Fabrizio Luccio

Mathematical issues in network construction and security

Dottorato 08

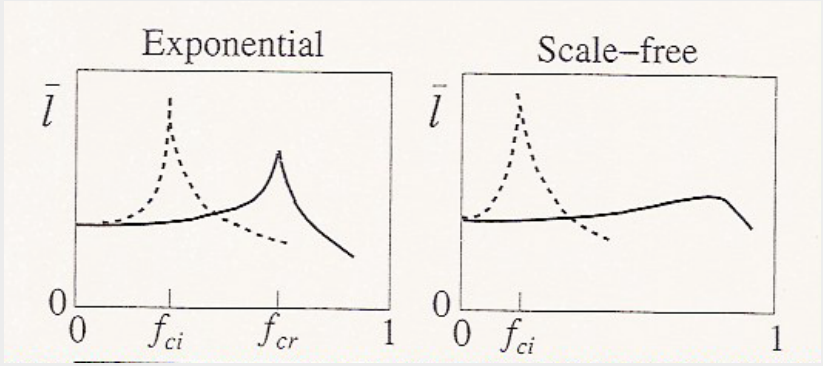
3. Failures, attacks, and spam farms

Failures and attacks

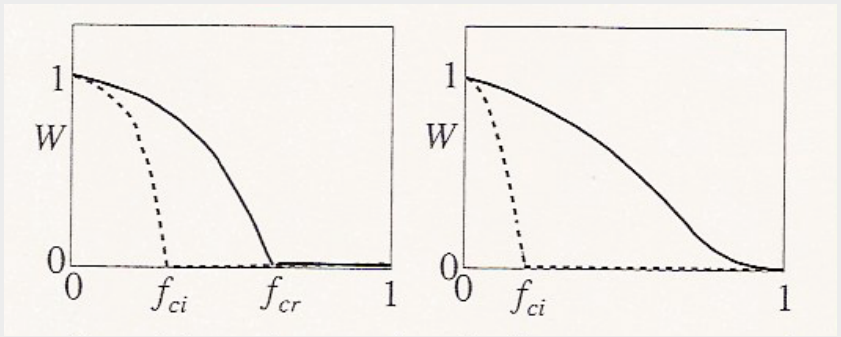
Internet and WWW have a power-law form of vertex degree with exponent between 2 and 3, so the second moment of the vertex degrees diverges as N grows. Then,for the Molloy-Reed criterion, the (big enough) Internet and WWW graphs have a Giant Connected Component. With power-law distribution the GCC tend to disappear for $\gamma > 3$.

> We report the simulation results of Albert et al. on the removal of a subset of vertices, done either at random (net failures) or intentionally (net attack). In particular we show the values of the average shortest-path length, and of the size of the giant component, as a function of the fraction f of vertices removed:
>
> > when the giant component disappears the network has been practically destroyed.

Failures and attacks on random and power-law nets ($\gamma < 3$)



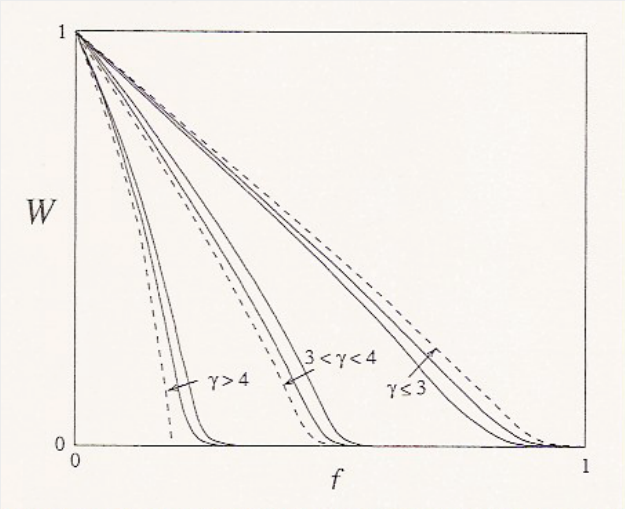
Average shortest-path length in the largest component, vs fraction of removed vertices,for intentional attacks (dotted) and random attacks (solid) on vertices, after Albert et al.



Relative size of the giant component for intentional and random attacks on vertices, after Albert et al.

Failures and attacks

The simulations show that scale-free networks like Internet and WWW are extremely robust to random attacks (maybe this is why scale-free networks are so common in nature).

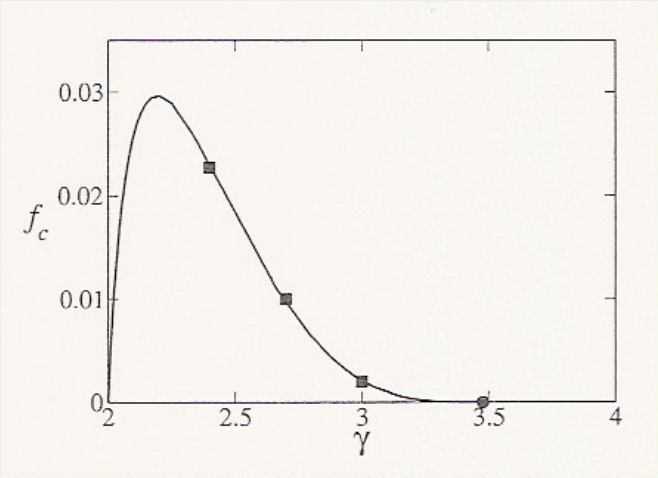


Relative size of the giant component for random attacks in scale-free networks with different exponents, after Dorogovtsev – Mendes

Arrows show increasing values of N, with dashed lines for N going to ∞

Failures and attacks

Scale-free networks are extremely robust to failures (random attacks) but can be destroyed with intentional attacks aimed at removing the vertices with highest degree.



Fraction f_c of the most connected vertices in a scale-free net, that must be removed for destroying the giant connected component, after Dorogovtsev – Mendes (GCC disappears for $\gamma = 3.48$).

The fact that scale-free networks are very robust to random attacks implies that they are very weak against infection spread (e.g., viruses in the Web). This phenomenon could be stopped automatically only by removing the vertices of highest degree.

Summarizing:

Internet and WWW are extremely robust against random attacks (failures). They can be destroyed only intentionally, attacking the vertices of highest degree.

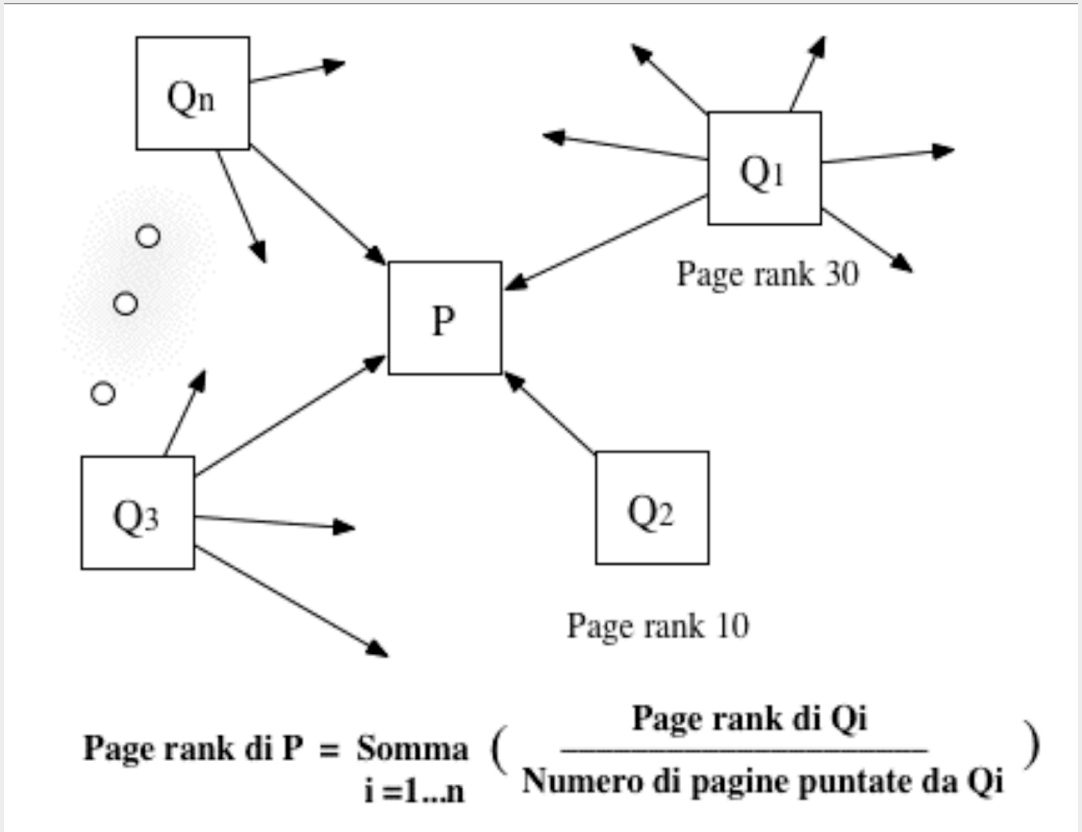
Internet and WWW are very weak against virus spread.

Google page ranking

Google became the most popular search engine partly due to the algorithm for deciding page popularity, through the computation of page rank.

Q1 has P-rank 30 and 6 outgoing links. It adds 30/6 = 5 to P-rank of P

Q2 adds 10/1 = 10 to P-rank of P



Google fooling

The answer of Google (or, for this sake, of any other search engine) can be manipulated by a malicious user?

Three main techniques

LINK AND TERM SPAM (for unduly boosting a page)

INVISIBLE LINKS (for misleading the engine)

MALICIOUS KEYWORDS (for misleading the user)

A good taxonomy of web span has been given by *Z. Gyongyi and H. Garcia Molina (2005).*

LINK SPAM AND TERM SPAM

The main theory has been laid down at Stanford University by Gyongyi and Garcia Molina in 2005. Then a good amount of work followed (possibly some of it has not been disclosed).

A basic definition is the one of *link spam* via a *link farm*, i. e. a group of boosting pages B inserted in the Web to promote page ranking of a target page T. To this end, a group of regular pages H are also corrupted with the insertion of new links, thus becoming hijacked.
The major known attacks have been designed for Google.
A link farm is optimal if, for a given number of pages B and H, the page rank of T is maximized.

Example of a hijacked page in a blog:
My great trip Napoliscarpe to Chiapas

LINK SPAM AND TERM SPAM

Term spam refers to inserting probably queried terms in the different fields of a page (title, body, header) to make the page relevant for a query. Spam terms may also be added in the anchor text of a hyperlink to the page: in this case the spamming term is not in the page itself.

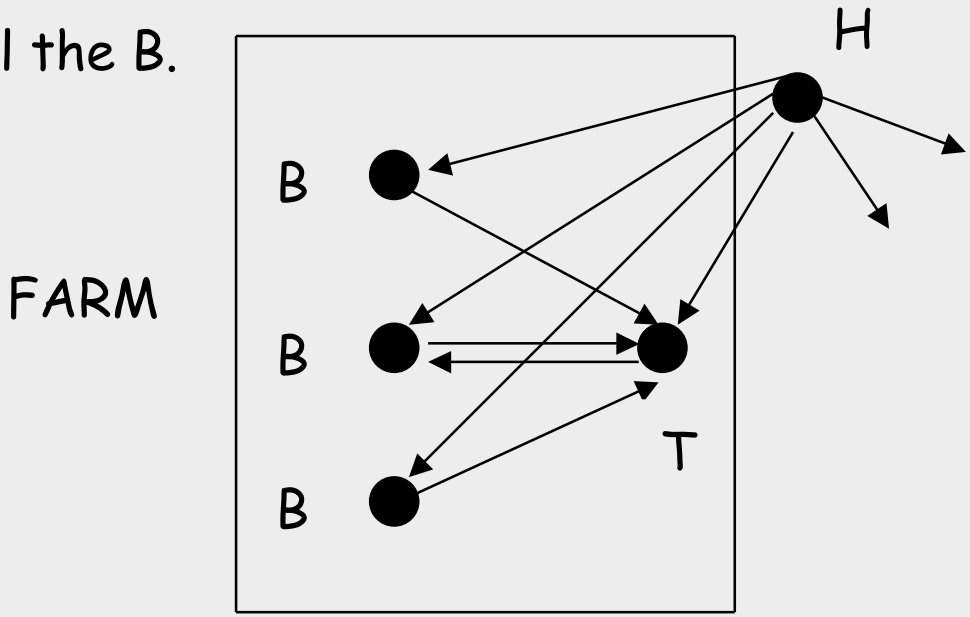
We give a brief account of work on link farming published in 2006-07, in particular by:

(1) Du, Shi, and Zhao: theoretical
(2) Becchetti, Castillo, Donato, Leonardi, and Baesa-Yates: experimental

Link farming

Theorem 1. If each page H points to at least two normal pages that: 1. are not hijacked, and 2. do not point to a page B, then the spam farm is optimal iff:

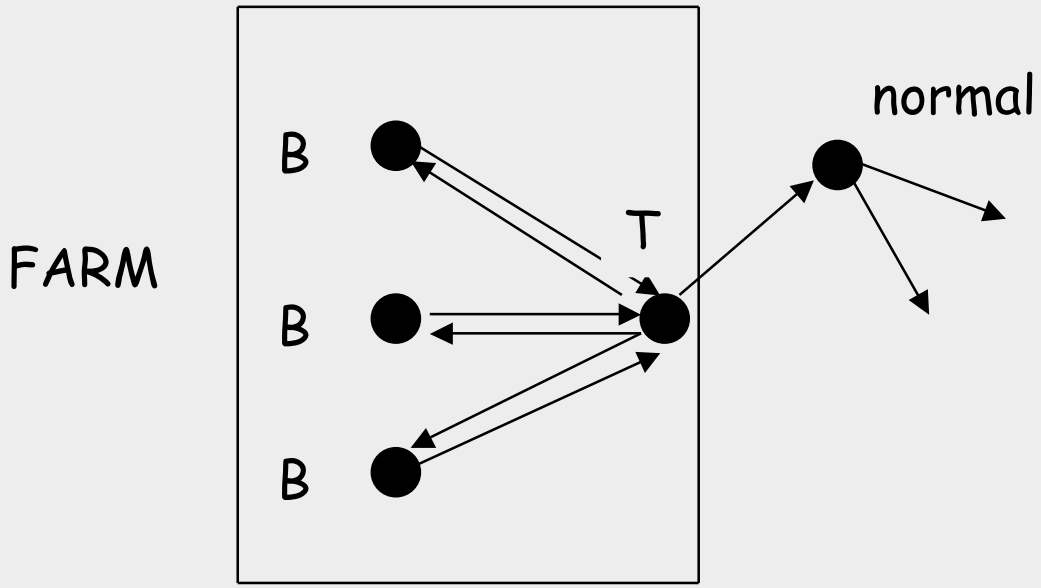
- pages B point only to T;
- page T points only to some pages B;
- pages H point to T and to all the B.



Link farming

Theorem 2. If T is required to point to some normal pages, then the spam farm is optimal only if:

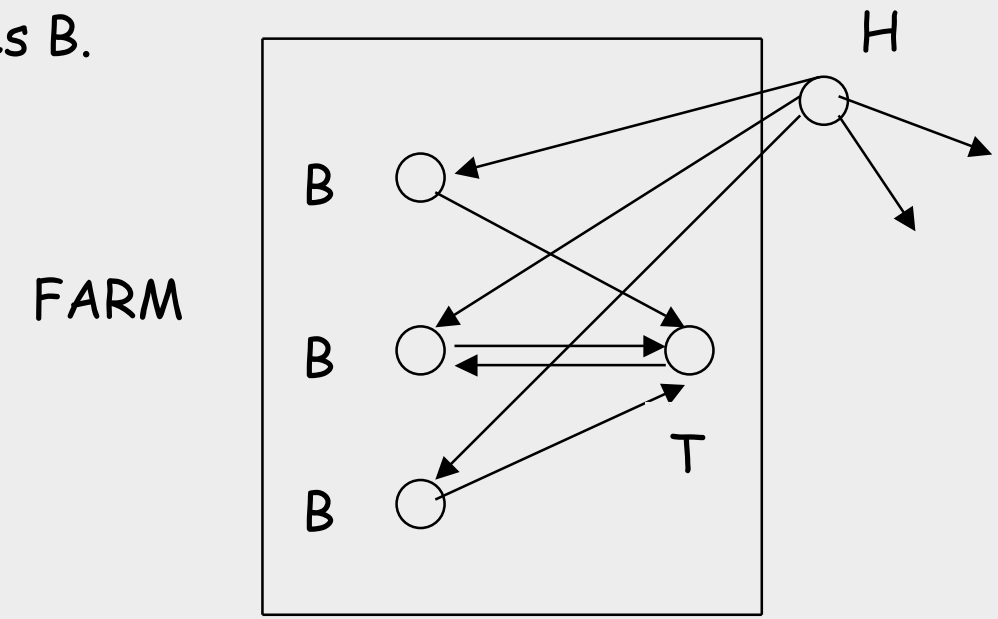
- pages B point only to T;
- page T points to all pages B.



Link farming

Theorem 3. If pages H do not point to T, then the spam farm is optimal only if:

- pages B point only to T;
- page T points to some pages B;
- pages H point to all pages B.



Link farming

Work (2) reports on various web spam classifiers for different types of page rank, together with an experimental analysis for detection of spam farms. Combining these classifiers, around 80% of Web spam can be detected, with a 1.1% of false positives.

Fooling search engines

INVISIBLE KEYWORDS

The code for a Web page P is written HTML, a language that allows specifying all pictorial details of the page, including the color of writing and background. Then a keyword K can be written in P, e.g. in white on a white background, as to remain invisible when the P shows up. Still K is there, and is discovered by the search engines that associate K to P, thus promoting the visibility of P while a user was looking for something completely different.

If you ask Google for "campionato di calcio" (soccer championship, in Italian) you will get in the first position a German Web page advertising hotel reservations at low rates. The visible page is written in German and English, but the invisible sentence, inserted several times in the page, is there for misleading Italian users.

Fooling users

MALICIOUS KEYWORDS

Keywords very similar to popular ones, possibly with some common misprinting. The aim is directing a misprinting user towards specific sites, for commercial purposes or even for fraud.

The situation is much more serious with malicious domain names, that may different from sound ones only for almost invisible details (e.g., number 0 instead of capital O, or number 1 instead of small l). The attack here is at browser level.

For example a commonly visited site in Italy is www.repubblica.it corresponding to an important newspaper. Misprinting the word "repubblica" with only one b, causes "e.bay" to appear.

A bibliography for starting

M. Gori, I. Witten. The bubble of Web visibility. Communications of the ACM N. 48, March 2006.

L. Becchetti et al. Link-based characterization and detection of Web spam (2006).

Y. DU et al. Using spam farm to boost page rank (2006).

Z. Gyongyi, H. Garcia Molina. Indexable Web spam taxonomy (2005).

Available on the Web.