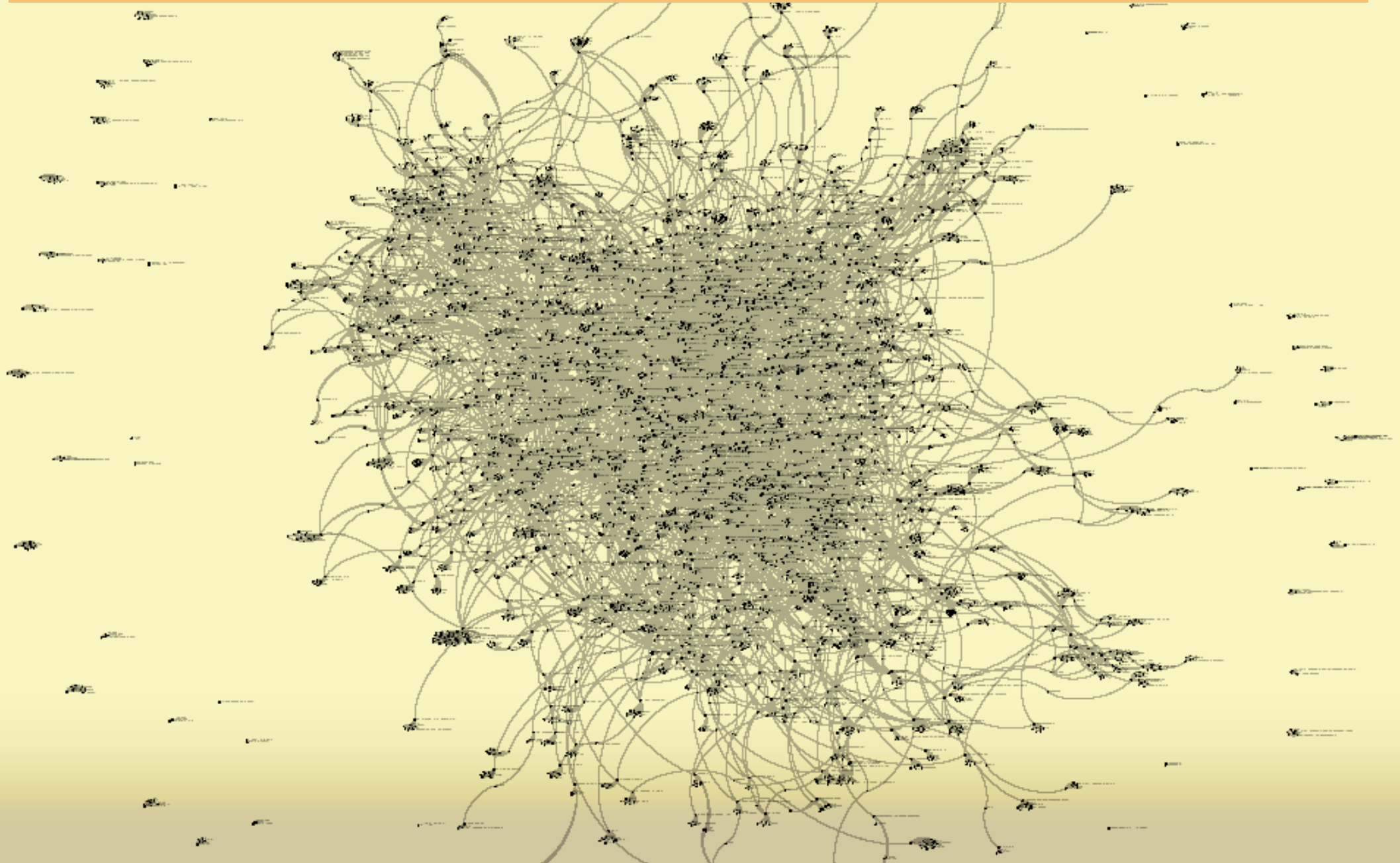
Deux classes :
Moins de liens internes
Plus de liens externes
-> Un seul cluster

Propriétés des graphes (15/25)

On peut avoir des **noeuds**
avec un **bon coefficient de clustering**

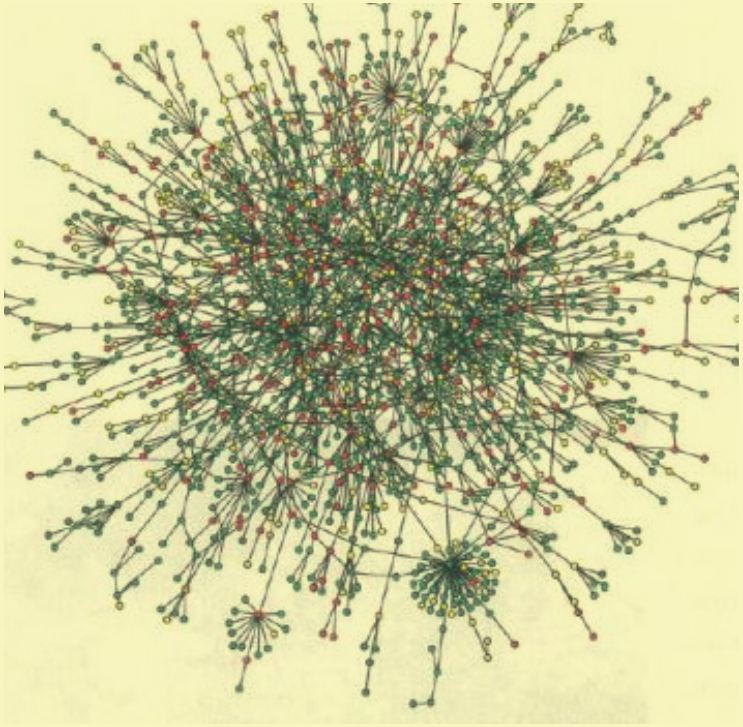
...Mais **pas de clusters**
identifiables...

Propriétés des graphes (16/25)

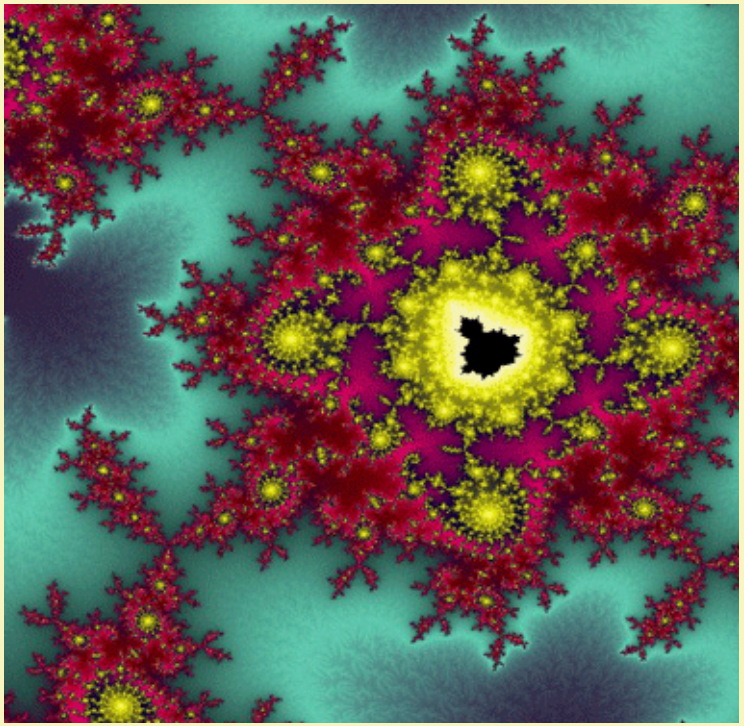


Propriétés des graphes (17/25)

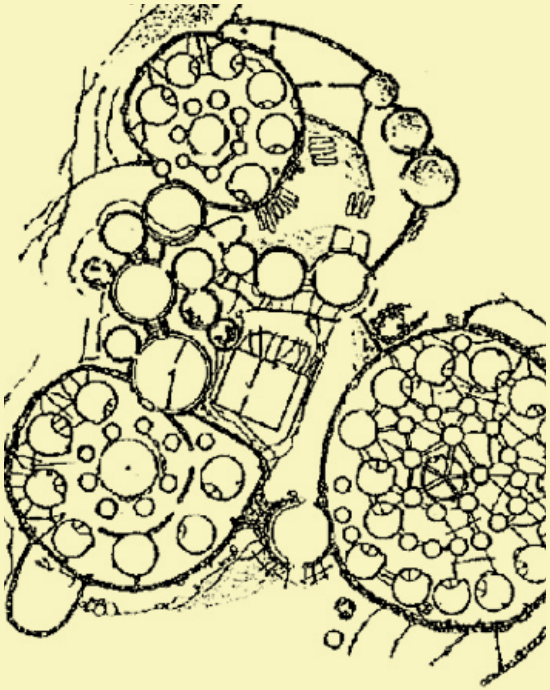
Portion du web



Fractale



Village africain



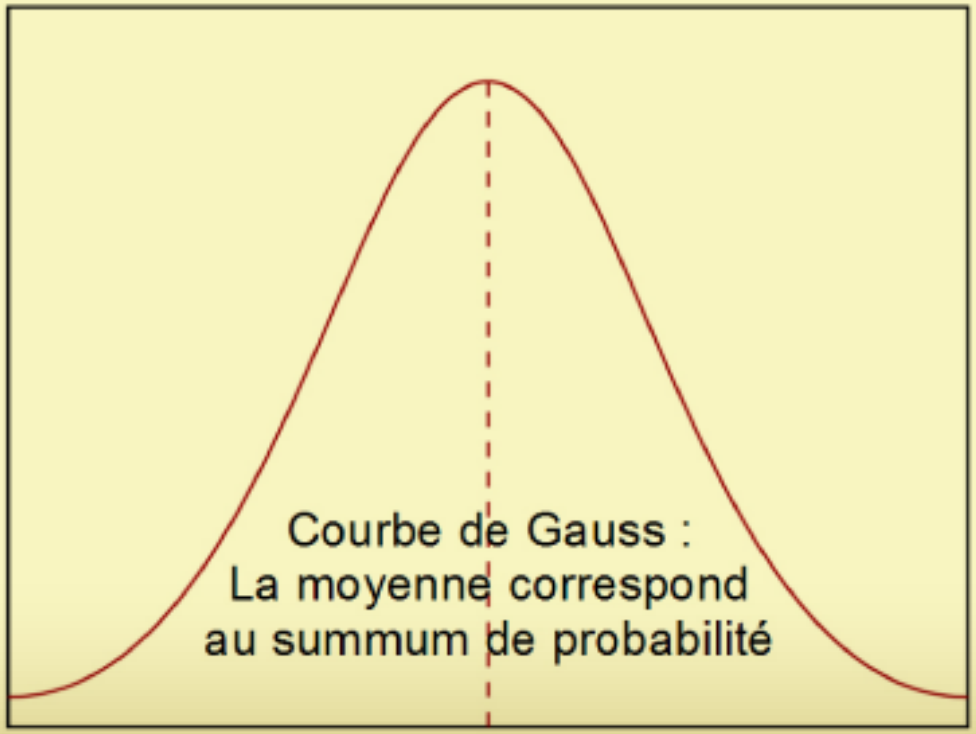
Car les clusters sont pris dans de plus gros clusters, qui sont pris dans des macro-clusters...

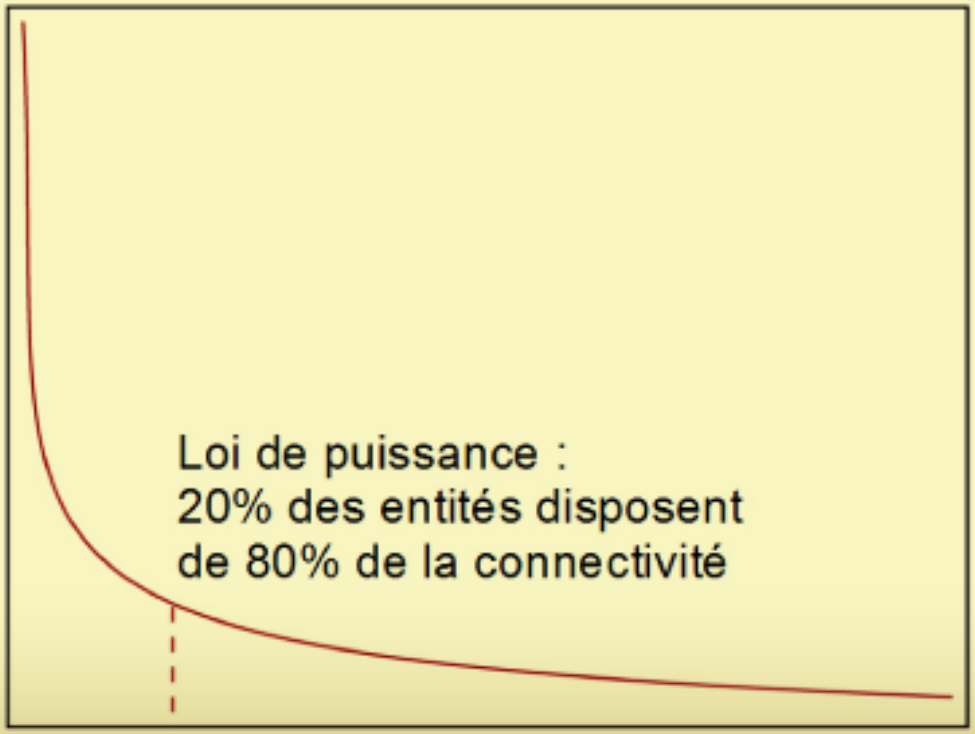
Propriétés des graphes (18/25)

Power Law

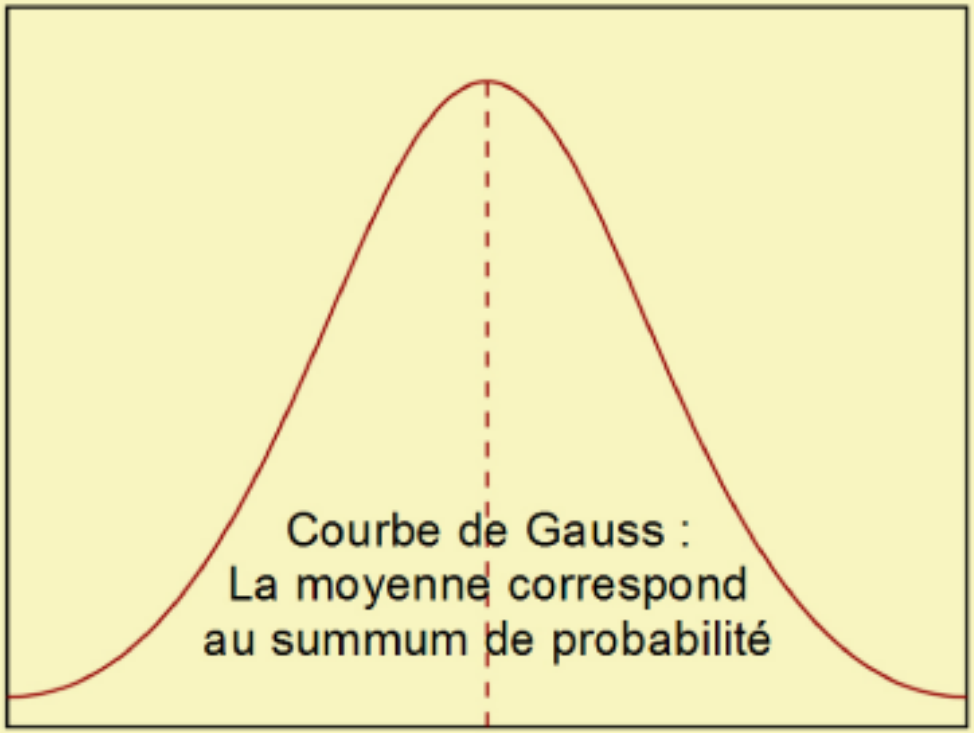
La distribution du nb. de liens par noeud suit une **loi de puissance :**

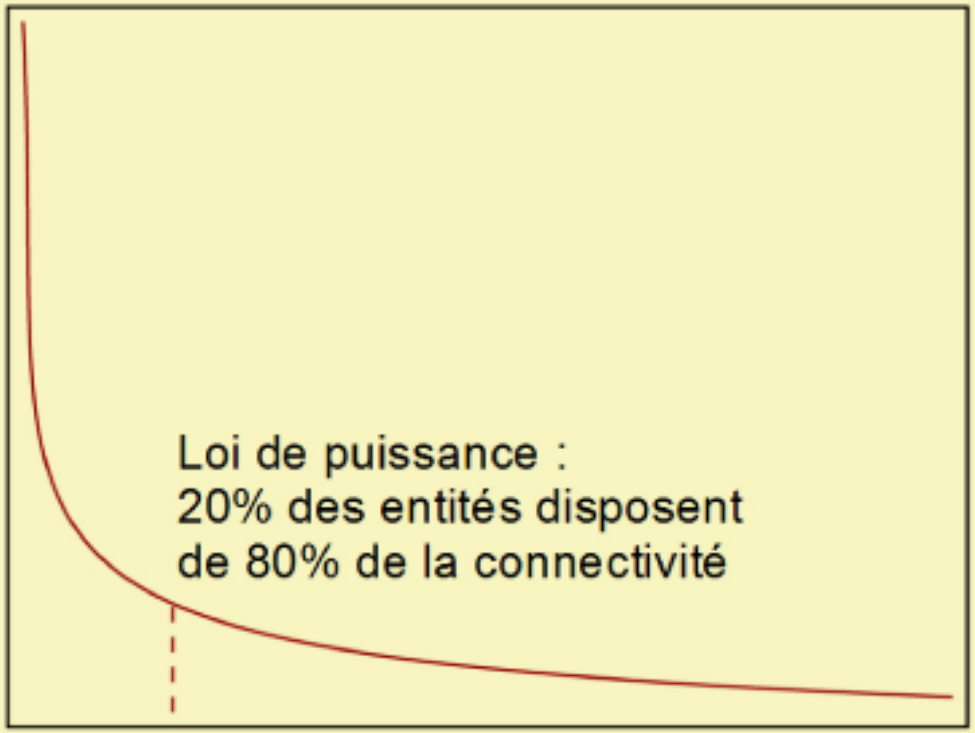
20% des noeuds ont 80% des liens



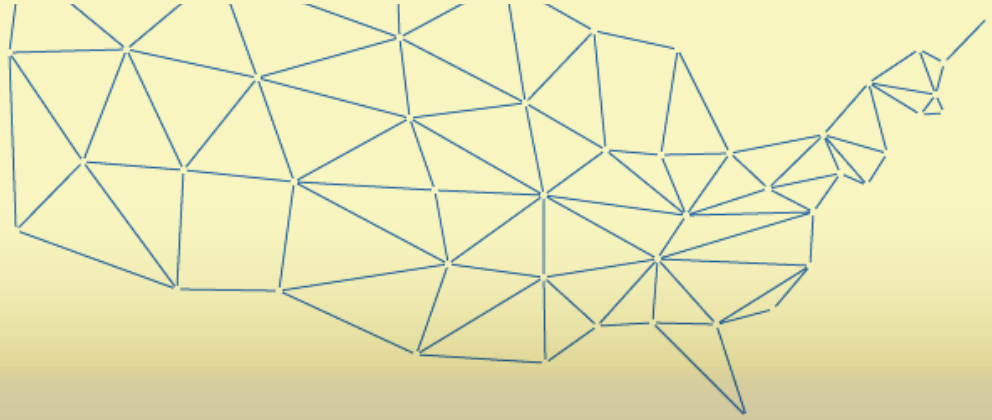


Propriétés des graphes (19/25)

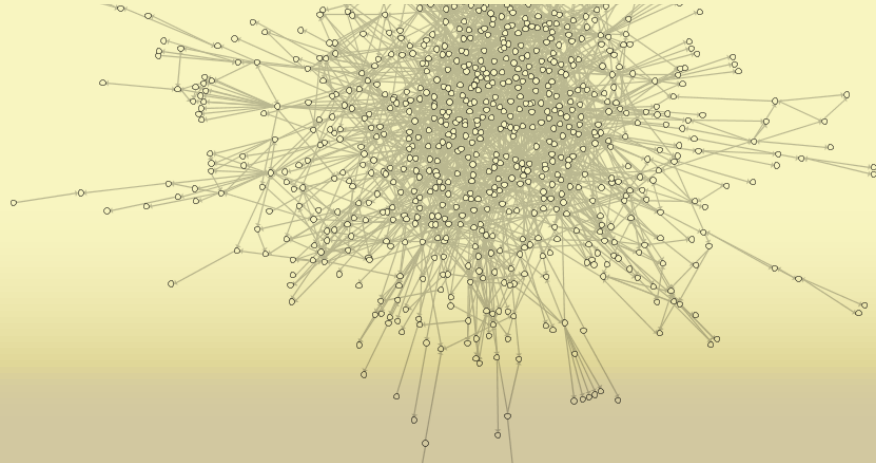




Echelle caractéristique



Pas d'échelle caractéristique



Propriétés des graphes (20/25)

Invariance d'échelle

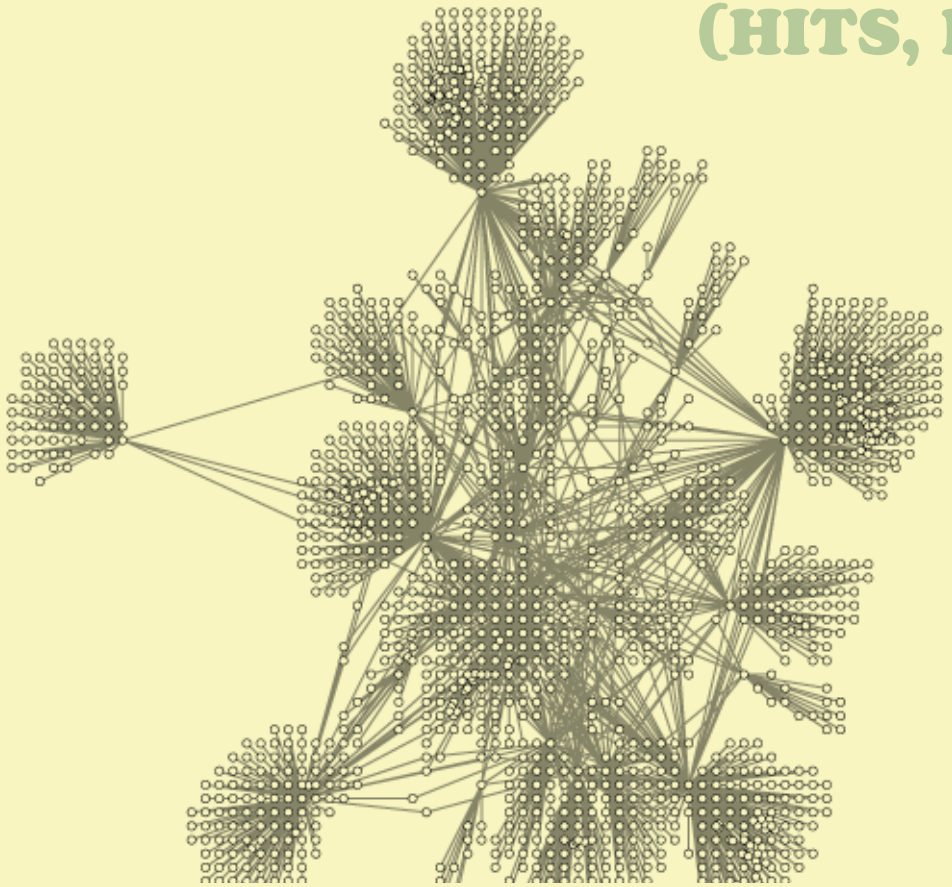
On appelle aussi ces réseaux
"à invariance d'échelle"
ou "**scale-free networks**"

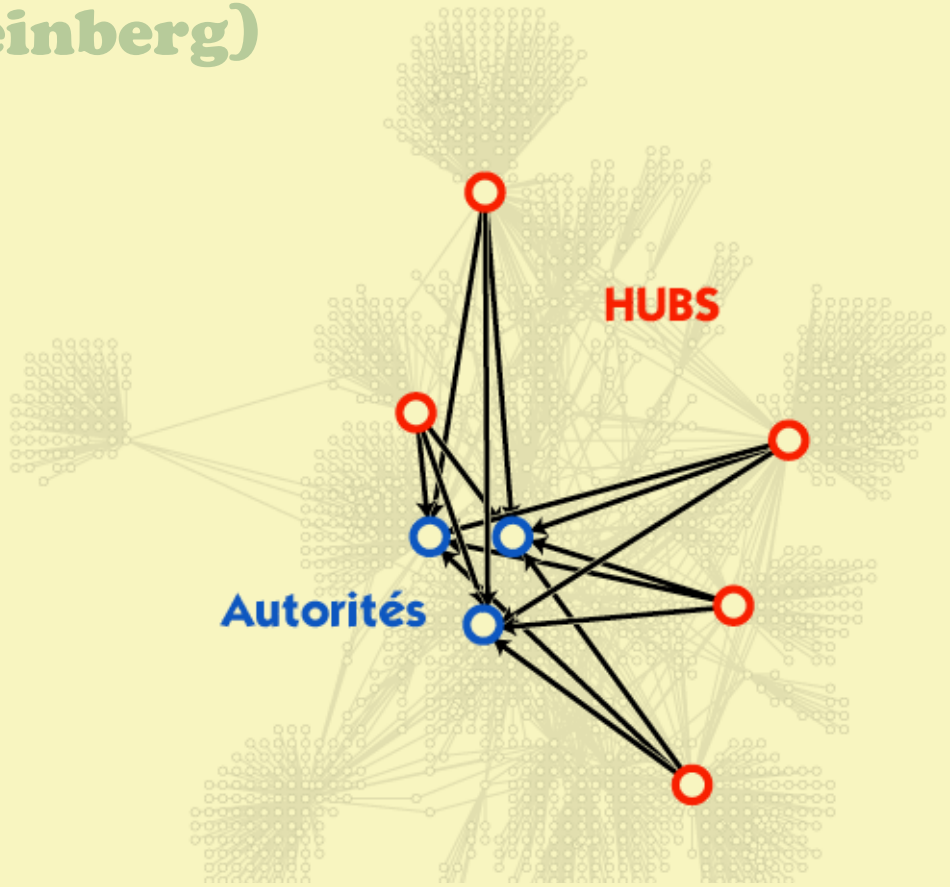
(application directe de la loi de puissance)

Propriétés des graphes (21/25)

Calcul de Hubs et Autorités

(HITS, Kleinberg)





Les hubs ont beaucoup de liens sortants

Les autorités, de liens entrants

Propriétés des graphes (22/25)

L'attachement préférentiel

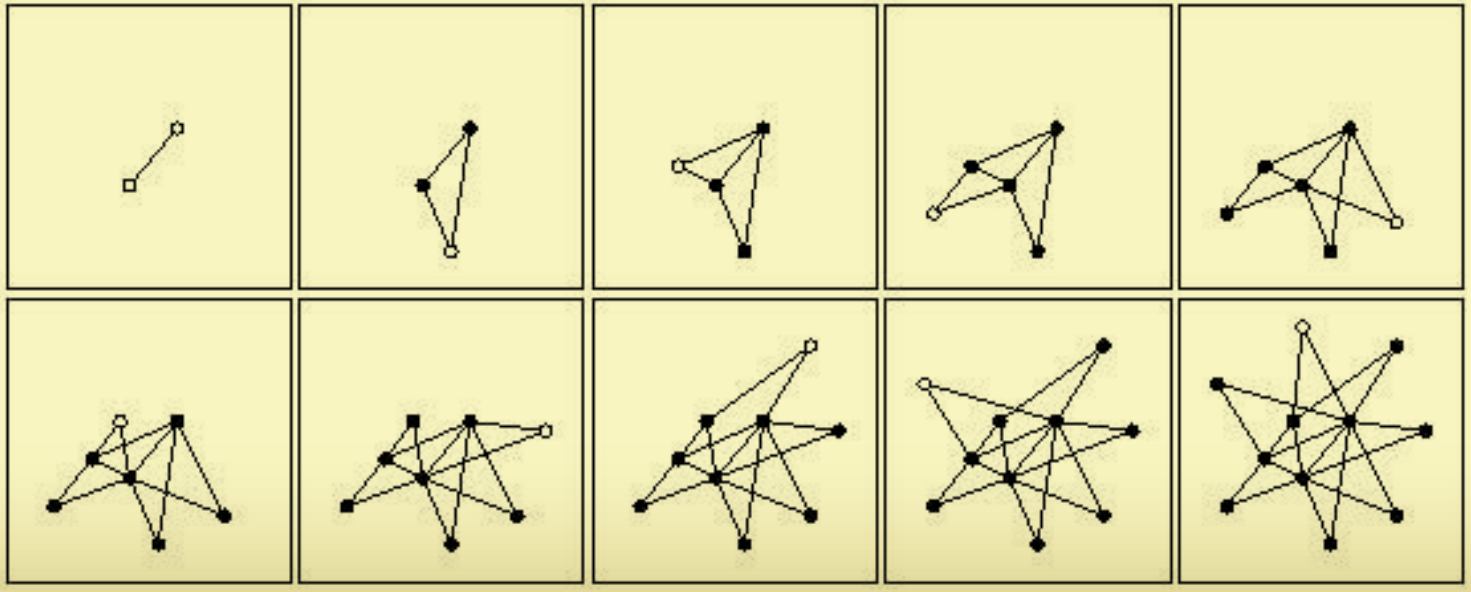
(Barabasi & Albert)

Les "petits" se lient d'abord aux "gros"

"The winner takes all"

"Rich get richer"

"Effet Monopoly"...



Propriétés des graphes (23/25)

Vulnérabilité

Le réseaux reste connexe
si 80% des noeuds tombent...
...s'ils sont pris au hasard

Mais en ciblant
les plus connectés,
détruire 5% des noeuds
suffit à défaire le réseau

**Les "gagnants"
sont aussi les "points faibles"**

Propriétés des graphes (24/25)

Lorsqu'on parle de "complex networks"

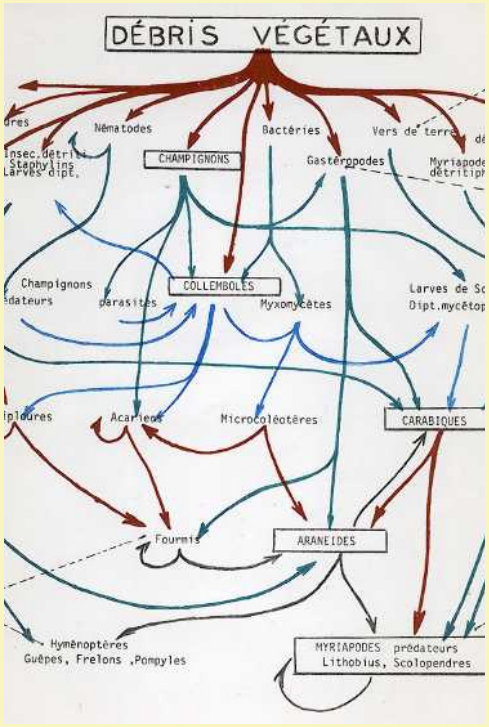
On parle de réseaux petit-monde
(tradition de Watts)

Ou encore de réseaux à invariance d'échelle
(tradition de Barabasi)

Ce sont les mêmes réseaux !

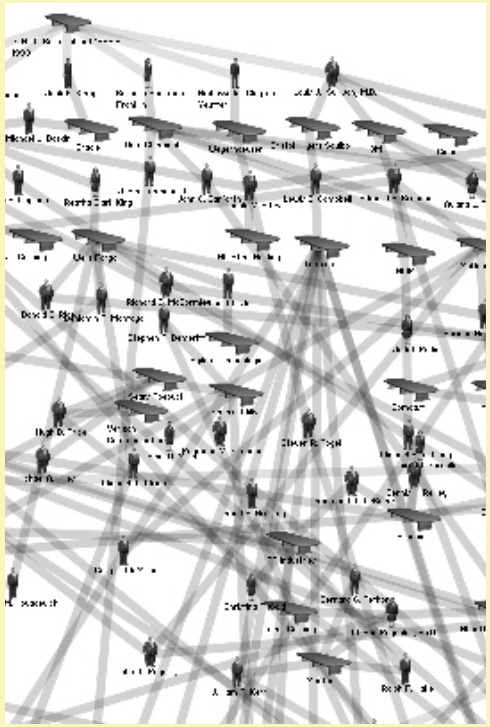
Propriétés des graphes (25/25)

Ecosystèmes



Réseaux sociaux

Economie



Web

On retrouve ces réseaux dans de nombreux domaines

Le web

Le web : définition

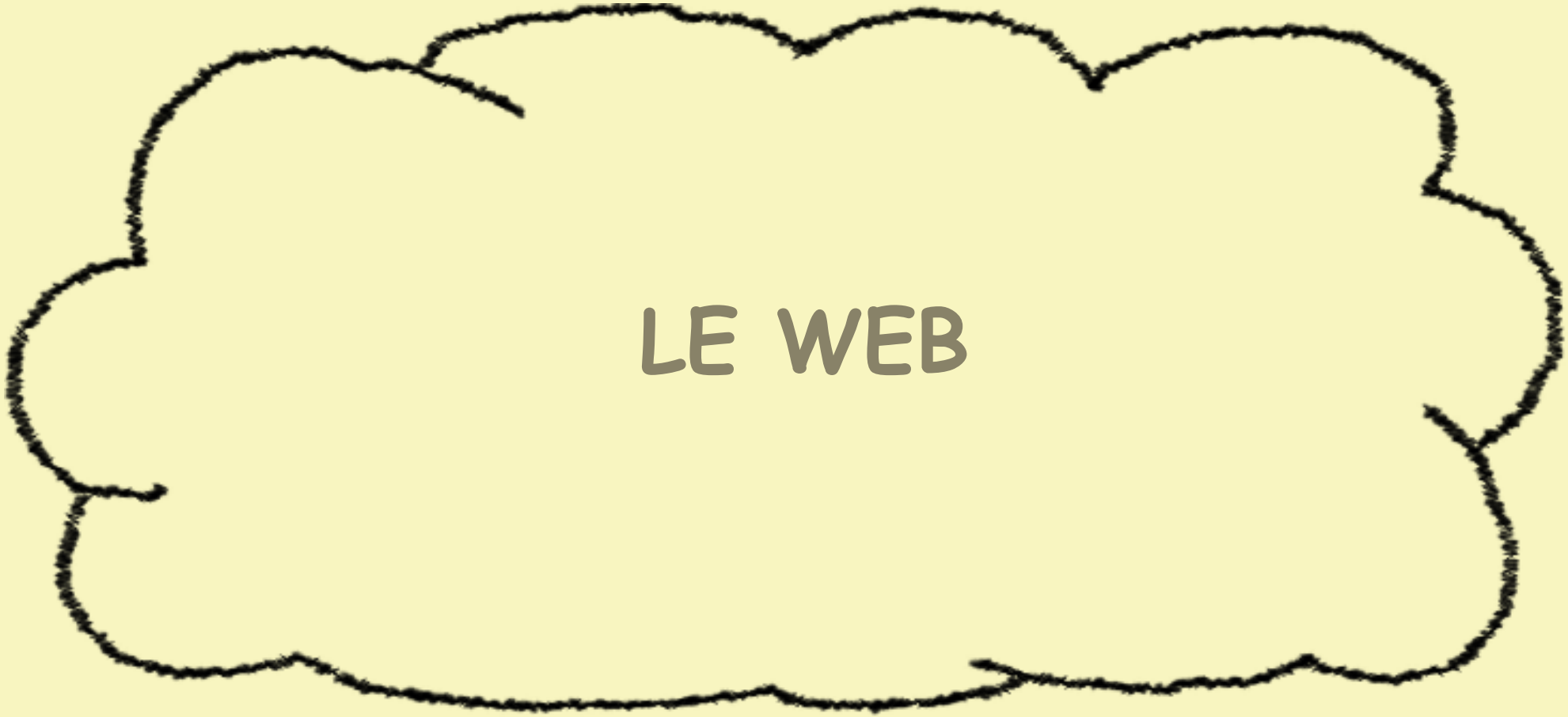
Internet est un réseau à invariance d'échelle
= les "tuyaux" : câbles, routeurs...

Nous nous intéressons ici au web :
Les pages reliées par des liens hypertextes
= tout ce à quoi on accède par un navigateur
(donc pas le mail, le P2P, la VoIP...)

Le web est aussi
un réseau à invariance d'échelle

(se) Représenter le web (1/4)

On a longtemps représenté le web ainsi :

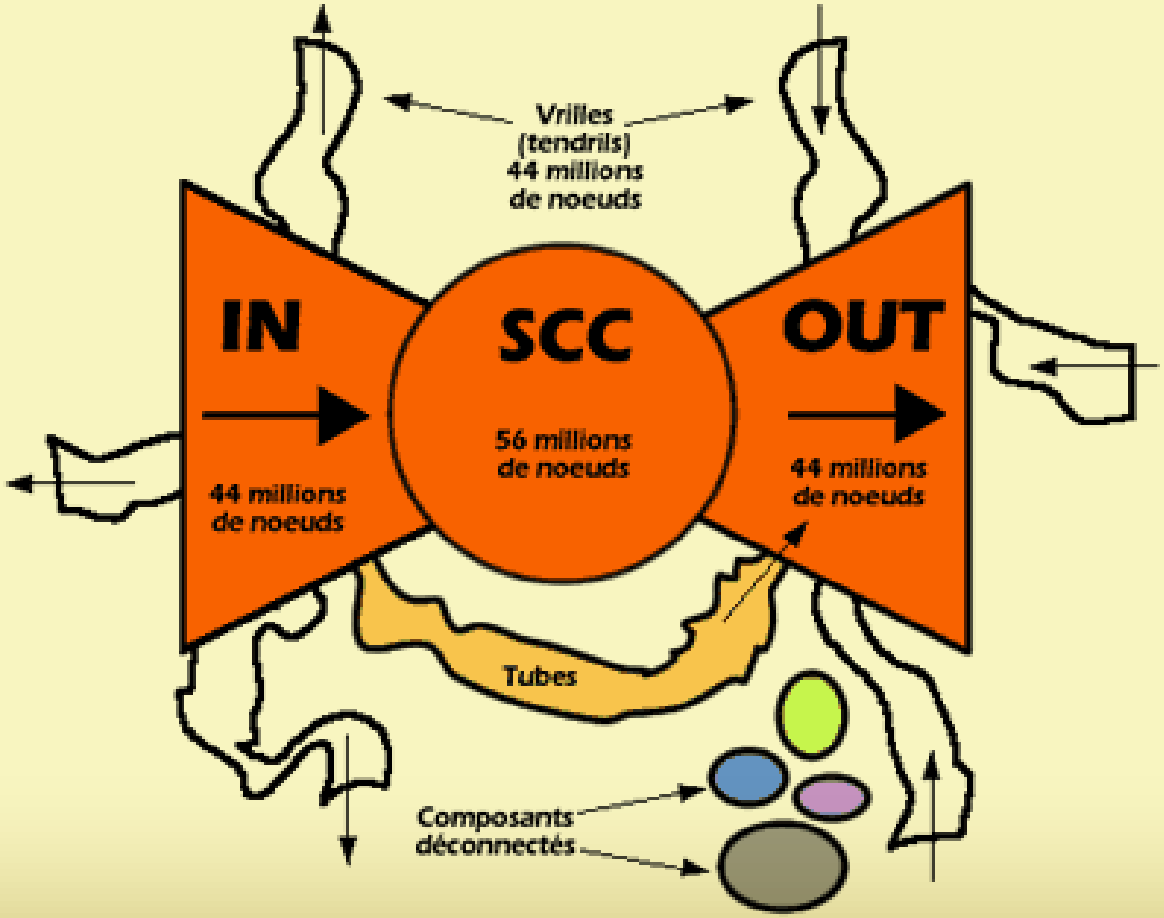


Parce qu'on n'y comprenait rien...

(et que les chercheurs ne savent pas dessiner)

(se) Représenter le web (2/4)

La première représentation scientifique est "la théorie du noeud papillon"



(se) Représenter le web (3/4)

Aujourd'hui on voit plutôt le web comme une immensité de contenus

divers

et

variés

reliés par une structure abominablement complexe

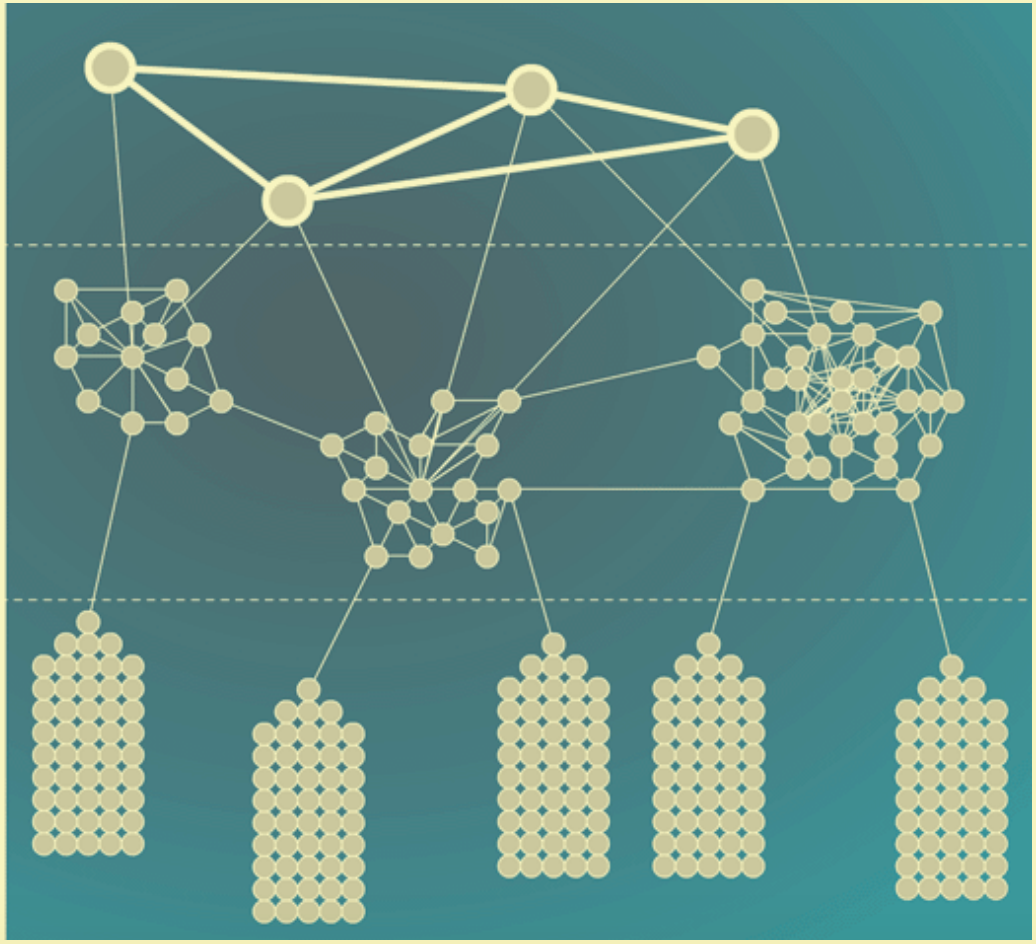
(se) Représenter le web (4/4)

Mais on peut pourtant en construire un modèle

Couche **la plus visible** du web : Google, Amazon, Voila, SNCF, etc...

Couche intermédiaire : **agrégats**, communautés en ligne

Couche profonde : **bases de données**



Le modèle du web en couches (Ghitalla)

Le web : quelques données

Taille : inconnue

Temps moyen passé sur une page :
Quelques secondes

Portion indexée :
(par TOUS les moteurs)
Quelques pourcents

Diamètre : ~17 clics

Le web est un monstre mal connu

Explorer le web en douceur

Mais il n'est pas si méchant !

Pour l'analyser il faut procéder pas à pas :

1) Un protocole **humble**
2) De bons **points d'entrée**
3) Avancer avec **méthode**
4) Etre **attentif aux données**

Trois niveaux d'exploration

Quels outils pour explorer le web ?

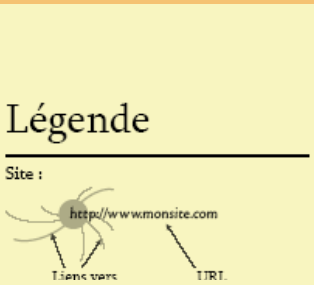
Débutant : à la main

Confirmé : Navicrawler

Expert : crawler professionnel

/!\ En dehors du graphe,
on peut tout faire à la main !

Protocole 1 : Expansion (1/2)



1) Choisir un petit domaine

Ex. : “Les blogs de migrants marocains”

2) Trouver quelques sites pertinents

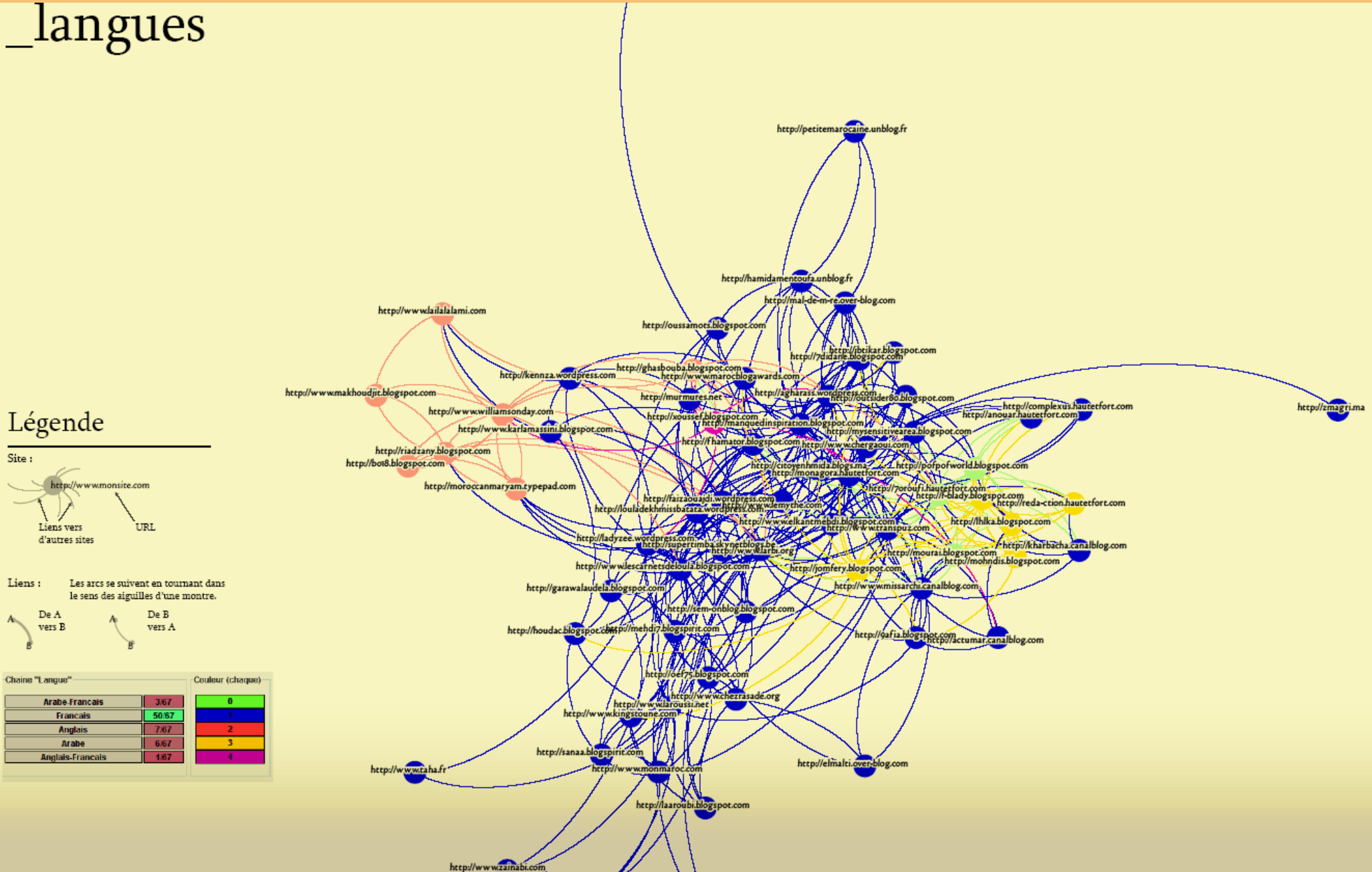
Demander à des “connaisseurs”

3) Explorer en suivant les liens

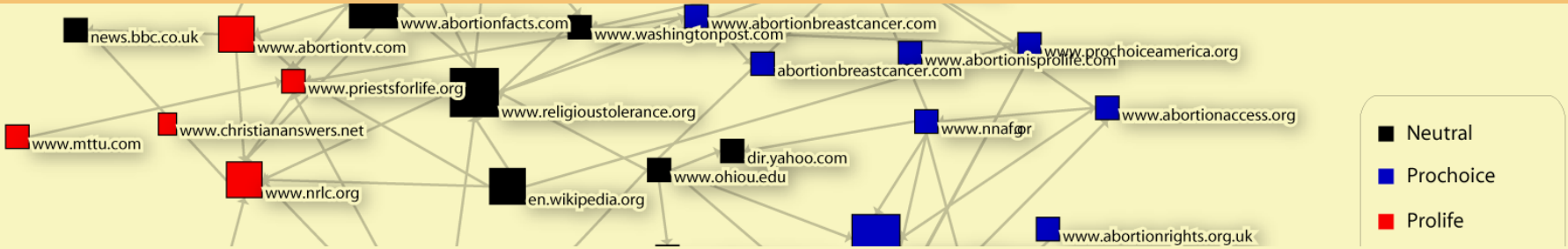
4) Eliminer au fur et à mesure les sites non-pertinents

Protocole 1 : Expansion (2/2)

_langues



Protocole 2 : Anaylse de requête (1/2)



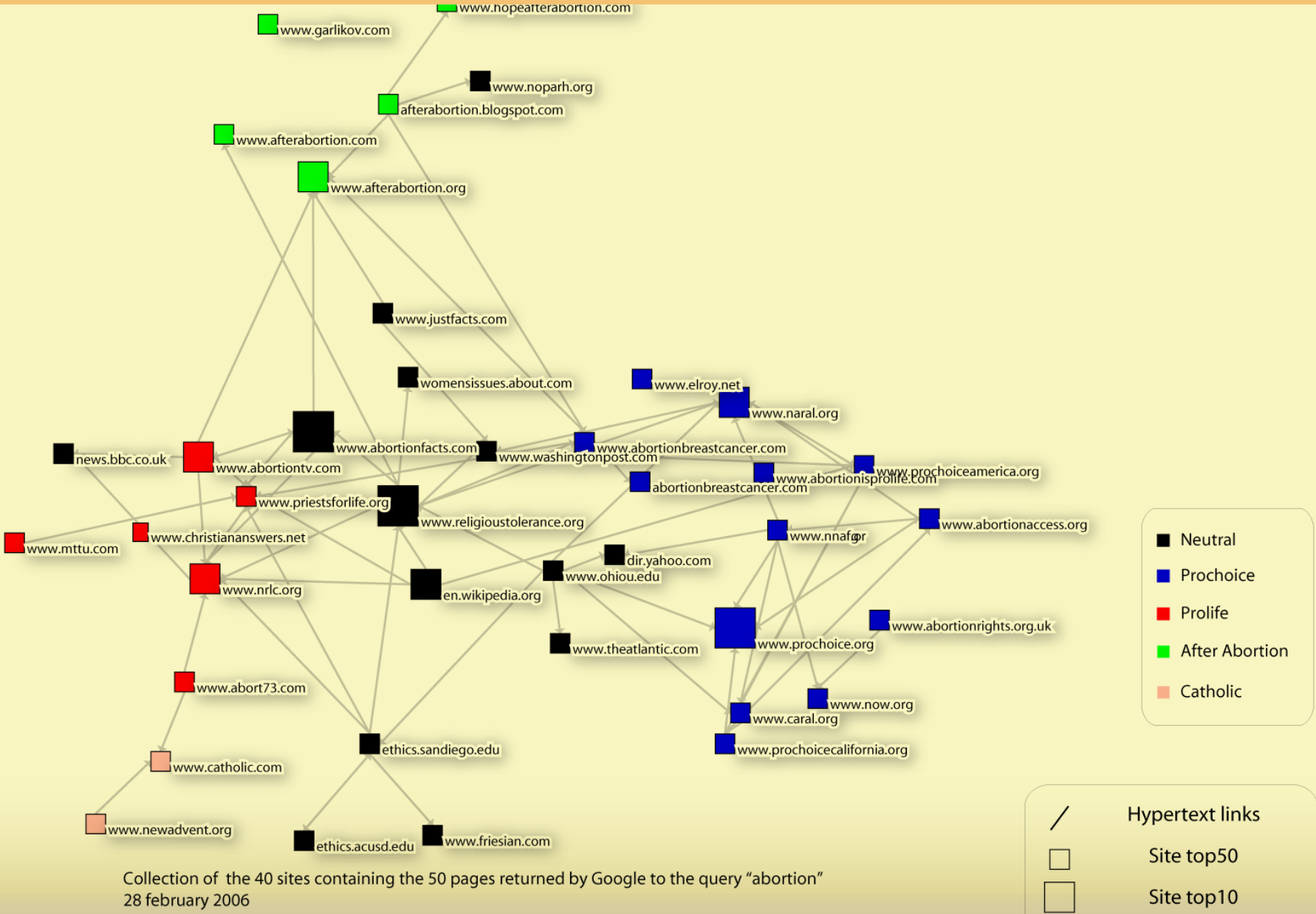
1) **Choisir une bonne requête dans un moteur de recherche**

Ex. : “abortion” dans Google anglophone

2) **Classer les résultats dans des catégories**

3) **Ne pas hésiter à revoir ses catégories en cours de route**

Protocole 2 : Anaylse de requête (2/2)



Collection of the 40 sites containing the 50 pages returned by Google to the query "abortion"
28 february 2006

Protocole 3 : Anaylse sito-centrée (1/2)

1) Choisir un site particulièrement **important**
Ex. : un portail actif

2) Explorer et trier systématiquement
tous ses voisins, et seulement ceux-ci

Protocole 3 : Anaylse sito-centrée (2/2)

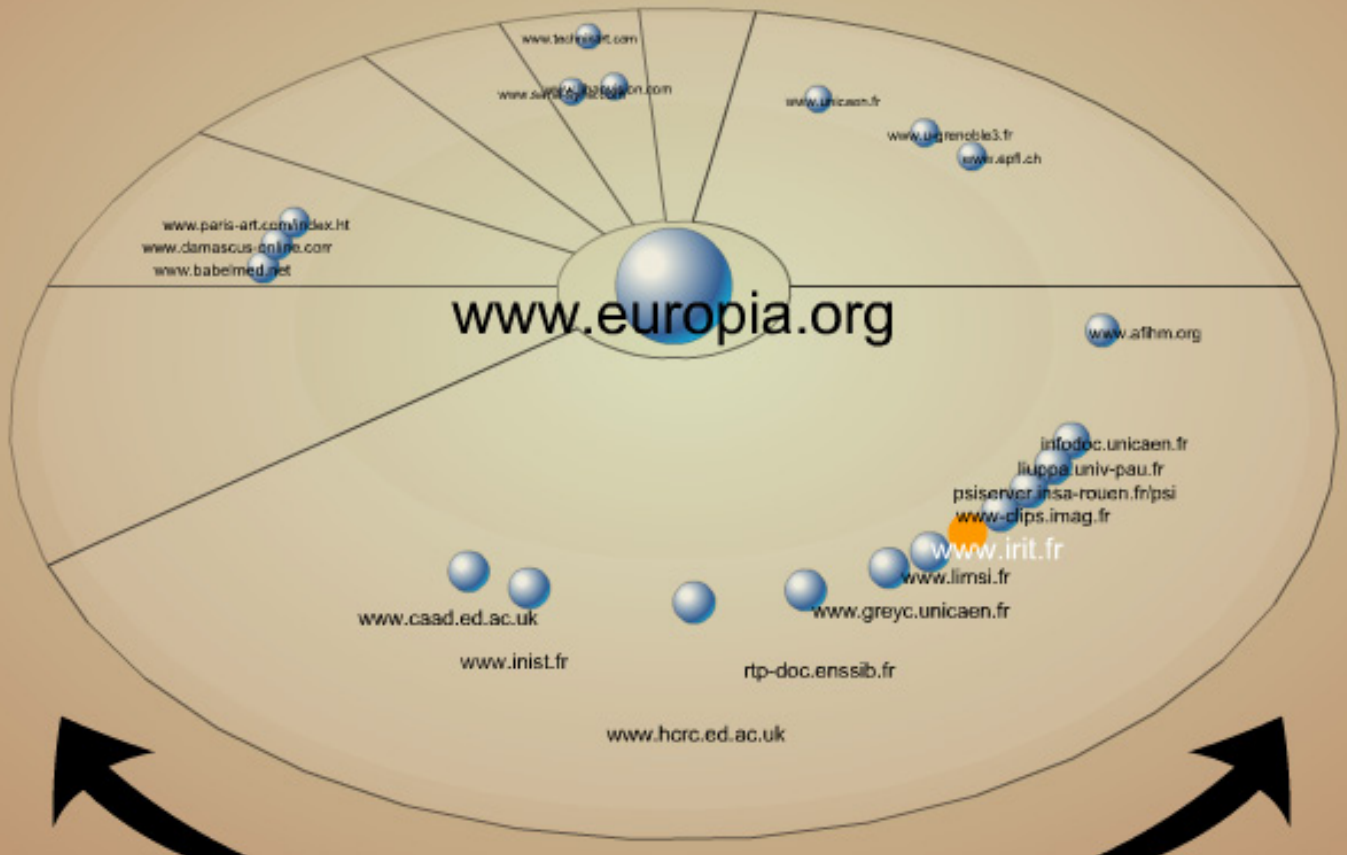
Voisinage EuroplA

aout 2005

RTGI

Recherche

Informatique, IA



IRIT, Institut de Recherche en Informatique de Toulouse

http://www.irit.fr

France

themes de recherche, partenariats, formations, evenements, valorisation

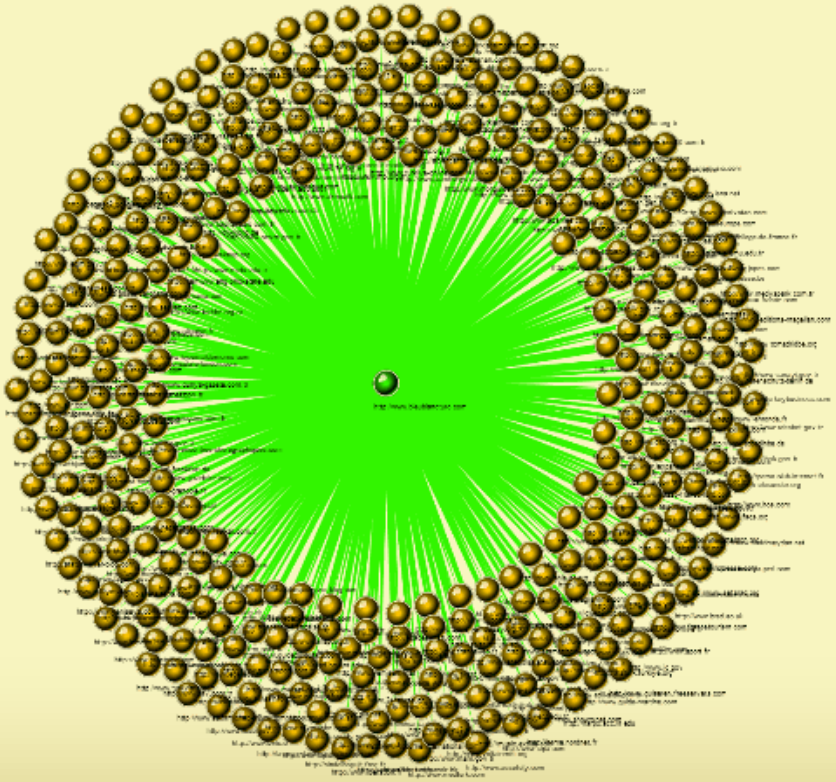
Les pièges du réseau (1/3)

Le principe d'une exploration pertinente,
c'est de **tirer profit des propriétés du réseau**

...tout en résistant aux pièges de celui-ci

Les pièges du réseau (2/3)

Danger 1 :
L'éparpillement dû aux hubs

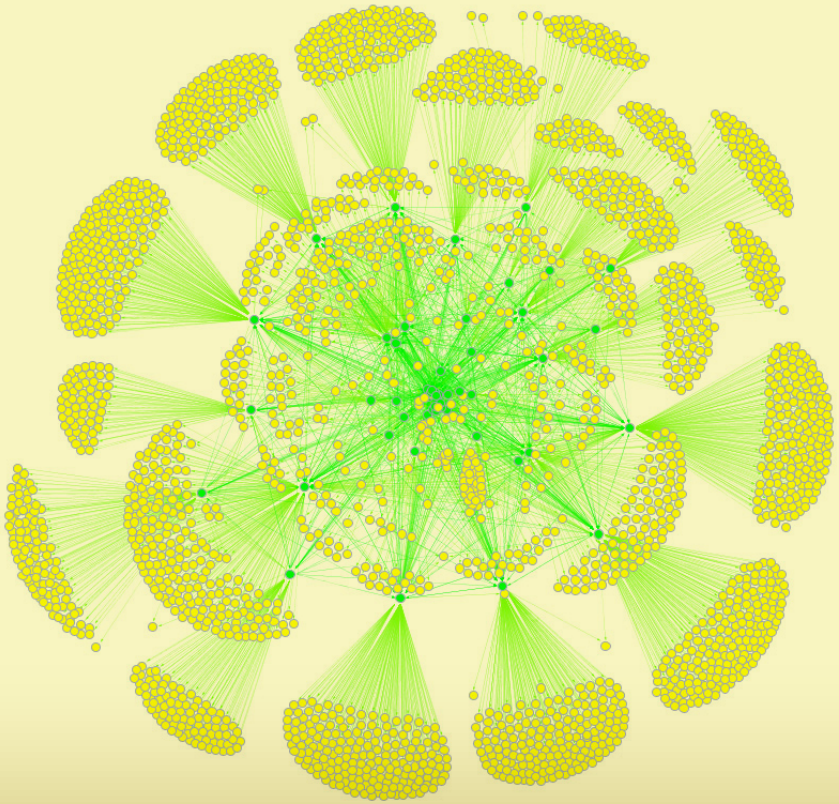


Voici un petit Hub :
Il a 200 voisins

S'ils ne sont pas majoritairement pertinents, il faut s'éviter de tous les trier

Les pièges du réseau (3/3)

Danger 2 :
Le crawl incontrôlé



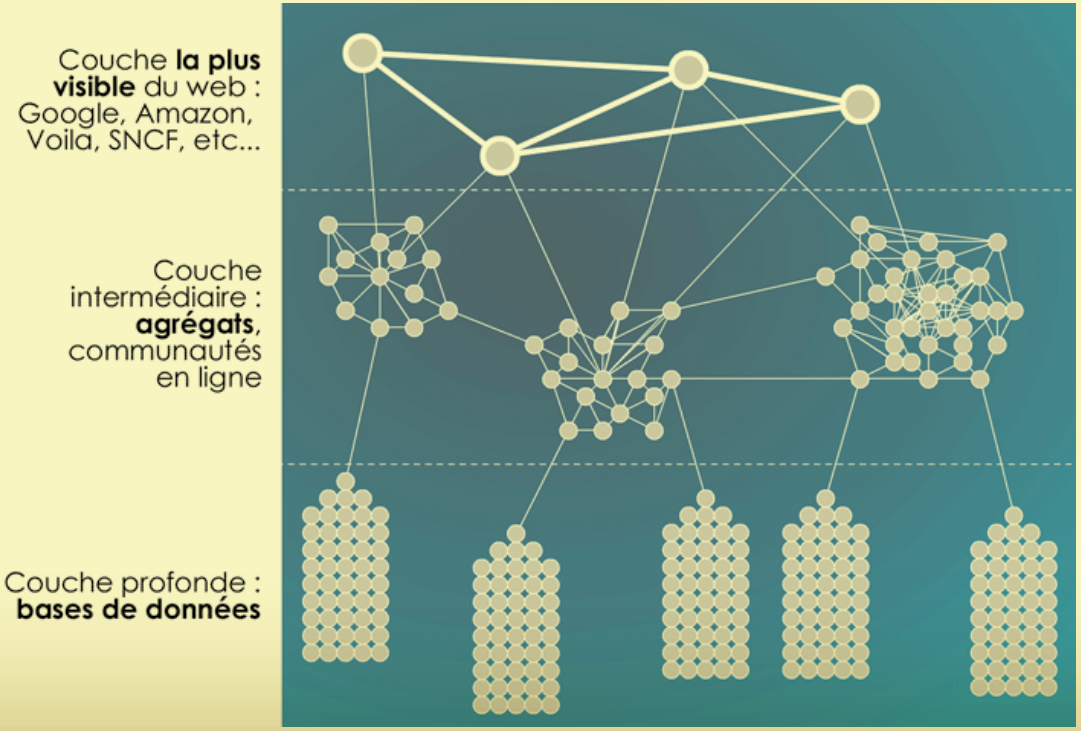
Exemple de crawl à 2 clics :

Facile à faire,
surtout si on aime
trier des milliers
de sites pour rien

Suivre les plis du web

Avantage du réseau :
Les agrégats cadrent l'exploration

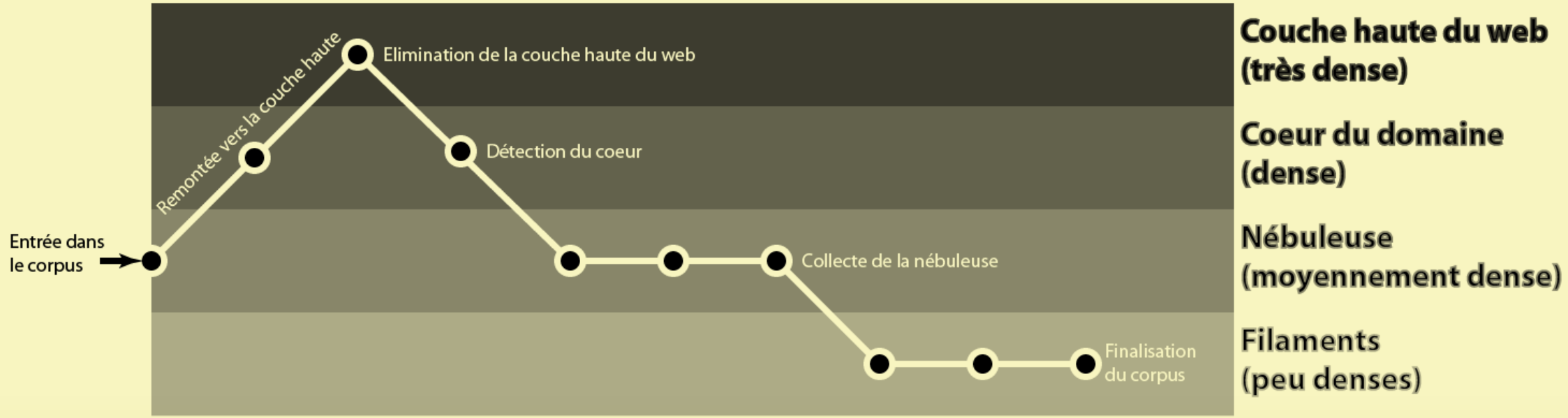
Inconvénient :
L'aspiration vers le haut



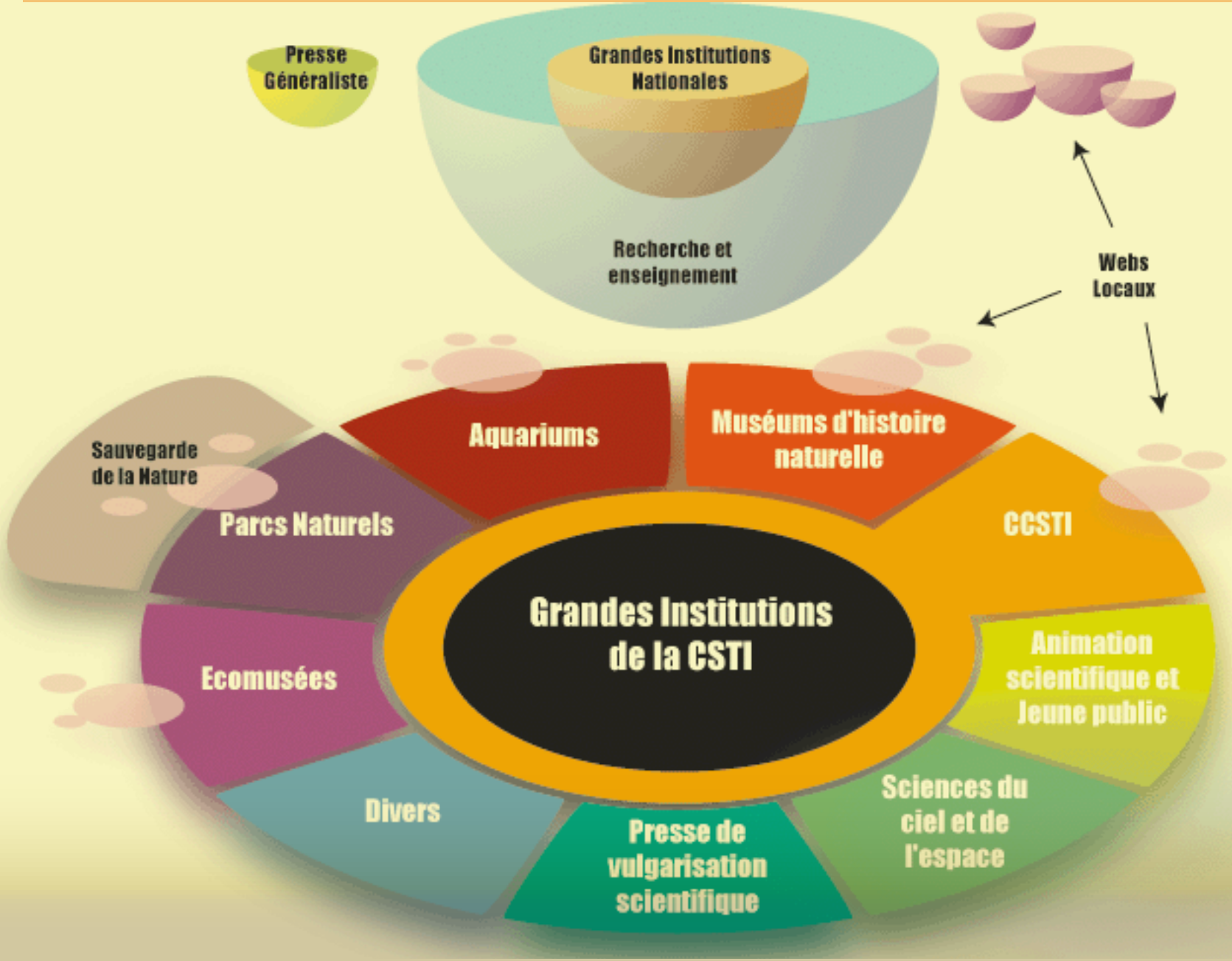
“La” technique

Remonter vers les couches hautes d’abord
pour endiguer l’aspiration vers le haut

Puis redescendre ensuite vers les couches basses



Visualiser un domaine



Domaines voisins de la couche haute du web

Domaine et sous-domaines de la CSTI, domaine voisin de la couche basse

La fonction “clef”

Exploration pas à pas :
Lutter contre l’éparpillement

La fonction essentielle :
Marquer les sites triés
pour ne pas repasser dessus

C’est à ça que sert le Navicrawler…
(entre autres)

Avant de parler du Navicrawler...

Le meilleur exercice
pour apprendre à explorer le web :

Prendre conscience
de sa propre stratégie
de navigation

La navigation

Un unique modèle :
L'extension de son territoire
en rayons
autour de points de repère

Combien de liens successifs
supportez-vous
avant de cliquer compulsivement
 **sur "back"** 
pour revenir à votre point de repère ?

La navigation exploratoire

Explorer, c'est parcourir les liens

Il faut lutter contre la désorientation

Le Navicrawler

Logiciel

Logiciel libre

Extension Firefox

Développé principalement par WebAtlas
avec le soutien de Sciences Po
...en continu !

Téléchargeable sur :
webatlas.fr > Navicrawler

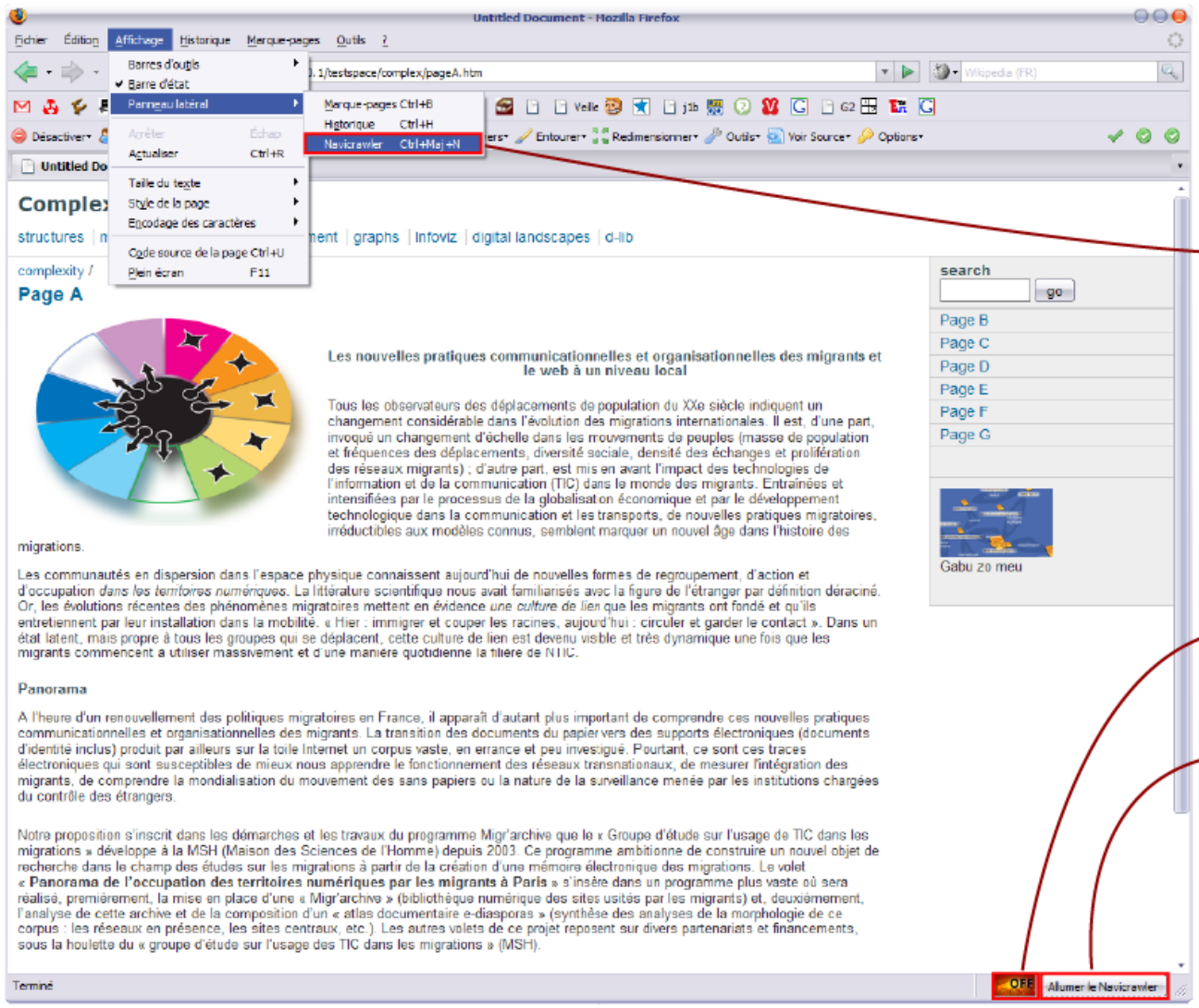
Un outil d'aide à la navigation

Le Navicrawler est comme un "bras bionique" :

Même principe que l'exploration à la main...

...en plus efficace

Interface de l'extension



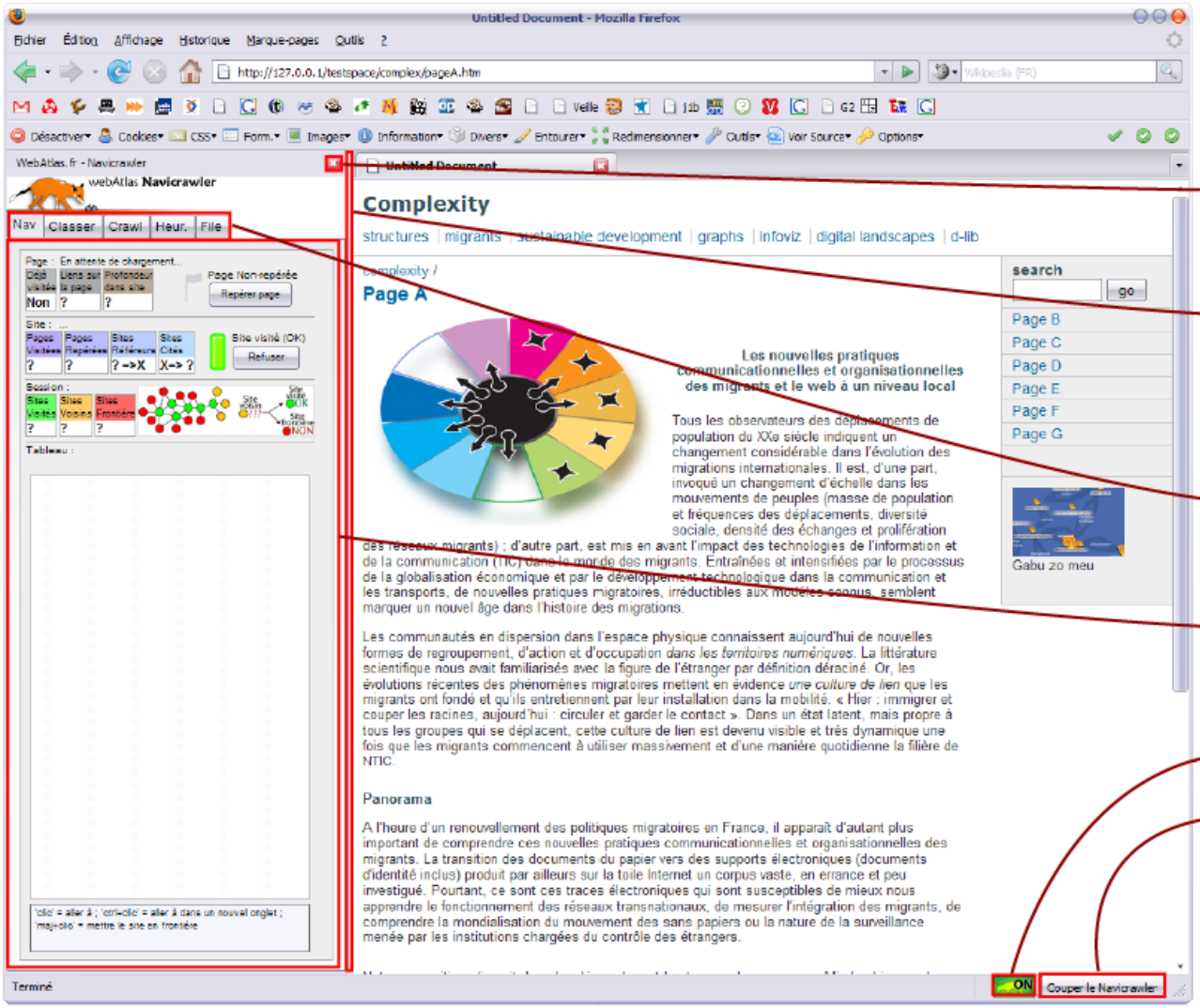
Navicrawler installé et inactif

Menu pour activer le Navicrawler

Voyant d'activité du Navicrawler

Bouton pour allumer le Navicrawler

Le Navicrawler activé



Navicrawler actif

Bouton pour fermer le Navicrawler

Séparateur :
vous pouvez modifier la largeur de l'interface

Onglets du Navicrawler

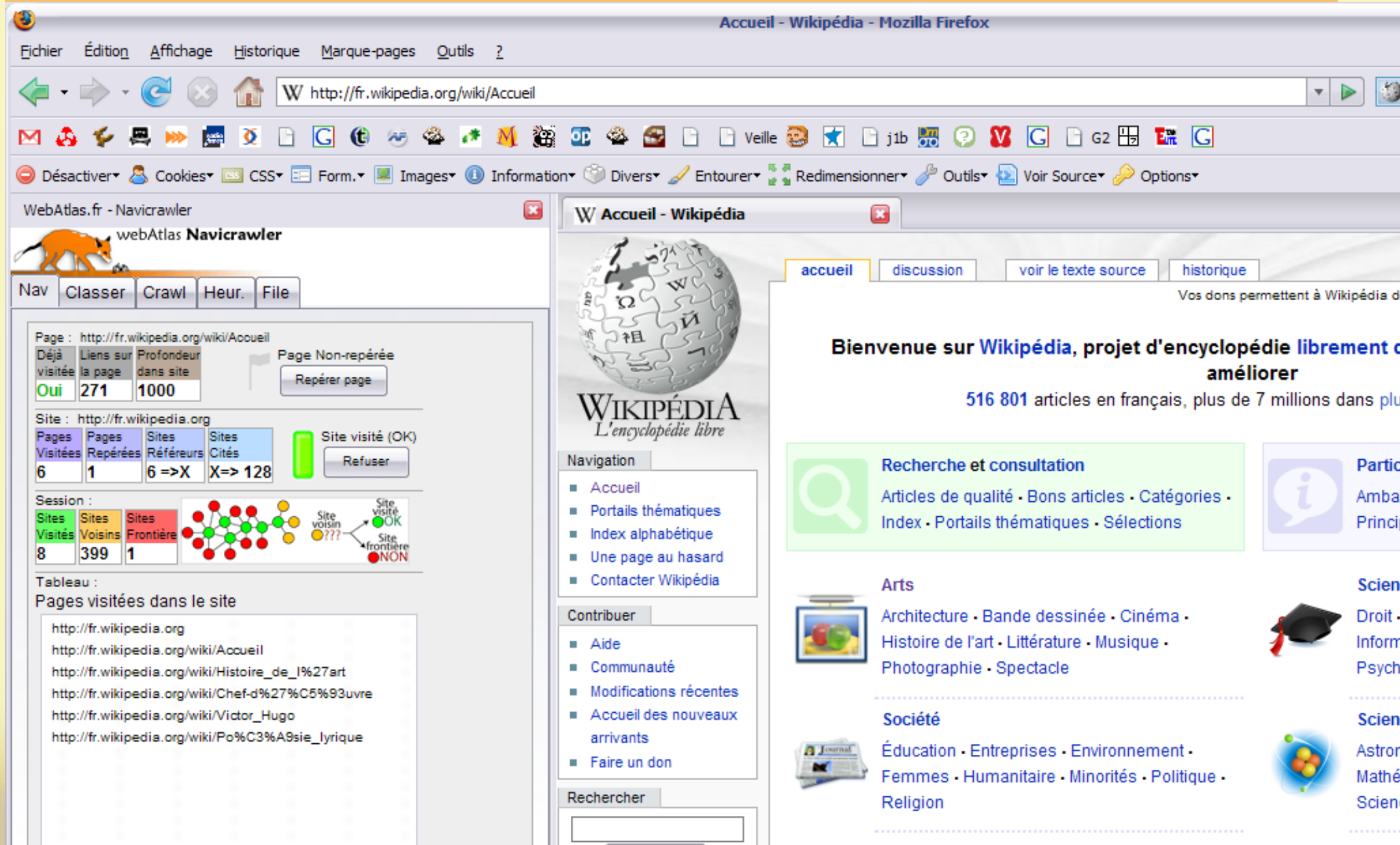
Interface du Navicrawler
(dépend de l'onglet sélectionné)

Voyant d'activité

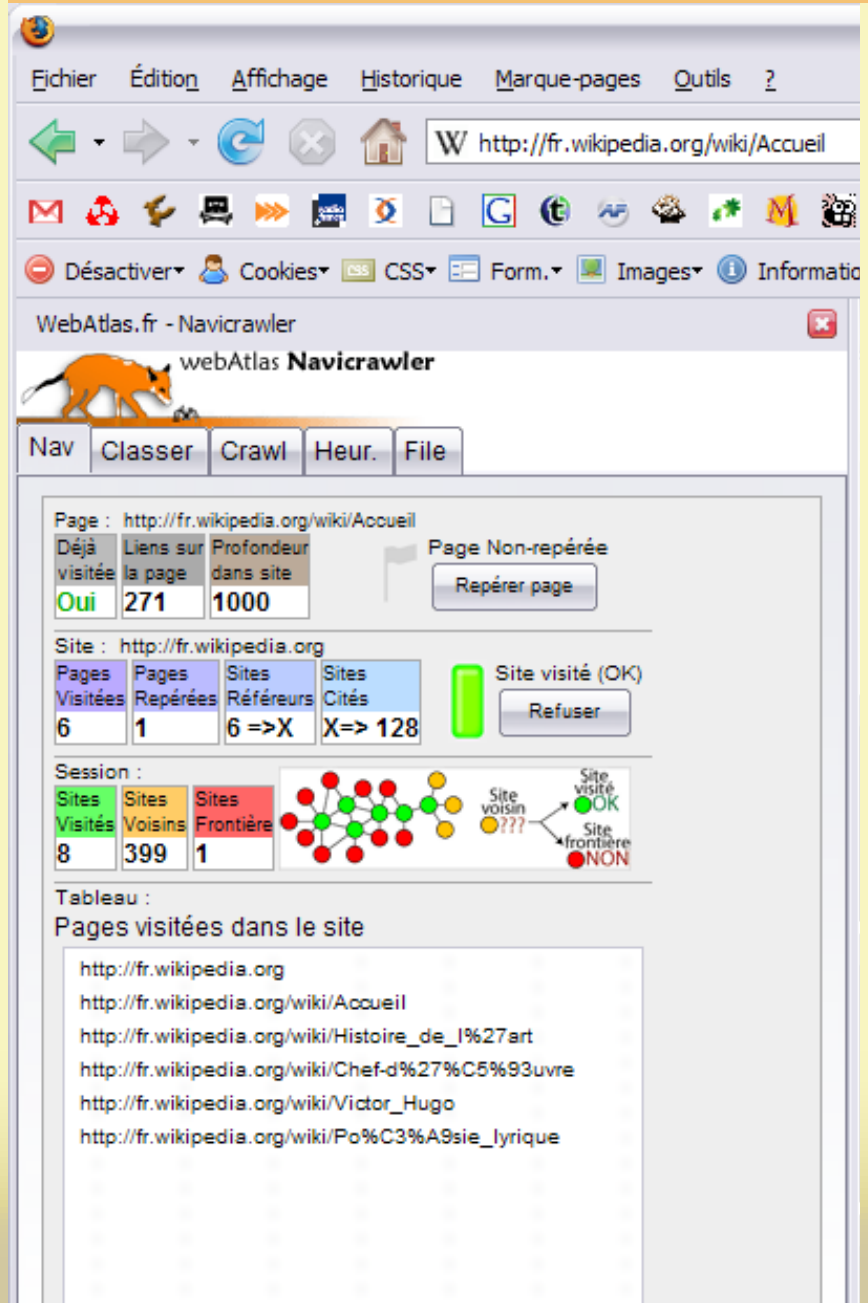
Bouton pour fermer le Navicrawler

NB : Une fois le Navicrawler activé, il faut charger une page pour le mettre en route.

Onglet principal



Onglet principal

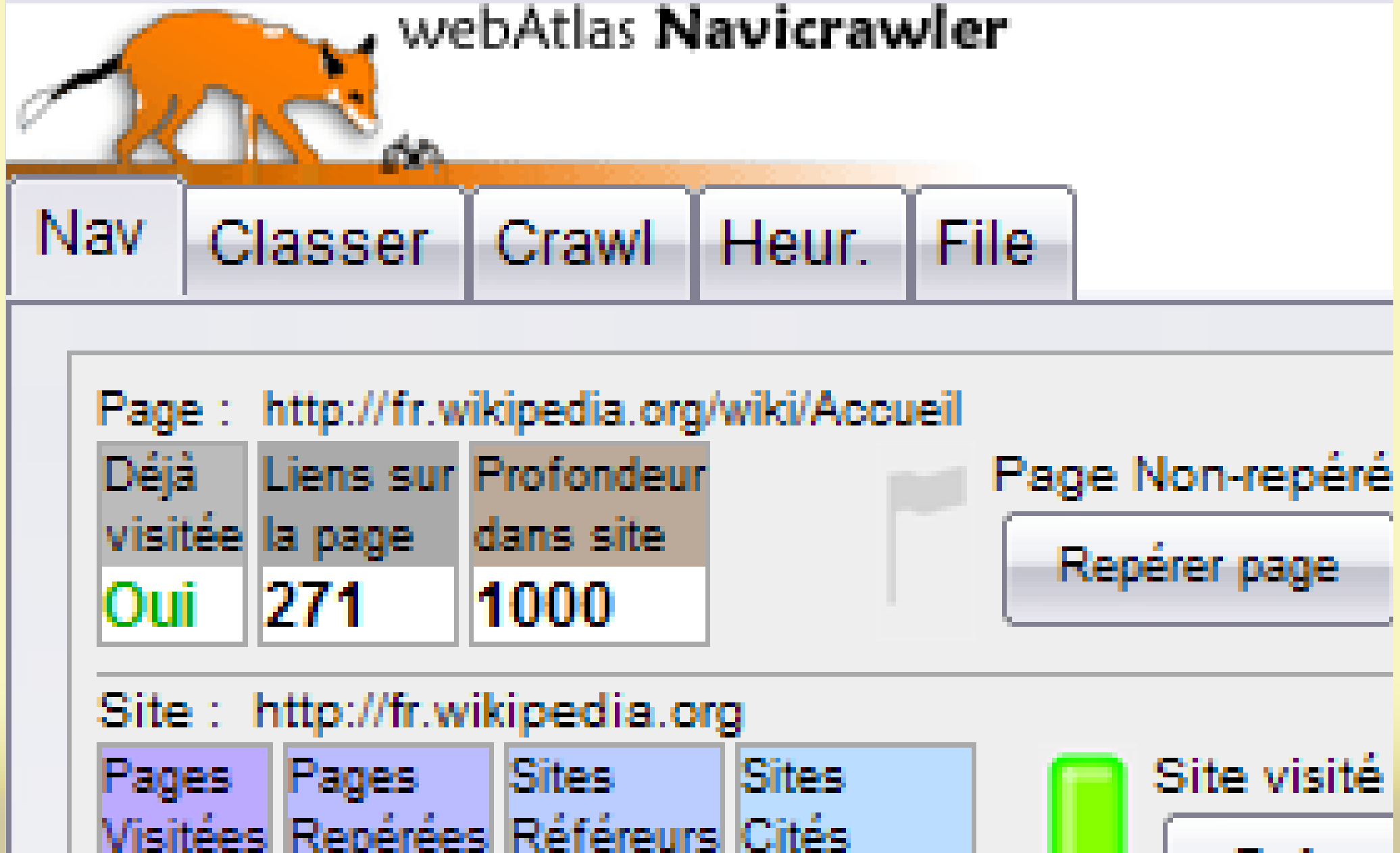


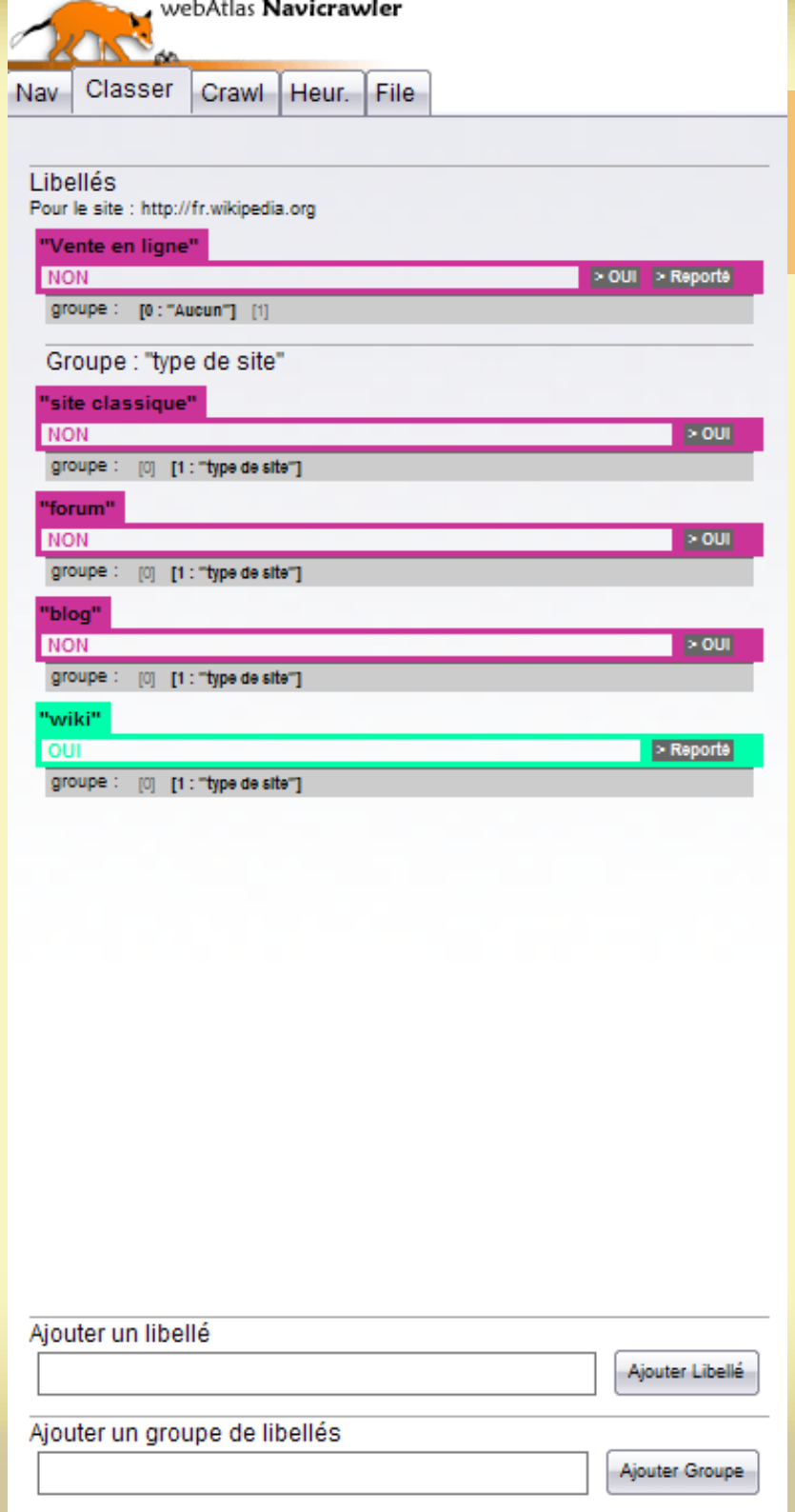
Visiter un site :

Par défaut, il est “incorporé”

Cliquer sur “refuser” : il devient “écarté”

Les onglets


webAtlas Navicrawler
Nav
Classer
Crawl
Heur.
File
Page : http://fr.wikipedia.org/wiki/Accueil
Déjà visitée
Liens sur la page
Profondeur dans site
Oui
271
1000
Page Non-repéré
Repérer page
Site : http://fr.wikipedia.org
Pages Visitées
Pages Repérées
Sites Référeurs
Sites Cités
Site visité



Classement

Cette interface sera bientôt revue de fond en comble

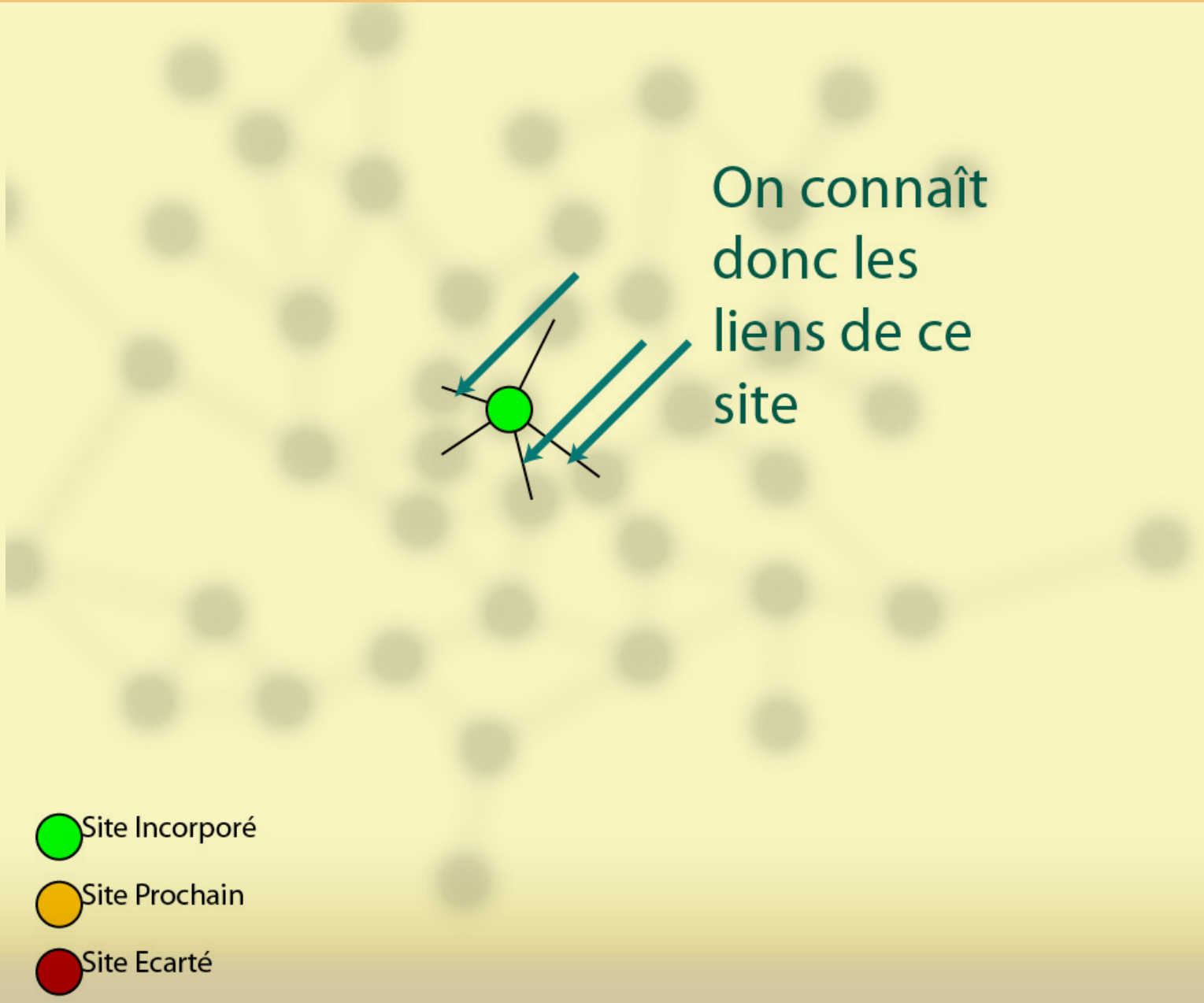
Principe


Nous allons visiter ce site
Site Incorporé
Site Prochain
Site Ecarté

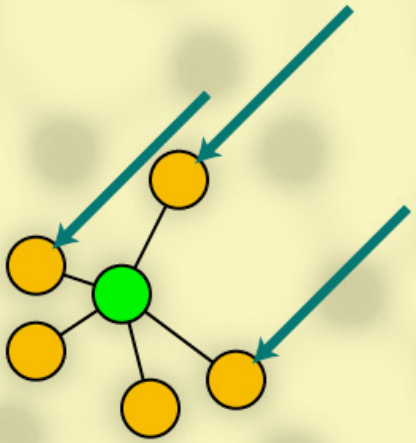
Principe


Ce site est visité.
Par défaut, il est
accepté.
Site Incorporé
Site Prochain
Site Ecarté

Principe


On connaît donc les liens de ce site
Site Incorporé
Site Prochain
Site Ecarté

Principe



Les sites au bout des liens sont détectés : ce sont les sites prochains.

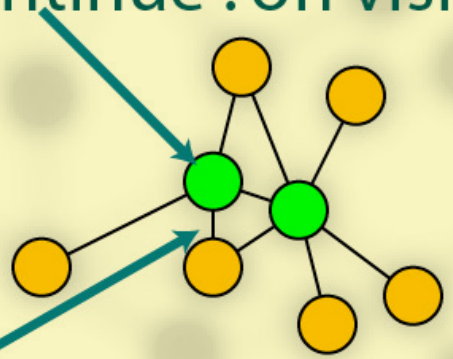
Site Incorporé

Site Prochain

Site Ecarté

Principe

On continue : on visite ce site.



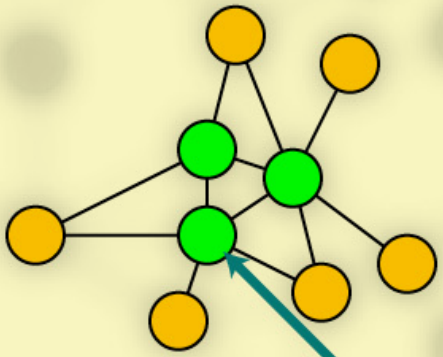
Ce lien est découvert
(on n'avait pas les liens entre sites prochains)

Site Incorporé

Site Prochain

Site Ecarté

Principe



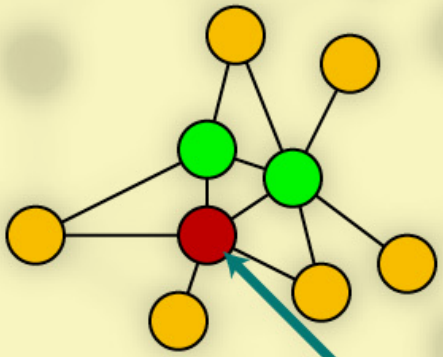
On viste encore ce site
Mais il ne nous
convient pas !

Site Incorporé

Site Prochain

Site Ecarté

Principe



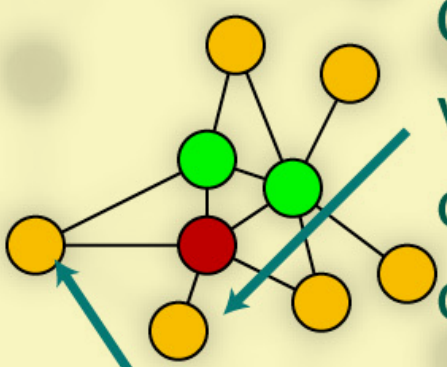
Ce site est donc écarté.

Site Incorporé

Site Prochain

Site Ecarté

Principe



Ce site prochain
va être effacé (car
on efface les voisins
des sites écartés)

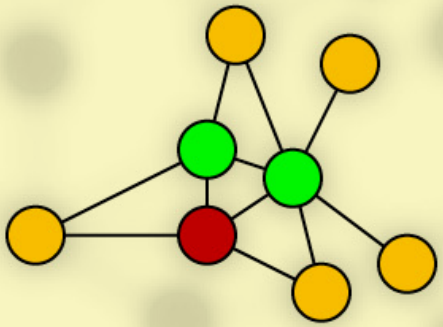
Ce site prochain est
gardé quand même car
il est lié à un incorporé

Site Incorporé

Site Prochain

Site Ecarté

Principe



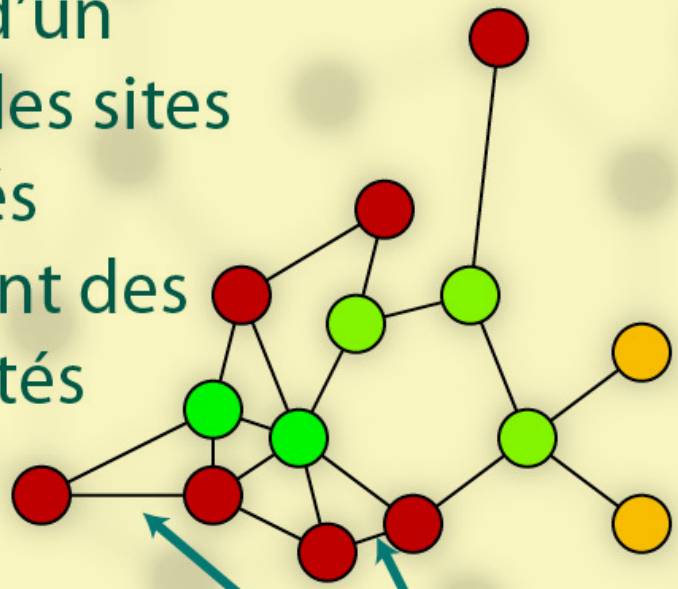
Site Incorporé

Site Prochain

Site Ecarté

Principe

Au bout d'un moment les sites incorporés s'entourent des sites écartés



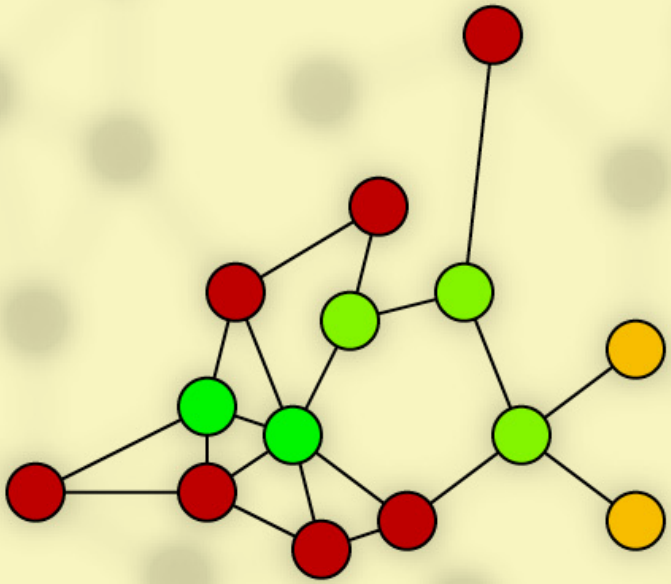
NB : on connaît les liens entre sites écartés

Site Incorporé

Site Prochain

Site Ecarté

Principe



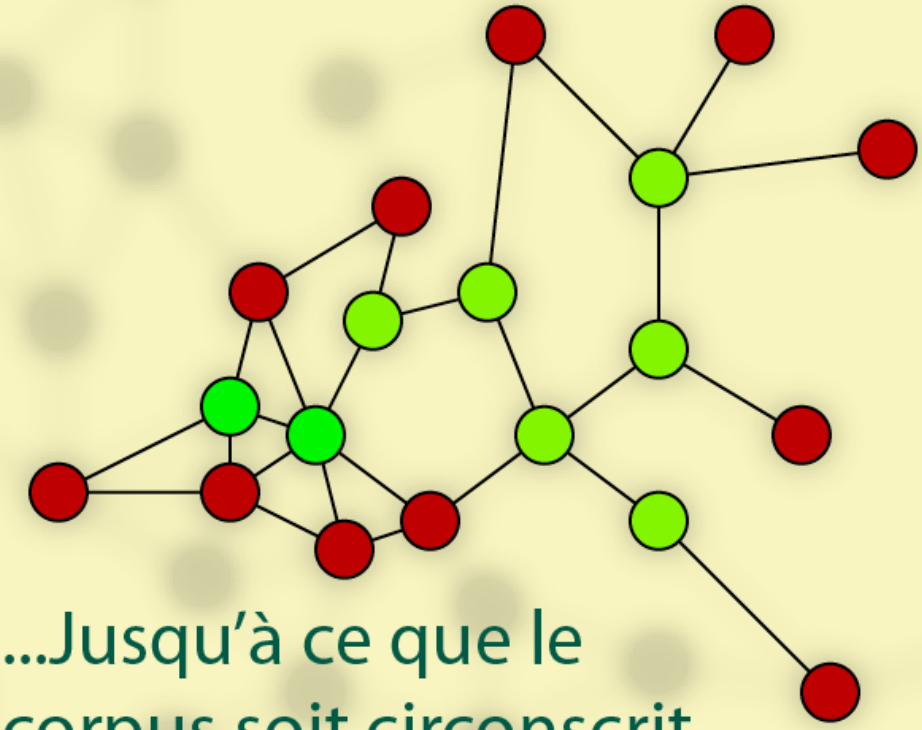
On continue
à traiter
les sites
prochains...

Site Incorporé

Site Prochain

Site Ecarté

Principe



...Jusqu'à ce que le
corpus soit circonscrit.
On a conservé les sites écartés
en tant que contexte.

Site Incorporé

Site Prochain

Site Ecarté

Gephi

Le logiciel

Logiciel libre

Supporté par WebAtlas
Développeur principal : Mathieu Bastian

Développé en continu

Plus d'informations :
Gephi.org

Demonstration

Présentation en direct

Merci de votre attention !