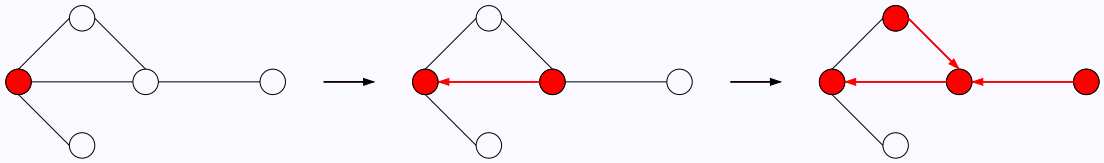
A Data-Driven Analysis to Question Epidemic Models for Citation Cascades on the Blogosphere

Abdelhamid Salah Brahim (LIAFA, Université Paris 7)
Lionel Tabourier (naXys, Université de Namur)
Bénédicte Le Grand (CRI, Université Paris 1)

ICWSM-13

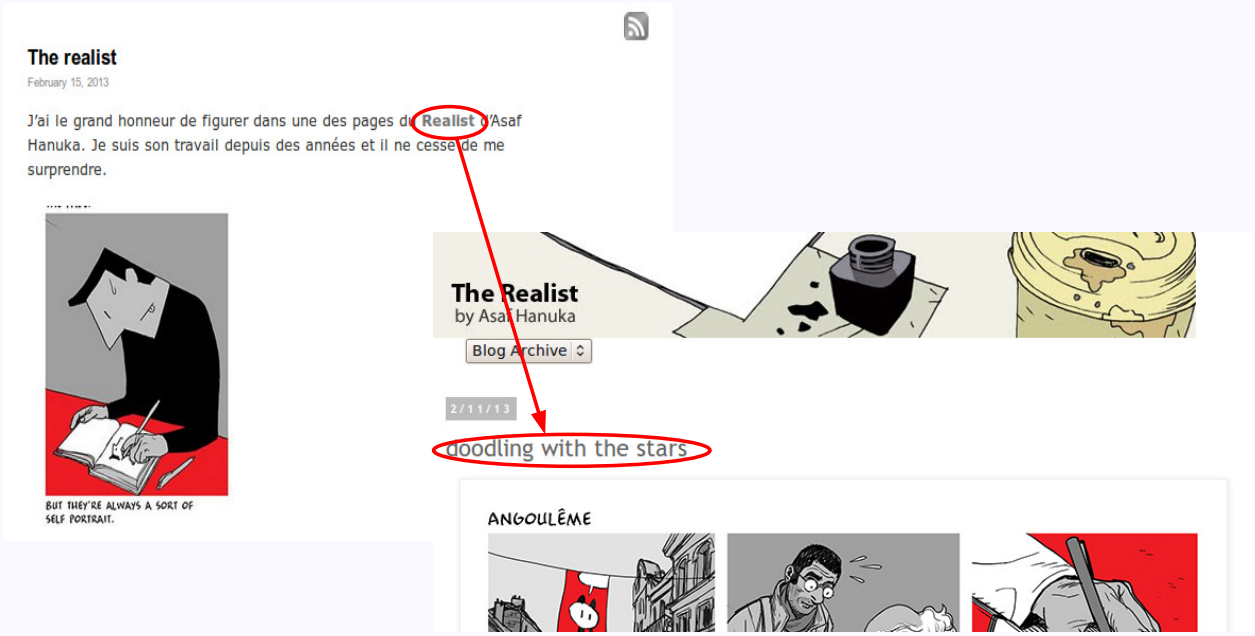
Context: epidemic-like spreading

Epidemic analogy



cascades, key-elements: **item** and **path**

Citations and cascades



Role of citations in information spreading?

Data

Raw dataset

Webfluence project

French-speaking blogosphere:

- $\sim$ 10,000 *A-list* blogs
- observed from February 1^{st} to July 1^{st} 2010
- $\sim$ 850,000 posts

Processing

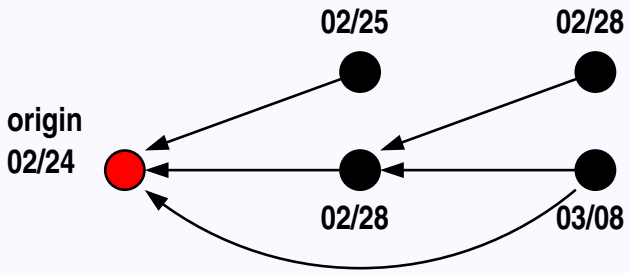
Among $\sim$ 1 million citations, selection:

- from the dataset to itself
- without self-citations

$\Rightarrow$ 3,199 blogs and 20,885 citations

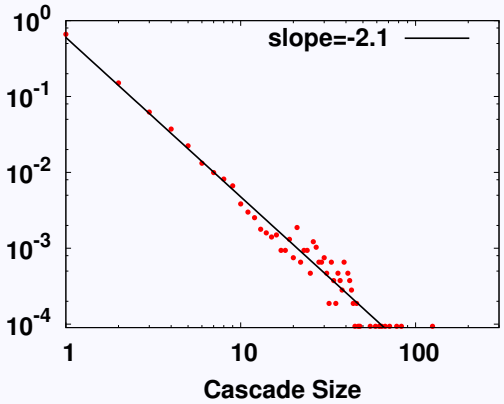
Structure of citation cascades

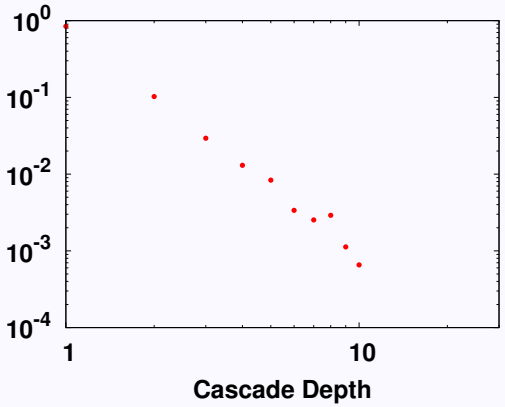
Directed Acyclic Graph (DAG)



characterized by size (number of posts) and depth (longest distance)

Basic features: comparison to literature



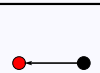
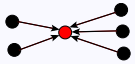


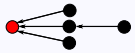
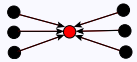
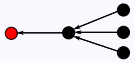
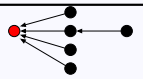
strong similarities with *Leskovec et al. (2007, 2008)*

ex: size distribution exponent 1.97 vs 2.10

Cascade frequencies

Comparing frequency rankings

Rank here	1	2	3	4	5	6	7
Rank in *L. et al.*	1	2	3	4	5	6	7
Shape							

8	9	10	11	12	12	14	15
10	11	8	9	13	14	15	12
							

Spearman coefficient : 0.95

What does it mean?

- similar information spreading mechanisms?
- ...
- consequence of trivial features (as posting activity)?

What do we need to observe such cascades?

Qualitative insights

Cascade topic definition

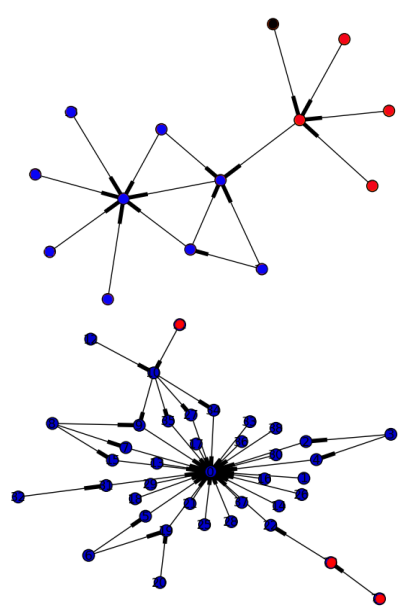
Expression or set of words such that:

- as many posts as possible address the topic
- two cascades may not have the same topic

Typical behaviors

topic mutations:

high topic unity cascades:



Frequent topic changes, particularly in chain-like patterns

Relation between structure and topic

Measuring topic-unity of a cascade

$$tu = \frac{|\mathcal{C}^*|}{|\mathcal{C}|}$$

$\mathcal{C}$: posts of the cascade

$\mathcal{C}^*$: posts of the cascade dealing with cascade topic

Measuring chain- or star-like cascades

$$sc = \frac{\sum_{x \in \mathcal{C}: n_i(x)=0} n_o(x) - 1}{\sum_{x \in \mathcal{C}} n_o(x) - 1}$$

$n_i(x)$ and $n_o(x)$: number of incoming and outgoing arcs from x

$$sc(chain) = 0,\ sc(star) = 1$$

Correlation tu-sc : 0.57 Spearman coefficient

(on a sample of 50 "large size" cascades)

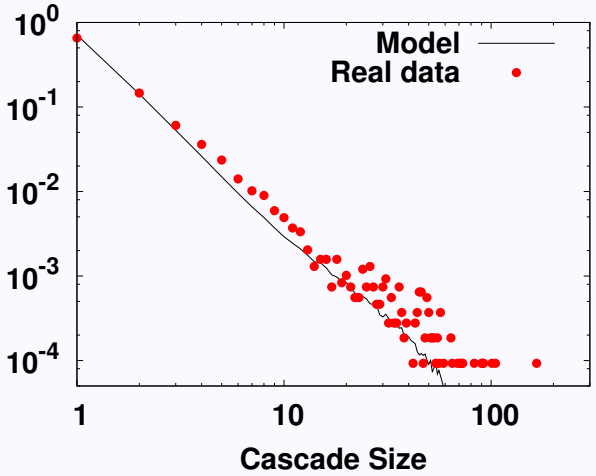
Model with uncorrelated citations

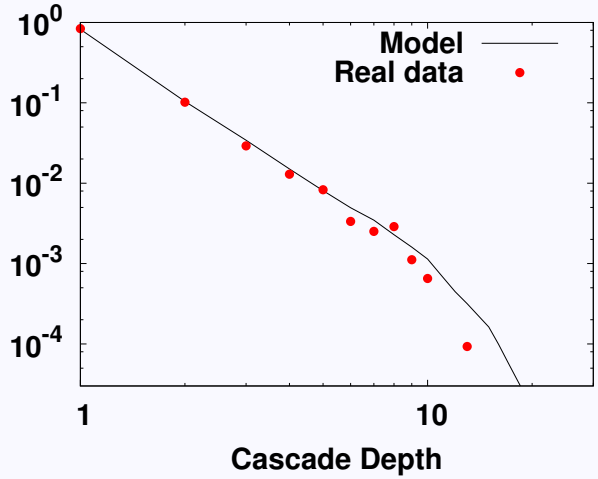
Ingredients

- same posting and citing activity of blog A
- but citing any post of blog B
- keep the same global distribution of latencies

No explicit reference to the content of posts

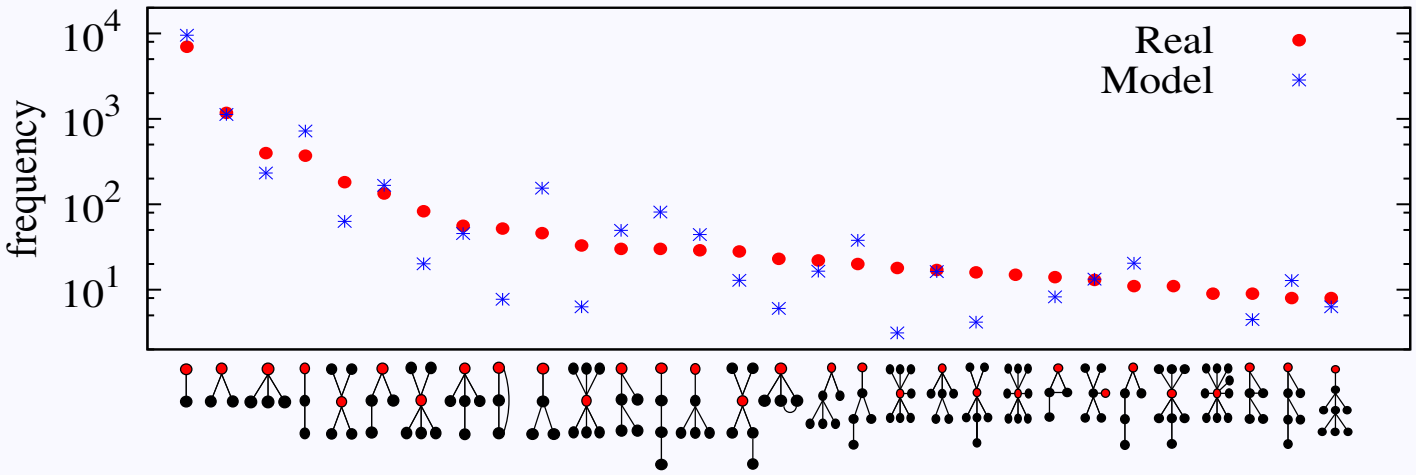
Results on basic features





Model with uncorrelated citations

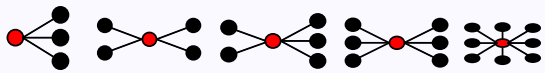
Results on cascades ranking



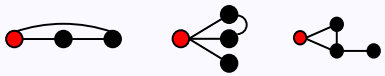
Spearman coefficient : 0.76

Underestimated cascades:

star-like cascades:



moral cycles cascades:



Conclusion

Questioning the epidemic analogy

- who cites whom and when mostly governs cascade shapes
- no clear notion of item spreading through citations

Clues to look for information spreading?

- specific patterns

  ex: star-like and cyclic cascades

- cross-check with other information sources

  ex: text-mining tools, multimedia contents

- constrained social networks

  ex: microblogging platforms, games

Thank you for your attention