Local networks, local topics: structural and semantic proximity in blogspace*

Jean-Philippe Cointet
INRA SenS - IFRIS / Univ. Paris-Est
5 bd Descartes, Champs/Marne F-77454 Marne-la-Vallée
cointet@poly.polytechnique.fr

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

Abstract

Blog networks are often described as "small world" social networks where links are primarily created towards similar-minded individuals and well-connected bloggers. Examining a portion of the US blogosphere on several months, we find that bloggers relate to each other in a relatively local fashion, overwhelmingly and preferentially establishing links towards a limited neighborhood rather than the whole network. Furthermore, while long-distance interactions may indeed be dominated by homophily and authority effects, we show that close neighborhoods feature significantly more horizontal and diversified interactions — thereby challenging the conjecture of a widespread balkanization of Internet communities. We shed further light on this issue by describing the dual evolution of social and semantic proximity before and after two individuals interact with each other. We discuss in particular whether interactions are preceded or followed by a structural "contraction" and/or by an increasing similarity of the surrounding local social network.

Introduction

While the Internet became an increasingly popular online agora, a debate emerged as to whether this discussion space indeed facilitates confrontation of opinions of individuals holding dissimilar viewpoints & coming from distant social circles, or channels conversations between similar-minded people within closed groups of interaction (Van Alstyne and Brynjolfsson, 1996; Sunstein, 2008). In the specific case of blog-based discussions, the existence of opinion-based or topic-based boundaries between bloggers has already been characterized in several works: Adamic and Glance (2005) for instance showed that Democratic- and GOP-leaning blogs were much less connected between each other than within each other, while Uchida et al. (2009) described the "blogspace" as a juxtaposition of various communities being both structurally and topically cohesive.

Focusing precisely on the determinants of link creation, several recent studies on a wide range of empirical social systems suggested that links were preferentially directed towards most connected individuals (Barabási and Albert, 1999, *inter alia*), as well as towards individuals who are closer structurally — i.e., at a smaller distance (Newman, 2001; Liben-Nowell and Kleinberg, 2003; Roth, 2006; White et al., 2006) — and closer semantically — i.e., with similar thematic profiles (McPherson, Smith-Lovin, and Cook, 2001; Roth, 2006; Cattuto et al., 2009). Reinforcing mechanisms could also make some websites significantly more accessible, notably the closest and/or most connected ones (e.g. Hindman, Tsioutsiouliklis, and Johnson (2003), and specifically for blogs Herring et al. (2005)).

Altogether, these various independent results tend to suggest that the blogspace as a social network could also be a thematically clustered system dominated by a relative small number of authorities, plausibly shaped by local and homophilic discussion and citation patterns. In this respect, the traditional claim that small worlds, such as blog networks, are facilitating the mix of remote and diverse kinds of actors would have to be confronted with the possibility that they could also be local worlds of similarly-minded individuals.

To our knowledge, we still lack an integrated perspective that would precisely appraise and decypher the *respective* and *combined* contributions of structural effects and semantic features to the formation and evolution of these online discussion circles. Moreover, little is known on the dynamics that precede, lead to and follow the creation of citation links and, more broadly, relationships between blogs.

In this short paper, we describe the socio-semantic structure and dynamics of these conversation circles on a portion of the US blogosphere on a period of several months. We first show that relationships in this network are overwhelmingly local, with seemingly little interactions with individuals located at a topological distance of three or more hops. This supports the idea of "local worlds" where everyone interacts within a very small radius, generally among neighbors or neighbors of neighbors, rather than "small worlds" where everyone should be at reach (and *therefore* reachable) from everyone. In a dual fashion and with respect to thematic diversity, we then show that links are essentially homophilic at an aggregate network level, yet, we also show that this result essentially holds *outside* the close circle of repeated interactions — whereas repeated interactions exhibit a more mixed profile where links are both preferen-

*This work has been partially supported by the French ANR through grant "Webfluence" #ANR-08-SYSC-009. We thank Guilhem Fouetillou from *Linkfluence* for providing us with the dataset.

tially created towards very similar and very dissimilar individuals. This latter finding is in relative constrast with the idea of a wholly balkanized blogspace. We finally examine the joint evolution of social and semantic proximity before and after two individuals first interact. We demonstrate notably that interactions are preceded *and* followed both by structural "contraction" and by growing semantic similarity within the surrounding local social network. Notwithstanding the above provision that local circles may be both homophilic and "heterophilic", this joint semantic homogenization and social integration process sheds a finer light on the emergence of local, cohesive thematic conversation circles.

Related Work

The above issues first connect with the question of "localism" in personal networks. In particular recent quantitative research on the topology of empirical social networks made universal observations regarding the so-called "small-world effect" and "scale-free structure". The former refers to the small diameter of those networks, where a short-length path exists between most pairs of nodes (e.g. Newman, 2001); i.e. *small worlds* where everyone could be "only" a few steps away from everyone else (Milgram, 1967; Watts and Strogatz, 1998; Dodds, Muhamad, and Watts, 2003) — the jury is still out regarding whether social interactions and processes indeed are or are not significantly influenced by this relative topological proximity. The latter observation indicates that network structure is very heteregenous, with few heavily connected/cited hubs/authorities, surrounded by many weakly linked individuals (see Herring et al., 2005; Leskovec et al., 2007, in the case of blog networks).

Several works shed light on these large-scale characterizations by focusing on the underlying ego-centered processes of link creation. For instance, the emergence of connectivity heterogeneity is often assumed to stem from reinforcing dynamics of strictly structural nature, where high degree nodes receive preferentially more links (Barabási and Albert, 1999). Similarly, topological distances appear to decrease over time (Leskovec, Kleinberg, and Faloutsos, 2005) while links are also preferentially created towards closer nodes (Roth, 2006; Raux and Prieur, 2009). As regards integrating topology and semantic homogeneity, Adamic, Buyukkokten, and Adar (2003) noticeably exhibit a decreasing similarity of agents with increasing social distances on a social networking site. In a dynamic setting, Crandall et al. (2008) observe a continuous increase of semantic similarity between two individuals *both* after and before they contribute on a same Wikipedia discussion page (semantic similarity being measured by the topics each actor had been dealing with on the platform); while Cointet and Roth (2009) describe a higher propensity to make links towards thematically more similar nodes in a blog network.

These findings often come from online communities at large, but not so frequently from blog networks specifically. Besides an abundant literature focusing on diffusion information issues in blog networks, a few morphogenesis models have recently been proposed (Leskovec et al., 2007; Goetz et al., 2009) which successfully connect normative

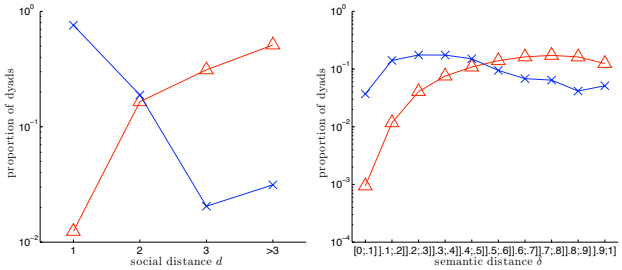

Figure 1: Distribution of new links *(blue crosses)* and pairs of blogs *(red triangles)* at a given structural *(left)* and semantic distance *(right)* — computed on the last period.

link creation processes with the emergence of several classical topological features. Nonetheless, the mechanisms proposed so far remain generally blind to semantic features.

Protocol

Social web content analysis company *Linkfluence* provided their dataset of 1066 active political blogs on 4 months from Nov 1, 2008, including dated post content and citation links. We divided the dataset into 16 weekly slices, considering the 8 first weeks as initialization period and carrying measures on the 8 last periods only. The social network consists of a growing network (possibly featuring repeated links) such that at a given period it aggregates all links up to that period. To characterize blog contents, we define a set of 79 "meaningful" *concepts* made of one- or two-words terms manually extracted from the list of most frequent non-trivial words. This list features concepts such as *"global warming"*, *"national security"*, or *"tax cuts"* and has been designed to cover a sizeable, sufficiently discriminating number of topics discussed during the last US presidential campaign. Bloggers are then described by semantic profiles as 79-dimensional vectors whose values are the classical *tf.idf* ratio for each concept for each blog — that is the concept occurrence frequency (*tf*) in a blogger's post production divided by its occurrence frequency in the whole corpus (*idf*). These profiles are also dynamic and are built by aggregating content published until a given week. It is eventually possible to define a semantic dissimilarity $\delta(i,j)$ between two blogs as 1 minus the cosine of their semantic profile. $\delta(i,j) = 1$ when two blogs have no content overlap at all, while $\delta(i,j) = 0$ blogs having exactly similar contents. We also classically define a social network structural distance d as the length of the smallest path connecting two nodes. By analogy (only) we may also call δ a semantic "distance".

Propensity

The connectivity distribution is heterogeneous in a power-law fashion, as expected, and not shown here. We rather focus on topological distance and semantic dissimilarity and gather the absolute number of newly-created links and existing dyads in both cases on Fig. 1. These graphs reveal that new citations are overwhelmingly close: new links at distance 3 and above are two orders of magnitude less frequent than repeated interactions, and even one order rarer

than links at distance two. Similarly, new citations occur at significantly smaller semantic dissimilarities when compared with potential citations between all possible dyads.

These computations do not reflect however the *relative preference* of bloggers towards specific kinds of other bloggers. To compute this, we introduce a basic measure of propensity Π which consists of a ratio between proportions of new links of a given type and of potential links of that kind. This ratio evaluates if some types of links are empirically more ($\Pi > 1$) or less likely (<1) when compared with a random baseline where all pairs of bloggers would appear by chance (i.e. irrespective of any individual preference/behavior). This measure is traditionally refered to as one of "preferential attachment" (Barabási and Albert, 1999). Here, it is computed with respect to several egocentered and dyadic properties: degree, social and/or semantic distance. In addition to the usual linear preference for more connected blogs, interaction propensities are monotonously decreasing with respect to each property, as can be understood from Fig. 1; i.e. there is homophily and preference for local and authoritative blogs (high degree).

How are those various dimensions related? We contrast authority and homophily with social distance by plotting the joint propensity with respect to social distance and (i) degree (Fig. 2) or (ii) semantic distance (Fig. 3). Profiles vary significantly depending on whether interactions are local or not. First, the effect of authority is all the more important that blogs are remote: propensities are relatively flat in the close vicinity of *ego* and progressively increasing for growing values of d. In other words, the traditional "rich-get-richer" effect is essentially valid for the wider blogspace, but not for the immediate neighborhood.

Second, and again, homophily is all the more marked that blogs are remote. Within the close circle of repeated interactions, we even observe that bloggers both tend to cite preferentially very similar blogs but also blogs with very dissimilar profiles. Put differently, repeated interactions are rather of a different nature than new links towards remote "strangers". This suggests two modes of relationships: on one hand, exploratory link creations on similar topics (distance 2,3, and more); on the other hand, exploitation of previous citation relationships on both similar and radically distinct topics (U-shaped propensity at distance 1).

Ex post and *ex ante* socio-semantic dynamics

To encompass the particular construction of these local social circles, we eventually focus on local social & semantic alignment processes. We designed $\rho(d)$ and $\rho(\delta)$ to measure the *contraction* of respectively social and semantic distances between two interacting blogs. Semantic similarity remains as above, while social proximity is appraised through numbers of paths of length 2 and 3 connecting i to j. We normalize these values by mean proximities from i and j to the whole network, to account for its simultaneous evolution. For example the semantic contraction of a pair of blogs (i, j) is $\rho_{ij}(\delta) = \delta(i,j)/\sum_k \frac{1}{2}(\delta(i,k) + \delta(j,k))$. Contraction thus measures how two blogs get relatively closer ($\rho > 1$) or more remote ($\rho < 1$) than towards the rest of the network.

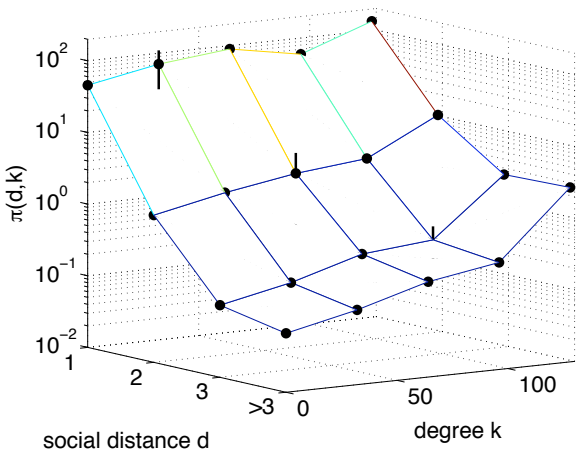

Figure 2: Interaction propensity $\Pi(d, k)$ with respect to social distance and degree (averaged over the 8 last periods).

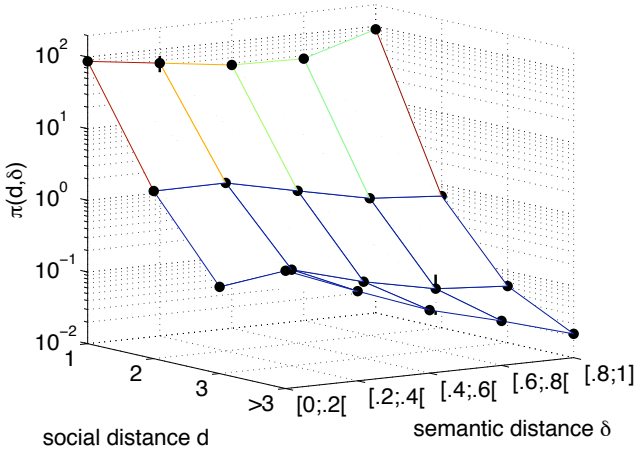

Figure 3: Interaction propensity $\Pi(d, \delta)$ for social distance and semantic similarity (averaged over the 8 last periods).

Then, we measure the evolution of ρ both before and after two blogs get connected for the first time. To provide a common evaluation basis, we aggregate averages over a moving referential such that all first interaction times are synchronized at the same $t = 0$ (a same point may thus correspond to data from various periods but always to the same delay with respect to first interaction). Results for social and semantic contraction evolutions are gathered on Figs. 4 and 5.

A strong contraction between pairs of interacting blogs can be observed both before and after interaction. Weeks before a first interaction, semantic distances are already almost 20% lower than average, which is consistent with previous similar observation made by Crandall et al. (2008) on Wikipedia contributors. Interacting blogs tend to align their semantic profile. The same pattern applies to social contraction, as the relative (normalized) number of 2- and 3-steps paths is rapidly growing before interaction and keeps on increasing after. Yet, while ρ for 2-steps paths intersects 1 at interaction time and reaches 2 about 5 weeks later, we observe that ρ for 3-steps paths is still below 1 even 5 weeks

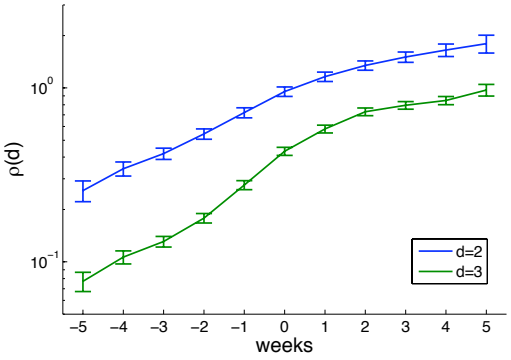

Figure 4: Evolution of the normalized number of paths of 2 and 3 steps between blogs getting first connected at $t = 0$.

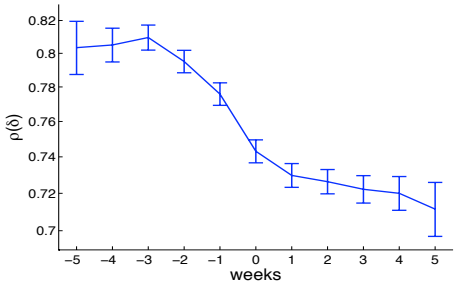

Figure 5: Evolution of the normalized semantic distance between pairs of blogs getting linked for the first time at $t = 0$.

after interaction. This indicates that the network contraction essentially concerns the very close vicinity of a dyad.

Concluding remarks

Bloggers overwhelmingly and preferentially establish local links within a limited neighborhood rather than navigate the whole network by profiting from its small diameter. Furthermore, local relationship patterns are distinct from those of the wider blogspace. While long-distance interactions are indeed dominated by homophily and authority effects, close neighborhoods feature significantly more horizontal and diversified interactions — plausibly suggesting a two-mode behavior opposing search engine-based or friends-of-friends exploration to local neighborhood exploitation. Finally, the surrounding network exhibits both structural and semantic contraction both after *and before* a link creation.

As such, these results tend to both nuance and confirm the existence of balkanization forces in blog communities. Besides highlighting the dynamics of link creation at both a local/global and social/semantic level, these results will certainly be useful in providing more empirical foundations to realistic morphogenesis models of blog networks.

Still, it should not be immediately possible to determine whether Internet communities are more balkanized than "real-world" contexts without a proper protocol aiming at comparing blogger behaviors with what happens in offline interaction. While the Internet certainly seems to exhibit reinforcing socio-semantic dynamics, the debate is still open as to whether these forces are stronger or weaker than in the offline world.

References

Adamic, L. A., and Glance, N. 2005. The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proc. 3rd LinkKDD Workshop*, 36–43. New York, NY, USA: ACM Press.

Adamic, L. A.; Buyukkokten, O.; and Adar, E. 2003. A social network caught in the web. *First Monday* 8(6).

Barabási, A.-L., and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286:509–512.

Cattuto, C.; Barrat, A.; Baldassarri, A.; Schehr, G.; and Loreto, V. 2009. Collective dynamics of social annotation. *PNAS* 106(10511).

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

Crandall, D.; Cosley, D.; Huttenlocher, D.; Kleinberg, J.; and Suri, S. 2008. Feedback effects between similarity and social influence in online communities. In *Proc. 14th SIGKDD*, 160–168.

Dodds, P.; Muhamad, R.; and Watts, D. 2003. An experimental study of search in global social networks. *Science* 301:827–829.

Goetz, M.; Leskovec, J.; McGlohon, M.; and Faloutsos, C. 2009. Modeling blog dynamics. In *Proc. 3rd ICWSM*.

Herring, S. C.; Kouper, I.; Paolillo, J. C.; Scheidt, L. A.; Tyworth, M.; Welsch, P.; Wright, E.; and Yu, N. 2005. Conversations in the blogosphere. In *Proc. 38th HICSS Intl Conf System Sciences*.

Hindman, M.; Tsioutsiouliklis, K.; and Johnson, J. A. 2003. Googlearchy: How a few heavily-linked sites dominate politics on the web. In *Annual Meeting Midwest Pol. Sci. Association*.

Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N.; and Hurst, M. 2007. Patterns of cascading behavior in large blog graphs. In *SDM 2007 Proc. 7th SIAM Intl. Conf. on Data Mining*.

Leskovec, J.; Kleinberg, J.; and Faloutsos, C. 2005. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proc. 11th ACM SIGKDD*, 177–187.

Liben-Nowell, D., and Kleinberg, J. 2003. The link prediction problem for social networks. In *CIKM '03: Proc. 12th Intl. Conf. Information and knowledge management*, 556–559.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27:415–444.

Milgram, S. 1967. The small world problem. *Psychology Today* 2:60–67.

Newman, M. 2001. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *PRE* 64:016132.

Raux, S., and Prieur, C. 2009. Liens proches dans les réseaux sociaux – dynamique de Flickr. In *Proc. Algotel 11*.

Roth, C. 2006. Co-evolution in epistemic networks – reconstructing social complex systems. *Structure and Dynamics: eJournal of Anthropological and Related Sciences* 1(3):article 2.

Sunstein, C. 2008. Architecture of serendipity. *Harvard Crimson*.

Uchida, M.; Shibata, N.; Kajikawa, Y.; Takeda, Y.; Shirayama, S.; and Matsushima, K. 2009. Identifying the large-scale structure of the blogosphere. *Adv. in Complex Systems* 12(2):207–219.

Van Alstyne, M., and Brynjolfsson, E. 1996. Electronic Communities: Global Village or Cyberbalkans? In *Proc. 17th Intl Conf Information Systems, Cleveland, OH*.

Watts, D. J., and Strogatz, S. H. 1998. Collective dynamics of "small-world" networks. *Nature* 393:440–442.

White, D.; Kejzar, N.; Tsallis, C.; Farmer, D.; and White, S. 2006. A generative model for feedback networks. *PRE* 73:016119.