

A data-driven analysis to question epidemic models for citation cascades on the blogosphere

Abdelhamid Salah Brahim
LIAFA, Université Paris 7 Denis Diderot
salah@liafa.jussieu.fr

Lionel Tabourier
naXys, Université de Namur
lionel.tabourier@fundp.ac.be

Bénédicte Le Grand
CRI, Université Paris 1 Panthéon-Sorbonne
benedicte.le-grand@univ-paris1.fr

June 3, 2013

Abstract

Citation cascades in blog networks are often considered as traces of information spreading on this social medium. In this work, we question this point of view using both a structural and semantic analysis of five months activity of the most representative blogs of the french-speaking community. Statistical measures reveal that our dataset shares many features with those that can be found in the literature, suggesting the existence of an identical underlying process. However, a closer analysis of the post content indicates that the popular epidemic-like descriptions of cascades are misleading in this context. A basic model, taking only into account the behavior of bloggers and their restricted social network, accounts for several important statistical features of the data. These arguments support the idea that citations primary goal may not be information spreading on the blogosphere.

Keywords: blog network, citation cascades, information spreading, statistical analysis, epidemic models

1 Introduction

During the last decade, the *World Wide Web* functions and usages have been widely impacted by the popularization of user-friendly content editors, most noticeably wikis and blogging platforms. Blogs in particular emerged as a public space for opinion broadcasting as well as a form of participative journalism or even a shop-window for commercial activities.

This new kind of media has focused much attention from the scientific community. In addition to the novelty of the phenomenon, the blogosphere is a huge dataset of rich and publicly available content, and thus comes across as a means to understand the principles of information spreading on social networks. The seemingly similar mechanisms at stake in epidemiology and information adoption made the tools of the former a popular source of inspiration. Both measurements and models have been developed in this line, often describing the information transmission as an infection-like process.

In this article we question the legitimacy of these approaches as a proper trace of information spreading, in the specific context of bloggers citing each other. Following a typical complex networks approach, we use statistical tools to describe large dynamical datasets, and observe some similarities with an unrelated dataset of the literature. We put forward the idea that usual measures seize mostly effects which stem from individual posting and citing behavior of bloggers. This assumption is given more credit by a content-oriented analysis, which helps to understand the origin of the properties captured by these statistical tools. Finally, we provide a content-independent model to further support this hypothesis; it also yields clues on where an observer should look for traces of information spreading in this dataset.

1.1 Related work

The rise of blogs among online social media has been discussed from a sociological point of view in numerous papers, e.g., [16] give some insights about bloggers demographics and cultural behaviors; in [10], the author describes their influence on society. Given the size and richness of the blog datasets, automatic classification and text-mining tools have been widely used to study the dynamics of trends and opinions in the blogosphere [12, 2, 15, 13, 21]. For example, some studies concentrate on the political blogosphere to understand the ties between political parties, in particular the way information spreads from a group to another [1, 9]. The question of trends is closely related to the definition of authority, influence and trust in social networks [10, 9]. As such, these text-mining tools are also used for various practical purposes as online advertising [27] or search engines, ranking blogs according to their spreading ability [3].

A large amount of works has been dedicated to discovering routes of information spreading in online social media such as products recommendation platforms [19], online games [6], Flickr [8], Twitter [25, 23], and of course the blogosphere. These routes are often inferred from the structure of the social network and the dynamical behavior of the agents (adopting a new technology, joining a group), especially when there is no concrete trace of the peer influence, e.g., [2, 8]. However, the observed contagion may be the result of direct adoption from one's neighbor, but also a consequence of homophily, as connected people are prone to behave similarly [5, 7], making route inference methods questionable.

In this work, we use data in which the connection between users is explicit and combine it with a popular cascade-like description of the spreading [2, 18, 22, 6, 11, 20, 24]. More precisely, we adopt definitions very similar to the ones developed in [18], and will therefore often refer to this study for comparison purposes. In addition to the statistical analyses of these datasets, models have been proposed to explain the observed features [12, 15, 18, 11]. Among them, many are inspired by virus spreading

models in epidemiology; in the following, we address the issue of the relevance of this particular class of models in the context of blog networks.

1.2 Outline

Section 2 is dedicated to the exposition of the studied blog dataset. After describing how the data have been collected, we achieve standard statistical measurements with frequent comparisons to similar measures in [18]. We then investigate the contents of a subset of posts both qualitatively and quantitatively, in regards to spreading processes phenomena. In Section 4, we propose a simple model of the data to mimic the statistical features aforementioned. It brings us to challenge usual depictions of information propagation in the blogosphere, and more generally on online social media.

2 Dataset

2.1 Description and filtering

The data corpus analyzed in this paper has been obtained by daily crawls of 10,309 blogs during five months (from February 1st to July 1st 2010), yielding 848,026 posts. These blogs have been selected according to their popularity, activity and their representativeness of the French-speaking blogosphere. It therefore excludes blogs which, even if public, aim at informing a small group of friends (such as teenage blogs of the `skyrock.com` platform, popular in France). These *A-list* blogs have been crawled by Linkfluence, a company specialized in online opinion watch (`linkfluence.net`) during a research project called *Webfluence*. In the following, it will thus be referred to as *Webfluence* dataset.

We focus here on *citation links*: consider a post P_a from blog A and a post P_b from blog B. If P_a contains the URL of P_b , then there is a citation link from P_a to P_b . Notice that other kinds of interaction (e.g. comments, blogrolls) are not taken into account here. We collected 1,079,195 citation links in this dataset, but after filtering out citations pointing to posts outside the dataset, citations from a blog to itself¹ and crawling artifacts (e.g. citation from an anterior post or different indexations of a same post), this came down to 3,199 blogs publishing 461,134 posts² and 20,885 citation links, for which both source and destination are in the crawling period.

2.2 Information items and routes

Various hypotheses have been made to track information circulating on the blogosphere. Most of them rely on the assumption that diffusion can be described as a piece of information moving on a network connecting blogs. For the sake of clarity, we will call this (hypothetical) element of information *item* and a *route* is the path followed by such an item.

¹The goal of this paper is to study diffusion phenomena and such citations do not contribute to the spreading process.

²Among them 24,938 are involved as source or destination in citation links contained in the crawling period.

In [3, 2, 9], the authors chose to consider the reference to a particular resource (picture, URL, ...) as an item. Such a definition is unambiguous, but there is in general no guarantee about where a member of the cascade has found the information, in other words the route is unknown. The authors of [3, 2] looked for subtle criteria to learn the routes from past behaviors, however their studies suppose that routes remain in the blog network, as opposed to information brought by other media. In [9], this point of view is refined by only taking into account resources associated to the citation of the source post, but it reduces drastically the set of usable citations in the dataset.

Here, we explore the approach developed in [11]. The authors consider that a citation from a post to another is a possible path for information spreading. We investigate the definition of an item propagating through such a route, i.e. we are looking for a piece of information that would remain unchanged when a blog is citing another. As we will show in this paper through our observations, in general it is not possible to define such an item.

2.3 Cascades of citations

Describing a route as a *cascade* is now a quite usual picture of information spreading on social networks [2, 18, 11, 20, 24]. In our study, a cascade is a subgraph of the post network where nodes correspond to posts and edges to citation links. It is built in the following way: its first node is a post with no outgoing link (the *origin*), posts citing it are included in the subgraph with the corresponding directed citation link, then we look for posts citing these posts in the dataset, and the process carries on recursively. Because of the temporal ordering of citations, such a definition implies that cascades are directed acyclic graphs (DAG); an example of cascade is given in Fig 1. According to this definition, we detect 10, 667 non-trivial cascades³ in the *Webfluence* dataset.

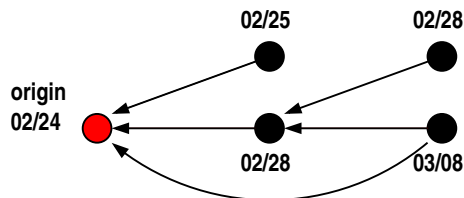


Figure 1: Example of cascade. Here, a node is a post and an arc is a citation link. The subgraph obtained is a DAG.

In a former paper, the structural properties of these cascades have been discussed relatively to a predefined community hierarchy [4]. We will now investigate in more details the cascades with regard to their content and to the way information may move from a post to another. It is worth noticing too that this definition is identical to the one in [18] and [22], allowing us to compare the features of other datasets to ours. Another possible convention consists in merging all cascades which feature the same post, see for example [20].

³i.e., not isolated posts

2.4 Comparison to existing results

We compare standard measurements on the *Webfluence* dataset to the features observed in [18], where the authors analyzed a 90-days crawl of about 45,000 active blogs connected by 205,000 citations in 2005 — thus larger than ours. So, considering the period and the fact that the French-speaking community is only a small part of the world-wide blogosphere, these datasets are unrelated. Yet, both datasets have been processed in a similar way in order to identify general properties of the blog networks.

Citation features

First, we plot on Fig. 2 some features characterizing our dataset’s activity and citation dynamics during the crawling period.

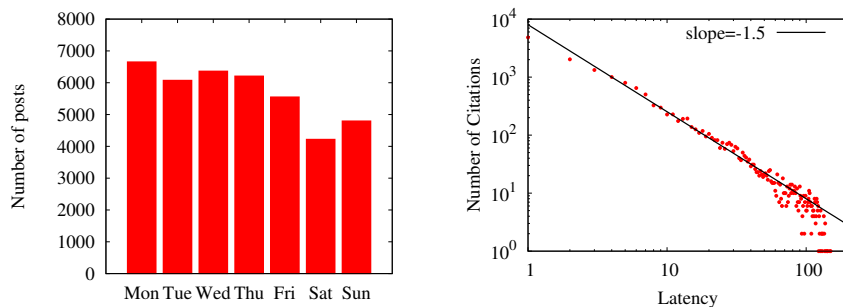


Figure 2: Activity and citation dynamics statistics. Left: average number of posts published per day of the week. Right: distribution of latencies (temporal gaps) between a post and its citations.

The number of posts published as a function of the day of the week exhibits a usual behavior: bloggers publish less during the week-end, reducing their activity by about 27% when compared to another day of the week (around 40% in [18]).

The distribution of latencies between a post and its citations looks roughly like a power-law with a cut-off. The cut-off is known to be an effect of the finite time-window, as observed for example in [29] in the context of email networks; the slope of the exponent is close to -1.5 (-1.6 in [18]). Notice that this shape is modulated by a week-cycle: for example it is slightly more likely than expected by the model to be cited within 7 days, and slightly less to be cited within 8 days.

Then, we report on Fig. 3 the distributions of in-coming and out-going links for each blog, as well as the correlation between them. Both qualitative and quantitative behaviors are quite similar to the ones in [18]: degree distributions are heavy-tailed, and may be approximated by power-laws — even if a 1 to 100 range is not sufficient to be affirmative about the quality of such a model. The exponents measured are quantitatively close too, as summarized in Table 2; however, the slope is not as steep as other values reported in the literature, which is not surprising, since the steepening is known

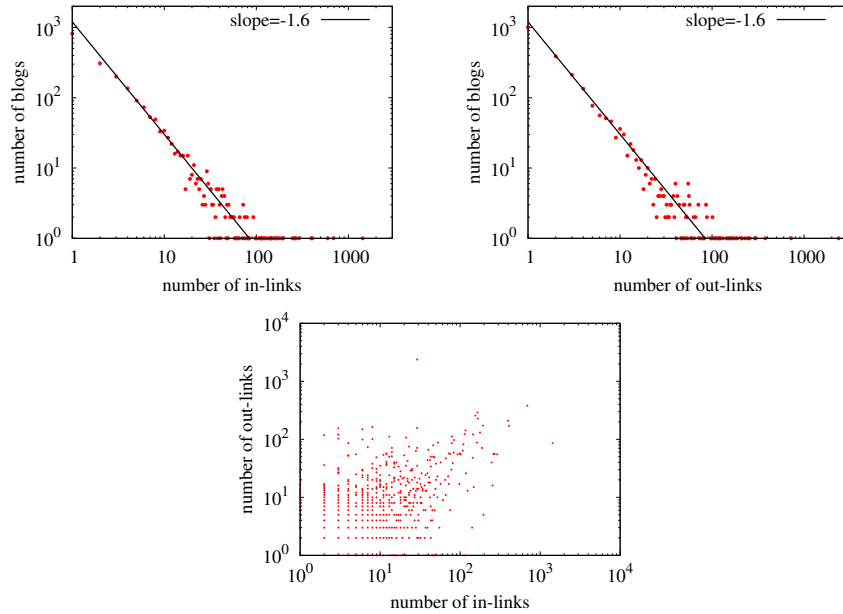


Figure 3: Blog network statistics. Top left: distribution of incoming links per blog. Top right: distribution of outgoing links per blog. Bottom: number of in-links as a function of the number of out-links.

to vary with the time-window of measurement [26]. The Pearson correlation coefficient computed from these distributions is about 0.18 (0.16 in [18]). If we consider that the number of citations is an appropriate measure of attention on the web, the value supports the idea that attention and activity are not much correlated in the blogosphere.

Cascade features

We plot the distribution of cascade sizes — the size is the number of posts in the cascade, origin excluded. It may be fitted by a power-law model with a slope close to the one in [18]. We considered the cascade depth too, defined here as the longest distance⁴ between any node of the cascade to its origin (Figure 4).

We then investigate cascades shapes by ranking them in a decreasing order according to their frequency in the dataset. Strikingly enough, we can see in Table 1 that the ranking of the most common shapes is very close to the hierarchy detailed in [17] — it should be interpreted cautiously with regard to the low number of patterns measured. Note that at equal size, star-shaped cascades (like the ones ranked 2, 3 or 5) are more frequent than chain-shaped (4 and 12). It is usual to compare the amount of patterns to

⁴The distance between two nodes is defined as the number of arcs of the shortest directed path from a node to the other

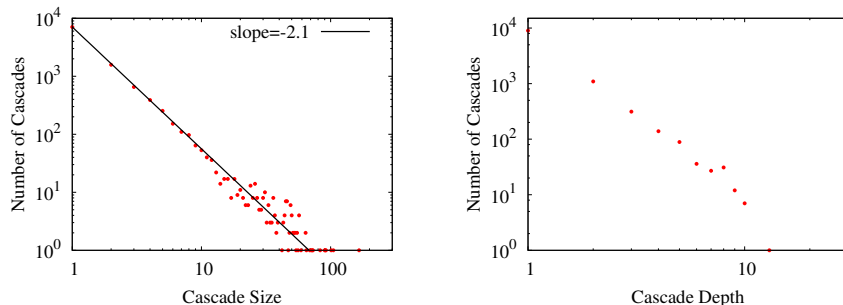


Figure 4: Left: distribution of cascade sizes. Right: distribution of cascade depths.

a suitable random model to evaluate which ones are over- and under-represented; this point will be developed in Section 4.

Comparative summary

In Table 2, we summarize the results reported above and compare them to the values in [18]. It enlightens striking similarities between the datasets despite the differences in origins and sizes. These observations bring us to think that the underlying process of cascade building may be similar. However, this is not necessarily evidence of the social mechanism of information spreading in this media: this may be the mere product of statistical properties of the network itself, an issue that we intend to address in the following.

3 First insights from a content-based analysis of large cascades

For an *item* to exist, it would be expected that the posts of a cascade deal with the same topic; however, identifying the topic of a cascade is a hard task. A standard method calls to the use of text mining and natural language processing techniques as well as strict statistical criteria such as the ones developed in [12, 2]; however the experimenter will still be required to make assumptions on parameters. Besides, it has been underlined that blog contents are difficult to process with such techniques, e.g., [13]. Even if our dataset is cleaned out from spam blogs and advertisements, we have to cope with semantic data with few structural conventions, as well as complex content, like lingo or hidden meaning.

Here, we want to get an overview of what different posts in a same cascade deal with. A manual investigation provides us with the possibility to do so, in a more controllable way than what automatic tools allow. We randomly chose 50 cascades among the top 5% larger cascades (in terms of number of posts), that is to say 50 of the 526 cascades involving 9 nodes or more. We draw benefit from this relatively low amount

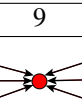
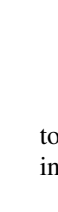
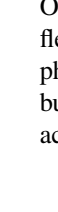


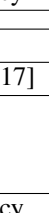
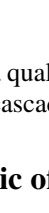
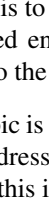
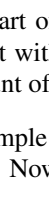


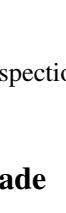
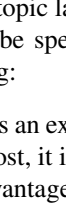


Rank	1	2	3	4	5
Rank in [17]	1	2	3	4	5
Shape					
Frequency	6992	1173	397	370	182
Rank	6	7	8	9	10
Rank in [17]	6	7	10	11	8
Shape					
Frequency	134	83	56	52	46
Rank	11	12	12	14	15
Rank in [17]	9	13	14	15	12
Shape					
Frequency	33	30	30	29	28

Table 1: Cascade shapes ranked by frequency.

to achieve a qualitative inspection and get some insights on the real processes involved in citation cascades.

3.1 Topic of a cascade

Our goal here is to perform a manual inspection of large cascades to define topics in a flexible way, which will give us insights to take our analysis one step further. The basic philosophy is to define a topic large enough to be shared by as many posts as possible, but restricted enough to be specific to the cascade. The so-called topic is identified according to the following:

- A topic is defined as an expression or set of expressions which can be considered as addressed in a post, it is not necessary that the expression itself appears in the text (this is one advantage of the manual investigation).
- From a practical point of view, after reading the posts, we establish a list of expressions that may be the most widely spread in the cascade, and which are not part of the set of existing topics. Then for each post, we decide if it deals or not with each expression; the topic is the one which is shared by the largest amount of posts.

For example *Nicolas Sarkozy* is a very usual topic of the French blogosphere during this period. Now, if we demand that the posts refer to *Nicolas Sarkozy & regional*

Dataset	Duration	B	N	L	r	α	β	τ	γ
<i>Webfluence</i>	151 days	3,199	461,134	20,885	0.18	-1.60	-1.60	-1.50	-2.10
in [18]	90 days	44,362	2,422,704	245,404	0.16	-1.70	-	-1.60	-1.97

Table 2: Comparative results between various datasets. B : number of blogs, N : number of posts, L : number of citations, r : Pearson correlation coefficient between number of in- and out-links of nodes. Fitting with power-law models, we report the following exponents: α : blog in-links distribution, β : blog out-links distribution, τ : latencies distribution, γ : cascade sizes distribution.

*elections*⁵, we narrow the topic down, but still several cascades may be defined by this expression. But there is only one large cascade defined by *Nicolas Sarkozy’s comments on the regional elections results*. Other examples of topics include: *Death penalty in the US: Hans Skinner case*, *Novelties revealed during Facebook conference* or *The Geneva international motor show*.

Obviously, some choices remain arbitrary, for example it may be tricky to settle whether a post actually deals with a topic, and other choices lead to slightly different estimations. However, we tested a strict versus a flexible convention on the sample, the former leading to lower topic-unity (see Sec. 3.2) evaluations than the latter, which is implemented in the following; this choice does not significantly affect our conclusions.

3.2 Topic mutations along cascades

The manual analysis presented in Section 3.1 strengthens our daily experience intuition that the topic of a citing post may be quite different from the cited post’s. We develop this idea more quantitatively in the following by means of measures connecting structural features to the post content.

Star-shaped and chain-shaped cascades

In this section we propose a metric, denoted sc , to seize to what extent a given DAG belongs to the star vs chain topology. Let $n_o(x)$ and $n_i(x)$ be the number of arcs going respectively from and to the node x of the cascade \mathcal{C} , then:

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

Long chains imply that nodes with outlinks also have inlinks, so that sc is smaller when compared to cascades with the same amount of arcs, but with shorter chains. This measure has been chosen to be normalized, so that chain-like patterns have $sc = 0$ and star-like $sc = 1$, as summarized on Figure 5. Several other definitions may be suggested in these lines, however this one gives a good grasp of the feature we want to account for in the context of cascades.

⁵In France, a *région* is the largest administrative subnational division, the council ruling them is elected every 6 years and the latest elections took place in March 2010, i.e., during the crawling period.

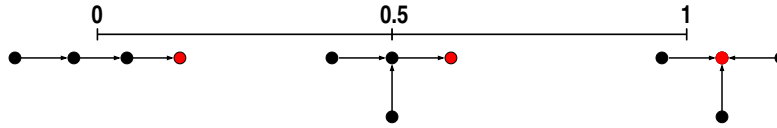


Figure 5: Behavior of sc coefficient for examples of 4-node cascades.

Topic-unity vs. sc coefficient

Topic-unity of a cascade is defined as the ratio of nodes in the DAG dealing with the topic attributed in accordance with the protocol of 3.1⁶. Here again, other definitions may be suggested (e.g. taking into account the fact that there may be more than one dominating topic in a cascade), however this estimate gives a first insight on the content of a cascade.

We ranked the sample of cascades by increasing order, using the sc coefficient on the one hand (R_1), and the topic-unity on the other hand (R_2)⁷. Both rankings are correlated: the Pearson correlation coefficient $r(R_1, R_2) = 0.57$, meaning that star-shaped cascades are more likely to exhibit a largely shared topic than chain-shaped ones. In other words, the topic of a cascade is likely to change along the citations. As the existence of an information item supposes that the posts deal with a unique topic, this analysis challenges the relevance of the notion of item in this context. Notice that this observation is sociologically consistent: as the sustainability of a blog partly stems from the originality of its content, bloggers have a strong incentive not to copy-paste information that they found elsewhere, but rather to add their “personal touch”.

A glimpse of typical user behaviors

To give the reader a more comprehensive grasp on the cascade content, we gather here some qualitative remarks with practical examples.

- Our manual investigation reveals that the blogosphere is flooded with information from external sources; in particular, references to mass media are extremely usual⁸ — especially, but not only, in the political blogosphere.
- Cascades with high topic-unity (close to 1) often deal with the blogosphere itself, for example the set-up of a meeting of the far-left blogger community, or the reactions to a libel action against a blogger. Among most noticeable examples, organizing events such as games among bloggers is a usual practice, generating star-like shaped cascades (see Fig. 6). In this case, a picture or a motto is often used to refer to the event — which can be seen as an item spreading.
- It can be observed that topic changes along a cascade are mostly brutal. For example, the cascade *Outcome of the ‘No-Sarkozy Day’* contains a secondary topic:

⁶Some posts were not online anymore when we checked their content; they are not taken into account for the topic-unity evaluation.

⁷We use rankings of sc and topic-unity values as they are not homogeneously distributed on $[0; 1]$.

⁸In [23], the authors attempted to take this phenomenon into consideration in the context of Twitter.

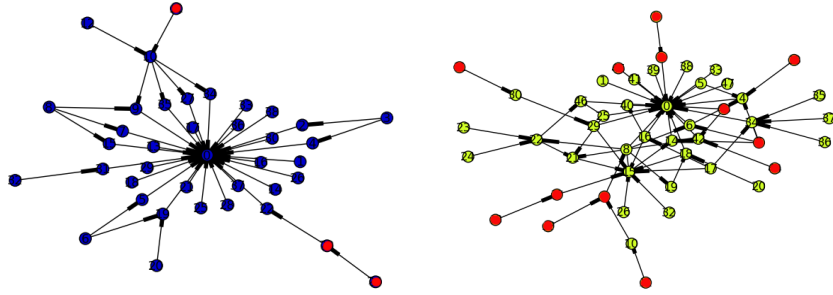


Figure 6: Examples of high topic-unity cascades: red nodes do not deal with the topic of the cascade, or their content is no longer available. Left: *Cooking game: an olive-oil based sweet recipe*. Right: *About the law challenging anonymity on the Internet*.

3 Suisses sexist advertisement, as can be seen on Figure 7 and the joint node reveals where and how the topic mutates. In this case, the citing blogger refers to the cited one as a representative of female-blogger community but without any relation to the content of her cited post, thus feeding the idea that some citations do not aim at spreading content but at acknowledging another blogger's status.

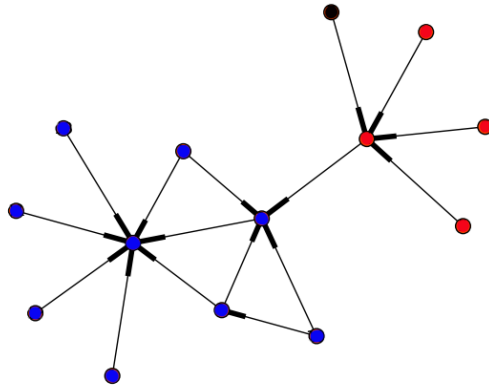


Figure 7: Example of mutating topic. Blue nodes topic: *Outcome of the 'No-Sarkozy Day'*; red nodes topic: *3 Suisses sexist advertisement*; black node: other.

- The topic-unity of cascades being questioned, it is not surprising to observe that the origin of a cascade, which is unique according to our definition, only rarely defines its topic. The ability of a node to trigger a discussion, e.g., [24] should thus be adapted to a formalism where a cascade size might be an inadequate measure of the intensity of a discussion.

- Depending on the domain of interest, usual writing behaviors are clearly different. For example, in the political blogosphere, the average post is quite long and often deals with several topics, it implies a higher topic mutation probability than posts dealing with individual hobbies, as cooking or knitting.
- In the same line, bloggers may have very different behaviors regarding their use of citations. Some give an overview of their community in the form of a heterogeneous review, other rather cite with regard to the very specific topic they are dealing with⁹. In [22], the authors measured that typical cascade type strongly depends on the community it belongs to; this observation and ours both suggest that the explanation is rooted in the typical citing habits of bloggers in a community.
- Citations are frequently associated to comments on other bloggers, which give evidence of tight relationships : regular readers, sometimes off-line friends etc. We could therefore think of detecting communities using cascade analysis. For example, it seems likely that a group of bloggers participating to several cascades of unrelated topics may be part of the same social group on the web.

4 Modeling cascades without items

Models describing citation data often call to an epidemic-like description of spreading [2, 11]. Hence, they aim at reconstructing some features of the observed citation structure using a detailed description of the behaviors of users in the network. While it may happen that a piece of information is duplicated without any change from a blogger to another, a closer examination of the content show that this behavior is not dominant. The analysis of Section 3 makes us think that different citations within the same cascade may not be strongly correlated. In this section, we propose a simple model consistent with this observation, that is to say independent from items, and show that it is sufficient to reconstruct most features of the cascades.

4.1 Model description and results

The central idea of this model is that citations may be considered without any reference to any item spreading on the network, so that citations are treated as much as possible as uncorrelated events. We thus focus on the structure of the underlying network and on the citing activity of bloggers to account for the citation cascades observed.

We built a very simple model according to the following description:

- If a post P_a of blog a refers to a post P_b (of b), we now consider that P_a cites *any* post already published by b .
- The post cited is selected randomly, but with a probability bias that enables to fit the real latency distribution. In more details, the process consists in choosing

⁹For examples illustrating these various behaviors, see respectively: www.geeek.org and falconhill.blogspot.fr.

a post randomly, and possibly discarding it according to a power-law function of the latency; the parameters of the power-law are set to fit the distribution on Fig. 2.

Without this second prescription, the distribution of latencies is much alike the one obtained for a poissonian process, exhibiting in average longer lapses of time between citations. The number of posts cited being roughly 2% of the total number of posts published in the dataset, this randomization process is supposed to break efficiently correlations between the events of a cascade.

We measure statistical features of the data generated and compare them to the original dataset (see Figure 8). First, we observe that sizes and depths follow heterogeneous distributions, and more precisely a power-law model seems appropriate to fit the size distribution. Both the size and depth distributions of the model are close to the real one, except for slightly smaller cascades (fits give a 2.3 slope for the model, versus 2.1 for the real data). So in spite of its great simplicity, this model is able to mimic these important features of the original dataset.

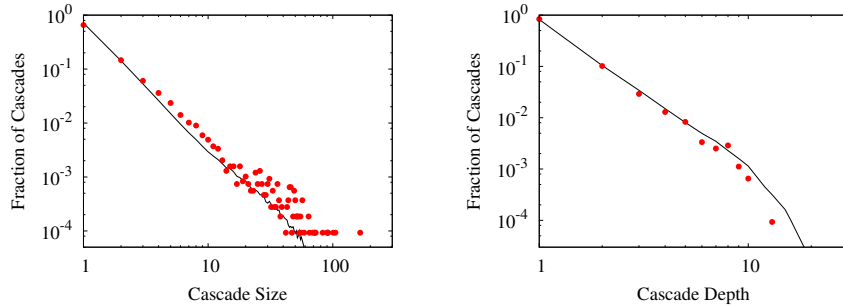




Figure 8: Comparison between real data (red circles) and an average of 100 realizations of the model (black line). Left: probability distribution of cascade sizes. Right: probability distribution of cascade depths.

We then compare in more details the patterns found in our model to real data. Figure 9 reveals that the number of patterns produced by the model is roughly of the same order of magnitude as in the original dataset. However, rankings are not identical, and there are significant differences between the real data and our model. Two types of cascades are clearly underestimated by the model (in some cases even nearly absent): the first exhibits several outgoing links from a node to other nodes of the cascade, and consequently displays cyclic patterns: . The second type represents star-like cascades: .

More generally, it can be observed on Figure 9 that most underestimated cascades have high sc values (typically $sc \in [\frac{2}{3}; 1]$) while overestimated cascades have lower sc .

As discussed in Section 3, star-shaped cascades may exhibit high topic-unity, while the random cascade generator is blind to such semantic effects. We may then assume

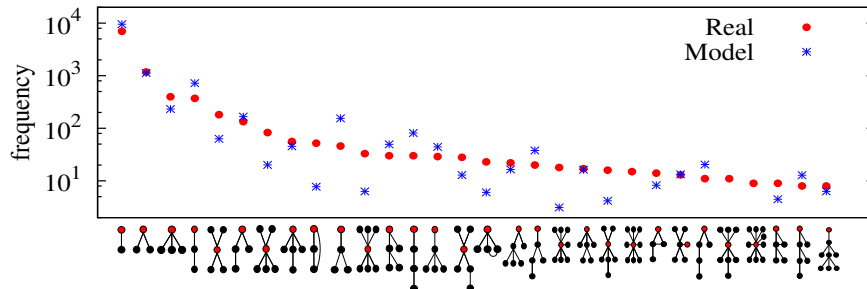


Figure 9: Frequency of most usual cascades. The model is averaged over 100 realizations, values absent from the picture are below the minimum of the frequency range (=2).

that a fraction of cyclic and star-like cascades result from correlated citations, and therefore cannot be rebuilt by the model. In other words, these patterns are salient characteristics of the real dataset. It can be related to other works [30, 28], in which the authors look for evidence of spreading phenomena on other types of large communication networks by comparing the number of dynamical patterns in real data to various dynamical benchmarks.

4.2 Discussion

The model described above gives clues that some citation patterns are more likely than others to carry information spreading — namely, cyclic and star-shaped cascades. Even in this case, it is more accurate to talk about correlated citations rather than information spreading; this assumption is supported by the former semantic analysis. As the model merely takes into account who cites who and when, we think that the structure of the cascades observed is mostly due to the structure of the ego network of bloggers when it comes to citation. From this perspective, citations often seem loosely related to the intent of spreading an information. A possible explanation is that a substantial fraction of citations may rather be a mean to acknowledge an acquainted blogger.

The model presented above is very simplistic, and we do not claim that it fully captures the mechanisms of citing behaviors on the blogosphere. Yet, the fact that such a model rebuilds some important features of cascades implies that we should be very cautious when using any other model in this context. More precisely, distributions of citation cascades sizes and depths are clearly not reliable traces of diffusion phenomena; the number of cascades per type may be more relevant in this context, but richer metrics are needed to prove information spreading. As we can hardly infer more information from a purely structural analysis, it calls for the use of other metrics, like content or detailed temporal information. These observations deter from using elaborate models describing complex agent behaviors: we would not be able to know if the outcome stems from the assumptions, or only from the lack of relevant measure to set-

tle if the synthetic data resembles the real one. It raises tricky questions: what is a good benchmark to compare real data with? And consequently, what is an *expected property* in the context of blog citation cascades?

5 Conclusion

In this paper, we tackled the problem of the origin of citation cascades in blog networks. Measurements indicate that our dataset shares many statistical properties with the one in [18], while focusing on an unrelated subpart of the blogosphere. We then put forward the assumption that the underlying process may be identical in both cases. As epidemic-like models are popular in this area, we explored the contents to see if we can identify items, i.e., propagating piece of information, throughout a cascade. But items are difficult to define in this specific context: we made the minimal choice of isolating a common topic of discussion, and even in this case we observed that topics often change along a cascade. A simple model based on independent citations, turns out to mimic well some features of the real dataset, namely the size and depth distributions.

It brought us to reconsider the results obtained with models in this context: if such a basic model rebuilds these observations, these features do not reveal information spreading. According to an epidemic-like description of information propagation, it can be compared to a virus spreading among a population while mutating. However, here we have no simple equivalent of the genetic distance, and even if we did, we should probably consider very high rates of mutation. Our belief is that diffusion in the sense of an epidemic spreading is not appropriate to model what happens on this network. In fact, we questioned the very idea that citations primary function is information spreading.

Another objection to an epidemic-like description comes from the fact that it relies on the assumption that the network is the only medium of propagation. But even if bloggers are known to be very active consumers of online content [16], there is no guarantee that the information spreading comes from the blogosphere itself. Content and structure analyses reveal that the blogosphere is flooded with information from external sources: the flow generated by mass media, other social networks, and off-line experience are presumably more influential on the posting activity than other blogs. It is consistent with the sociological literature, which describes information adoption as a complex interplay between mass media and local opinion dynamics [14]. Furthermore, recent works acknowledge the existence of external sources of information in the context of online social media, observing “jumps” of information and trying to both measure and model this phenomenon [23]. In that sense, the epidemiological metaphor may be misleading when transposed to social media.

Contagion models are attractive as they propose a simple, versatile way to describe subtle processes; besides, it is a fact that word of mouth spreading of information on social media does sometimes happen, but that does not make it a measurable and/or dominating means of information spreading. Such description has little large-scale empirical support, as tracking spreading in social media is a very challenging task. Our study deals with specific patterns of a particular family of datasets, other fields may be more suited to epidemic tools. We pointed out a few ingredients that would contribute

to more reliable analyses: a strict definition of the item spreading in the system, and a network where sources of information are identified and controllable. Efforts should also focus on defining an adequate set of measurements to locate information spreading traces in blog data, and more broadly in social datasets. Cyclic and star-like patterns are unexpectedly numerous with regard to the benchmark that we proposed. It suggests that they play a specific role in the *Webfluence* dataset and could be good candidates to fingerprint correlated events in various contexts.

Acknowledgements

We would like to thank Christian Borghesi and Renaud Lambiotte for their useful comments.

This work has been partially supported by the FNRS and the City of Paris *Emergence* program through the *DiRe* project.

The data has been collected as part of the French National Agency of Research *Webfluence* project #ANR-08-SYSC-009.

This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors.

References

- [1] L.A. Adamic and N. Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, pages 36–43. ACM, 2005.
- [2] E. Adar and L.A. Adamic. Tracking information epidemics in blogspace. In *Proceedings of the 2005 International Conference on Web Intelligence*, pages 207–214. IEEE, 2005.
- [3] E. Adar, L. Zhang, L.A. Adamic, and R.M. Lukose. Implicit structure and the dynamics of blogspace. In *WWW 2004 Workshop on the Weblogging Ecosystem*, 2004.
- [4] Anon. Anonymized for review process.
- [5] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.
- [6] E. Bakshy, B. Karrer, and L.A. Adamic. Social influence and the diffusion of user-created content. In *Proceedings of the 10th Conference on Electronic Commerce*, pages 325–334. ACM, 2009.

- [7] E. Bakshy, I. Rosenn, C. Marlow, and L.A. Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web*, pages 519–528. ACM, 2012.
- [8] M. Cha, A. Mislove, and K.P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World Wide Web*, pages 721–730. ACM, 2009.
- [9] J.P. Cointet and C. Roth. Socio-semantic dynamics in a blog network. In *Proceedings of the 2009 International Conference on Computational Science and Engineering*, volume 4, pages 114–121. IEEE, 2009.
- [10] K.E. Gill. How can we measure the influence of the blogosphere. In *WWW 2004 Workshop on the Weblogging Ecosystem*, 2004.
- [11] M. Götz, J. Leskovec, M. McGlohon, and C. Faloutsos. Modeling blog dynamics. In *International Conference on Weblogs and Social Media*, 2009.
- [12] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM, 2004.
- [13] A. Joshi, T. Finin, A. Java, A. Kale, and P. Kolari. Web (2.0) mining: Analyzing social media. In *Proceedings of the NSF Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation*, 2007.
- [14] E. Katz. The two-step flow of communication: An up-to-date report on an hypothesis. *Public Opinion Quarterly*, 21(1):61–78, 1957.
- [15] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. volume 8, pages 159–178. ACM, 2005.
- [16] A. Lenhart and S. Fox. Bloggers: A portrait of the internet’s new storytellers’. In *Pew Internet and American Life Project*, 2006.
- [17] J. Leskovec. *Dynamics of Large Networks*. PhD thesis, Carnegie Mellon University, 2008.
- [18] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. *Arxiv preprint arXiv:0704.2803*, 2007.
- [19] J. Leskovec, A. Singh, and J. Kleinberg. Patterns of influence in a recommendation network. *Advances in Knowledge Discovery and Data Mining*, pages 380–389, 2006.
- [20] H. Li, S.S. Bhowmick, and A. Sun. Blog cascade affinity: analysis and prediction. In *Proceedings of the 18th Conference on Information and Knowledge Management*, pages 1117–1126. ACM, 2009.

- [21] S.A. Macskassy. Contextual linking behavior of bloggers: leveraging text mining to enable topic-based analysis. *Social Network Analysis and Mining*, pages 1–21, 2011.
- [22] M. McGlohon, J. Leskovec, C. Faloutsos, M. Hurst, and N. Glance. Finding patterns in blog shapes and blog evolution. In *the 1st International Conference on Weblogs and Social Media (ICWSM-07)*, 2007.
- [23] S.A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining*, pages 33–41. ACM, 2012.
- [24] M. Papagelis, N. Bansal, and N. Koudas. Information cascades in the blogosphere: A look behind the curtain. In *Proceedings of the 3rd International Conference on Weblogs and Social Media (ICWSM-09)*, 2009.
- [25] D.M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 695–704. ACM, 2011.
- [26] X. Shi, B. Tseng, and L.A. Adamic. Looking at the blogosphere topology through different lenses. In *the 1st International Conference on Weblogs and Social Media (ICWSM-07)*, 2007.
- [27] A. Stewart, L. Chen, R. Paiu, and W. Nejdl. Discovering information diffusion paths from blogosphere for online advertising. In *Proceedings of the 1st International Workshop on Data Mining and Audience Intelligence for Advertising*, pages 46–54. ACM, 2007.
- [28] L. Tabourier, A. Stoica, and F. Peruani. How to detect causality effects on large dynamical communication networks: a case study. In *Proceedings of the 4th International Conference on Communication Systems and Networks (COMSNETS)*, pages 1–7. IEEE, 2012.
- [29] A. Vazquez, B. Rácz, A. Lukács, and A.L. Barabási. Impact of non-poissonian activity patterns on spreading processes. *Physical Review Letters*, 98(15):158702, 2007.
- [30] Q. Zhao, Y. Tian, Q. He, N. Oliver, R. Jin, and W.C. Lee. Communication motifs: a tool to characterize social communications. In *Proceedings of the 19th International Conference on Information and Knowledge Management*, pages 1645–1648. ACM, 2010.