Some Insight on Dynamics of Posts and Citations in Different Blog Communities

Abdelhamid Salahbrahim, Bénédicte Le Grand and Matthieu Latapy
LIP6 – CNRS and Université Pierre et Marie Curie (UPMC – Paris 6)
104, avenue du Président Kennedy 75016 Paris – France
Email: firstname.lastname@lip6.fr

***Abstract*—This paper explores new approaches and methods to characterize post and citation dynamics in different blog communities. In particular, evolution of post popularity over time is studied, as well as information spreading cascades. This methodology goes beyond traditional approaches by defining classes of dynamic behaviors based on topological features of the post network, and by investigating the impact of topical communities on post popularity dynamics and on information spreading cascades. This methodology has been applied to a corpus of active French blogs monitored during 4 months.**

I. Introduction

Over the last years, activity and popularity of online social networks have been increasing continually. These virtual networks are extremely popular as they allow the production, sharing and spreading of information and opinions. A blog consists in a set of web pages (posts) frequently updated and commented. Each blogger can publish a post which may refer to other posts and cite other blogs. A blog network, made of blogs interconnected by citation links, is a very interesting object for the study of complex systems dynamics as it is easy to extract and provides precise temporal information.

This paper explores new approaches and methods to characterize blogs dynamics. In particular, evolution of post popularity over time is studied, as well as information diffusion cascades. We aim at going beyond traditional approaches by defining classes of dynamic behaviors based on topological features of the post network, and by investigating the impact of topical communities on blog dynamics.

This article is organized as follows. After a description of selected related works in Section 2, Section 3 describes the studied blog dataset, with a focus on methods to minimize measurement biases. Section 4 presents the results of traditional blog and post analysis, and shows consistent observations with the literature on the subject. Finally our contributions related to the characterization of post and citation dynamics in different blog communities are described in Section 5, before concluding and presenting perspectives of this work.

II. Related Work

Studying and modeling blogs as social networks has been raising an increasing interest [1], [2], [3], [4]. Blogs, like other social networks extracted from the Internet and the web, represent a rich content as semantic complex networks and social behavior networks. Previous works have studied the process of post and link creation and showed that their activity was bursty; results show that the network dynamics follows a power law. This burstiness results from human behavior [5], [6] and may be observed in different contexts. Power law distribution is indeed observed in many social network properties [7].

The spreading of information through a social network can be seen as diffusion of information on a network. The two main families of models from the literature describing how nodes adopt new ideas are threshold models [8], [9] and cascade models [10]. In threshold models, one node is 'infected' if a given number (or proportion) of its neighbors are infected. In this context, it can be seen as an opinion adoption model. In cascade models, a node spreads information towards a given proportion of its neighbors. In this paper, we study information spreading based on cascade extraction. The statistics computed from our dataset and presented in Section 4 are based on Leskovec et al. methodology [11] which uses post citations to extract the cascades rather than content analysis methods, which produce less realistic cascading behaviors [12]. Our paper goes beyond the state of the art by proposing original approaches to define posts classes based on topological features of the post network.

During the 2004 U.S. Presidential election, authors of [4] studied discussion topics and citation patterns of political bloggers. They showed that blog behaviors could be different according to blog type or topical community. For example news blogs do not behave as personal diaries. The diffusion of topics over time has also been addressed in [12]. This study concluded with the existence of two types of topics: *chatter* topics, with a stable popularity over time, in opposition to *spiker* topics with varying popularity values. In this paper, we investigate the impact of a blog theme on its dynamics by correlating topological dynamics to community activity; this is possible as the studied blog dataset contains manually-defined classes of blogs, as described in the following Section.

III. Dataset Description and Bias Correction

A. Definitions

In a blog, a publication is called a *post*. A post can, in addition to its own content, make a reference to a previous post (from the same blog or from another blog) by quoting the corresponding url, which we call a *citation link*. Consider a post *Pa* from blog *A* and a post *Pb* from blog *B*. If *Pa* contains a reference to *Pb*, then there is a citation link from *Pa* to *Pb*, i.e. *Pa* cites *Pb*. Post *Pb* has an incoming link pointing to it (noted *in-link*) while post *Pa* has an outgoing link starting from it (noted *out-link*). In terms of information spreading, we can say that *Pa* has 'adopted' *Pb*'s content or that *Pb*'s content has been spread towards *Pa*. The citation relationships between posts may be extended to the blogs they belong to: we may then say that blog *A* cites blog *B* and consider that *A* has adopted *B*'s information (or that *B*'s information has been spread towards *A*).

Citation links should not be mistaken with *comments*. If someone comments an existing post, this contribution is not a post (as it does not start a new discussion). The corpus analyzed in this paper was obtained by daily crawls of 6344 blogs ($1,230,692$ posts) during 4 months from November 1st, 2008 to March 1st, 2009. These blogs have been chosen according to their popularity and activity in the French-speaking blogosphere. They have been selected by a company specialized in blog and opinion analysis (RTGI, http://linkfluence.net) as being the most active and productive blogs which provide richer information for activity and dynamics study. All blogs provide a description of their publications by RSS (Rich Site Summary); RSS represents a key feature to capture temporal features as all posts published by the 6344 blogs are recorded as they are published. Post information is composed of two types of data:

- metadata which contain general information like the date of publication as well as the manual community classification of the blog publishing the post,
- the core of the post containing the full text and the citation links.

Moreover, this dataset has a major advantage as blogs have been clustered manually into labeled classes (which we will call *communities* in this paper); this will allow us to study the impact of topical communities on post dynamics (see Section 5). This manual classification provides three abstraction levels: Continent, Region and Territory (from the most general to the most specific). In the context of this paper, we will focus mostly on the Continent layer, which is composed of three communities: 'Leasure', 'Society' and 'Individuality' and the Regional layer composed of 16 sub-communities.

B. Data Cleaning

A preliminary step consists in cleaning the dataset, in order to remove errors and ensure that data represent the actual blogs dynamics.

The data consist of post and citation links. Each link connects a post to another post at a given time (indicated by a timestamp). An example of error in the initial dataset is when a post cites a more recent post (i.e. a post which has a higher timestamp); this would mean that it refers to a post which has not been created yet! This may happen if the corresponding blogs are hosted on servers in different time zones; this may also be due to a human manipulation. Such 'impossible' links are filtered as the information they provide is erroneous (at least with regard to the temporal information). In order to study spreading cascades between different blogs, self-citation links (i.e. citations of posts within the same blog) have been removed, as they do not provide any information about diffusion towards other blogs.

Out-links can point to a post inside the dataset (if the post has been published by one of the 6344 blogs of the corpus), or outside the dataset if the post refers to a post belonging to another blog, a resource (picture, video...) or any web page. During the cleaning process those links were removed for two reasons: first, resources like pictures cannot 'cite' any post and therefore cannot contribute to the growth of cascades. Second, no temporal information is available about blogs outside the dataset, so they are unusable for our study of the blogosphere dynamics.

When building the blogs graph, blogs with no in-link and with out-links only towards blogs outside the initial dataset were removed as these blogs are not connected to any blog from the initial dataset. After this filtering process, the blogs network contains 4907 nodes (blogs) and $28,258$ edges (citation links among blogs).

C. Biases

Trying to study the dynamics of a complex network is usually biased since only a partial view of the studied object may be observed [13].

First, a bias is introduced by the choice of blogs corpus. Blogs have been selected according to their activity and popularity. However, in order to represent the 3 Continents fairly, the same number of blogs has been chosen in the 3 corresponding communities, as shown in Table 1[1]. Note that the three continents have differences in term of post and link production, even though the number of blogs is comparable. A second bias is induced by missing information, typically all links that appear before or after the capture. Biases have been studied in [14], [13] in Peer-to-Peer networks. We consider here each (post, out-link) couple at publication time, so there is no incoming-link before the measurement begins. Border effect also appears when links arrive after the end of the measurement. This creates a bias toward posts appearing at the end the capture as they cannot be cited during as many days as earlier posts. The solution we propose consists is observing the popularity of all posts during the same period of time. In this paper, each post created during the first 40 days of the capture has been monitored (i.e. the citations towards it has been recorded) during 40 days after its publication (see Figure 5a).

In this Section, we have described the blogs corpus which is used throughout this paper. In the following Section, traditional blog and post analysis is performed in order to check data consistency with existing work on the subject.

[1] In this table, all filtering steps have been applied except removing posts referring to resources.

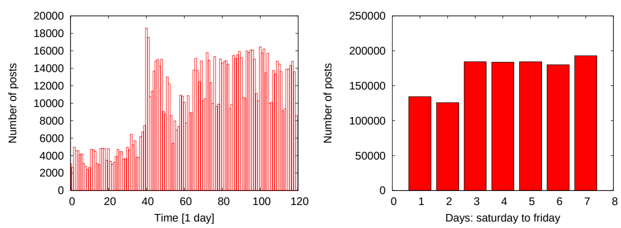
TABLE I
ACTIVITY BY CONTINENT

Continent	nbr of posts	nbr of blogs	nbr of links
Society	564535	2245	736324
Leasure	469896	2045	429748
Individuality	150927	2054	326228

IV. TRADITIONAL BLOG AND POST ANALYSIS

A. Blogosphere Activity

This Section presents traditional statistics on blogosphere activity. Figure 1a shows the number of posts published daily during 4 months. Instead of starting from a low value and increasing gradually, the number of posts is high from the beginning, which is due to the crawling technique, as the activity of all blogs of the corpus are monitored (as opposed to a 'snow ball' approach starting from a single blog and discovering new blogs through citation links).



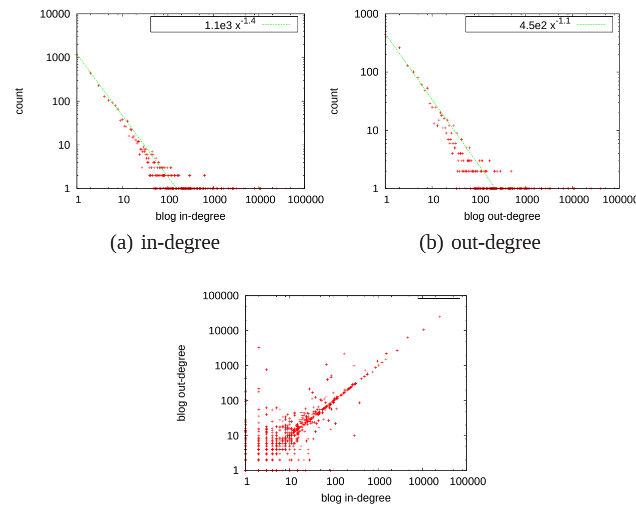
(a) (b)

Fig. 1. Number of posts a) per day during 4 months, b) for each day of the week

Another observation is that the number of posts increases suddenly after the 40th day. This is also due in part to the crawling method, as a new set of blogs was introduced in the corpus at that time. To avoid any bias due to measurement the first 40 days have not been taken into account. We can also observe an activity decrease between the 50th and the 63th days. This corresponds to the period of Christmas holidays, during the two last weeks of December, which suggests that bloggers are less active during holidays. Figure 1b displays the number of posts created for each day of the week from Saturday to Friday. A weekly periodicity may be observed on this figure, as week-end days show a lower production activity.

B. Blog and Post Network Characterization

The blog network is a graph representing blogs connected by inter-blog citations. Nodes of the graph are blogs and edges, corresponding to citation links, are directed and weighted. Blog *A* is linked to blog *B* with a weight *n* if there are *n* posts from blog *A* which have cited posts from blog *B*. The blog graph is composed of 4907 nodes and $28,258$ edges. The number of citation links is $666,191$ which represents the sum of all edge weights. The in-degree and out-degree follow similar distributions (Figures 2a, 2b). They are very heterogeneous and well fitted by power laws with exponents 1.4 and 1.1. The correlations between in and out-degrees (Figure 2c) show that these degrees are correlated for high



(a) in-degree (b) out-degree

(c) in out correlation

Fig. 2. In-degree and out-degree distributions in the blogs network.

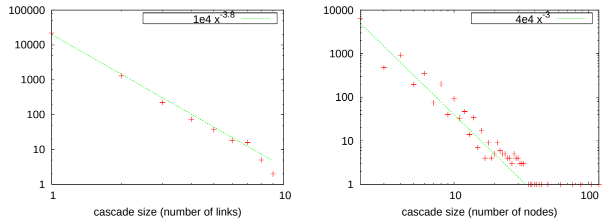
degree values. However, as the number of dots with low degree values is high the correlation coefficient between in and out degrees is only equal to 0.36. A small value of this coefficient contradicts the intuition that the popularity of a blog (indicated by its in-links) is related to its activity (its out-links). However, this result is in line with outcomes of previews studies [11] who observed an even smaller correlation coefficient.

The activity of a blog can be measured through the number of posts it sends (publishing activity) or through the number of citations by others blogs (in-linking activity). We have calculated the distribution of the number of citations per blog. In the blog graph this represents the sum of weights of outgoing edges. It is close to a power-law distribution with exponent 1.2. The total number of citations is $666,191$ which represents an average of 100.5 citations per blog during 4 months. The number of posts per blog is also close to power law with exponent 1.25. The distribution of edge weights in the blog network, which corresponds to the number of blog-to-blog links also follows a power-law distribution with exponent 2.27. The post network has similar characteristics. We have found that distributions of posts in- and out-degrees is close to a power law with exponents 2.6 and 3.1 respectively. All these results are consistent with the state of the art.

C. Cascades

Cascades are subgraphs of the post network, where each node corresponds to a post and edges to citation links. In order to compute post cascades, we start by posts which do not cite any other post; each of them represents the end of a cascade. Consider one such post; if it is cited by one or several posts, the process carries on recursively: posts which have cited this citing post (or posts) are looked for and so on. Each post can belong to several cascades, represented as trees. The root of each tree is the initiator of the cascade, with in-coming edges from the nodes that cited it. Information is therefore spread from the root to the leaves of the tree. Propagation flows represent influence between blogs through their posts. As explained in Section 3, as we focus on information diffusion among different blogs, links between two posts from a same blog were removed. Indeed, self-citations can make cascades longer but do not represent a diffusion process from one blog to another.

The total number of cascades (with at least two nodes) is $13,247$. The most common cascade is composed of two posts and represents 75% of the whole.



(a) (b)

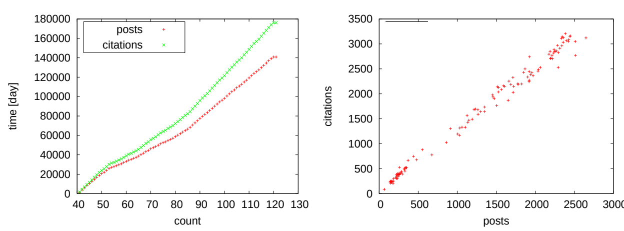
Fig. 3. Distribution of the number of a) links and b) nodes per cascade.

We observe in Figure 3 that the distributions of the number of links and nodes in cascades have a heavy-tailed distribution with respectively 2.5 and 3 as exponents. Most cascades are made of only 2 nodes while only a few contain a large number of nodes.

In this Section we computed basic statistics to characterize post and blog networks, as well as information diffusion cascades. Obtained results are consistent with the state of the art. The following Section goes further in the study of post and citation dynamics and presents our approach to classify post according to their popularity and study the impact of communities on this popularity and on cascades.

V. Dynamics of Posts and Citations in Different Blog Communities

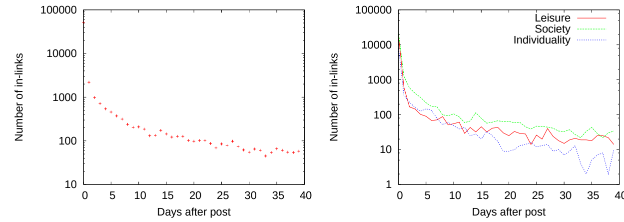
A. Dynamics of Post and Citation Arrivals



(a) (b)

Fig. 4. a) Evolution of the number of posts and citations. b) Citation and post correlation.

We investigated how the overall number of citations, posts and blogs evolved over time. Figure 4a presents this evolution during the 80 days of measurement (the first 40 days have not been taken into account as explained in Section 4). Figure 4a shows that the number of posts and citation grows in similar ways. Figure 4b presents the correlation between the number of citations and posts per days. The daily ratio between total number of citations and post remains approximatively constant during the 80 days, with average value 1.28. This high correlation of post and link numbers per day means that the ratio between total number of citations and post is constant even when blogosphere topology changes.



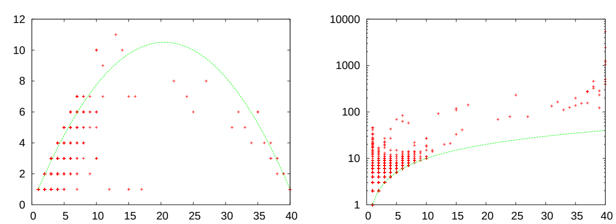
(a) (b)

Fig. 5. Evolution of post popularity a) overall evolution b) according to post community.

Another interesting aspect is the evolution of post popularity after publication, in terms of number of links each post gathers from its first day of appearance (i.e. its posting day) until 40 days later, according to the methodology to correct bias described in Section 3.3. In Figure 5a we observe that most citations are made within the first 24 hours which is in accordance with previous studies [10]. Results differ from [11] were popularity decreases significantly at the end of the measurement, but this is due to the bias we mentioned earlier. Here, popularity is divided by half after the first 24 hours but later on it decreases at a much slower pace.

Figure 5b represents the evolution of post popularity according to their community (Leasure, Society, Individuality). We observe that popularity of posts from Individuality community decreases more than the popularity of posts from Leasure and Society after 13 days after post creation. Moreover, although the number of citations of posts from Society community is higher than the number of citations of posts from Society, both plots tend to converge after 27 days after post creation. As a conclusion, we may say that post popularity evolves quite similarly in the 3 studied communities.

B. Two types of Citation Dynamics



(a) (b)

Fig. 6. Post classification - for both plots, the value on the x-axis (denoted by X) represents the number of days during which the posts have been cited. a) The value on the y-axis (denoted Y) corresponds to the number of distinct citation periods (e.g. for a post cited on days $1, 2, 3, 5, 7, 8$, X=6 as it is cited on 6 days, and Y=3 as these citations occured during 3 periods of time: $\{1, 2, 3\}$, $\{5\}$ and $\{7, 8\}$). b) The value on the y-axis represents the sum of all citations of a post for the corresponding number of days. So for example a dot with coordinates $(3; 34)$ represents a post which has been cited on 3 distinct days and 34 times in total.

An interesting approach to study posting behavior is to classify posts according to incoming citations. Thanks to our bias correction method, all incoming citations for each post are known during 40 days after its creation. Figures 6a and 6b explore this information.

On Figure 6a, an additional curve is displayed, corresponding to a stochastic behavior. We can observe that the dots are very close to this stochastic behaviour. Apart from posts which are cited only a few days (i.e. for which X is low), the citation of posts is neither oscillating (as Y values are not higher than those of the stochastic curve) nor continuous (as there are very few dots with low Y values when X is high).

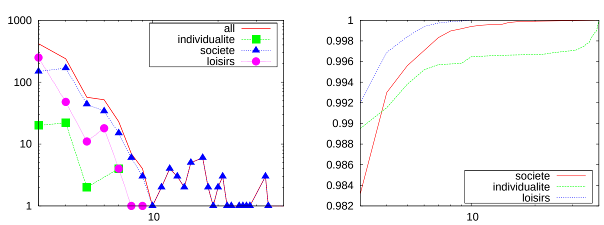
One limit of Figure 6a is that the number of citations of posts on each day is not visible. Figure 6b completes the view. We observe that most posts are cited only few days. However a group of posts can be identified which are cited more than 30 days; they correspond to a specific class of behavior. Moreover, posts which are cited on many days (i.e. with high X values) are cited more than once a day, as Y values tend to be higher than the minimum bound corresponding to the pink curve defined as y=X. We may therefore conclude that the higher the number of citation days, the higher the frequency of citations per day.

Figure 6b indicates a statistically significant classification of posts into two groups:

- posts which are cited on at most 10 distinct days; they represent 99% of all posts. Within this class, the number of posts cited only one day represents 89% of posts.
- posts cited on more than 10 days.

Going further, we study in the next Section how posts from the three communities are distributed among these two types of popularity.

C. Types of Citation Dynamics and Communities



(a) (b)

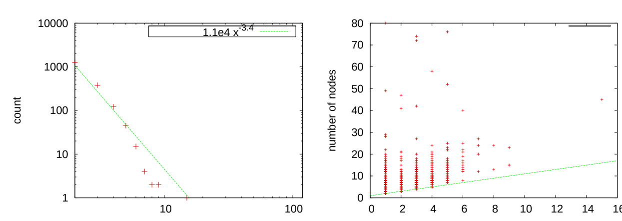
Fig. 7. a) Distribution of citations of posts cited only one day. A dot with coordinates (x, y) means that there are y posts which have been cited x times on their single citation day (as X=1). b) Cumulative number of citations per total number of citation days. This plot represents the cumulative probability density function (PDF) of the number of citations per number of days for X greater than 3.

Let us first focus on posts which are cited only one day. Figure 7a shows the distribution of the total number of in-links for all posts cited only one day and for each community. The reasons of this restriction to posts cited only one day are the following: first, the number of these posts represents 89% of the total number of posts in the dataset as said earlier; second, taking into account all posts of the first class would be confusing e.g. a post cited 10 times on a single day would appear like a post cited once on 10 distinct days. The distributions of the posts of each class are heterogeneous, however, the behaviors within each community differ: the Individuality continent has the lowest total number of citations (corresponding to the surface below its curve) which indicates that personal blogs have a smaller popularity than posts from other blogs. The popularity of posts from the Leasure community is much higher but these posts are not cited many times as the maximum x value is 9 (i.e. posts cited 9 times on the only day they have been cited). On the other hand we observe that all posts having a $y > 10$ belong to the Society continent. Those posts have a high popularity and the corresponding topics can be considered as *spikers* because there are cited a lot on a short period of time (here, only one day). After studying these posts in more detail, we found out that they belonged mainly to blogs related to news sites, thus confirming their interpretation as spikers.

We have focused so far on posts cited only one day. We now study the citation dynamics of posts cited on at least 3 different days, illustrated on Figure 7b. The curve representing the popularity of the Individuality community increases more slowly than the 2 other curvesfor low X values and increases significantly for the maximum values of X. This rapid increase at the end of the measurement is a border effect due to the fact that each post citations are monitored during only 40 days. Therefore the maximum value (1) has to be reached at the end of the plot, and it is not possible to know whether posts cited 40 days have actually been cited 40 days, or 41 days (or more). An isolated group of posts cited on more than 30 days may be identified in this community; they belong to blogs with an important forum-like activity. These blogs may be considered as ’different’ from traditional blogs and this representation is therefore interesting to detect outliers. On the other hand, the two other communities (Leasure and Society) show a different behavior in terms of popularity evolution. In particular, no specific increase can be observed at the end of the measurement on these curves as the maximum value is reached around $X = 10$ and $X = 20$ for Leasure and Society communities respectively. The conclusion is that posts of these communitites are no longer cited after 10 (resp. 20) days since their publication and that these posts therefore do not need to be monitored more than 10 days (resp. 20 days).

In this subsection, a deeper study of each type of post popularity has shown an impact of post communities on the evolution of their popularity. In the following subsection, we investigate their potential impact on information spreading cascades.

D. Cascades and Communities



(a) (b)

Fig. 8. a) Distribution of communities per cascade. A dot with coordinates (x, y) means that there are y cascades which contain posts from x different communities. b) Number of communities per cascade. A dot with coordinates (x, y) means that there is a cascade containing y nodes which involves x communities.

Cascades may be composed of posts from different communities, but one may expect that they mostly occur among posts of a same community. The data considered here makes it possible to explore this question, for the first time to our knowledge. This is the focus of this Section.

86% of cascades contain only posts from the same Continent (Society, Leasure or Individuality). However this community view may be too high-level and a more detailed analysis is necessary.

For the following figures, we have used the second level community classification which is composed of 16 Regions (i.e. sub-communities). Figure 8a represents the distribution of the number of Regions per cascade. This distribution shows that 75% of these cascades are made of posts from only one sub-community. This means that at the Region level cascades are homogeneous in term of community composition, which confirms our intuition. Figure 8b represents the size of cascades (i.e. the number of nodes they contain) as a function of the number of communities. It shows that few cascades are composed of posts from more than 6 communities. We also observe that the cascades which contain the highest number of nodes are not those which involve the highest number of communities. These results indicate that there is an impact of topical communities on information spreading behaviour.

VI. Conclusion

This paper proposed new approaches and methods to characterize post and citation dynamics in different blog communities. In particular, the evolution of post popularity over time has been studied, as well as information spreading cascades. We have gone beyond traditional approaches by defining classes of post popularity evolution based on topological features of the post network, and by investigating the impact of topical communities on citation dynamics. We have proposed a new representation of posts' popularity using daily incoming citations, in order to classify posts. Moreover, we showed that distinct patterns related to both the duration and the frequency of citations could be observed in the various communities.

This methodology has been applied to a corpus of active French blogs which contains a rich information with regard to blog communities. Section 4 showed that the results of traditional statistics were consistent with previous work on this subject, which somehow validated our dataset.

One of the contributions of this paper is the methodology defined in Section 3 to correct biases towards most recent posts and blogs in the dataset. An interestingly high correlation was found between the total numbers of posts and links per day, which means that the ratio between total number of citations and post is constant even when the blogosphere topology changes. We also proposed in Section 5 an original representation to identify classes of post popularity based on the structure provided by incoming citation links. In particular, the topology of the post network has allowed us to identify two main classes of post popularity. A deeper study of each class has shown that blogs related to spiker topics could be detected easily as well as blogs with forum-like activity. Finally, the impact of topical communities on the evolution of post popularity over time and on information diffusion cascades has been investigated. In particular, it showed that the measurement's duration may be reduced for specific blog communities. The study of the impact of community information on post and citation dynamics has shown promising results and we will carry on working on this in the future.

In this study, the spreading process has been studied only through the analysis of citation links between posts. A new dataset containing both these citation links and comments about existing posts will be available shortly. This enriched blog information will allow us to compare the results presented in this study with statistics based on comments only and on the joint use of comments and citation links.

Another perspective of this work is to apply this methodology to other types of complex networks in which diffusion processes take place in order to assess the interest of this approach in other contexts.

Acknowledgements: This work is partially funded by the *WebFluence* ANR and Ville de Paris *DiRe* projects.

References

[1] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose, "Implicit structure and the dynamics of blogspace," in *Workshop on the Weblogging Ecosystem, WWW2004*, New York, NY, 2004.

[2] E. Cohen and B. Krishnamurthy, "A short walk in the blogistan," *Comput. Netw.*, vol. 50, no. 5, pp. 615–630, 2006.

[3] S. C. Herring, I. Kouper, J. C. Paolillo, L. A. Scheidt, M. Tyworth, P. Welsch, E. Wright, and N. Yu, "Conversations in the blogosphere: An analysis "from the bottom up"," 2005, p. 107b. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1385453

[4] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 u.s. election: divided they blog," in *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*. New York, NY, USA: ACM, 2005, pp. 36–43.

[5] M. Wang, N. H. Chan, S. Papadimitriou, C. Faloutsos, and T. Madhyastha, "Data mining meets performance evaluation: Fast algorithms for modeling bursty traffic," *Data Engineering, International Conference on*, vol. 0, p. 0507, 2002.

[6] A. Vazquez, J. G. Oliveira, Z. Dezso, K.-I. Goh, I. Kondor, and A.-L. Barabasi, "Modeling bursts and heavy tails in human dynamics," p. 036127, 2006. [Online]. Available: http://link.aps.org/abstract/PRE/v73/e036127

[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*. New York, NY, USA: ACM, 1999, pp. 251–262.

[8] M. Granovetter, "Threshold models of collective behavior," *The American Journal of Sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.

[9] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing Letters*, 2001.

[10] E. Adar and L. A. Adamic, "Tracking information epidemics in blogspace," in *Web Intelligence*, Compiegne, France, September 2005. [Online]. Available: http://www.hpl.hp.com/research/idl/papers/blogs2/trackingblogepidemics.pdf

[11] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading behavior in large blog graphs," in *In SDM*, 2007.

[12] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *WWW '04: Proceedings of the 13th international conference on World Wide Web*. New York, NY, USA: ACM, 2004, pp. 491–501.

[13] L. Benamara and C. Magnien, "Estimating properties in dynamic systems: the case of churn in p2p networks."

[14] D. Stutzbach and R. Rejaie, "Understanding churn in peer-to-peer networks," in *IMC '06: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*. New York, NY, USA: ACM, 2006, pp. 189–202.