Finding the semantic-level precursors on a blog network

Telmo Menezes*

CREA and ISCPIF,
CNRS, ISC-57-59, rue Lhomond,
F-75005 Paris, France
E-mail: telmo@telmomenezes.com
*Corresponding author

Camille Roth

CAMS and ISCPIF,
CNRS-EHESS, 54, bd Raspail,
F-75006 Paris, France
E-mail: roth@ehess.fr

Jean-Philippe Cointet

INRA-SenS and ISCPIF,
INRA-SenS, Université Paris-Est, Bois de l'Etang,
5, Bd Descartes, Champs sur Marne,
F-77454 Marne-la-Vallée, France
E-mail: jean-philippe.cointet@polytechnique.edu

Abstract: In this work, we study semantic-level precedence relationships between participants in a blog network. Our methodology has two steps: a process to identify units of discussion at the semantic level and a probabilistic framework to estimate temporal relationships between blogs, in terms of the order in which they arrive at those units of discussion. We propose dyadic precursor scores that can be used to construct semantic-level precedence networks. From these scores, we derive global precursor and laggard scores. Dyadic precursor scores are compared with URL linking to show that the semantic-level temporal relationships we estimate are an indicator of influence. Global scores are compared to traditional link degree and PageRank metrics, and we uncover relationships between semantic-level temporal behaviour and popularity. We show that our method reveals information about the network that could not be obtained from structural links alone.

Keywords: social networks; complex networks; blog networks; network analysis; network metrics; topic detection; temporal relationships; social computing.

Reference to this paper should be made as follows: Menezes, T., Roth, C. and Cointet, J-P. (2011) 'Finding the semantic-level precursors on a blog network', *Int. J. Social Computing and Cyber-Physical Systems*, Vol. 1, No. 2, pp.115–134.

Biographical notes: Telmo Menezes obtained his PhD in Computer Science from the University of Coimbra, Portugal. He is a Post-doctoral Researcher at the Center for Research in Applied Epistemology and the Complex Systems Institute in Paris, France. His research interests include complex systems and networks, multi-agent simulations, evolutionary computation and biologically-inspired computational intelligence.

Camille Roth is a Permanent Researcher in Computer Science at CNRS in Paris, after being Associate Professor of Sociology at the University of Toulouse, France. He obtained his PhD in Social Science from Ecole Polytechnique, Paris, with a background in general engineering ('ing enieur des Ponts', Paris) and cognitive science (MSc, EHESS, Paris). His interests focus on socio-semantic dynamics and mathematical modelling of social systems of knowledge production, including scientific networks and online communities (blogs, wikis, tagging systems, etc.).

Jean-Philippe Cointet is a Researcher at INRA-SenS and ISC-PIF. He holds a PhD in Sociology from the Ecole Polytechnique, Palaiseau France. He is interested in the reconstruction, visualisation and modelling of social and knowledge dynamics from empirical systems such as scientific communities or blog networks.

1 Introduction

For cultural anthropologists, understanding fads, trends, or, generally, cultural similarity, essentially comes to explaining "the capacity of some representations to propagate until becoming precisely cultural, that is, revealing the reasons of their contagiosity" (Lenclud, 1998). This type of research programme admittedly assumes the possibility of, on one hand, describing representations in a consistent manner, and, on the other hand, apprehending processes of social mediation. Defining consistent cultural items is indeed crucial to describe adoption of similar ideas, behaviours, opinions, topics, etc. – the literature proposes here a large variety of concepts, such as using same bags of terms, having identical opinion vectors, duplicating references (for instance to digital content such as online video or news articles, tagged by the same URL) or, more loosely, being 'infected' by spreading 'memes'. Second, describing social mediation requires to understand jointly how some types of social network configurations and some types of interactions may or may not favour the transmission, reproduction or adoption of behaviours, ideas, etc. Again, a vast amount of research has been concerned with normative models or descriptive protocols aimed at understanding which kind of individuals were more or less likely to pass on some pieces of information, and which type of network positions could favour the diffusion of some items.

By relying on large-scale datasets on which individuals talk about what and when, specifically in online communities, social computing has recently contributed to this broad research programme by intensively developing two pragmatic streams of study: detection of 'topics', and characterisation of 'informational cascades'. Studies focused on topic detection explore bursts and regularities of behaviour or term use (e.g., Kleinberg, 2002), sometimes in order to infer trends in the general population (Ginsberg et al., 2009, Asur and Huberman, 2010). In all these studies, cultural representations are assumed to be extremely atomic, i.e., based on a single behaviour (a vote), item (a reference, a URL), apprehending cultural contagion pretty much similarly to disease contagion – to the notable exception of (Leskovec et al., 2009) who gather similar sentences into clusters of quotes, getting closer to the polymorphism of cultural representations emphasised by anthropologists.

On the other hand, studies on informational cascades currently adopt a structural stance, migrating from the 'two-step-model' to more recent arguments underlining the importance of more horizontal, less hierarchical patterns (Watts and Dodds, 2007; Cha et al., 2010). Importantly, in this perspective, information flows and diffusion paths are characterised along a given social network, available *a priori*. In many cases, however, and certainly in blogs in particular, much of the information regarding the whole underlying interaction infrastructure is simply missing (be it in terms of news media readership, e-mail exchanges and broadly any type of non-blog-based online conversation, phone calls, etc.). Additionally and, in part, as a result, current methods for appraising the contribution of nodes in blog networks usually do not account for temporal relationships.

In this paper, we aim at bridging these rather separate streams by adopting

1 a looser view on representations, as stories or cultural attractors (Sperber, 1996; Sperber and Claidiére, 2006) rather than atomic items

2 by considering information sources, in our case bloggers, as sensors in a social system – in particular as representatives of topics discussed in the society – so as to suggest possible/implicit information diffusion flows or, at least, precedence relationships.

As an aside, the current contribution also considers *observed* social networks as effects rather than just causes of information propagation.

We thus propose to identify topic classes, exhibit temporal precedence relations between sources based on *significant plausibility* for an individual to address a topic before others do, and eventually compare this structure with the partial network of interactions constituted by explicit links among bloggers. Classical authority measures are found to have only a weak correlation with our approach, which rather exhibits potential online whistleblowers. The next section presents an overview of the relevant literature, while Section 3 details the empirical protocol used to identify topics. Section 4 then describes our approach to compute probable precedence relationships; results are discussed and reframed in Section 5.

2 Related work

2.1 Temporal detection of topics/bursts

Topic characterisation from (online) text corpora generally relies on *terms*, *n-grams* (i.e., a basic linguistic unit of *n* terms) or sentence segments. Once basic text units have been defined and extracted, topics are appraised both quantitatively and temporally, essentially by describing "how much on which period of time they are being discussed". This led to distinguishing bursts of interest ('spikes') (Kleinberg, 2002), as opposed to continuous discussions ('chatters') around topics (Gruhl et al., 2004). Models of the temporal (Balog et al., 2006) or spatial (Lloyd et al., 2006) regularities in the usage of topics have been subsequently developed, up to inferring and predicting accurate information regarding the whole population behaviour (Ginsberg et al., 2009; Asur and Huberman, 2010).

Another stream of research has focused on improving the qualification of topics: for instance, by detecting whether issues are addressed in a positive light or not [the so-called field of 'sentiment analysis', see Mishne and de Rijke (2006), among others]; or, closer to our issues, by managing to group portions of text into classes of similar content (Leskovec et al., 2009) – thereby implicitly addressing one common critique among social scientists regarding the atomism of 'memes' as cultural items.

2.2 *Precedence and influence*

Empirical studies of influence generally rely on interaction networks. They use relational information to characterise contagion paths and follow a relatively long tradition of social network-based models of information diffusion. As regards blogspace in particular, after initial descriptions of the underlying social network structure [e.g., Kumar et al. (2005), who also discuss bursty behaviour in link creation], Leskovec et al. (2007) has been one of the first studies to specifically focus on the structure of link cascades. In a previous work, Cointet and Roth (2009) describes more precisely local influence patterns – such as the relationship between, e.g., holistic patterns and the weakness of links, in Granovetter's sense. In Java et al. (2006), on the other hand, various social network structures are used to show that possible influence of a given blog is best described by strictly structural page-rank-style measures.

Since influence is obviously related to precedence relationships, several papers focus rather on temporal behavioural precedence. For instance, Kossinets et al. (2008) exhibit explicit temporal dependencies on a e-mail transmission network by characterising possible shortcuts in information paths, because a dyad (A, B) could communicate less quickly than (A, C) and (C, B) separately do.

In terms of intertwining social network structure and precedence/influence, the relationship between topology and precursors or laggards had also been explored in Valente (1996), but with the assumption that the social network is known a priori, and by monitoring the adoption of a unique yes-or-no behaviour. As said before, it is likely that a lot of information about the social structure is missing in most of the above studies, which consider the (given) social network as the substrate of information propagation. By assuming that the social structure describes only a non-significant fraction of all possible interaction links and contagion paths in the context of (for instance) political discussions, we basically wish to suggest that, here, the social network could just be a secondary material in the study of contagion.

Some studies do exactly so and exhibit influence relationships from usage information only: for instance in Zhou et al. (2006), a Markov chain model is used to characterise which topics are most likely to transition into others, using data extracted from scientific bibliographic databases. Back to blogs, 'probable' content diffusion paths could be exhibited in Adar et al. (2004) by using classifiers based upon blog features: for instance, having similar citing and content posting patterns; however, the analysis does not seem to make use of topic dynamics *per se*. Another reference, Java (2006) introduces an analysis which integrates more semantics – essentially in order to design automatic feed recommenders – which appears nonetheless to be still based on structural features (in-degree statistics), even if a filter is applied over general topics (politics vs. IT, etc.).

On the whole, and in the context of partial social network information, the issue of the detection of implicit, non-structural influence flows using temporal precedence in addressing topics remains a pending question.

3 Unit of activity detection

We are interested in identifying topics of discussion for which we can later analyse the temporal relationships of their participants. Such topics must have two characteristics to be relevant to our analysis: to have well defined time boundaries within our observation period and to be maintained by the participation of several blogs. If these two constraints are respected then we are observing what we will call a well-defined 'unit of activity'. We empirically define a method that identify bursty topics which meet these constraints.

In Leskovec et al. (2009), research related to the problem of topic detection is classified into two main categories: probabilistic models to identify long-range trends in general topics and the use of rare named entities to study short information cascades. We are not interested in long-range, general topics, nor in having to rely on the occurrence of very specific, rare strings. Instead, our goal is to identify topics that can identified by a set of n-grams and a well-bounded period of time, and that represent self-contained units of activity.

We propose a holistic approach that takes advantage of both the textual content of blogs posts, and the times at which these posts where published.

The process of topic detection we propose consists of a sequence of treatments on the dataset:

1 part-of-speech tagging and lemmatisation of each post's title and content in order to enumerate every relevant n-grams in the corpus

2 detection and filtering of n-gram temporal bursts

3 merging of redundant n-gram bursts into unique topics.

3.1 Linguistic treatment

In the first step we use the TreeTagger (Schmid, 1994) tool to generate a new version of each post's title and textual content, where each word is lemmatised and augmented with a part-of-speech tag.

We then divide the corpus of text generated by the previous step into chunks, delimited by punctuation marks. Afterwards, we find all the n-grams that occur in these chunks. This search is constrained by a set of rules, as to not generate an intractable amount of n-grams, and explore only cases we believe are likely to lead to meaningful topics. The rules are the following:

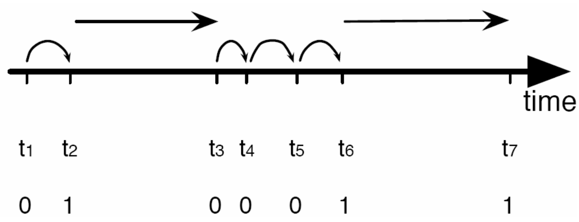
- n-grams must have two or more words
- an n-gram must contain at lease one noun
- all words that are not nouns, verbs, adjectives or numbers are discarded
- all n-grams that contain words in a special set called *stop-words list* are rejected.

These rules are empirical, having been obtained by experimentation with real datasets. The word set in the last rule contains words that have a strong temporal meaning, and that would later on lead to the detection of meaningless temporal bursts of usage. We used a set containing names of months, days of the week and holiday seasons (like Christmas), in both French and English.

3.2 Temporal bursts detection

In the second phase, we analyse the pattern of occurrence of each n-gram, dividing the period of observation into bursts of activity. For this purpose, we devised an algorithm that iteratively divides the timeline into intervals, aiming at the maximisation of a value we will call the *burst ratio*. Let us consider an ordered set $T = \{t_1,\ldots,t_n\}$ (in ascending order), where each element is the time of an occurrence of the n-gram. Furthermore, any two consecutive elements of T must originate from different blogs. This guarantees that a burst can only be maintained by the participation of multiple blogs.

Figure 1 Example of a sequence of occurrences of a given n-gram



Notes: The ordered sets T and Θ are depicted. Inter-bursts and intra-burst intervals are represented by arrows (respectively straight and curved).

We are interested in partitioning T into subsets which correspond to temporal bursts. Let us consider the ordered set $\Theta = \{\theta_1,\ldots,\theta_n\}$ where $\theta_k = 1$ if element t_k is the last element of a burst, and $\theta_k = 0$ otherwise. Each time θ_k equals 1 it means that the burst ends at t_k (see Figure 1). Given a partition Θ of the sequence of a n-gram into bursts, it is straightforward to compute the time-lag between the end of a burst and the beginning of the next burst or the time-lag between two occurrences inside the same burst. We can compute the average time-lag between two consecutive bursts or the average interval inside each burst on the whole timeline as follows:

$$V_{\mapsto}\vert(T,\Theta)=\frac{\sum_{i=1}^{|T|-1}(t_{i+1}-t_i)\theta_i}{\sum_{i=1}^{|T|-1}\theta_i}, \quad \text{if} \sum_{i=1}^{|T|-1}\theta_i > 0,0 \text{ otherwise} \tag{1}$$

$$V_{\curvearrowright}(T,\Theta)=\frac{\sum_{i=1}^{|T|-1}(t_{i+1}-t_i)(1-\theta_i)}{\sum_{i=1}^{|T|-1}(1-\theta_i)}, \quad \text{if} \sum_{i=1}^{|T|-1}\theta_i < 0,0 \text{ otherwise} \tag{2}$$

We also define the minimum inter-burst interval $m_{\mapsto}(T,\Theta)$ as:

$$m_{\mapsto}(T,\Theta) = \min_{\{i<|T|,\, \theta_i=1\}} \left(t_{i+1} - t_i\right) \tag{3}$$

We then define the *burst ratio*, $\rho(T, \Theta)$ as:

$$\rho(T,\Theta) = \frac{V_{\mapsto}(T,\Theta)}{V_{\curvearrowright}(T,\Theta)},\ \text{if } V_{\curvearrowright}(T,\Theta) > 0, 0 \text{ otherwise} \tag{4}$$

Simply put, $\rho(T, \Theta)$ is the ratio of the mean time interval between bursts to the mean time interval between elements inside bursts.

Algorithm 1 Pseudo-code of algorithm to perform temporal clustering of n-gram occurrences into bursts

```
stop ← False
while stop = False do
    best_burst_ratio ← –1
    best_postion ← –1
    for pos = 1 to |T| do
        if Θ_pos = 0 then
            aux_Θ ← Θ
            $\Theta_{pos}$ ← 1
            burst_ratio ← $\rho(T, aux\_\Theta)$
            min_inter_interval ← $m_{\mapsto}(T, aux\_\Theta)$
            if burst_ratio < $\alpha$ or min_inter_interval < $\beta$ then
                burst_ratio ← 0
            end if
            if burst_ratio > best_burst_ratio then
                best_burst_ratio ← burst_ratio
                best_pos ← pos
            end if
        end if
    end for
    if best_pos > 0 then
        $\Theta_{best\_pos}$ ← 1
    else
        stop ← True
    end if
end while
```

On Algorithm 1, we present the pseudo-code that describes the clustering method. The process is started with all the elements of Θ set to 0, meaning that in the initial state, all n-gram occurrences are considered to belong to a single burst. The algorithm iteratively tries to add new divisions to Θ, keeping the ones that increase the *burst ratio*, until no further improvement is possible.

Parameters α and β determine, respectively, the minimum *burst ratio* and interval between bursts (in days) that are accepted. These parameters allow us to prevent the formation of bursts that are not sufficiently separated, both in relation to the average interval between n-gram occurrences and in absolute value. For our purposes, we experimentally determined $\alpha = 5$ and $\beta = 5$ to be good values.

We devised our own burst detection algorithm instead of using one of the available ones, due to the specific requirements of our approach. For example, the weighted automaton model described in Kleinberg (2002) is suitable for detecting bursts at quantifiable levels of intensity, but does not lend itself to the detection of bursts with well defined limits. For the probabilistic model we are going to describe in the following section, it is crucial that we consider bursts with well defined limits, as not to lose initial or late arrivals. Our algorithm detects cases where the activity on a certain n-gram set can be characterised by intervals with a sufficient level of activity, separated by large enough intervals of no activity.

Finally, we filter the n-gram bursts, only accepting the ones that meet the following criteria:

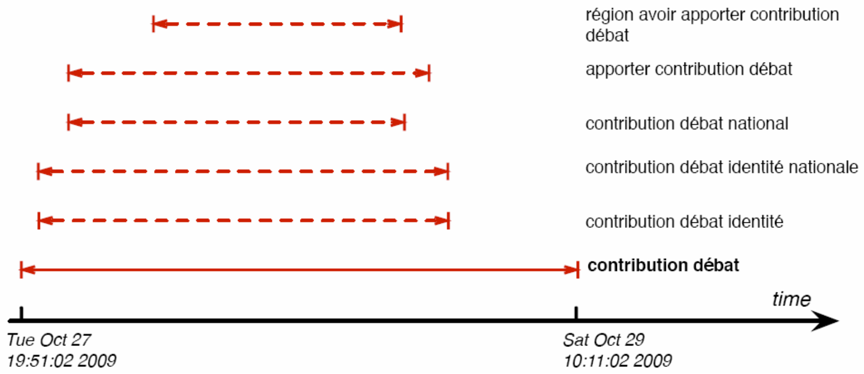
- a minimum number of blogs participating in the burst of four
- a minimum average time between posts participating in the burst of one hour
- a maximum average time between posts participating in the burst of one day
- a minimum burst duration of three days
- a maximum total duration of all the bursts of the n-gram of one month.

The purpose of these rules is to end up with n-gram bursts that are more likely related to a real topic. We discard bursts that are too sparse, too dense, too short lived or defined by an n-gram that is too common.

3.3 Merging n-gram bursts into topics

Finally, on the last phase, we extract a set of topics from the set of n-gram bursts that resulted from the previous step. We define a topic as a tuple $(\{g_0, g_1, \ldots, g_n\}, t, t')$, consisting of a set of n-grams occurring between times t and t'. Topics are defined with the minimum possible set of n-grams for maximum generality. Figure 2 illustrates on a real example how the n-gram bursts are selected to define a topic. The underlying idea is the following: consider two n-gram bursts, defined by n-grams g_a and g_b, occurring over time intervals $[t_a, t'_a]$ and $[t_b, t'_b]$. Furthermore, consider that the sequence of words in n-gram g_b is a sub-sequence of the sequence of words in n-gram g_a, and that $t_a \geq t_b$ and $t'_a \leq t'_b$. Referring to Figure 2, this could be exemplified by g_a = 'région avoir apporter contribution débat' and g_b = 'apporter contribution débat'. We assume that, in this kind of situation, it is very likely that both bursts belong to the same topic. g_b is more general than g_a, because it includes all the cases covered by g_a, while the opposite is not necessarily true.

Figure 2 Example of selection of n-gram bursts to define a topic (see online version for colours)


région avoir apporter contribution débat
apporter contribution débat
contribution débat national
contribution débat identité nationale
contribution débat identité
contribution débat
time
Tue Oct 27 19:51:02 2009
Sat Oct 29 10:11:02 2009

Note: Bursts in sold line are selected for the topic definition, while bursts in dashed lines are discarded.

We transverse the entire set of n-gram bursts, in descending order of the number of words contained in their n-gram. For each burst, we look for bursts ahead in the set with n-grams that are a sub-sequence of the first one, and with time intervals that contain the interval of the first one. If such bursts are found, the original burst is discarded. If one of the bursts found is already assigned to a topic, we also assign the other bursts found to that topic, otherwise we assign all bursts found to a new topic.

4 Probabilistic precedence scoring

After the process described in the previous section, we now have a set of topics, and know which blogs participated in each topic and at what time. We are now in the position of defining a probabilistic model that estimates the tendency that blogs have to participate in topics before other blogs.

We will start by defining a *dyadic precursor score* from blog b to blog b'. We will call this score $\gamma(b, b')$. Let us define A as the set of all topics where both blogs participate, and Y as the subset of A where the first participation of b precedes the first participation of b'. We also define C as a vector of probabilities. Each element of C is the probability that b participates on a topic before b' by chance. We will detail later how these probabilities are computed. We know define the likelihood of $\gamma(b, b') = p$, given A, Y and C:

$$\lambda(\gamma(b,b') = p \mid A,Y,C) = \sum_{\substack{Z \cup R = Y \\ Z \cap R = \emptyset}} \lambda(\gamma(b,b') = p \mid A,Y,C,Z,R) \tag{5}$$

The likelihood in equation (5) is defined as the sum of the likelihoods for all possible hypothesis of the appearances of b before b' being caused by a temporal relationship or by chance. The set Y of topics where the first participation of b precedes the first participation of b' can be decomposed as the union of the set Z of topics where b is assumed to display a behaviour of precedence over b', and the set R of topics where b is assumed to precede b' by chance. We define the likelihood of each hypothesis as:

$$\lambda(\gamma(b,b') = p \mid A,Y,C,Z,R) = P_Z(A,Z,p) \cdot P_R(A,R,C) \tag{6}$$

$P_Z(A, Z, p)$ is the probability that b precedes b' in the topics in Z and not in the topics in A\Z, given a probability of a precedence relationship of b over b' of p. $P_R(A, R, C)$ is the probability that b precedes b' by chance for the topics in R, and not for the topics in $A \backslash R$, given C. These probabilities are defined as:

$$P_Z(A,Z,p) = p^{|Z|}(1-p)^{|A|-|Z|} \tag{7}$$

$$P_R(A,R,C) = \prod_{r \in R} C_r \prod_{r \in A \backslash R} 1 - C_r \tag{8}$$

Now, we have to define how to compute the probabilities C_r that topic r is mentioned by b before b'. We compute these probabilities by taking into account the total number of posts published by each blog during the time interval of the topic, in the following way:

$$C_r = \frac{Np\big(b,[t_s(r);t_e(r)]\big)}{Np\big(b,[t_s(r);t_e(r)]\big) + Np\big(b',[t_s(r);t_e(r)]\big)} \tag{9}$$

$t_s(r)$ is the time of the beginning of topic r and $t_e(r)$ is the time of its end. $Np(j, t, t')$ gives the number of posts published by blog j between times t and t'. Simply, this expression reflects the idea that, the higher the number of posts of blog b as compared to the total number of posts from both blogs in the time interval, the more likely b is to publish the first post on the topic by chance. We do not consider the overall posting rates of the blogs, as these changes over time.

The computation of the likelihood expressed in five suffers from combinatorial explosion. In fact, the number of computations that have to be performed to calculate $\lambda(\gamma(b, b') = p|A, Y, C, Z, R)$ scales exponentially with $|Y|$. For this reason, when $|Y|$ is above 15, we resort to an estimation based on sampling.

Finally, we estimate $\gamma(b, b')$ by calculating the mean of the possible values it can take ($\gamma(b, b') \rightarrow [0, 1]$), weighted by their likelihood:

$$\gamma(b,b') = \frac{\int_0^1 l(\gamma(b,b') = p \mid A,Y,C) \cdot p \cdot dp}{\int_0^1 l(\gamma(b,b') = p \mid A,Y,C) \cdot dp} \tag{10}$$

Not having an analytical solution for equation (10), we use Monte Carlo integration.

Having a way to compute dyadic precursor scores, we are now interested in scoring the blogs according to their overall precursor/laggard behaviours over the entire network. For this purpose, we will define two metrics: the global precursor score (P) and the laggard score (L).

A dyadic precursor score $\gamma(b, b')$ can be interpreted as the probability that a post from blog b' participates in a topic under a temporal relationship with blog b, where b precedes b', given that both blogs are known to participate in that topic. We can remove the topic co-participation assumption using Bayes' theorem. Considering M to be the event of the post participating in the topic under the temporal relationship, and H to be the event of the post for blog b' participating in a topic where blog b also participates:

$$\gamma(b,b') = P_r(M \mid H) \tag{11}$$

$$P_r(M \mid H) = \frac{P_r(H \mid M)P_r(M)}{P_r(H)} \tag{12}$$

$$\omega(b,b') = P_r(M) = P_r(M \mid H)P_r(H) = \gamma(b,b')P_r(H) \tag{13}$$

We will call $\omega(b, b')$ the adjusted dyadic precursor score. Notice that $P_r(H \mid M) = 1$, because if the post participates in a topic under a temporal relationship with the other blog, the blogs will necessary co-participate in that topic.

We define the global precursor score for a blog $b(P(b))$ as the mean of all adjusted dyadic precursor scores where b is the origin, and the laggard score ($L(b)$) as the mean of all adjusted dyadic precursor scores where b is the target. Being B the set of all blogs in the network:

$$P(b) = \frac{1}{|B|-1} \sum_{b' \in B \setminus \{b\}} \omega(b,b') \tag{14}$$

$$L(b) = \frac{1}{|B|-1} \sum_{b' \in B \setminus \{b\}} \omega(b,b') \tag{15}$$

Notice that these scores measure temporal relationships and not influence, and are thus robust to the existence of external influences to the network.

5 Results and discussion

The above protocol was applied to a dataset generated from a crawl of the French political blogosphere, consisting of 916 blogs, between the days of October 1st 2009 and February 11th 2010. These blogs were selected by human experts to be a good representation of the French political blogosphere. During this period, 40,191 posts were published, containing 16,909 citation links to other blogs in the network. We applied our topic detection process on this data and identified 2,619 different topics.

We then computed the dyadic precursor and adjusted dyadic precursor scores, as well as the global precursor and laggard scores according to the process described in the previous section for each blog that published at least seven posts during the whole observation period. We discarded nearly 300 blogs with very low posting rates because of the noise they may introduce into the computation of the scores.

In the next subsection, we will discuss results pertaining to dyadic scores, and in the following one the ones relating to global scores.

5.1 Dyadic scores

In Table 1, we present a number of metrics related to the dyadic precursor scores and their relationship to the URL linking between blogs.

Table 1 Dyadic precursor scores and internal URL linking

Directed blog dyads (b, b')	367,842
Dyads with URL linking $u(b, b') > 0$	466
Dyads with $\gamma(b, b') > 0$	41,050
Dyads with $\gamma(b, b') > 0$ and $\gamma(b, b') > \gamma(b, b')$	20,525
Link/precursor overlaps $\gamma(b, b') > 0$ and $u(b, b') > 0$	293
Reverse link/best precursor overlaps $\gamma(b, b') > \gamma(b', b)$ and $u(b', b) > 0$	174
Direct link/best precursor overlaps $\gamma(b, b') > \gamma(b', b)$ and $u(b, b') > 0$	119
Reverse link/best adj. precursor overlaps $\omega(b, b') > \omega(b', b)$ and $u(b', b) > 0$	160
Direct link/best adj. precursor overlaps $\omega(b, b') > \omega(b', b)$ and $u(b, b') > 0$	133
$\mathrm{P}(u(b, b') > 0)$	0.00127
$\mathrm{P}(u(b, b') > 0 \mid \gamma(b, b') > 0)$	0.00714
$\mathrm{P}(u(b, b') > 0 \mid \gamma(b, b') > \gamma(b', b))$	0.00848
$\mathrm{P}(u(b, b') > 0 \mid \gamma(b, b') > \gamma(b', b))$	0.00580
$\mathrm{P}(u(b, b') > 0 \mid \omega(b, b') > \omega(b', b))$	0.00780
$\mathrm{P}(u(b, b') > 0 \mid \omega(b, b') > \omega(b', b))$	0.00648
$\mathrm{P}(\gamma(b, b') > 0)$	0.11160
$\mathrm{P}(\gamma(b, b') > 0 \mid u(b, b') > 0)$	0.62876

We consider directed dyads (b, b') of blogs in the network. The precursor score $(\gamma(b, b'))$ and adjusted precursor score $(\omega(b, b'))$ can be known for each one of these dyads. If these scores are greater than zero for a certain dyad, we consider that b is a precursor of b' with a strength determined by the scores. Similarly, it is possible to know the URL link weight for a dyad. We consider this weight to be given by the function $u(b, b')$, and it corresponds to the number of URL links from posts of blog b to posts of blog b'.

Both the precursor scores and the link weights can be used to define networks where the blogs are the vertices. In any case, edges are considered to exist for dyads where the score or weight is greater than zero. One interesting initial observation from Table 1 is that the *precursor network* has a much greater number of edges that the *URL link network*. It is thus reasonable to assume that the former network has the potential to reveal information that is not present in the latter.

After computing the scores and weights that define the networks, we wanted to find out if there is a relationship between these networks. Taking a random dyad (b, b'), there is a probability $P(\gamma(b, b') > 0)$ that this dyad corresponds to an edge in the *precursor network*, and a probability $P(u(b, b') > 0)$ that this dyad corresponds to an edge in the *URL link network*. Consider the null hypothesis that the two networks are unrelated. If the null hypothesis is correct, then these two probabilities should be independent. That is to say: $P(u(b, b') > 0) \approx P(u(b, b') > 0 \mid \gamma(b, b') > 0)$ and $P(\gamma(b, b') > 0) \approx P(\gamma(b, b') > 0 \mid u(b, b') > 0)$. As can be seen in Table 1, this is not the case. In fact, the conditional probability $P(u(b, b') > 0 \mid \gamma(b, b') > 0)$ is more than five times higher than $P(u(b, b') > 0)$ and, conversely, the conditional probability $P(\gamma(b, b') > 0 \mid u(b, b') > 0)$ is also more than five times higher than $P(\gamma(b, b') > 0)$.

Given the nature of the dyadic precursor score, as described in Section 4, if the score is greater than zero for the directed dyad (b, b'), it is also greater than zero in the reverse direction (b', b). That is because if two blogs share a topic, there is always a probability, no matter how low, that any of the blogs is the one with the strongest temporal precedence. The scores for each direction can be, and indeed usually are, different. This difference will be greater the more likely it is that one of the blogs is the strongest precursor. Let us now consider a precursor network which only includes edges where the precursor score is the highest of both direction. We will call the *best precursors network*. An edge (b, b') belongs to this network if $\lambda(b, b') > \lambda(b', b)$. This condition is sufficient, because it also implies that $\lambda(b, b') > 0$, since all $\lambda(b, b') \geq 0$.

Comparing the *best precursors network* to the *URL link network*, we can see in Table 1 that, as with the *precursor network*, there is a higher probability for an edge to be an URL link if it is previously known to have a best precursor score. The interesting distinction is that there is an even higher probability that there will exist a reverse URL link (from b' to b) if $\lambda(b, b') > \lambda(b', b)$. This is in accordance with what we would expect from simple intuition. If blog b has a high probability of being a precursor of blog b', then it will have a higher than average probability of being an influencer of blog b', which in turn leads to a higher than average probability that b' will cite a post in b through an URL link. This result indicates that temporal precedence is in fact related to influence. Given the large amount of precedence edges as compared to URL link edges, precedence score might be used to indicate potential influence relationships that would otherwise go undetected.

Similarly to the *best precursor network*, we can construct the *best adjusted precursor network* using the adjusted dyadic precursor scores. In fact, this network presents a similar, although less accentuated relationship to the *URL link* network.

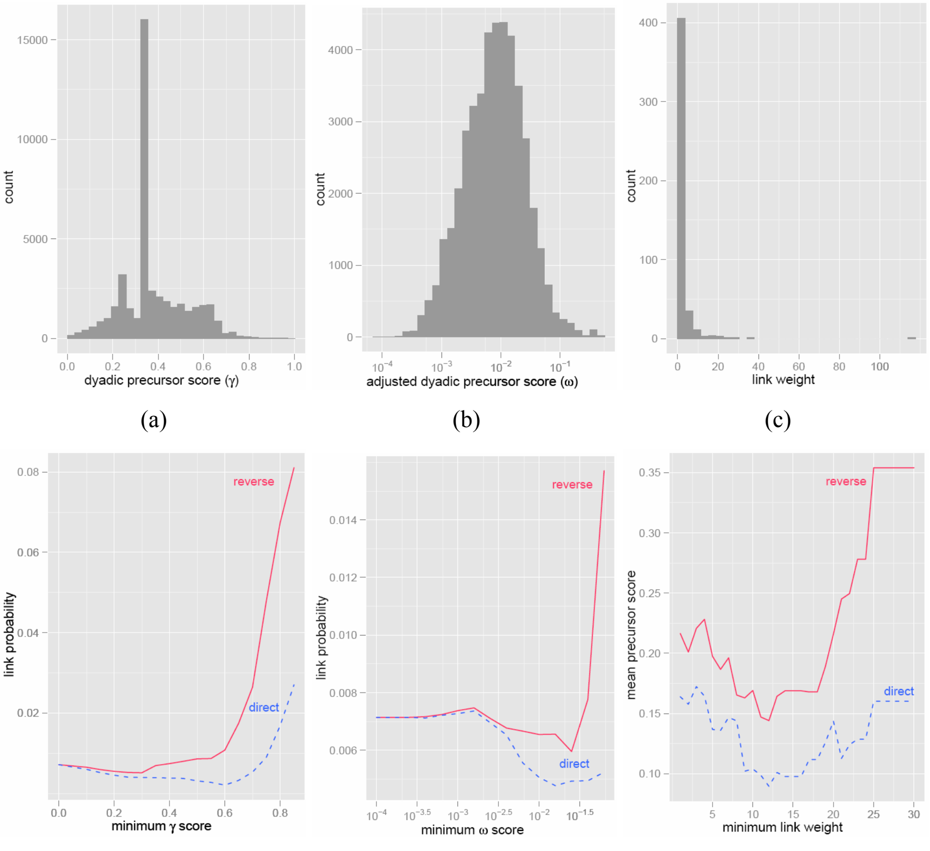
In Figure 3, we further analyse the relationships between precursor scores and URL links.

In the first column, we have a histogram showing the distribution of precursor scores on top. On the bottom, we have a plot of the probabilities of existence of direct and reverse URL links for dyads with a minimum λ score. As can be seen, both probabilities increase with minimum λ, but the reverse link probability is always equal or higher than the direct link probability. Also, the difference between the two probabilities also increases with larger minimum λs. This gives more credence to our previous hypothesis: the higher the probability of precedence, the higher the probability of influence, eventually manifested in reverse URL links.

The middle column presents a similar analysis for adjusted dyadic precursor scores and the results are analogous, although much less accentuated. The standard λ score seems to be a better predictor of reverse URL linking than the adjusted ω.

Finally, in the third column of Figure 3, we compare mean direct and reverse precursor scores to minimum URL link weights. The curves obtained are more noisy than the previous ones. Nevertheless, it is clear that the reverse mean score is always higher than the direct one, and that the mean reverse score becomes clearly higher for link weights above 25. These results are aligned with the previous ones and our proposed explanation.

Figure 3 (a) Distribution of dyadic precursor scores (γ) (b) distribution of adjusted dyadic precursor scores (ω) (c) distribution of URL link weights, where the weight is the number of links from one blog to another (d) probabilities of reverse (red, solid) and direct (blue, dashed) links, given a minimum γ score (e) probabilities of reverse (red, solid) and direct (blue, dashed) links, given a minimum ω score (f) mean reverse (red, solid) precursor score and direct (blue, dashed) precursor score given a minimum URL link weight (see online version for colours)



(a) (b) (c)

(d) (e) (f)

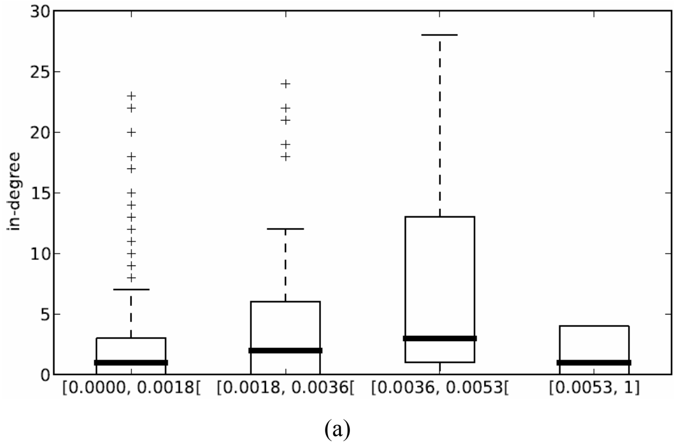
5.2 Global scores

We now turn or attention to the global precursor and laggard scores. A blog may score low on these metrics for different reasons. It could be that it does not tend to participate in popular topics (which also means that the topics it discusses are not spread through the network), or that it maintains relationships of influence with other blogs which are close to being symmetrical. This type of relationship between two blogs makes it approximately equally likely that each blog influences the other to enter a topic. Our global scores are not capable of distinguishing a symmetrical influence relationship from an indirect relationship.

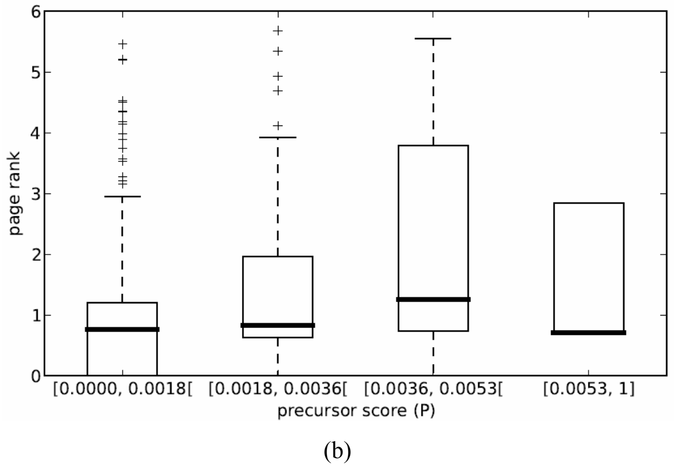
In the study of blog networks, it is common to establish popularity metrics based on the URL links that point to a blog. We compute the in-degree of a blog as the number of blogs that link to it at least once during the observation period, as well as the classical page rank. Our goal is to compare those metrics based on the topology of the hyperlinks network with our temporal semantic-based scores.

Figure 4 shows box plots of in-linking and page rank per interval of precursor score. The two plots present similar shapes, showing an increase in both in-link degrees and page ranks up to the third bar. On the fourth bar there is a clear decrease, suggesting that the precursor behaviour is positively correlated with blog popularity only up to a certain point.

Figure 4 (a) Box plots of in-linking distributions for intervals of precursor scores (b) box plots of page rank distributions for intervals of precursor scores



(a)



(b)

In Figure 5, we plot in-linking per interval of laggard score. This plot is more noisy and the pattern is less clear than the previous one. Higher laggard scores appear to have a detrimental effect on link popularity. Although not shown, a similar pattern was found when comparing page ranks to laggard scores.

Figure 5 Box plots of in-linking distributions for intervals of laggard scores

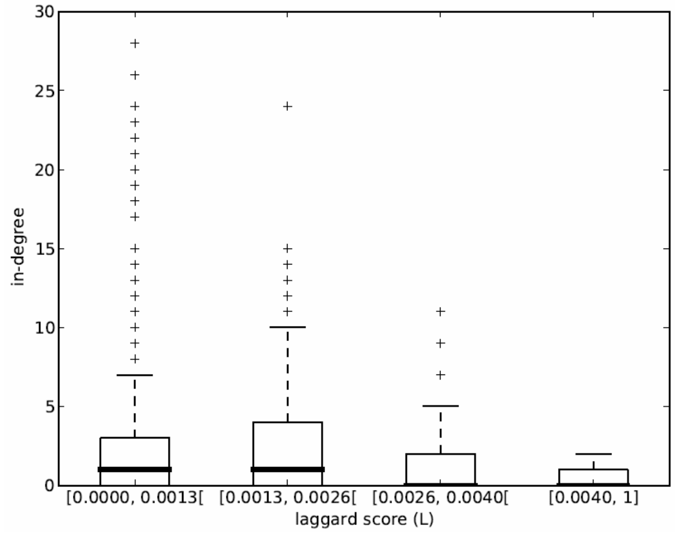


Table 2 Significance of mean in-degree relationships for classes of blogs determined according to precursor and laggard score intervals

		2.08	*6.19*	*1.59*	*3.50*
		pl	**P***l*	*p***L**	**PL**
2.08	*pl*				
6.19	**P***l*	**			
1.59	*p***L**	*	***		
3.50	**PL**			***	

In order to derive general principles, we divided the blog set into four classes. Each class is characterised by a high or low precursor score and a high or low laggard score. A precursor score is considered low if it is equal or lesser than the mean precursor score for the entire set $\left(P \in \left[0, \bar{P}\right[\right)$, and high otherwise $\left(P \in \left]\bar{P}, 1\right]\right)$. Laggard scores are classified in an analogous fashion. We use the notation *p* for low precursor, **P** for high precursor and so on. The class **P***l*, for example, is the one containing blogs with a high precursor score and low laggard score.

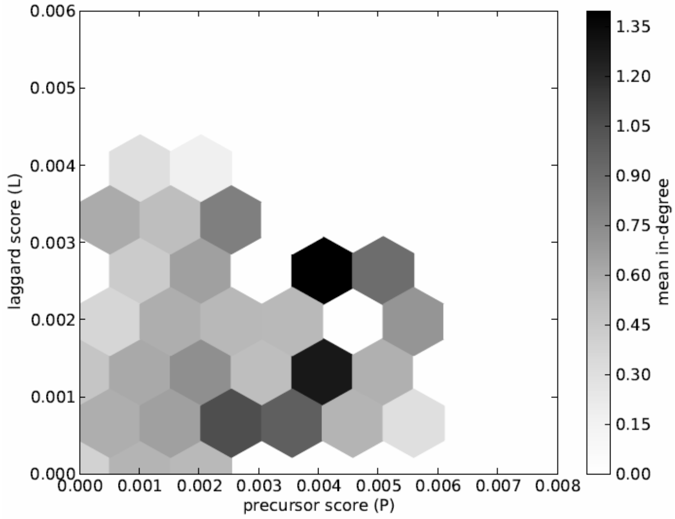
In each cell of Table 2, we perform a comparison between the mean in-link degree of each class. The statistical significance of the differences was determined using *Wilcoxon* rank sum tests. We use a number of * symbols to denote the level of significance found. One * if *p-value* < 0.05, two if *p-value* < 0.01 and three if *p-value* < 0.001. The mean in-degrees for classes are shown in row and column headers.

When comparing the two classes with low laggard scores, the one with a high precursor score has a higher mean in-degree. The same is true of the two classes with high laggard scores. When comparing the two classes with a low precursor score, the one with the low laggard score has the higher mean in-degree. In the two cases where no significance was found, the *p*-value was very close to 0.05, suggesting that the relationships are likely true, but we have insufficient data to be certain. This confirms that higher precursor scores and lower laggard scores have a positive effect on in-linking. These results also show that the two scores are not just reflecting the effect of participating in discussions. In fact, both scores require higher participation for higher values, but have opposite effects.

It is clear, however, that these general principles do not tell the whole story. The box plots show that, despite the general principles, blogs with high precursor scores are not necessarily rewarded with high in-link degrees.

This becomes more obvious by observing the hexagonal binning plot, shown in Figure 6. It displays the mean in-linking per region of precursor and laggard scores. The darker the colour, the higher the in-linking mean. It clearly confirms for example that a higher precursor score does not guarantee higher in-degree.

Figure 6 Hexagonal binning plot displaying mean in-linking per region of precursor and laggard scores



Note: The darker the colour, the higher the in-linking.

From this spatial distribution, list 1 contains the blogs closest to point (0, 0) – low precursors, low in-degree; list 2 the blogs closest to (0, 1) – low precursors, high in-degree; list 3 the blogs closest to (1, 0) – high precursors, low in-degree and list 4 the blogs closest to (1, 1) – high precursors, high in-degree.

We then provided these four lists to an expert on the French blogosphere. She had no prior knowledge of our classification process. We simply asked her if she could notice any significant pattern inside groups. She described blogs of list 1, which belong to the category of low precursor and low in-degree, as very ‘small’ blogs essentially concerned with regional or local issues. According to her, list 2 (low precursors, high in-degree) is typically composed of experienced bloggers who emerged during the last presidential election in 2007 and now gather together despite their political differences. As such their pattern of linking is similar to a ‘rich-club’ which may explain their high in-degree in spite of their low precursor score. Blogs which have high precursor score and low in-degree (list 3) are exclusively made of copycats. These sites are basically systematically relaying the media or making reviews of regular papers on the web. The presence of such behaviour in the dataset incidentally explains the sharp decline of mean in-degree and page rank among blogs with highest precursor scores that we observed previously (Figure 4). The fourth list is composed of high precursors and high in-degree blogs. All of them have been described by the expert as very active in political contestation, both from the left and the extreme right, against the government policy and, more broadly, against the current political balance.

6 Conclusions

In this work, we strived to extract quantifiable metrics from the wealth of semantic information contained in blogs. We presented a method for the detection of bursts of activity at the semantic level that was tested on a real dataset and shown capable of identifying topics characterised by n-grams and time intervals. We then described a probabilistic model to quantify temporal relationships between blogs. Dyadic precursor scores are able to quantify temporal relationships between pairs of blogs, where one tends to enter a topic before the other, discounting the effects of asymmetrical posting rates.

From the dyadic precursor scores, we generated a network that we then compared with the citation network. This comparison revealed interesting results. The existence of a precursor relationship between two blogs is a significant indicator of a greater probability of the existence of a citation link in any of the directions, but higher in the reverse direction. Considering only the direction of the precursor score with the highest value in a dyad, the relationship with the reverse direction citation probability is even more accentuated. This result is intuitive: the higher the probability that blog A precedes blog B in addressing topics, the more likely that blog B will link to A. This result is particularly relevant because no information about URL linking is used when computing the precedence scores. This result is both a validation of the methods we propose and an indication that our methods can reveal additional information about the blog network, since the precedence network is much more connection-dense than the citation network.

From the dyadic scores we derived two scores to classify blogs according to their overall precursor and laggard behaviours. The comparison of these semantic temporal metrics with the more traditional in-link degree-based popularity metrics revealed non-trivial relationships between the two. The expert assessment indicates that the scores we proposed lead to relevant distinctions that could not be derived from classical structural-based methods only. Search engine ranking algorithms, like the well-known PageRank (Page et al., 1999) used by Google, are more sophisticated than simple reliance on URL link in-degrees. However, they are still based on structural aspects of the web, deriving their estimations from the analysis of the network of URL links. We found that the precursor/laggard scores are able to identify blogs that have a high tendency to be precursors in topics under discussion, but that would likely not be distinguishable from other blogs with similar page ranks or in-degrees by relying only on this later type of metric. It is conceivable that search engine ranking algorithms could be improved with the approach we propose. Including precursor scores in ranking metrics could help improve the quality of searches, for example the ones related to time sensitive events. It could also reward blogs that generate influential content, but that are not especially popular in the sense of receiving many in-links.

Acknowledgements

This work has been partially supported by the French National Agency of Research (ANR) through grant 'Webfluence' #ANR-08-SYSC-009. The authors would like to thank Clémence Lerondeau and Guilhem Fouetillou from Linkfluence for their help in qualitatively assessing the data.

References

Adar, E., Zhang, L., Adamic, L.A. and Lukose, R.M. (2004) 'Implicit structure and the dynamics of blogspace', *Workshop Weblogging Ecosystem, 13th WWW.*

Asur, S. and Huberman, B.A. (2010) 'Predicting the future with social media', arXiv.org e-print archive: 1003.5699.

Balog, K., Mishne, G. and de Rijke, M. (2006) 'Why are they excited? Identifying and explaining spikes in blog mood levels', *Proc. 11th Meeting Eur. chapter of the Association for Comp. Ling. EACL*, pp.207–210.

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, K. (2010) *Measuring user Influence in Twitter: The Million Follower Fallacy.*

Cointet, J-P. and Roth, C. (2009) 'Socio-semantic dynamics in a blog network', *IEEE Intl. Conf. Social Computing*, pp.114–121.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2009) 'Detecting influenza epidemics using search engine query data', *Nature*, Vol. 457, pp.1012–1014.

Gruhl, D., Guha, R., Liben-Nowell, D. and Tomkins, A. (2004) 'Information diffusion through blogspace', *WWW2004: Proc. 13th Intl Conf on World Wide Web*, 17–22 May, NYC, NY, USA, pp.491–501.

Java, A. (2006) 'Tracking influence and opinions in social media', *Ebiquity*, November, p.69.

Java, A., Kolari, P., Finin, T. and Oates, T. (2006) 'Modeling the spread of influence on the blogosphere', *Proceedings of the 15th International World Wide Web*, May, p.7.

Kleinberg, J. (2002) 'Bursty and hierarchical structure in streams', *Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, ACM, New York, NY, USA, pp.91–101.

Kossinets, G., Kleinberg, J. and Watts, D.J. (2008) 'The structure of information pathways in a social communication network', *Proc. SIGKDD 08.*

Kumar, R., Novak, J., Raghavan, P. and Tomkins, A. (2005) 'On the bursty evolution of blogspace', *World Wide Web*, Vol. 8, pp.159–178.

Lenclud, G. (1998) 'La culture s'attrape-t-elle?', *Communications*, Vol. 66, pp.165–183.

Leskovec, J., Backstrom, L. and Kleinberg, J. (2009) 'Meme-tracking and the dynamics of the news cycle', *Proc. ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining.*

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N. and Hurst, M. (2007) 'Cascading behavior in large blog graphs', *Proc. 7th SIAM Intl. Conf. on Data Mining (SDM).*

Lloyd, L., Kaulgud, P. and Skiena, S. (2006) 'Newspapers vs. blogs: who gets the scoop?', *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs (AAAI-CAAW)*, Palo Alto, California, USA.

Mishne, G. and de Rijke, M. (2006) 'Capturing global mood levels using blog posts', *AAAI 2006 Spring Symp. on Computational Approaches to Analysing Weblogs.*

Page, L., Brin, S., Motwani, R. and Winograd, T. (1999) 'The PageRank citation ranking: bringing order to the web', Technical Report 1999-66, Stanford InfoLab.

Schmid, H. (1994) 'Probabilistic part-of-speech tagging using decision trees'.

Sperber, D. (1996) *Explaining Culture: A Naturalistic Approach*, Blackwell Publishers, Oxford.

Sperber, D. and Claidiére, N. (2006) 'Why modeling cultural evolution is still such a challenge', *Biological Theory*, Vol. 1, No. 1, pp.20–22.

Valente, T. (1996) 'Social network thresholds in the diffusion of innovations', *Social Networks*, Vol. 18, pp.69–89.

Watts, D.J. and Dodds, P.S. (2007) 'Influentials, networks, and public opinion formation', *Journal of Consumer Research*, Vol. 34, No. 4, pp.441–458.

Zhou, D., Ji, X., Zha, H. and Giles, C.L. (2006) 'Topic evolution and social interactions: how authors effect research', *Proc. CIKM'06*.