SEO
Search Engine Optimization
-1-

Prof. Romina Eramo

University of Teramo

Department of Communication Sciences

reramo@unite.it

Acknowledgement: Alessandro Celi (University of L'Aquila)

Introduction to SEO

- **SEO** is the activity of **optimizing the site for search engines,** i.e. the application of techniques, tools and knowledge to make a website better *digestible* for search engines and improve its positioning in relation to certain keywords.
- **SEO is not an exact science** , due to the variability of search technologies.

Introduction to SEO

SEO methods can be divided into 3 main areas:

- **SEO on language**
- **SEO on architecture**
- **SEO on reputation**

And on a site everyone is needed.

Taken from Web Usability 2.0 – Jacob Nielsen

SEO on language

- **Express yourself in a language suited to the target audience we are addressing**

- it is difficult for a user to carry out a search using our slogans

- We need to get into the heads of our possible users/customers and write on the site what they want to find.

SEO on

What does it mean?

- **Ensure that your web pages can be easily indexed**
- **Ensure an adequate link structure to guide spiders through content**

A crawler (also called a spider or robot), is a piece of software that analyzes the contents of a network (or database) in a methodical and automated way, usually on behalf of a search engine. A crawler is a type of bot (program or script that automates tasks).

SEO on reputation

- Search engines pay great attention to the reputation of sites.
- Sites considered authoritative will have a good positioning in the search engine results.
- **To gain a good reputation you need to be linked from other sites.**
- **Not all links are the same** : those that come from sites with a good reputation are worth more.

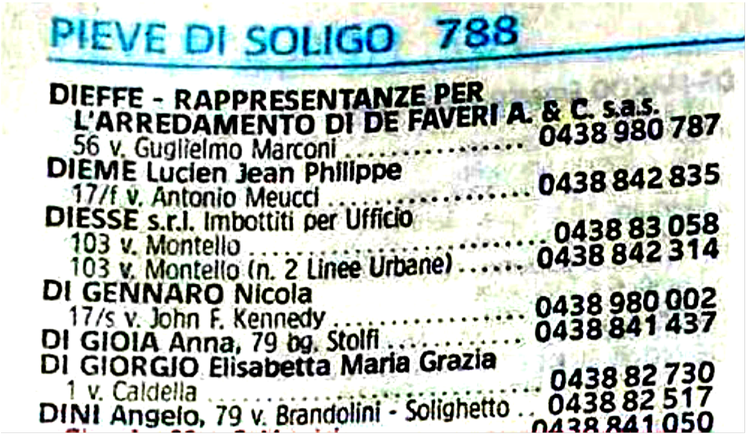
SEO and Google

- Google has also dedicated space to the definition of SEO
  http://www.google.com/support/webmasters/bin/answer.py?hl=it&answer=35291

  It has also published a guide to the SEO technique:
  http://www.google .com/webmasters/docs/search-engine-optimization-starter-guide.pdf

Search Engines

- **a** classic example: **the paper telephone directory**



PIEVE DI SOLIGO 788

DIEFFE - RAPPRESENTANZE PER L'ARREDAMENTO DI DE FAVERI A. & C. s.a.s. 56 v. Guglielmo Marconi	0438 980 787
DIEME Lucien Jean Philippe 17/f v. Antonio Meucci	0438 842 835
DIESSE s.r.l. Imbottiti per Ufficio 103 v. Montello	0438 83 058
103 v. Montello (n. 2 Linee Urbane)	0438 842 314
DI GENNARO Nicola 17/s v. John F. Kennedy	0438 980 002
DI GIOIA Anna, 79 bg. Stolfi	0438 841 437
DI GIORGIO Elisabetta Maria Grazia 1 v. Caldella	0438 82 730
DINI Angelo, 79 v. Brandolini - Solighetto	0438 82 517
	0438 841 050

Book with a list of names inserted in alphabetical order and classified by location.
What do we do to search for a number of interest to us?

Search Engines

- Another example is the **classic dictionary** .
  There are the terms, with their definitions, in alphabetical order.

  To search for a term we must leaf through the book, following the correct order of the letters until we reach the 'word' of our interest.

Search Engines

- When the archive is very large, or there is a lot of data to be searched, the search is undoubtedly slow and complex.

- To solve this problem, **search engines exist** .

What are Search Engines?

- A search engine is an automatic system that stores a large amount of data and classifies this data based on mathematical formulas, making the search for information within this large archive simple, fast and efficient.

Why do Search Engines exist?

- Search engines exist **to solve information search problems** , especially when the archived data is large in number, or is such that alphabetical, alphanumeric and sequential search would be too complex.

Web search engines

- These are **large data archives, which contain detailed information on a large number of web pages and which allow rapid searching.**
- Clarification: individual documents that make up a website are stored in search engines. It is therefore possible that a specific website is present numerous times in the archive , in relation to the number of pages it is made up of.
  In Google: *site:libero.it* -> indexed pages

Web search engines

- **The insertion of Web pages into the archives of search engines** can take place in two ways:
  - manual (through the user 's report)
  - automatic (through a particular software that manages to visit millions of Web sites a day, inserting new pages and updating those already present in the archive).

The impossible job of MoRs

- Trying to survey all the existing web pages on the internet is an impossible task!

- Why is it impossible?

The impossible job of Search Engines

- **Because the web is fragmented** : we can imagine it as a jagged forest made up of clusters of contiguous trees, but also of isolated plants on the top of a mountain or in an impervious valley, unknown and unattainable, unless there is a specific indication of their exact position
- **Because it is constantly evolving and growing:** every day millions of new documents are created and put online, or those already present on the internet are updated

The Science of Networks

- A bit of history.

  Taken from the Book:
  Link, The science of networks - Albert-Làszlò Barnabàsi
  Einaudi Editore

  The WWW (World Wide Web) is the largest network ever built by man :
  a vast set of documents, multimedia contents and services that can be rendered available from Internet users themselves.

The WWW

- **The Web is a virtual network where the nodes are the web pages that contain any type of information:** documents, images, videos, ...

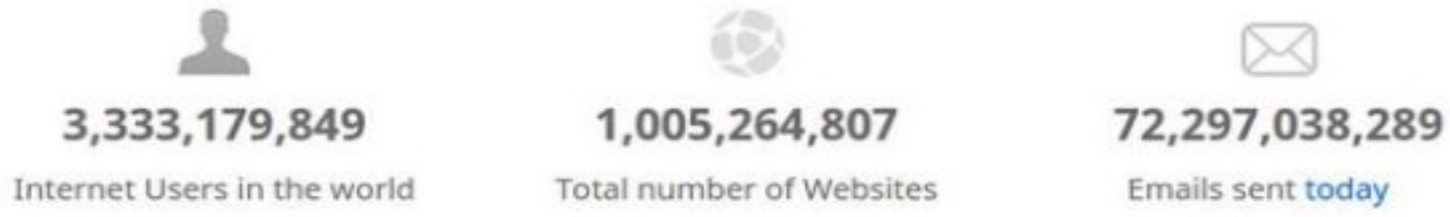
- **The power of the web lies entirely in Links** or URLs (Uniform Resource Locators), thanks to which it is possible to move between connected web resources.

How big is the Web?

- How many documents does it contain? How many links?

Until a few years ago no one knew this, not even approximately.

- Two NEC researchers took up this challenge. According to their survey, already in 1999 the web contained 1 billion documents.



3,333,179,849
Internet Users in the world

1,005,264,807
Total number of Websites

72,297,038,289
Emails sent today

How big is the Web?

- **The factor that matters is not the size of the Web, but the distance from one resource to another.** How many clicks are needed to move from one document to another, even if it is on the other side of the ocean ?

- If web pages are thousands of clicks away from each other , without a search engine we could never find a document.

The connectors

- There is therefore a handful of people who have the ability to make an exceptional number of friendships called **' Connectors ' .**
- Connectors have great importance in our society. They are those who hold the interconnecting threads of social relationships.
- These are **nodes with an incredibly high number of links.**
  Their discovery revolutionized all the knowledge already acquired about networks.

Is the web really democratic?

- If this were the case, everything published on the web would have the same opportunity to be read by millions of users.
- If the law that regulates this network were actually random, it would probably be true. But the reality is very different.

Is the web really democratic?

- Once we publish our ideas on the web, everyone has the opportunity to read them
- They become accessible to anyone, anywhere in the world, with a simple internet connection.

- **The right question to ask is another.**

Will anyone notice me on the web?

- **If I post something online will anyone notice it?**

- **To be read you have to be visible** : a banal truth that applies equally to writers and scientists. **On the web, the measure of visibility is links: the number of links and the weight of each link** .

Will anyone notice me on the web?

- Each web page has on average 5-6 links pointing to one of the billions of existing web pages. **The probability that the creator of a document will insert a link to my web page is close to zero** .
- For example, if the homepage of my website is linked to 50 web pages on the internet, and the web is made up of 10 billion web pages, the probability that my page will be visited is 50/10,000,000,000 ~ 0.0000005 % .

Will anyone notice me on the web?

- From the mapping of the web carried out by researchers at Notre Dame University, on a sample of 200 million web pages, a very interesting characteristic emerged: **90% of web pages did not receive more than 3 external links . Instead,**
  there were 2 or 3 web pages that actually received millions.

- Let's talk about portals and search engines: Yahoo, Altavista, Amazon.

The Hubs

- As happens in human society, a few individuals know an unusually high number of people (the connectors), **the WWW architecture is dominated by very few highly connected nodes called Hubs .** It is therefore clear that **all the little-known, poorly visible nodes with a small number of links are held together by these rare highly connected sites.**

The Hubs

- **Collectively we create hubs in some way.**
  These are the sites that everyone connects to, very easy to find, traceable from any point, typically search engines.
- However, questions remain:
  How are hubs formed?
  How many can a given network accommodate?

Power Law

- **The network that brought the robot back had many nodes with very few links and very few nodes with many links (hubs).**
- The result: **the distribution of links on various web pages followed** a very precise mathematical expression, called the **scaling law or power law** .

The growth factor

- **Networks in the real world tend to have one common characteristic: growth.**
  All the networks we know were born from a handful of nodes and then expanded further and further: the Web is a clear example.
  The assumptions in Erdos' models instead considered the existence of a finite number of nodes.

The rich are getting richer

- Growth alone still cannot explain the presence of hubs or connectors and therefore we cannot yet explain the power laws .

Preferential connection

- **The most connected nodes win!**
- Anyone, almost unconsciously, tends to connect with nodes they already know, almost always with the most connected ones on the web.
  Hubs tend to be preferred . Having to choose between two pages, we tend to visit the one that receives double the links, in other words the most connected one.

Preferential connection

- **The evolution of the network is governed by the law of preferential linking: we unconsciously tend to add links to already super-connected nodes.**

Why are Hubs created?

- The most important role is played by growth: early nodes have more time to acquire links, compared to younger nodes. **Growth offers a clear advantage to older nodes that will become the richest in links.**

- **Furthermore, with preferential connection, new nodes prefer to connect with nodes with more links.**

Why are Hubs created?

- **With the arrival of new nodes that continue to choose the most connected nodes, the oldest nodes will accumulate a very high number of links, distancing themselves from the group, until they form Hubs.**
- We can say that **the speed with which each new node attracts new links is proportional to the number of links it already has.**

Why are Hubs created?

- However, there are aspects of networks in the real world that cannot be explained by power laws and the scale invariance model:
  - internal links between old nodes
  - the replacement of one link with another
  - the sudden removal of links
  - the aging of a node with consequent loss of interest
  - etc…

How do the new ones survive?

- **How do newcomers survive in this unforgiving world?**
- In almost all complex systems, each node has special characteristics that go beyond its degree of connectedness. Web pages, companies and actors have intrinsic properties that influence the speed with which, in a competitive environment, I can attract new links.

How do the new ones survive?

- Some nodes, despite appearing very late, quickly win most of the links in the network.
  Others, despite arriving very early, do not win any.
- InkTomi case, Yahoo's partner, replaced by the newcomer Google...
- De Havilland case annihilated just one year after the start of Boing flights. Boeing itself lost an important market share to Airbus many years later .
- We can therefore say that the nodes are not all the same.

Competition in complex systems: fitness model

- Although it is impossible to find the universal key to success, we can study the process that separates the losers from the winners: competition in complex systems.
- We then introduce **fitness: the competitive ability of each node.**
  The ability of a web page to make us return to its content every day , rather than to that of billions of other pages competing for our attention.

Competition in complex systems: fitness model

- Between two nodes that have the same fitness, the older one still remains the favorite in the link race.

Competition in complex systems: fitness model

- **How did Google become a Hub overnight?**
- Last arrived, but with great technology behind it as a search engine, it gained new links much faster than its competitors, to the point of outclassing them.
  “ Beauty ” wins over age .

Winner takes all

- **It is possible that in some networks ' winner takes all ' .**
  The behavior of a network does not depend on the nature of its nodes and links, but on how its fitness is distributed.
  An example? Microsoft Windows which has practically conquered the entire market, despite not being the first operating system invented.

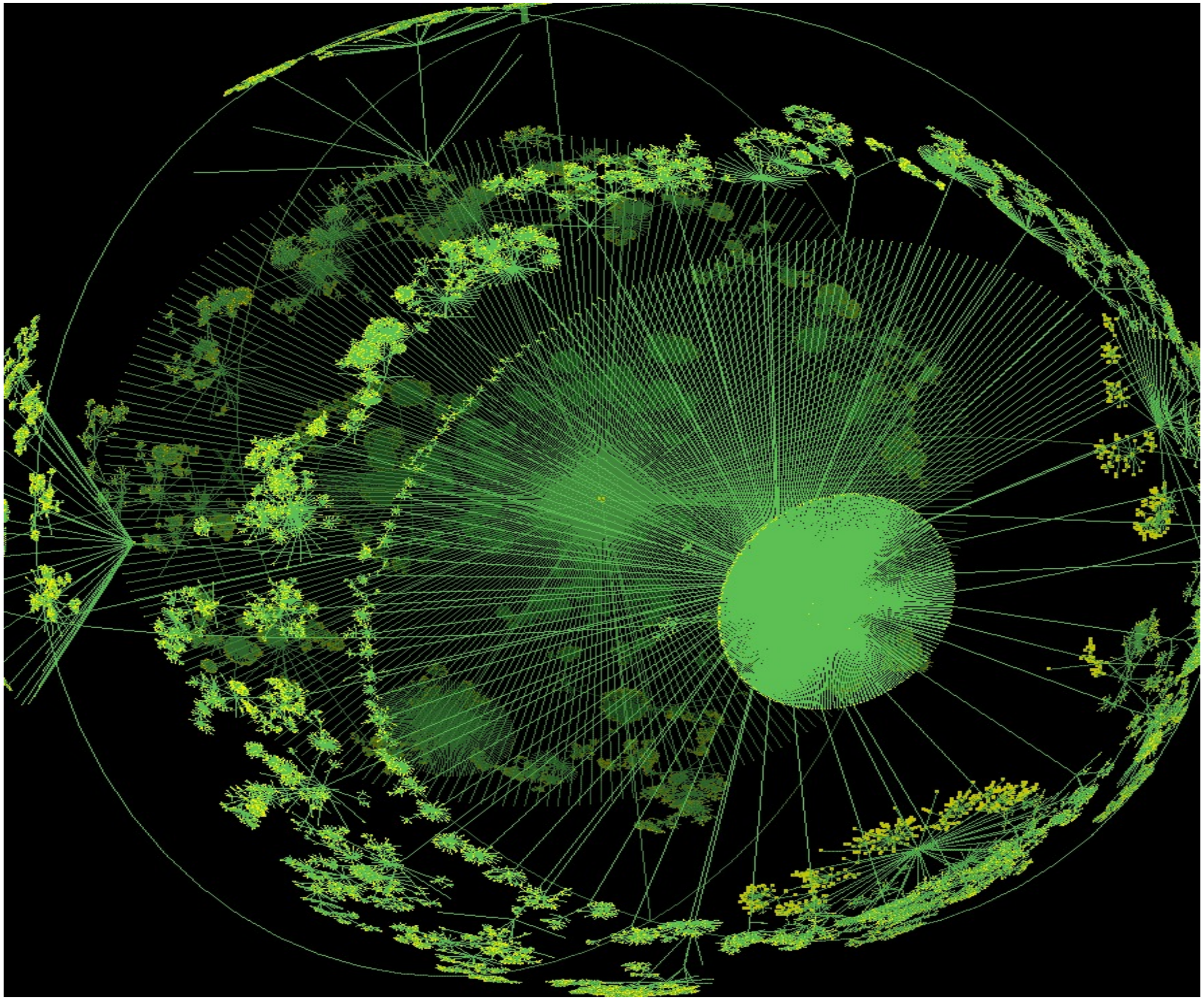
Winner takes all

- **Nodes are in constant competition because links represent the main source of survival in an interconnected world.**
  As long as we continue to choose nodes and discard others there will always be Winners and Losers.
  These are competitive systems where nodes fight forcefully to obtain new links.

Internet

- Understanding the topology of the Internet is the fundamental requirement for designing tools and services that offer a fast and reliable communication infrastructure.
- On the internet there is a close relationship between node density and population density. New nodes arise where there is more demand, i.e. where there is a certain number of interested people.

Internet



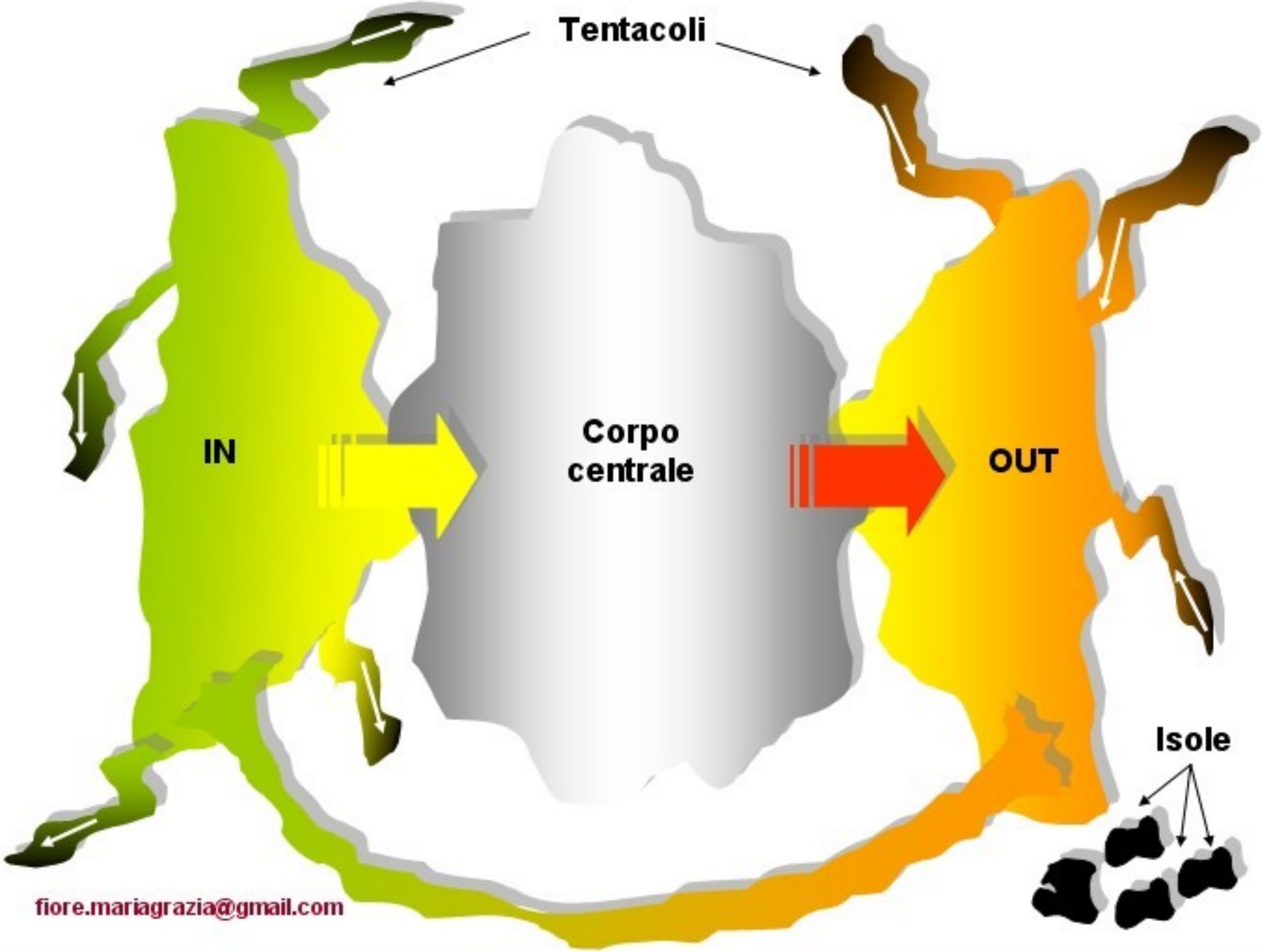
The Fragmented Web

- What does the fragmented web imply and why is this the case?
- - **Difficulty for spiders to index the entire web** .
- - **Difficulty returning the most relevant results** . 20 excellent results would be more useful than 100,000 of dubious usefulness.
- - **Links on the web are <<oriented>>:** along a given URL you can travel in only one direction.

The Fragmented Web

- **The web, having direct links, does not form a single homogeneous network** .
- It is divided into **4 large continents** , each with its own navigation rules (discovery by Andrei Broder of Altavista with other collaborators from IBM and Compaq, 1999).

The Continents



The Continents: Central Body

- The first of these continents comprises **a quarter of all web pages** . Called ***central body*** **hosts all the biggest websites. Highly connected, easy to navigate.** This does not mean that all documents are connected to all others, but simply that it is **very simple in this area to find a path to reach two random nodes** .

The Continents: IN and OUT

- The second and third continents, called IN and OUT, have the same dimensions as the central body but are more difficult to navigate. **From the pages of the *IN continent* it is simple to reach the central body, but once there there are no paths to go back.**

The Continents: IN and OUT

- Conversely, **the nodes belonging to the *OUT continent* are easily reachable from the central body, but there are no links to go back.** We can see, for example, the OUT continent as a **series of company sites that can be easily reached, but once entered there is no going back** .

The Continents: Tentacles and Islands

- The fourth continent is made up of ***tentacles*** and ***islands :*** **groups of pages connected to each other but not reachable from the central body.** Groups of pages connected to each other that can lead from the IN continent to the OUT continent, and vice versa, or remain completely inaccessible to those who do not know the precise address.
- About ¼ of web pages are located on this continent.

The Continents

- It is important to understand that **the position of a page on the web depends, more than on its content, on its relationships with other documents, through its external and internal links .**
  If your page is located in an island, search engines will never discover it unless you report the URL manually.

The Continents

- The possibility of tracking the entire web for a search engine is therefore not just a question of algorithms, technologies and economic resources. Search engines **can only map the central part and the OUT continent, therefore 50%. The rest remains hidden.** No matter how much effort the robots or spiders may make, they will never be able to find the documents located there.

The Continents

- **As the size of the Web increases , will it be possible in the future to form a single continent that absorbs all 4?**
  The answer is simple: **NO!**

- **As long as the links are direct this will never happen .**
  And links on the web will never become bilateral ☺

The Continents

- Continents are not exclusive to the World Wide Web.
- Even the network of scientific articles with their citations have a directionality and therefore they also have the 4 continents.
- The food web also has a direction: the lion eats the antelope , but the antelope does not eat the lion.
- **All direct networks are divided into the same 4 continents.**

Communities or Social Networks

- However, the 4 continents are not the only areas that can be defined on the web.
- On a smaller scale, small towns and metropolises proliferate within them. They are the sites that, united by an idea , a hobby or a habitat, create communities of shared interests: fans of rock music or carp fishing or opera. We can call them **communities or social networks** .

Communities or Social Networks

- **Every time we create a link to another web page, we highlight its importance with respect to our area of interest** . So presumably the links of a five-a-side football fan can lead us to others of the same type, allowing us to reconstruct the community of five-a-side football fans. **The identification of these virtual communities has enormous application potential** : social, economic, organisational.

Communities or Social Networks

- A problem arises: **among billions of web pages, is it possible to identify these communities?**
- The NEC researchers had formulated their hypothesis **: documents that have more links between them than links outside the community belong to the same community .**

Communities or Social Networks

- The NEC researchers' deduction was an excellent starting point for creating algorithms that identified different groupings within the topology of the World Wide Web.
- Years ago this would have been impossible.
- Google, on the other hand, has shown that this is absolutely possible, expensive algorithmically, but possible.

Communities or Social Networks

- **The web is** therefore **divided** into various **continents** , each of which hosts various villages and territories that have the form of **overlapping communities. Why overlapping?**
  A community that talks about English literature could be identified in multiple groupings or subsets:
  sites in English, literary sites, sites dedicated to Shakespeare, etc…

Democracy on the Internet

- We have seen that in the scale-invariant topology the documents have an average low visibility, because a minority of highly popular sites holds almost all the links.

- **There is freedom of speech on the web , but our voices may be too weak to be heard.**

Democracy on the Internet

- **Pages with few external links will never be found by a casual search** , because hubs tend to focus all the attention on them.
- It seems impossible for robots to escape this popularity trap.
- Pages that have only one external link have a 10% chance of being indexed. Those that receive between 21 and 100 external links have a 90% chance of being indexed.

Democracy on the Internet

- **Whether someone finds my Web page depends on only** one factor **: my location on the Internet** .
- If many visitors connect to my web page, find it interesting, and create a link from their pages, then it will gradually become a hub, which search engines will necessarily have to notice.
- If my web page is ignored, it will end up in the oblivion of the forgotten and isolated pages of the web.