

The blogosphere as a network: an empirical laboratory

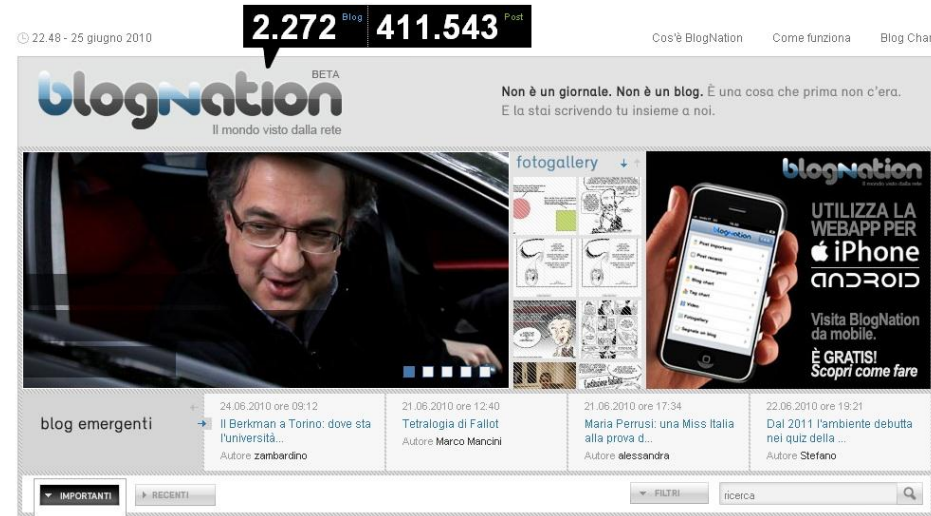
Kang Zhao and Massimiliano Spaziani

June 25

SFI CSSS 2010

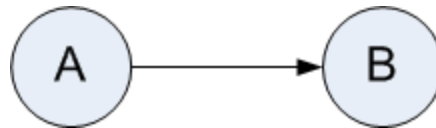
Introduction

- The dataset
 - Blognation
 - An Italian blogosphere
 - RSS feeds
 - Web crawling
- Basic statistics
 - 2,000+ authors/bloggers
 - 370,354 blog posts.
 - Topic classification:
 - News, cars, culture, entertainment, food, sports, tech



Post-Post Citation Network

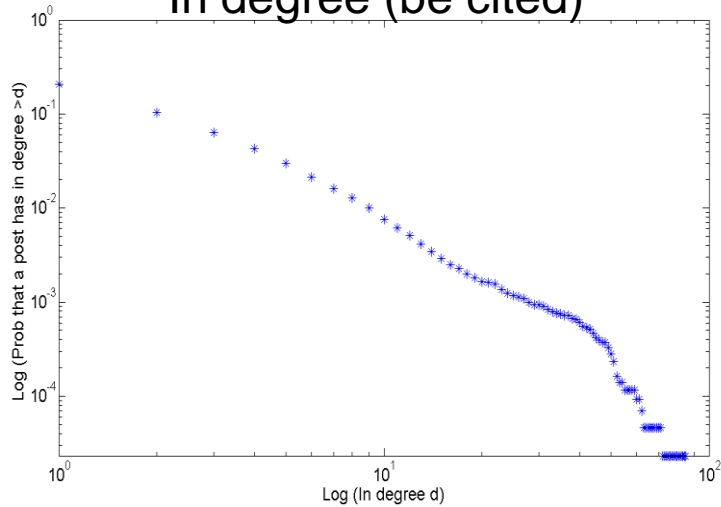
- Nodes-- blog posts
- Edges -- citations between posts
 - E.g. post A contains a link that points to post B



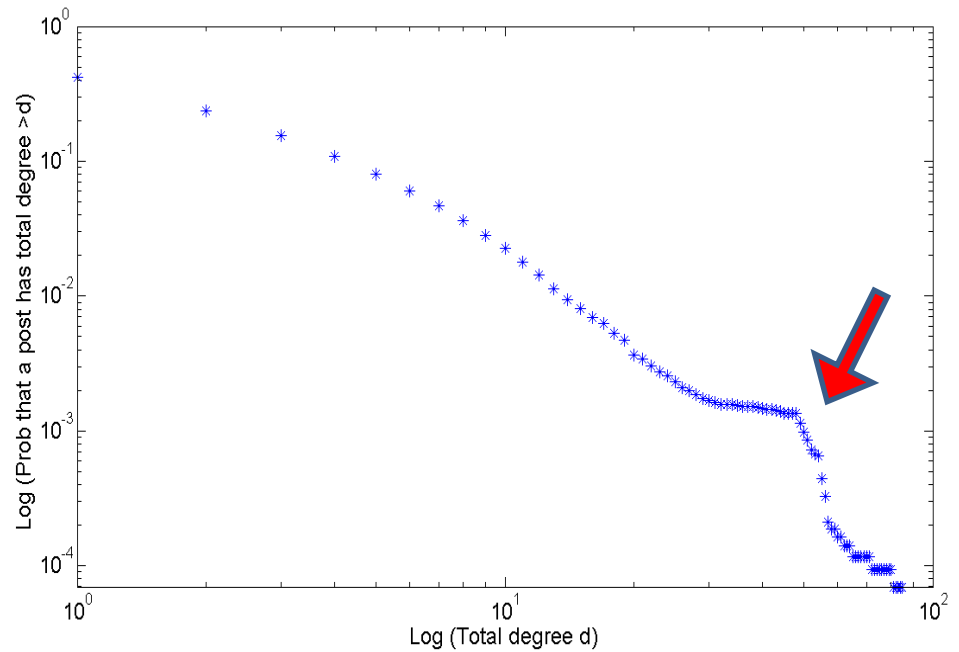
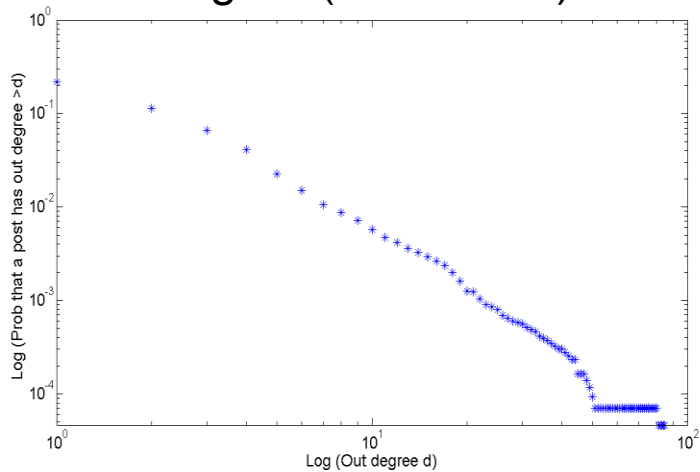
- Citation network statistics
 - 43,047 nodes.
 - 50,434 edges.
 - 906 authors

Degree distribution

In degree (be cited)



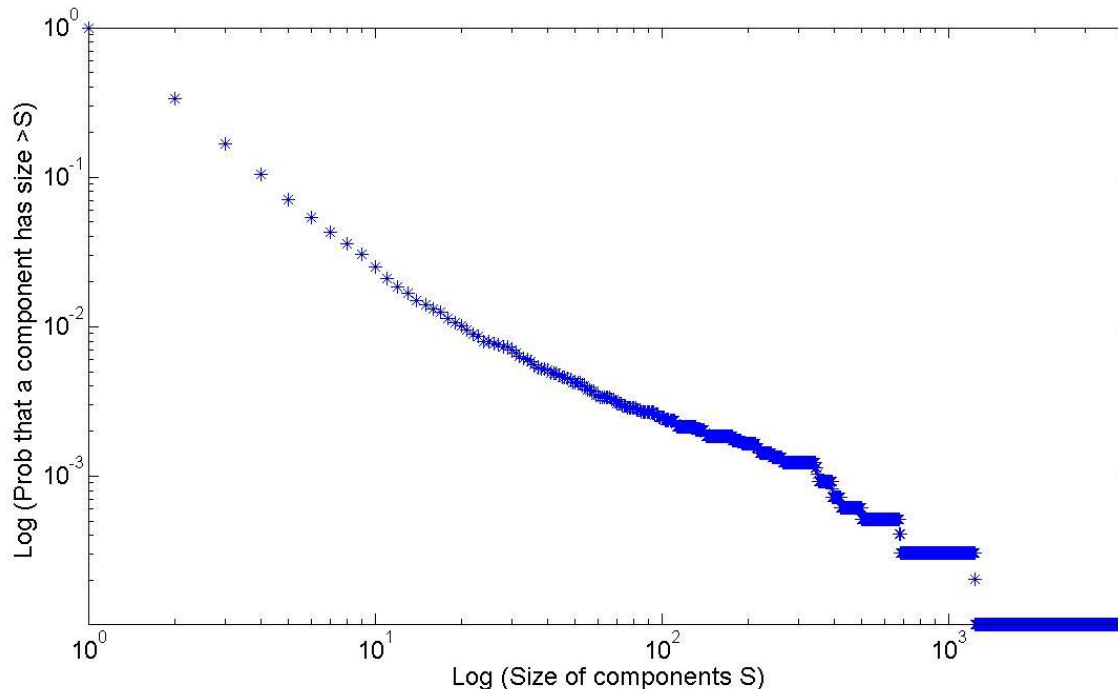
Out degree (cite others)



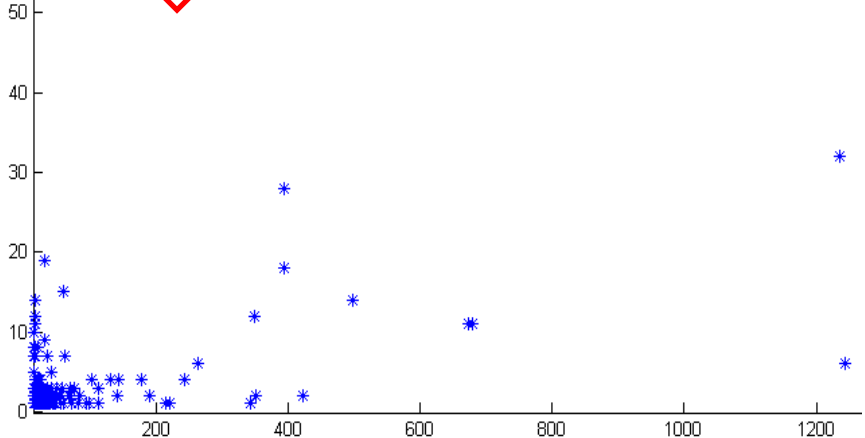
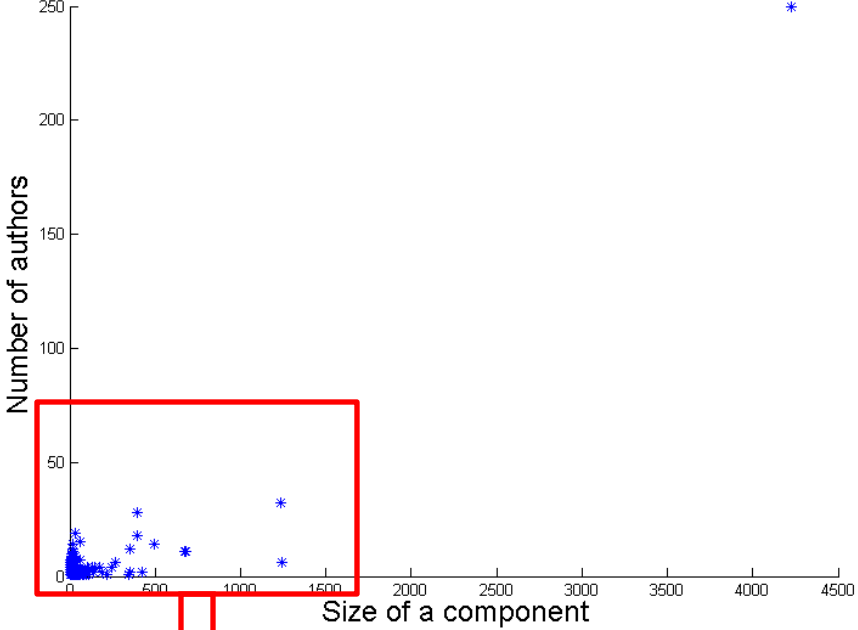
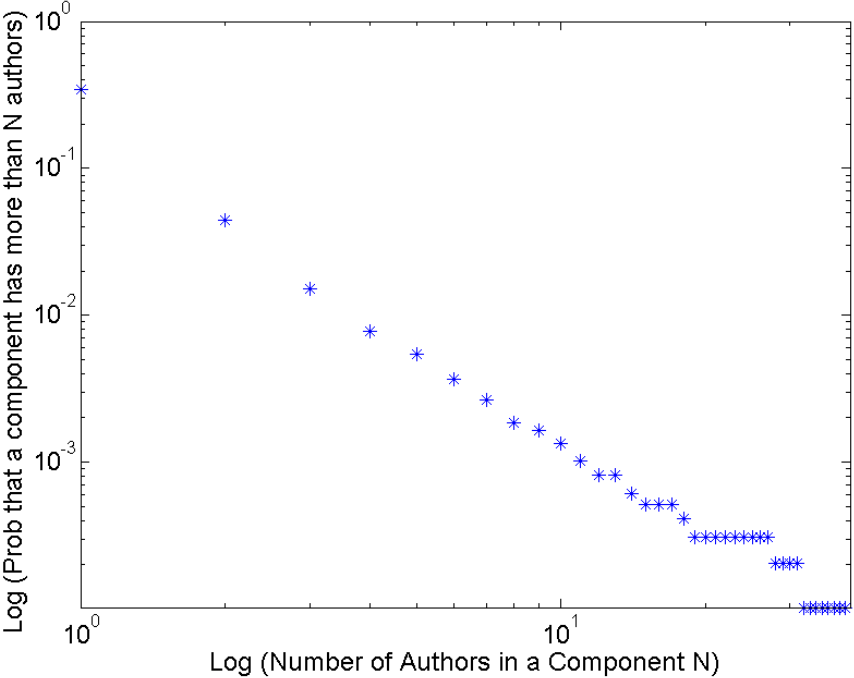
Total degree

Connected components

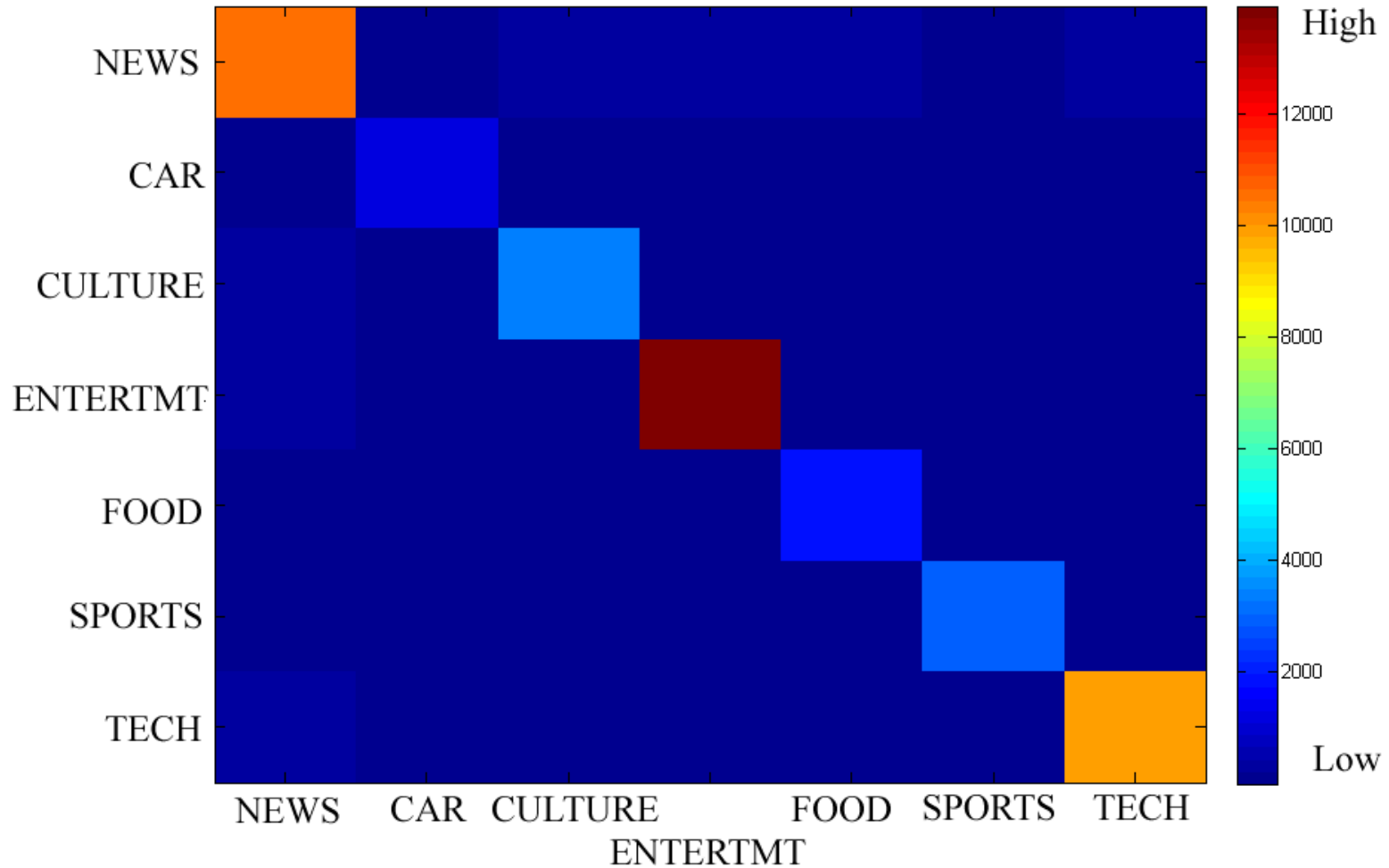
- No giant component
- 9754 components that don't connect to each other
- The largest connected components have 4228 nodes
- Component size distribution



Number of authors in components



Citation count map

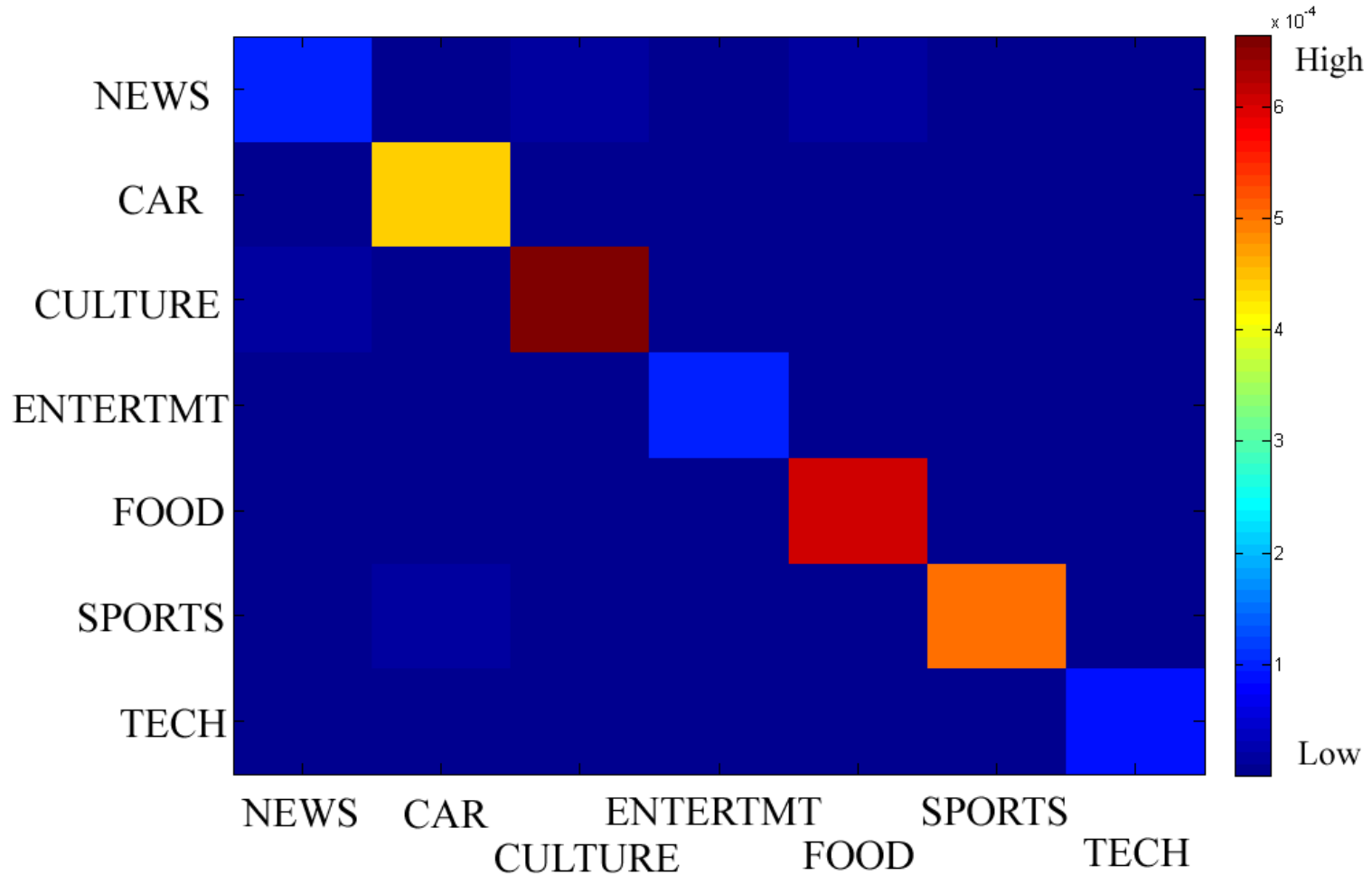


Highly assortative

The bias of citation count

Tags/Topics	Number of posts
NEWS	10276
CAR	1683
CULTURE	2231
ENTERTAINMENT	12072
FOOD	1745
SPORT	2417
TECH	10652

Citation density map



Self-citation

- In academia, we often cite our own papers for good or bad reasons

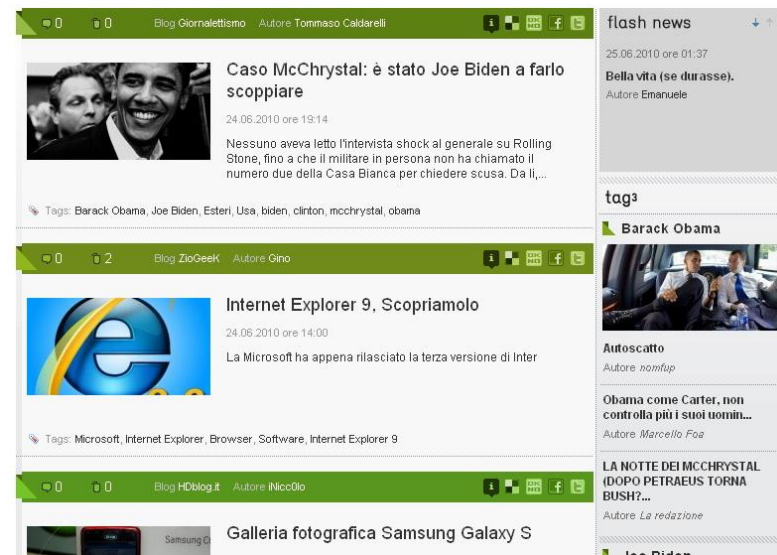
Your (real) Impact Factor

$$\text{Impact Factor (corrected)} = \frac{\begin{array}{l} \# \text{ times your work is cited} \\ - \# \text{ citations that actually trash your work} \\ - \# \text{ times you cited yourself (nice try)} \\ - \# \text{ times you were cited just to pad the introduction section} \\ - \# \text{ citations the editor pressured the author to include to increase the journal's impact factor} \end{array}}{\begin{array}{l} \# \text{ original articles you've written} \\ + \# \text{ articles you were included in out of pity or politics} \\ + \# \text{ not-so-original articles you've} \\ \quad \text{written} \\ \quad \text{copied and pasted} \end{array}}$$

of times you cited yourself (nice try)

Well, we are not alone...

- Bloggers take self-citation to a new level
- 86% of the edges/citations are between posts from the same author!
- Why?
 - Boost their rankings?
 - Focus on one or two topics?
 - ...



Future work

- More network properties
- More granularities of analysis
 - The level of connected components
 - The level of individuals: professional/amateur
- More networks
 - E.g. the author-author network
- More dynamics
- Identify influential bloggers and posts

Thank you!