Socio-semantic dynamics in a blog network

Jean-Philippe Cointet, Camille Roth

Socio-semantic dynamics in a blog network

Jean-Philippe Cointet
CREA & TSV
CNRS-Ecole Polytechnique & INRA
ISC - 57-59, rue Lhomond
F-75005 Paris, France
cointet@shs.polytechnique.fr

Camille Roth
CAMS
CNRS-EHESS
54, bd Raspail
F-75006 Paris, France
roth@ehess.fr

***Abstract*—The blogosphere can be construed as a knowledge network made of bloggers who are interacting through a social network to share, exchange or produce information. We claim that the social and semantic dimensions are essentially co-determined and propose to investigate the co-evolutionary dynamics of the blogosphere by examining two intertwined issues: first, how does knowledge distribution drive new interactions and thus influence the social network topology? Second, which role structural network properties play in the information circulation in the system? We adopt an empirical standpoint by analyzing the semantic and social activity of a portion of the US political blogosphere, monitored on a period of four months.**

I. The "blogosphere" as a socio-semantic system

The blogosphere essentially gathers individuals who share, exchange and produce information and interact online by posting comments or referencing each other. As such, it is a socio-semantic network, in the sense that each blog can be characterized both by a relational profile, determined by its position in the underlying social network, and by a semantic profile, which describes cognitive attributes.

Adopting a dual perspective on these knowledge networks is likely to provide a better knowledge of the key mechanisms underlying their organization and evolution: essentially, both dimensions co-evolve, for instance network dynamics is likely to be affected by the distribution of knowledge, if we assume that semantic homophily is a driving force behind network evolution. Structural features of the implicit social network may also give rise to some specific patterns regarding knowledge distribution. Put differently, by supporting diffusion processes social networks may diversely affect information circulation among bloggers.

We propose to investigate empirically the coevolutionary dynamics of a portion of the blogosphere by examining the two intertwined following issues:

(i) how does knowledge distribution influence new relationship appearance, thereby influencing the topology?
(ii) how, in turn, do structural network properties play a role in the way information circulates in the system?

Related work

Blogs attracted much attention as an empirical goldmine for quantitative social science and, more theoretically, as a rich instance of social and semantic complex system (1; 2; 3; 4; 5; 6; 7; 8). This recent effort is part of a broader interest in online knowledge-based networks, including for instance wikis (9) or content-sharing websites such as Flickr (10), which fundamentally are virtual spaces dedicated to production, sharing, and circulation of opinion, multimedia resource and more broadly information; and where various kinds of social interactions and collaborations are channeled by so-called "web 2.0" technologies.
Political blogging itself is also the focus of a decent part of the literature in that it allows investigation of multiple current issues, including influence of bloggers over media coverage or over the general political debate (11; 12; 13; 14).

Many of these studies focus on the *blogosphere* or *blogspace* with a social network perspective, aiming at measuring and characterizing topological properties including link configurations, cohesiveness phenomena and existence of groups or communities (15; 16; 17). Beyond a strictly structural approach, static descriptions of the joint distribution of topics and social configuration of a blogosphere has been achieved by (11); however the dynamic interrelations of these two dimensions remains a current problem.
Further, studies considering the blogosphere as an informational system have mainly focused on investigating topic and opinion evolution — thanks to the fine-grained dynamics of the underlying data — thereby developing automatic trend detection methods (3), characterizing opinion dynamics using sentiment analysis (18), or exploring the coexistence of chatters and spikes in blog conversations, and their cyclic behaviors (19; 20; 21), *inter alia*.

Blogs and more generally Internet-based communication systems have provided a novel opportunity for diffusion studies, through the in-vivo observation of what is generally refered to as "cascade dynamics" (22; 23).
This feature is common notably to viral marketing studies on large-scale online datasets exhibiting diffusion phenomena; including (6) which explores the distribution of the probability of purchasing a cultural consumer good when a large on-line retailer user receives a certain number of recommendations sent by her friends; and (24) computes the probability for one to join a Livejournal community when she already has some friends in it. More specifically cascades in blog networks

978-0-7695-3823-5/09 $26.00 © 2009 IEEE
DOI 10.1109/CSE.2009.105 114

have been extensively described by considering chains of posts citing each other as information pathways (7; 25; 17). In these cases as well as in other studies not restrained to blog networks (26) the focus has been put on influence spread through the study of the *topological* properties of cascades (such as typical patterns of cascade, distribution of cascade sizes, etc.)

Eventually, little is known yet on the dynamic underpinnings of content distribution over agents with respect to topology and on the processes underlying the actual formation of heterogeneous topical communities; or, more broadly, on the very intertwining of social and semantic dimensions and their effect on information propagation, with the notable exception of (1).

Most often, one only of the social and the semantic dimensions is considered. On one hand indeed, link creation patterns are generally essentially appraised through structural attributes rather than cognitive/semantic properties of blogs. As for diffusion, either *content* evolution is studied independently of the topology, or topology is the only reference frame for diffusion (one observes the propagation of links *of* the social network *along* the social network — i.e. some sort of structural transitivity). On the other hand, endeavors at understanding what triggers or increases diffusion have given a prevailing role to ego-centered characterizations (i.e. diffusion is often seen as stemming from individual properties, rather than the shape of the network at large).

Put shortly, in terms of diffusion, taking into account both the network structure and a transmission process on objects *distinct* of this structure is so far a current challenge. In this respect, we also aim at assessing how actual content diffusion pathways can be correlated with the (mostly distinct) underlying social network that supports such information circulation.

Outline

The paper is organized as follows: in the next section we first introduce the empirical protocol. Section III focuses on the dynamics of link creation in the comment and post networks according to both structural and semantic features, while Sec. IV investigates the dynamics of information propagation according to the underlying topology.

II. Experimental framework

A. A bounded subset of the US blogosphere

Our study is based on the observation of the activity of a medium-sized yet topically well-bounded portion of the US political blogosphere which has been gathered by LINKFLUENCE under the "PresidentialWatch08" project.[1]

The dataset consists of $1,066$ blogs, hereafter denoted by $\mathcal{B}$, monitored over the course of four months, from Nov 1, 2007 to Feb 29, 2008. For each blogger we crawled the date and full-text content, including hyperlinks, of each post published during the observation period, totaling $71,376$ posts.

[1]http://linkfluence.net, http://presidentialwatch08.com

B. A dynamic network

The couple $(\mathcal{B}, C)$ is the blog network, where C denotes post citation links as an adjacency matrix of size $|\mathcal{B}| \times |\mathcal{B}|$. This data is additionally *dynamic*, with a temporal granularity of one day: we deal with C_t, where t ranges from 1 to 121: $C_t(i,j) = 1$ if i cites j in a post at time t; 0 otherwise. In the remainder, t may be omitted in the notations when it is implicit.

We extracted $229,736$ dated edges in C, of which $15,032$ are unique (non-repeated links). We eventually define an aggregated weighted network as $\mathbf{C}_t = \sum_{t'=1}^{t} C_{t'}$.

C. An epistemic network

Aside of this structure, content defines a semantic dimension: posts are traditionally dealing with specific issues, sometimes broadcasting particular documents. Although the existence of a clear-cut distinction between high-level topics and specific cultural items may be debated, we assume that (i) textual contents broadly define the various issues a blogger addresses, whereas (ii) explicit URLs (which refer to hyperlinked documents and which are not citation links) define the various specific digital resources a blogger spreads around.

Subsequently, we distinguish:

- a set of high-level topics $\mathcal{W}$ relevantly linked to political commentary in our context, among the most frequent in the corpus (thus excluding rhetorical terms). $\mathcal{W}$ is thus made of 79 syntagms ranging from names of politicians to issues which kept the blogosphere busy during the presidential campaign, such as "*climate change*", "*national security*", "*super Tuesday*", "*tax cuts*", "*human rights*", etc.
- and a set of URLs, noted $\mathcal{U}$, which are not confusable with a link in the citation network — these are simply online videos, news media article, etc. $\mathcal{U}$ is a selection of $96,637$ URLs (of length larger than 10 characters). Note that these URLs are taken from the limited content of *posts only*, not webpages, so that $\mathcal{U}$ should exclude banners and platform-related links and ads, *inter alia*; it only covers links explicitly cited by bloggers in their posts.

More precisely with respect to $\mathcal{W}$, we introduce a temporal matrix W_t which tracks the contents published by bloggers: $W_t(i,w)$ equals 1 if term $w \in \mathcal{W}$ appears in a post published on blog i at time t, 0 otherwise. Eventually, the $|\mathcal{W}|$-dimensional vector $\mathbf{W}_t(i)$ defined as a the sum of rows $W_{t'}(i)$ for $t' \leq t$ denotes the aggregation of all topics addressed by blog i until t.

$\mathbf{W}_t(i)$ can be seen as the semantic profile of i at t.

In a similar fashion for $\mathcal{U}$, we introduce a temporal matrix $_t$ such that $\_t(i,u)$ equals 1 if blog i explicitly refers to a URL $u \in \mathcal{U}$ in a post published at time t. Since this matrix will mostly be used for diffusion purposes, we need not define in the present study an aggregated quantity for URL usage.

III. Evolution of topology:

THE CONTENT-BASED DYNAMICS OF LINK CREATION

We first study the link creation dynamics with respect to the configuration of the blogosphere, both on a social and semantic level. In particular, we examine the constraint induced by the current socio-semantic network on future citation patterns. The structure of both social and semantic configurations may, at least partially, determine link creations. Quite straightforwardly, remoteness in both spaces is likely to modify the landscape of potential relations and, subsequently, modify the likelihood of interaction. In what follows we focus notably on citation propensity with respect to simple notions of topological as well as semantic distances: how do proximity, increased attention or homophily processes actually impact authority attribution in this portion of the blogosphere?

A. Proximity and distance

To begin with, we define a series of simple distances which are all based on aggregated data at t, denoted by "bold" notations ($\mathbf{C}$ and $\mathbf{W}$); that is, we assume each notion of distance between two blogs to depend on the whole history of posting and linking at t.

1) Dissimilarity as a semantic distance: To semantically compare a pair of bloggers i and j at t, we adopt a classical cosine-based measure of dissimilarity on their profile vectors $\mathbf{W}_t(i)$ and $\mathbf{W}_t(j)$. We denote this **semantic distance** by δ: concretely, identical profiles yield a δ of 0, whereas strictly disjoint/orthogonal profiles are separated by a δ of 1; intermediate values from 0 to 1 indicate increasing levels of dissimilarity.[2]

2) Topological distances: Because network links are oriented, topological distances will be asymmetric measures, contrarily to the semantic distance δ.

We first classically define the **social distance** $d_t(i,j)$ between two blogs i and j in $\mathbf{C}$ as the length of the shortest path linking i to j in that network, irrespective of link weights. This basically refers to the number of steps one has to follow to reach another blog. On the example of Fig. 1-left, $d_t(b,f) = 3$.

[2]To this end, we first need to carry a normalization procedure to weight term occurrences properly, following the "tf·idf" canonical approach used extensively in information retrieval, famously introduced in the vector-space model (27). This approach more precisely consists in weighting the "term frequency", "tf" (so that most used terms in a given blog are more important) with the so-called "inverse document frequency", "idf", or frequency of the term in the whole corpus of blogs (so that rarer terms in the blogosphere are weighted more: this takes into account the discriminating power of terms which, while usually rare in the corpus, are being abnormally mentioned by a given blog).

For this computation, profiles $\mathbf{W}_t(i)$ are thus actually replaced by tf·idf-adjusted profiles $\hat{\mathbf{W}}_t(i)$ such that:

$$\hat{\mathbf{W}}_t(i,w) := \frac{\mathbf{W}_t(i,w)}{\sum_{w=1}^{|\mathcal{W}|} \mathbf{W}_t(i,w)} \cdot \log \frac{|\mathcal{B}|}{|\{j, \mathbf{W}_t(j,w) > 0\}|}$$

where the "log" part of the formula is the inverse ratio of the number of blogs where term w appears over the total number of blogs. Then, we obtain the dissimilarity between blogs i and j by dividing the scalar product of their adjusted profiles by the product of their norm:

$$\delta_t(i,j) = 1 - \frac{\hat{\mathbf{W}}_t(i) \cdot \hat{\mathbf{W}}_t(j)}{\|\hat{\mathbf{W}}_t(i)\| \|\hat{\mathbf{W}}_t(j)\|} \quad (1)$$

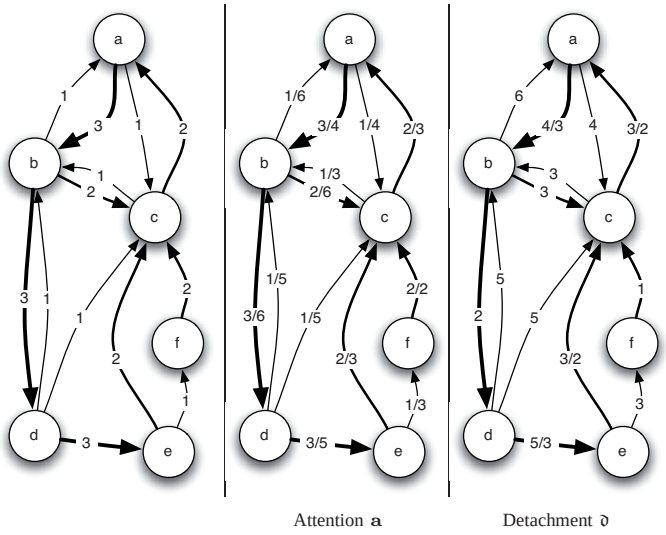

Fig. 1. *Left:* An example of weighted citation network $\mathbf{C}_t$: weights trivially correspond to the number of observed links between blogs at some time t. *Middle* and *right:* corresponding attention and detachment values, respectively. For example, blog b cited c twice out of a total of $1+2+3=6$ citation links, its attention toward c is thus $\mathbf{a}_t(b,c) = \frac{2}{6}$. *Detachment* $\mathfrak{d}_t(b,c)$ equals the inverse of the attention from b to c, it is 3. *Detachment-based distance* $\partial_t(b,c)$ is also 3 (since $b-c$ is the shortest weighted path from b to c), while $\partial_t(b,e) = 2 + 5/3 = 11/3$.

Since influence effects relate to attentional features, we suggest that a notion of remoteness based on "attention" may also be relevant. In this respect, we define a dyadic attention $\mathbf{a}_t$ by normalizing every row of $\mathbf{C}_t$:

$$\mathbf{a}_t(i,j) = \frac{\mathbf{C}_t(i,j)}{\sum_{j=1}^{|\mathcal{B}|} \mathbf{C}_t(i,j)}$$

$\mathbf{a}_t(i,j)$ is thus simply the proportion of links going from i to j among all outgoing links from i. Higher values indicate higher focus by i on j. Note that a similar notion is called "influence matrix" in (15).

Now, we can define an opposite notion to attention by considering inverse values of $\mathbf{a}$, defining a measure of *detachment* as $\mathfrak{d}(i,j) = \frac{1}{\mathbf{a}(i,j)}$. In other words, $\mathfrak{d}(i,j)$ can be compared with a relative cost for information to reach i directly from j. It is equal to infinity if there is no link from i to j, it is decreasing when attention of i towards j is growing. Basically, for instance, if i has three times more links towards j than towards k, then i's detachment to j is three times lower.

Eventually, we define a **detachment-based distance** as the minimal weighted distance (28) in a weighted graph $\mathfrak{D}$ where link weights from i to j are non-infinite $\mathfrak{d}(i,j)$ values. We denote this detachment-based distance $\partial(i,j)$ — as such, it can be considered as a measure of attentional remoteness, i.e. lightweight attentional paths will correspond to higher detachment-based distances. See an illustration on Fig. 1.

B. Method for appraising preferential link creation

While sophisticated regression models have been developed in mathematical social science to measure the preference of

link creation (29), we stick here to a basic yet insightful framework for comparing (i) the number of links actually received by some kinds of nodes during a period of time, with (ii) the potential number of such links — i.e. a kind of "preferential attachment" measurement (30), here with respect to any kind of property (31).

More precisely, we define $f(x)$ as the propensity of formation of new citation links (i,j) such that their social distance is $d(i,j)=x$. *Put simply, higher propensity values indicate stronger likelihood for dyads at a certain distance to form, all other things being equal.*

We concretely compute the propensity $f(x)$ as the proportion of new links appearing in $\mathbf{C}$ during a given time period $[t+1,\, t+T]$ and which were at social distance x at t, among the whole set of possible such pairs at distance x:

$$f(x) = \frac{\left|\left\{ \begin{array}{r}(i,j) \text{ such that } \mathbf{C}_{t+T}(i,j) > \mathbf{C}_t(i,j) \\ \text{and } d(i,j)=x \end{array} \right\}\right|}{|\{(i,j) \text{ such that } d(i,j)=x\}|} \quad (2)$$

Empirically, we estimate various propensities for a series of time steps $\left\{[t_k+1, t_k+T] \text{ such that } t_k = 60+kT, T=7\right\}_{k\in\{0,\ldots,7\}}$ — basically estimating the propensity at a weekly rate, given all previous observed interactions, with the exception that we start the computation only after an initialization period of two months ($t_0=60$).

Propensities with respect to the social, detachment-based and semantic distances are respectively denoted f, f^{∂} and g. In the figures, all propensities are normalized for comparison purposes.

C. Linking and social distance

We first analyze the effect of topological distances on new citation creation by using the plain social distance d and the detachment-based distance ∂. Figure 2 depicts the results for the social-distance-based propension f; the trend for f^{∂} is essentially similar, although not depicted here due to length constraints. On the whole, both propensity profiles are strongly and generally exponentially decreasing with higher distances,

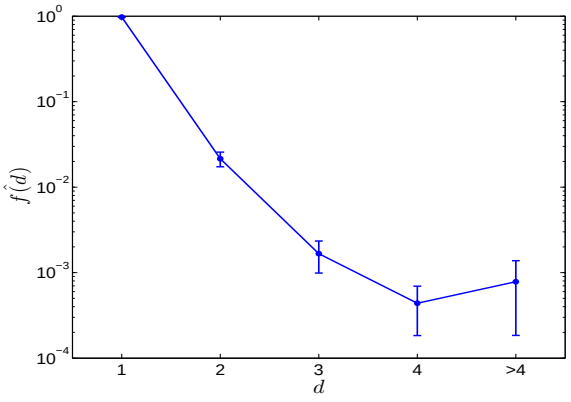

Fig. 2. Propensity f for new post citation in $\mathbf{C}$ as a function of social distance d. Error bars indicate 95%-confidence intervals on means.

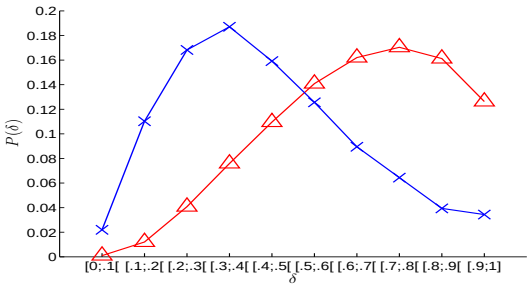

Fig. 3. Semantic distance distributions. *Triangles:* distribution computed over the whole set of possible pairs of blogs. *Crosses:* distribution computed on pairs of blogs actually linked in the citation network $\mathbf{C}$.

reflecting the effect on link creation likelihood of structural/topological and attentional remoteness: link creation basically occurs in the topological neighborhood, often not much farther than a couple of clicks away. Interestingly, above a certain threshold propensities stop decreasing: in other words, below a certain level of closeness, all bloggers are equally remote.

Specifically in terms of social distance, propensities are about at least one order of magnitude larger at distance 1 than other distances, for all networks. Links at distance 1 are actually repeated links, indicating that most relationships, by large, tend to occur between already connected bloggers; then, secondarily, towards friends of friends. Rather than speaking of a "small-world", in this case, one would rather talk of a "narrow-world" (32). When new links are established outside this close circle, the propensity to cite decreases particularly steeply with respect to social distance. Eventually propensities relative to detachment-based distances, while indicative of weighted attention-related processes, still mostly exhibit the same behavior and confirm these topogical effects.

D. Linking and semantic distance

Topology thus self-influences topology, yet content distribution may admittedly play a role in further shaping network structure; (11) demonstrated for instance how partisan divides corresponded to structural ones in the political blogosphere prior to 2004 US elections.

Here, we can first appraise homophily *statically*, or *a posteriori*, by observing the configuration of links already present at t. We therefore measure the semantic distance δ between blogs, distinguishing the whole blogosphere from the immediate neighborhood of blogs. We observe on Fig. 3 that the immediate neighborhood is very significantly closer semantically, when compared with the overall semantic distance between pairs of blogs of the whole set $\mathcal{B}$, indicating a very strong *a posteriori* homophily.

This fact suggests a strong homophilic behavior in link creation itself; in other words, it indicates a *dynamic*, or *a priori*, homophily. To check this, we compute propensities for link creation with respect to the semantic distance. The results are plotted on Fig. 4 and clearly confirm the above hypothesis. For instance, blogs at a semantic distance less than .2 will have a likeliness to cite each other about 10 times higher than blogs

at an average semantic distance and 100 times than couple of blogs strongly differing semantically.

Topological coevolution: To appraise how the social and semantic effects mix together, we finally compute propensities in a two-variable setting based on both social and semantic distances, as shown on Fig. 5. The main conclusion is that, outside of the close circle of repeated citations ($d = 1$), the above-mentioned homophilic behavior has a sensible effect, even stronger with increasing social distances. In the case of neighbors however (i.e. repeated citations), the semantic distance has a mixed role. Citations are indeed more likely towards very similar blogs, again ($\delta \in [0; 0.2[$), yet, it is also more and even much more likely towards very dissimilar blogs ($\delta \in [0.8; 1]$).

IV. Evolution of Content: The Topology-based Dynamics of Diffusion

Topology thus evolves with respect to content distribution. Yet, in a dual manner, how does the dynamics of content circulation depend on topological features? To assess this, we first need to introduce a notion of diffusion subgraphs (Sec. IV-A) and some specific characteristics of the underlying citation networks which may be likely to influence the diffusion phenomena, particularly attention-related features (Sec. IV-B).

A. Diffusion subgraph

More precisely, we focus on explicit diffusion events, which correspond to simultaneously posting some content and referring to another blog which already posted about this same content. We therefore define the notion of **diffusion subgraph**, which gathers every blog which mentioned a given URL in a post, and every directed link (i, j) between these blogs *such that* i simultaneously mentions the URL and refers to j which had already, previously, mentioned that URL.
Technically, given a resource $u \in \mathcal{U}$, we define σ_u the *diffusion subgraph of* u as a pair of:

- blogs mentioning u in a post, and

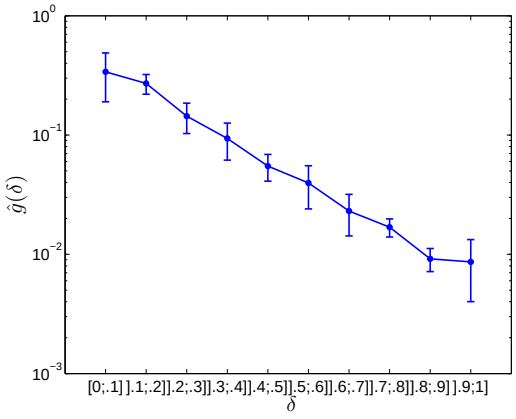

Fig. 4. Propensity for new post citation with respect to semantic distance δ.

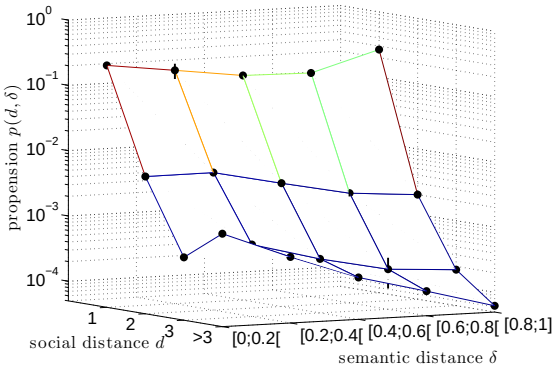

Fig. 5. Two-dimensional propensity with respect to social and semantic distances.

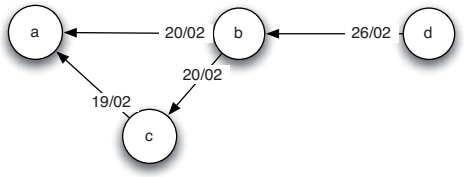

Fig. 6. Illustration of a diffusion subgraph σ_{u_0}. Date labels indicate the time when the origin blog both mentioned u_0 and did a post citation to the destination.

- directed edges (i, j) of $\mathbf{C}$ such that i simultaneously both cited j and mentioned u, *after* j mentioned u.
  Formally, these **transmission links** are edges (i, j) such that $\mathbf{C}_t(i, j) > \mathbf{C}_{t-1}(i, j)$ (i.e. there is a new link in $\mathbf{C}_t$ from i to j at t), $\_t(i, u) = 1$ (i mentions u at t) and $\exists t' < t$, $\_t'(j, u) = 1$ (i.e. j had mentioned u strictly before t).

We denote such subgraphs $\sigma_u \in \mathcal{P}(\mathcal{B}) \times \mathcal{P}(\mathcal{B} \times \mathcal{B})$.

We say that a diffusion subgraph is *trivial* if its edge set is empty, i.e. if the corresponding URL is not involved in any explicit diffusion event between two blogs. Of the $96,637$ URLs of $\mathcal{U}$, only $11,709$ correspond to non-trivial diffusion subgraphs over the whole collection period. In the remainder, we only focus on these non-trivial subgraphs.

Figure 6 provides an illustration of a real, non-trivial diffusion subgraph, whose underlying post citation network has previously been illustrated on Fig. 1. In this case, a given URL u_0 is first mentioned in blog a. It is then mentioned by c on Feb 19, who cites a on the same day. It then "diffuses" to b both from a and c on the next day. Eventually, blog d mentions u_0 along with a reference to b on Feb 26.

We plotted on Fig. 7 the size distributions of the $11,709$ non-trivial diffusion subgraphs. Sizes are sensibly heterogeneous both in terms of nodes and links, with a large number of small subgraphs (this observation is consistent with the shape of the cascade size found in (6)). Most of these subgraphs $(7,016)$ consist of a unique transmission event — 2 blogs and one link — while there are $39,540$ transmission events, over a total of $229,736$ citation links, i.e. slightly more than one

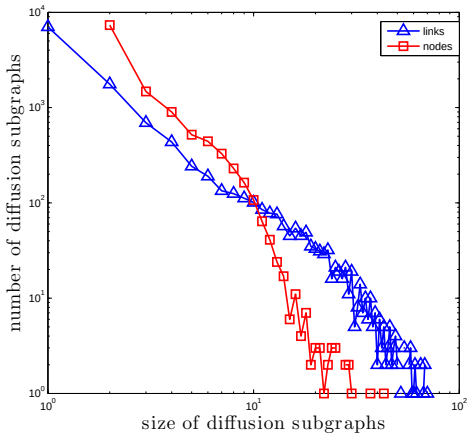

Fig. 7. Size distributions of diffusion subgraphs, in terms of nodes (*red squares*) and links (*blue triangles*).

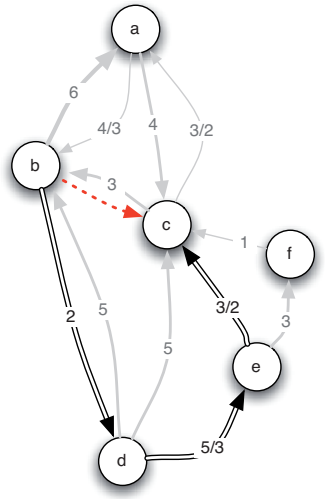

Fig. 8. *Edge-range calculation:* we compute for instance edge-range $\mathbf{r}(b,c)$. The link between from b to c is first removed before computing the minimal-cost path from b to c, using detachment values $\mathfrak{d}$ computed on $\mathbf{C}$. On this example, the path is $(b-d-e-c)$ and we have $\mathbf{r}(b,c) = \frac{31}{6}$. Note that paths with less steps such as $(b-a-c)$ may happen to be actually more expansive (with a cost of 10 in this very case).

sixth of post citations are also transmission links.

B. Diffusion-driven topological features

1) Total attention: A quite simple ego-centered measure likely to be relevant to study diffusion relates to the notion of **total attention** exerted by a blog j, defined as the sum of attentions exerted on all "attentive" blogs i:

$$\alpha_t(j) = \sum_i \mathbf{a}_t(i,j)$$

.

On Fig. 1, the total attention $\alpha_t(c)$ exerted by blog c aggregates attentions from blogs b, a, d, e and f towards c, it is equal to 2.45.

2) Edge-range distance: In addition, we now need a notion of structural distance that captures a feeling of remoteness between nodes *already* connected, obviously because the study of explicit diffusion is based on blogs which explicitly link towards and are thereby connected to other blogs. To this end, we use the notion of **edge range**, which has been notably recently used in diffusion studies in (33) and which had been initially defined in (34) for a link (i,j) as the distance between i and j if link (i,j) were removed.

We extend this notion to the case of a graph weighted with detachment values. Formally, we define edge range $\mathbf{r}(i,j)$ of link (i,j) in the weighted detachment-based graph $\mathfrak{D}$ as the minimal weighted distance between i and j when link (i,j) is removed.

In other terms, it is the minimal sum of detachment values along the "best" indirect path from i to j; or, so to say, the minimal total attentional cost an information requires to travel from a blog j to i if the edge from i to j were removed. More simply, it is also the detachment-based distance in a graph where edge (i,j) has been removed.

See an example of edge range calculation on Fig. 8.

C. Information relaying and attention

The likeliness of a blog to be influent, by inducing content diffusion, is often said to be directly related to the number of links which flows into it (35; 15) — influential bloggers being those who have more incoming links or those who have the largest audience. Following this standpoint, one can check the influence of ego by examining how ego-centered measures correlate with actual diffusion.

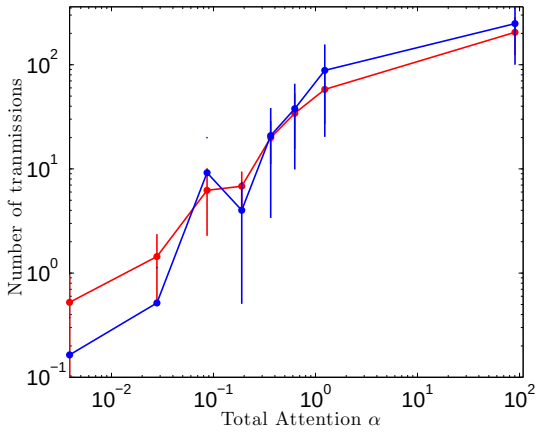

Fig. 9. Mean number of first (blue dots) and second (red dots) transmission links produced by initiating blogs depending on their total attention α. (NB: distributions are plotted using 8 quantiles of α values to accommodate for their sensibly heterogeneous spread).

As a first step, we check the correlation between the *total attention* of a blogger using a URL for a first time, and the transmission links s/he induces, i.e. as an *originator* in the corresponding diffusion subgraph. Figure 9 therefore depicts in blue the mean absolute number of such *first* transmission

links provoked by blogs having a given total attention α.[3]

Higher total attention values are indeed correlated with a larger number of transmission events. In other words, more "influential" blogs seem basically and unsurprisingly to be those with larger active readership, broadly speaking. However, influence appears to increase more than linearly for *total attention* values in the range of $5 \cdot 10^{-2}$ to 1, compared with total attentions below $5 \cdot 10^{-2}$. This suggests that there is an accumulative benefit of having a larger total attention; however, this effect seems to be bounded as it vanishes for even higher values: above a certain threshold, the increase in influence is flatter, although still relatively increasing. On the whole, this "broken" shape suggests that the influence of an initiator, as measured by the number of first transmission links, is not a direct, linear result of attention.

D. Information shortcuts and edge range

Beyond underlining immediate readership effects, i.e. somehow emphasizing that information transmission through citation is more frequent among regularly cited blogs, this kind of strictly ego-centered indicators is likely to provide little knowledge on a wider picture of information pathways; i.e. of propagation flows in terms of what makes an information propagate *more broadly*, in a *wider* arena.

1) Second transmissions: To explore this, we choose to focus on "second transmissions" in diffusion subgraphs. In what precedes, we indeed exhibited that first transmissions were likely to be initiated by blogs having a large *total attention*. First transmissions are relative to a given initial source — i.e. an initiator of a diffusion subgraph, who mentions a resource u without citing another blogger who mentioned u beforehand — while second transmissions are relative to a blog which already relays a resource. In other terms, it relates to the *longevity* of the diffusion phenomenon. Put simply, once a resource has been transmitted, how likely is it to pursue its way into the blogosphere?

As can be infered from the red curve on Fig. 9, the effectiveness of second transmissions are determined by the attention of the initiator in roughly the same way as first transmissions were; in other words, attention does not inform us more on the longevity of the informational cascade.

2) Weak ties and edge range: Rather, information spreading could depend on more holistic features related to the position of the pairs of individuals in the network: consistently with the vast amount of sociological literature on diffusion, information propagation could be more efficient along "weak ties" connecting remote areas of a network (36). In this respect, we use edge range values as they provide a less local information than ego-centered attentional profiles. Higher edge range values are indeed typical of pairs of blogs which would otherwise be relatively far apart within the network, in terms of informational and attentional pathways, if the link between them were absent. As such, higher values loosely indicate weak ties (37).

[3]Although not depicted here, we found similar correlations between the number of transmission links and the audience size in the broad sense, as measured by the number of *incoming links*. We nonetheless suggest *total attention* measures more precisely audience-related effects as it considers individual attentional landscapes, by weighting the number of links the referred blog receives *with the relative importance* it bears for the referring blogger.

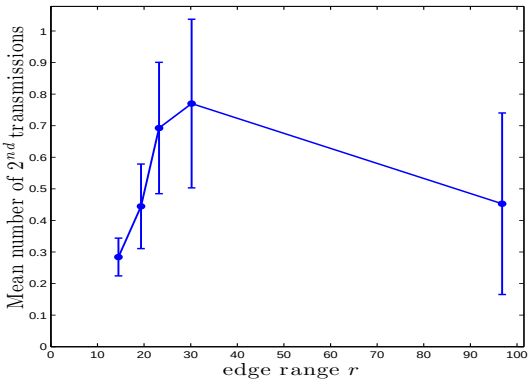

Fig. 10. Mean number of second transmission links (k, j) with respect to the edge range value $\mathbf{r}(j, i)$ of the first transmission. Data scarcity led us to bin $\mathbf{r}$ into five quintiles.

In particular, we examine the hypothesis that an information which has been channeled through a weak-tie as a "shortcut" may be more "contagious" for further diffusion. To test this hypothesis, we measure the number of *transmission links* in each diffusion subgraph with respect to the edge range of the edge from which the original resource was cited. In other words, if i is an initiator in subgraph σ_u, j cites i for u, we then examine the number of blogs k in σ_u such that (k, j) are edges of σ_u, with respect to $\mathbf{r}(j, i)$.

The corresponding statistics, plotted on Fig. 10, shows that resources which transited through edges of higher $\mathbf{r}$ generally tend to propagate to a greater number of blogs than for lower $\mathbf{r}$. This is however valid below a certain threshold, after which links seem to be too weak to efficiently provoke second transmissions. As such *weak ties*, i.e. with higher edge range, proportionally act more as catalyzers for ongoing diffusions in that they connect otherwise relatively remote areas.

To sum up, blogs (i) connected through a medium edge range to (ii) a "high attention" blog realize higher numbers of second transmissions.

V. Conclusion

Social and semantic dimensions are essentially co-determined in this blog network: first, both social and semantic topologies drive new interactions, specifically through a strong homophilic behavior and link creation within the structural neighborhood. Second, information circulation is shaped both by social and attentional topology, in a broad framework where influence in understood in relatively holistic terms. In particular, we showed how specific structural features may be associated to information pathways: while an ego-centered property such as total attention indicates a higher capacity to disseminate particular online resources, a non-ego-centered property such as edge range indicates that weaker links generally tend to bring richer diffusion in the longer term.

Higher attention combined, later on, with higher edge range significantly enhance the capacities for an online resource to be further diffused.

More broadly, we see this whole framework as a preliminary to a deeper understanding of the joint, coevolving dynamics of social and semantic structures, or the joint evolution of topology and information distribution, notably in the case where both dimensions evolve at comparable timescales.

VI. Acknowledgments

This work has been partially supported by the French National Agency of Research (ANR) through grant "*Webfluence*" #ANR-08-SYSC-009. We warmly thank Franck Sajous for NLP assistance; and RTGI and Guilhem Fouetillou for providing the original dataset and relevant feedback.

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, 13th WWW*, 2004.

[2] E. Cohen and B. Krishnamurthy, "A short walk in the blogistan," *Computer Networks*, Jan 2006.

[3] N. Glance, M. Hurst, and T. Tomokiyo, "Blogpulse: Automated trend discovery for weblogs," *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation*, Jan 2004.

[4] 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"," in *Proc. 38th Hawai'i International Conference on System Sciences (HICSS-38)*, 2005.

[5] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "Structure and evolution of blogspace," *Commun. ACM*, vol. 47, no. 12, pp. 35–39, 2004.

[6] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *portal.acm.org*, Jan 2007.

[7] J. Leskovec, A. Krause, C. Guestrin, and C. Faloutsos, "Cost-effective outbreak detection in networks," *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, Jan 2007.

[8] X. Shi, B. Tseng, and L. Adamic, "Looking at the blogosphere topology through different lenses," in *Proc. ICWSM Intl Conf Weblogs Social Media*, 2007.

[9] F. B. Viegas, M. Wattenberg, J. Kriss, and F. van Ham, "Talk before you type: Coordination in Wikipedia," in *Proceedings of the 40th Hawaii Intl Conf on System Sciences*, 2007.

[10] C. Marlow, M. Naaman, danah boyd, and M. Davis, "Position paper, tagging, taxonomy, flickr, article, toread," in *Proceedings of Collaborative Web Tagging Workshop, WWW 06*, 2006.

[11] L. Adamic and N. Glance, "The political blogosphere and the 2004 US election: divided they blog," in *LinkKDD '05: Proc. 3rd Intl. Workshop on Link discovery*. New York, NY, USA: ACM Press, 2005, pp. 36–43.

[12] J.-P. Cointet, E. Faure, and C. Roth, "Intertemporal topic correlations in online media," in *Proc. ICWSM Intl Conf Weblogs Social Media*, Boulder, CO, USA, 2007.

[13] D. Drezner and H. Farrell, "Web of influence," *Foreign Policy*, Jan 2004.

[14] K. Wallsten, "Political blogs and the bloggers who blog them: Is the political blogosphere and echo chamber," *American Political Science Association's Annual Meeting. Washington, DC September*, Jan 2005.

[15] A. Java, P. Kolari, T. Finin, and T. Oates, "Modeling the spread of influence on the blogosphere," *Proceedings of the 15th International World Wide Web*, p. 7, May 2006.

[16] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "On the bursty evolution of blogspace," *World Wide Web*, Jan 2005.

[17] M. McGlohon, J. Leskovec, C. Faloutsos, and M. Hurst, "Finding patterns in blog shapes and blog evolution," *Proceedings of ICWSM*, Jan 2007.

[18] G. Mishne and M. de Rijke, "Capturing global mood levels using blog posts," *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, Jan 2006.

[19] K. Balog, G. Mishne, and M. de Rijke, "Why are they excited? identifying and explaining spikes in blog mood levels," *11th Meeting EACL*, 2006.

[20] N. Bansal and N. Koudas, "Blogscope: a system for online analysis of high volume text streams," *Proc. of the 33rd Intl Conf. on Very large data bases*, Jan 2007.

[21] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," *Proc. 13th Intl Conf. on World Wide Web*, Jan 2004.

[22] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146, 2003.

[23] J. Kleinberg, "Cascading behavior in networks: Algorithmic and economic issues," *Algorithmic Game Theory*, 2007.

[24] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: membership, growth, and evolution," in *12th ACM SIGKDD*. New York, NY, USA: ACM Press, 2006, pp. 44–54.

[25] J. Leskovec, M. McGlohon, C. Faloutsos, and N. Glance, "Cascading behavior in large blog graphs," *SIAM Intl Conf on Data Mining (SDM 2007)*, Jan 2007.

[26] J. Iribarren and E. Moro, "Information diffusion epidemics in social networks," *eprint arXiv: 0706.0641*, Jan 2007.

[27] G. Salton, A. Wong, and C. S. Yang, "Vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[28] E. Dijkstra, "A note on two problems in connection with graphs," *Numerische Mathematik*, vol. 1, no. 269-270, p. 6, 1959.

[29] T. A. B. Snijders, "The statistical evaluation of social networks dynamics," *Sociological Methodology*, vol. 31, pp. 361–395, 2001.

[30] A.-L. Barabási, H. Jeong, E. Ravasz, Z. Neda, T. Vicsek, and A. Schubert, "Evolution of the social network of scientific collaborations," *Physica A*, vol. 311, pp. 590–614, 2002.

[31] C. Roth, "Generalized preferential attachment: Towards realistic socio-semantic network models," in *ISWC 4th Intl Semantic Web Conference, Workshop on Semantic Network Analysis*, ser. CEUR-WS Series (ISSN 1613-0073), vol. 171, Galway, Ireland, 2005, pp. 29–42.

[32] S. Raux and C. Prieur, "Liens proches dans les réseaux sociaux - la dynamique des commentaires de flickr," in *Proc. Algotel 11e Rencontres Francophones Aspects Algorithmiques Telecommunications*, 2009.

[33] G. Kossinets, J. Kleinberg, and D. J. Watts, "The structure of information pathways in a social communication network," *arxiv*, vol. physics.soc-ph, Jun 2008.

[34] D. J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*, ser. Princeton Series in Complexity. Princeton, N.J.: Princeton University Press, 1999.

[35] K. E. Gill, "How can we measure the influence of the blogosphere?" in *Proc. of WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, May 17-22 2004.

[36] E. M. Rogers, "New product adoption and diffusion," *The Journal of Consumer Research*, vol. 2, no. 4, pp. 290–301, 1976.

[37] M. S. Granovetter, "The strength of weak ties," *The American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.