Citations among blogs in a hierarchy of communities: method and case study

Abdelhamid Salah brahim, Bénédicte Le Grand, Lionel Tabourier, Matthieu Latapy

LIP6 – CNRS and UPMC (Université Pierre et Marie Curie)
4 place Jussieu 75252 Paris cedex 05, France.
email: First-name.Last-name@lip6.fr

Abstract

How does the structure of a network (e.g. its organization into groups or *communities*) impact the interaction among its nodes? In this paper we propose a generic methodology to study the correlation between complex networks interactions and their community structure. We illustrate it on a blog network and focus on citation links. We first define a *homophily* probability evaluating the tendency of blogs to cite blogs from the same communities. We then introduce the notion of *community distance* to capture if a blog cites (or is cited by) blogs distant or not from its community. We analyze the distribution of distances corresponding to each citation link, and use it to build maps of relevant communities which help interpreting blogs interactions.

This community-oriented approach allows to study citation links at various abstraction levels, and conversely, enable us to characterize communities with regard to their citation behaviour.

keyword: complex network, community structure, citation links, blog

1 Introduction

Understanding interaction patterns in real-world networks is an important topic with both fundamental and practical implications [9]. However the volume and complexity of these networks make this task very challenging. Intuitively, nodes with common features i.e. which belong to a same *community* [15] tend to interact preferentially with each other, but limited knowledge is available on this topic for real-world data [6]. In a previous study of a blog network, we have shown the impact of topical communities on citation behaviour [18]. In this paper we go further and propose a generic methodology in Section 2 to study

interaction links in complex networks with regard to their community structure through new original measures: link *homophily* and *community distance*. This approach consists in studying interaction links at various community scales, and thus at various granularity levels rather than considering nodes individually. Moreover, it allows to identify new classes of communities and to cartography them with regards to their interaction behaviour. We also study variations between incoming and outgoing links. We apply this methodology to the same blog network [18] in Section 3. This approach allows us to study interactions with regard to community structure and conversely to characterize communities according to links homophily and distances.

2 Framework

Our methodology consists in studying interaction links in a network with regard to its community structure. The construction of this structure is not the focus of this paper; we will thus suppose that it exists; this is a realistic hypothesis as we explain in Section 2.1 how a hierarchical community structure may be obtained for any complex network. This Section is organized as follows. We first introduce definitions and notations related to the hierarchical community structure we will use in the paper. We then explain our 2-steps methodology : we define in Section 2.2 two metrics which allow to evaluate whether interaction links relate nodes from a same community (at all levels of the hierarchical structure) and avoid bias related to the number of links in each community. We explain how these metrics are complementary to another measure (*modularity*) used traditionally to evaluate partitions quality. In Section 2.3 we introduce the notion of *community distance* to characterize interaction links between nodes according to the distance between these nodes communities. In other words, this distance allows to evaluate whether these links relate "close" or "distant" communities.

2.1 Hierarchical community structure

Let a graph $G = (V, E)$, with V a set of nodes and E a set of edges. Our methodology requires a hierarchical community structure, i.e. a tree of communities such that the set of communities at each level of the tree is a partition of V. Communities may be based on nodes features, e.g. groups of web pages dealing with similar topics, subtopics and so on. The hierarchical community structure may also be built automatically with community detection algorithms, with generally rely on topological information [2].

Definition 1 ***Partition of a graph into communities***
A partition $P = \{C_1, C_2, ..., C_l\}$ of the graph G into communities is a collection of disjoint subsets (called communities and noted C_i) such that $\bigcup_i C_i = V$ and $\forall i, j,\ C_i \cap C_j = \emptyset$

Definition 2 ***Hierarchical Community Structure***
Given a community partition $P = \{C_1, C_2, ..., C_l\}$ of G, a sub-partition $P' = \{C'_1, C'_2, ..., C'_l\}$

of P is a partition of G such that $\forall\ C'_i \in P'$, $\exists\ C_j \in P$ s.t. $C'_i \subseteq C_j$. This is denoted $P' \sqsubseteq P$.

A hierarchical community structure of G is defined as a series of partitions $P_k \sqsubseteq P_{k-1} ... \sqsubseteq P_2 \sqsubseteq P_1 \sqsubseteq P_0$ with $P_0 = V$, i.e. P_0 contains only one community which is the whole set of nodes and $P_k = \{\{v\}, v \in V\}$, i.e. P_k contains n communities containing each only one node. Given a partition P_i, i is called the level of the partition P_i within the global tree of communities with $(k+1)$ levels.

Let $C \in P_i$; we denote $D_j(C)$ the set of descendent communities of C in the community tree, i.e. $D_j(C) = \{C' \in P_{i+j}, C' \subseteq C\}$, with $(i+j) < k+1$.

Definition 3 ***Community function***
As each node in V belongs to exactly one community at each level of the hierarchical community structure (i.e. in each partition P_i) we may define a function denoted $\mathcal{C}_i$ identifying a node's community at level i of the community structure.

Let $v \in V$; $\mathcal{C}_i(v) = C \in P_i$, s.t. $v \in C$.

2.2 Homophily

Our approach requires two distinct objects: on the one hand an interaction network and on the other hand a hierarchical community structure (or a community tree), formally defined in Section 2.1. The first step of our methodology to study the relationships between interaction links within a network and its hierarchical community structure consists in evaluating the proportion of intra-community links. We therefore study, at all levels of the community tree, the probability (that we call *homophily* probability) that a link exists between two nodes from the same community.

Definition 4 ***Interaction link homophily probability***

Let C a community from the partition P_i of the hierarchical community structure. Let $G' = (C, E')$ be the subgraph induced by $G = (V, E)$ i.e. $C \subseteq V$ and $E' = E \cap (C \times C)$.

We define Δ_j the probability that an edge of E' connects two nodes from the same community at the jth level of the community tree, with $j > i$.

$$\Delta_j(C) = \frac{|\ \{(u,v) \in E', \mathcal{C}_j(u) = \mathcal{C}_j(v)\}\ |}{|\ E'\ |}$$

Δ_j is called *homophily probability* as it measures the proportion of links between nodes from a same community at level j in the community hierarchy.

Note that, in the previous definition, the value of $\Delta_j(C)$ may be biased by the number of links in communities at the j-th level; for example, if there is one very large community, $\Delta_j(C)$ is likely to be higher than if all communities have comparable sizes. In order to avoid such a bias, we consider the value of $\Delta_j(C)/\psi_j(C)$, where $\psi_j(C)$ is the probability

that a link exists between two nodes (chosen randomly) from the same community among the descendents of the community C at the j-th level of the hierarchy:

$$\psi_j(C) = \frac{\sum_{C' \in D_j - i(C)} \mid E' \mid \cdot (\mid E' \mid -1)}{\mid E \mid \cdot (\mid E \mid -1)}$$

High values of $\Delta_j(C)/\psi_j(C)$ indicate a high homophily, i.e. a significant fraction of links between nodes from a same community at the j-th level of the hierarchy, independently of the number of edges in these communities. The *modularity* function [13, 5] has been defined to evaluate the quality of a partition and is also based on link density. A high value of modularity means that there is a high density of links within the communities of the partition and a low density of links between distinct communities of this partition. However, our metrics Δ and ψ do not have the same goal: they measure the proportion of internal links with regards to a random distribution.

Given the subgraph $G' = (C, E')$ induced by G, we will compare the value of $\Delta_j(C) \div \psi_j(C)$ with the value of modularity $Q_j(C)$, with:

$$Q_j(C) = \sum_{s=1}^{card(Dj-i(C))} \left[\frac{l_s}{\mid E' \mid} - \left(\frac{d_s}{2 * \mid E' \mid} \right)^2 \right]$$

where l_s is the number of links between nodes within community s, d_s is the sum of the degrees (total number of links) of nodes in s, and i is the level of community C in the community tree. Two communities may have very close $Q_j(C)$ values but different $\Delta_j(C) \div \psi_j(C)$ values. This will be illustrated in Section 3.2.

2.3 Community distance

To characterise interaction links (e.g. to distinguish links between close and distant communities) we define the *community distance* which is half of the distance in the community tree.

Definition 5 ***Community distance***
Given an interaction link denoted (u, v) between a couple of nodes u and v, there exists a minimal integer t such that there is a community C in P_t with $u \in C$ and $v \in C$. We then define the hierarchical distance of the spreading link (u, v) as: $d(u, v) = k - t$

Among interaction links involving nodes of the community C, we distinguish links which start from C (outgoing links) denoted $out(C)$ and links which arrive to C (incoming links) denoted $in(C)$.

We then define the fractions of incoming links $in_\kappa(C)$ (resp. outgoing links $out_\kappa(C)$) at distance κ involving community C:

$$in_\kappa(C) = \frac{\mid \{(u, v) \in in(C) \ s.t. \ d(u, v) = \kappa\} \mid}{\mid in(C) \mid}$$

$$out_\kappa(C) = \frac{\mid \{(u,v) \in out(C) \ s.t. \ d(u,v) = \kappa\} \mid}{\mid out(C) \mid}$$

The distribution of distances associated to incoming and outgoing citation links will allow us to identify categories of blogs and to map communities according to their interactions (see Section 3.3).

3 Application to a real-world case

In this section, we use the formalism introduced in Section 2 to analyse a real-world interaction network consisting of blogs, described in the following section.

3.1 Dataset

The dataset we used for our experiment was obtained by daily crawls of 6007 active blogs in the French-speaking blogosphere (1,074,315 posts) during 4 months from November 1st, 2008 to March 1st, 2009. These blogs have been selected by experts in blog and opinion analysis (`http://linkfluence.net`).

A blog is a website containing publications called *posts*. A post can, in addition to its own content, make a reference to a previous post (from the same blog or from another blog) by quoting the corresponding URL, which is called a *citation link*. Consider a post *Pa* from blog *A* and a post *Pb* from blog *B*. If *Pa* contains a reference to *Pb*, then there is a citation link from *Pa* to *Pb*, i.e. *Pa* cites *Pb*. Post *Pb* has an incoming link pointing to it (noted *in-link*) while post *Pa* has an outgoing link starting from it (noted *out-link*). In this paper, we consider citation links at blog scale, i.e. blog *A* cites blog *B*.

The classification of the studied blogs into *communities* has been built manually by professional blog analysts according to blogs topics. This topical classification in structured in three hierarchical levels: *continent*, *region* and *territory* (from the most general to the most specific, see Figure 1).

For instance, the blog *http://www.sailr.com* belongs to the *leisure continent*, the *sport region* and the *sailing territory*.

The hierarchical community structure we consider for this dataset therefore comprises 5 levels: level 0 corresponding to a single community (with all blogs), level 1 with 3 continents (*Leisure*, *Individuality*, *Society*), level 2 with 16 regions, level 3 with 96 territories and finally level 4 with the 6007 individual blogs.

To refer to the formalism of Section 2, we therefore consider the directed graph $G = (V, E)$ where V is the set of blogs and E is the set of citation links.

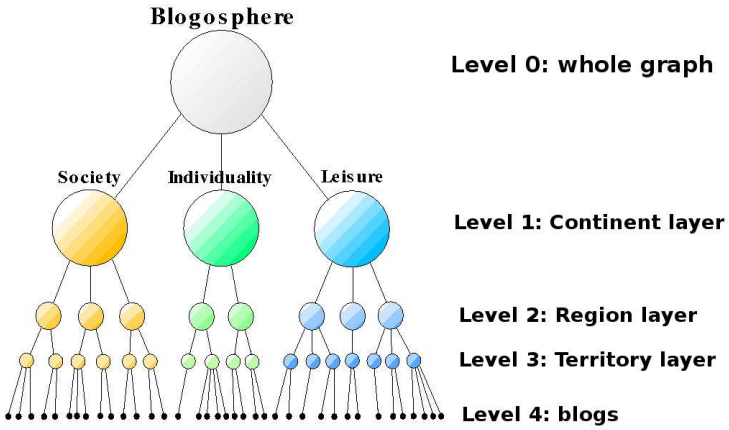


Figure 1: Blog network community structure (for readability reasons, not all communities are represented)

Table 1: Probability to link blogs from the same community at various levels.

Level i	$\Delta_i(G)$	$\Delta_i(G)/\psi_i(G)$	$Q_i(G)$
1-Continent	0.89	1.53	0.313
2-Region	0.74	1.95	0.363
3-Territory	0.21	2.62	0.167

3.2 Citation links homophily

First, we measure the Δ_j probabilities over the whole graph $G = (V, E)$ in order to evaluate the impact of the level j in the hierarchical community on homophily. [1]

The results presented in Table 1 show that $\Delta_1(G) = 98\%$ which means that 98% of blogs cite blogs from the same continent. Moreover, 74% of blogs cite blogs from the same region (and therefore same continent). The value of $\Delta_2(G)$ is inferior to $\Delta_1(G)$, but if we consider $\Delta_1(C)/\psi_1(C)$, layer 2 appears to be more significant (as homophile links are less expectable). In terms of modularity, layer 2 has a greater quality than layer 1. On the other hand, layer 3 has the lowest modularity. $\Delta_3(G)$ is also lower than $\Delta_2(G)$ and $\Delta_1(G)$ but the high value of $\Delta_3(C)/\psi_3(C)$ indicates that links at territory level are significantly more homophile than expected in a random case.

After considering the whole graph, we now focus on links within each continent at the region layer, i.e. the tendency of blogs from the same continent to cite within the same region (Table 2). We see that homophily values are very high. In particular, *Individuality* and *leisure* continents have Δ_2 values greater than 98%. However, $\psi_2(Individuality)$ is

[1] Since posts from the same blog have by definition the same classification we have removed auto-citation links, as they represent a different kind of citations, 24% (or 114, 261) of the total number of links remains.

Table 2: Probability to link blogs from the same region in each continent and associated modularity

Continent	# of link	$\Delta_2(Ci)$	$\psi_2(Ci)$	$Q_2(Ci)$
Individuality	43949	0.98	0.97	0.442
leisure	12811	0.99	0.56	0.667
Society	39579	0.78	0.13	0.0401

very high (97%) which means that the high value of $\Delta_2(Individuality)$ is more expectable than the value of $\Delta_2(Leisure)$.

Table 3: Probability to link blogs from the same *Territory* for each *Region*

Region	# of link	$\Delta_3(Ci)$	$\psi 3(Ci)$	$\Delta_3(Ci)$ / $\psi 3(Ci)$	$Q(Ci)$
agora	36878	0.149	0.178	0.837	-0.013
appearance	1047	0.820	0.335	2.448	0.382
automobile	1653	0.015	0.383	0.041	-0.252
notebook	208	0.995	0.454	2.188	0.0455
cooking	2591	0.948	0.884	1.072	0.002
culture	2114	0.507	0.500	1.013	-0.038
home	864	0.528	0.364	1.450	0.114
video_games	2619	0.026	0.352	0.074	-0.259
house	53	0.811	0.428	1.893	0.372
marketing_comm	170	0.747	0.848	0.880	0.003
human-resources	83	0.506	0.370	1.365	0.0242
health	19	0.263	0.317	0.828	-0.065
sports	3798	0.951	0.315	3.012	0.523
technology	2429	0.421	1.445	0.122	0.105
traveling	4	0.75	0.406	1.844	0.093
x-sports	32	0.531	0.381	1.393	-0.060

It is interesting to notice the very low value of modularity at the region layer for the *society* continent, which means that the quality of the partition is not good in terms of intra community links density with regards to inter community links density. However, the high value of $\Delta_2(society)/\psi_2(society)$ shows that although homophily of blogs among regions from *society* is lower than the two other continents, it is much higher than it would be in a random case, and this continent is therefore also relevant.

We now study citation links homophily within each region at the territory level, i.e. the tendency for blogs from a same region to cite blogs from the same territory (Table 3). Let us notice the low values of modularity (only three are higher than 0.114). This means that the quality of partitions of region into territories is not good. This is indeed not suprising as the classification into continents, regions and territories is based on blog topics and not

on their citation links.

We also observe in Table 3 a very low homophily probability with regards to random values, for example, $\Delta_3(automobile)/\psi_3(automobile) = 0.041$. Blogs in this region cite blogs outside their territories much more than in the random case. This suggests that the classification of *automobile* region into territories is not relevant from the citation point of view (although distinct topics are actually addressed in each territory and the topical classification is therefore).

Conversely, the homophily probability $\Delta_3(sports)$ is high = 0.95 (with also a high Δ_3/ψ_3 value). This result suggests that in this region the division into territories based on blogs topics is also relevant from the citation links point of view (i.e. there is consistency between the topical and the topological community structures).

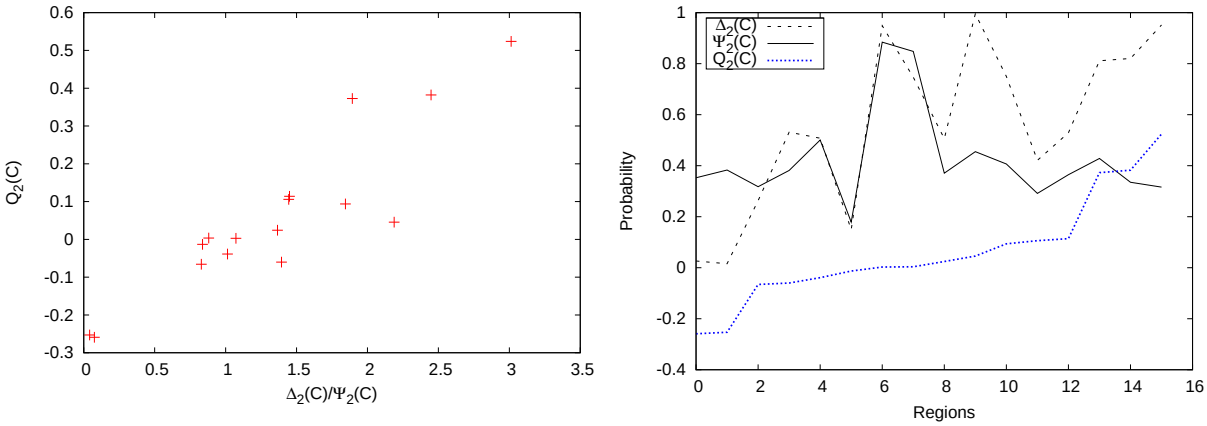
Moreover, this result indicates that *sports* sub-communities at the territory layer (for example *basketball*, *cycling* and *diving*) do not cite one another. Therefore, *sports* region is considered as a highly homophile community, which is also the case of *appearance*, *notebook* and *cooking* regions. We have seen that the homophily probability is very relevant to give an indication of blogs citation behaviour within a community. It is important to note that we cannot deduce this citation behaviour using only the modularity function.

Now we study in more detail the correlation between modularity and homophily at region layer. Figure 2a shows that $\Delta_2(C)/\psi_2(C)$ and modularity $Q_2(C)$ are correlated for almost all regions as most of them are close to the diagonal. We interpret this correlation by the fact that both homophily and modularity consider the density of links within a community. $\Delta_j(C)/\psi_j(C)$ could therefore be considered as a kind of unbiased modularity (with regards to the number of links in each community). More precisely, we illustrate the differences between both functions in Figure 2b which plots the values of $\Delta_2(C)$, $\psi_2(C)$ and $Q_2(C)$ for each region. The regions are sorted by increasing modularity value $Q_2(C)$. It is interesting to note that for regions 5, 6 and 9 (respectively *agora*, *cooking* and *notebook*) the values of modularity are very close but the values of $\Delta_2(C)$ and $\psi_2(C)$ are not. For *agora* region $\Delta_2(C)$ and $\psi_2(C)$ are almost equal but with a high value (0.9), while *cooking* region has $\Delta_2(C)$ and $\psi_2(C)$ almost equal with a a small value (0.17). For *notebook* region we observe a significant difference between $\Delta_2(C)$ and $\psi_2(C)$ values which will be interpreted differently from the two previous cases. This confirms that in order to study homophily we have to consider at the same time the value of $\Delta_j(C)$ and $\Delta_j(C)/\psi_j(C)$ and not only one of them.

3.3 Citation links community distance

Now we study the citation behaviour more precisely by characterising links with regard to their community distances.

Table 4 gives the distribution of community distances for all edges of G. Distance 1 links connect blogs from the same *territory* (and thus the same *region* and *continent*). Distance 2 links connect blogs from the same *region* but not the same *territories*. Distance 3 links connect blogs from the same *continent* but not the same *regions*. Finally, distance 4 links connect blogs from different *continents*.



(a) (b)

Figure 2: Delta vs modularity at region layer

Table 4: Distribution of community distances in G

Distance k	# links	% of links
1	15523	21%
2	38793	53%
3	11012	15%
4	6857	9.4%

We observe that distance 2 is the most frequent (53%), which means that most links are between blogs from the same *region* but not the same *territory* as one may suppose. The region layer is therefore significant from the citation point of view and in the following we start by studying region layer.

3.3.1 Community profiling based on links distance

We would like to compare the citation behaviour of the communities at the region layer, but as 78% of links come from the *agora* region (Figure 3), we will focus on *fractions* (rather than numbers) of links at each distance.

We distinguish *incoming* (*in*) and *outgoing* (*out*) links because they have different meanings. In links measure the attention raised by a community while out links reflect its centers of interest (i.e. the blogs it refers to). For example, a community can cite blogs from close communities (at a small distance) and be cited by blogs from far communities (at a high distance).

In Figure 4 we characterise blogs from each region according to their fraction of out links $out_\kappa(C)$ (Figure 4a) at each distance (resp. in links $in_\kappa(C)$ in Figure 4b): blogs from

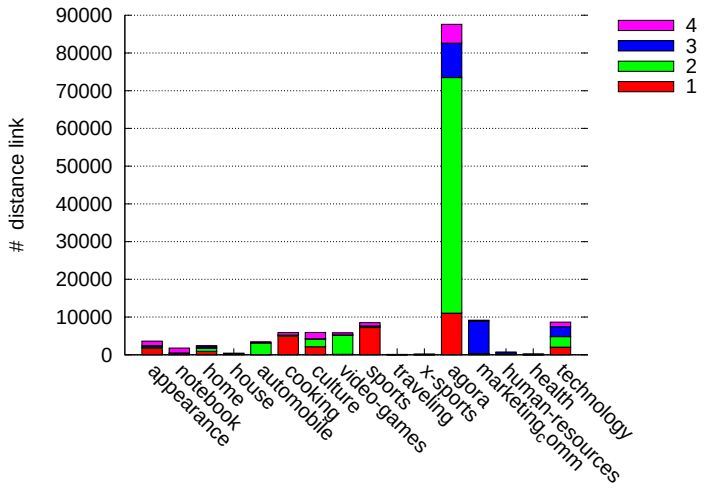
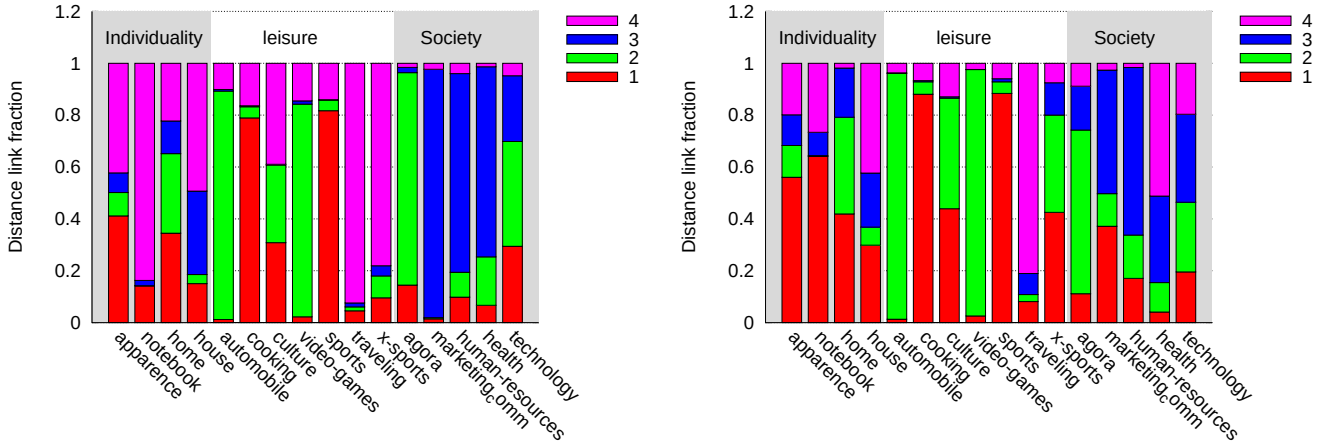


Figure 3: Number of link distances by region



(a) out link

(b) in link

Figure 4: Fraction of in and out link distances by region

notebook region cite distant blogs (i.e. from different continents as most out links are at distance 4). On the other hand, these blogs are mostly referred to by close blogs (from the same territory as distance 1 is the majority for in links). When we now compare regions, some communities with very similar profiles appear, for example, *sport* and *cooking* or *automobile* and *video-games*.

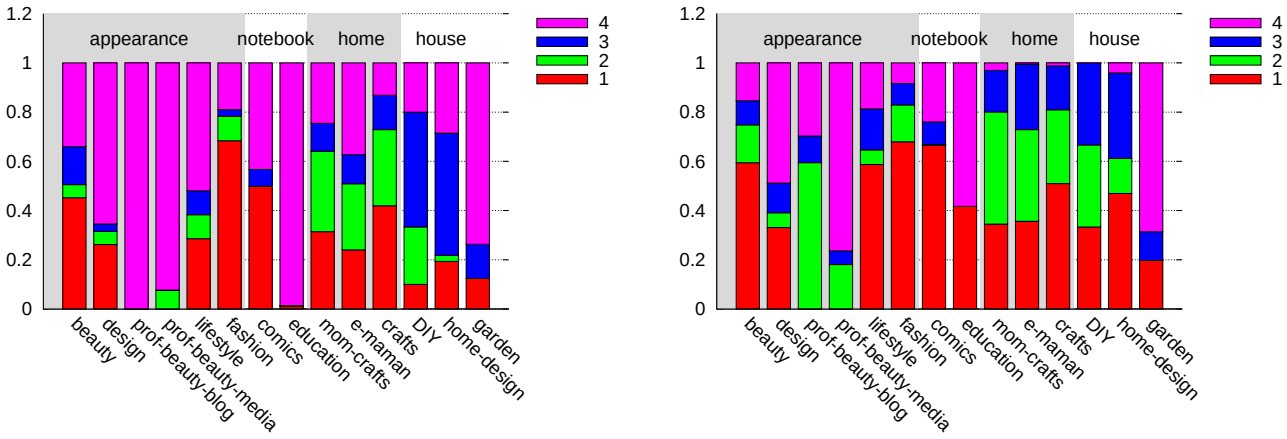
More precisely, *sports* and *cooking* have most of their links at distance 1 and others mainly at distance 4. This means that most in and out links in *sports* region are made within the same territories (e.g. *football*, *basketball*). *Sports* and *cooking* may thus be

classified as *self-centered* communities.

We may note that out links tend to have a dominating distance (which is less often the case with in links), e.g. *travelling*: distance 4, *health*: distance 3, *agora*: distance 2 and *cooking*: distance 1.

X-sport region incoming links come at 80% from the same territory or region (distance 1 and 2) while its outgoing links point to blogs from a different continent at 80% (distance 4). This is rather logical as a blog can have a "policy" with regard to the blogs it cites, but it cannot control who cites it, which leads to various in links distances.

We also observe that 8 regions out of 16 have a very small number of distance 3 links. This is not surprising if the fraction of distance 4 is also low for these regions (e.g. *video-games*), as it means that they are self-centered. However some of the regions with low fractions of distance 3 outlinks have a very high percentage of distance 4 outlinks (*notebook*, *traveling*, *x-sports*). This is non intuitive and suggests that for those specific communities, distances 1 and 2 links are more significant and can be considered as *strong community links*. Being able to distinguish this type of links could be used to improve network analysis.



(a) out link

(b) in link

Figure 5: Fraction of in and out link distances by territory in individuality continent.

One may deepen these observations by considering the territory layer. Figure 5 represents the fraction of incoming and outgoing link distances for *individuality* territories. *Individuality* continent is partitioned into 4 regions (*appearance*, *notebook*, *home* and *house*) and 14 territories (listed in Figure 5). The first observation is that in and out links distances distributions are more similar than it was the case at the region layer (on Figure 4). However, outgoing links have a more important proportion of links at distance 3 and 4 than incoming links. This means that all *individuality* territories cite blogs which are more distant than the blogs which cite them. Moreover, we may classify *individuality* territories into three classes. The first class gathers territories which have a significant fraction of

links at each distance i.e. which have a balanced links distances distribution. There are 6 territories in this class which belong to *home* and *house* regions (the 6 last territories in Figure 5). This citation pattern behaviour indicates that those blogs interact (both through incoming and outgoing citations) with a large variety and number of communities in the blogosphere at each level of the community tree. The second class contains self-oriented communities (*fashion* and *comics*). The third class is made of territories with a high distance majority. A deeper investigation shows that the topics of those blogs are related to *society* continent as they deal with everyday life topics.

3.3.2 Community mapping based on community distance

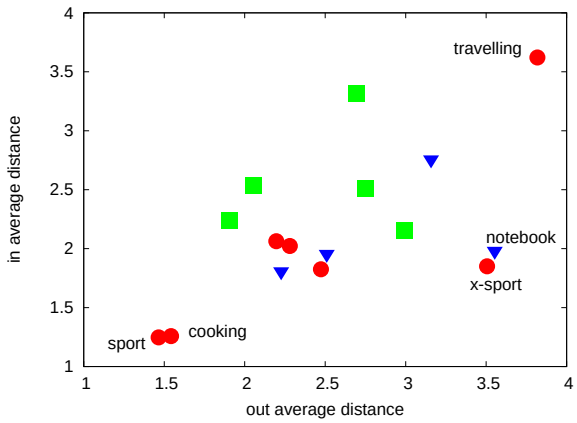


Figure 6: Average in and out links distance correlation at region layer. square=society, circle=leisure, triangle=individuality.

So far, we have analyzed the fraction of link distances in each region. In order to provide an overall picture, we have used average distances of in and out links as the coordinates of each region on a 2D map (Figure 6). Each region is colored according to its continent. We note that the patterns found in Figure 4 are confirmed here despite the use of an average value. For example we clearly see that *x-sport* and *notebook* are grouped together. *Self-centered* communities also appear, e.g. *cooking* and *sport*. On the contrary "*travelling*" has high in and out average community distances which means that the citation behaviour of these blogs is not related to their topical classification (they always cite and are being cited outside their topical community).

Now we focus on more specific communities at territory level. We first consider territories within *sport* region (Figure 8). In this example we study only *sports* territories which deal with one sport in particular and not blogs related to sport news. First we note that incoming links average distance is smaller than 2.4 for all communities, which means that in average incoming links come from *sports* region. *Cycling* and *yachting* have almost an average of 1 for incoming and outgoing links, so they are at the same time self-centered (no

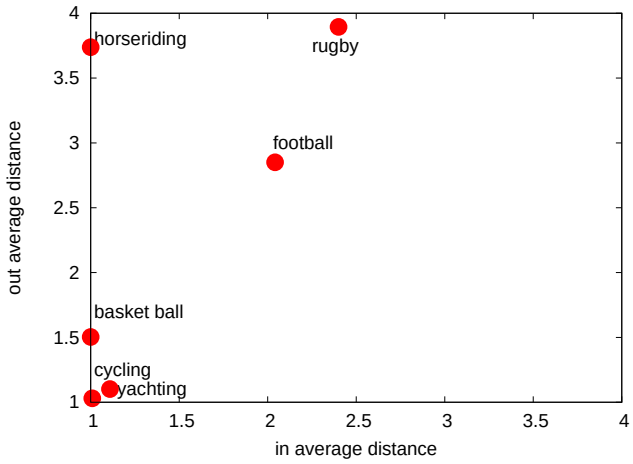


Figure 7: Average in and out links distance correlation in *sport* community

outgoing links with high distance) and do not get any attention from the rest of the blogosphere even from *sports* blogs. *Basket ball* community has the same profile for incoming links but is less self-centered and tends to cite blogs within its community as the average links distance is close to 1. On the other hand, *Horse-riding* and *rugby* communities cite blogs from other regions and continents and do not interact with close communities even in the same continent (distance greater than 3).

The second example is related to political blogs within *agora* region (Figure 8). First we observe that all political blogs have incoming and outgoing average links distances lower than 2.3. Consequently political blogs interactions globally remain within *agora* region. The territories correspond to political groups in France. It is interesting to observe that *europe* political group has fewer interactions outside its community than the other political groups while its activity in terms of number of interactions is high($\simeq 4000$). The *ecology* political group has a different profile, its audience (incoming links) remains local while its centers of interest tend to be outside the territory but still inside the political sphere. All other political groups have an average out links distance comprised between 1.9 and 2. They have the same citation behaviour which consists in a local citation activity within the community and at the same time a high interest in other political groups publications. On the other hand, the attention is different from a territory to another. *Right-wing* political group receives a rather local attention while *extreme left-wing* group has the largest audience in the blogosphere.

Until now we have studied community citation profiles in terms of proportion of links and average distance rather than number of links. We complete the study in Figure 9 where we plot the number of in links with regards to the number of out links at distances 3 and 4 at the region layer.

It shows an important correlation of incoming and outgoing links at distance 3 meaning that there is a high reciprocity between blogs at region level. For distance 4, correlations

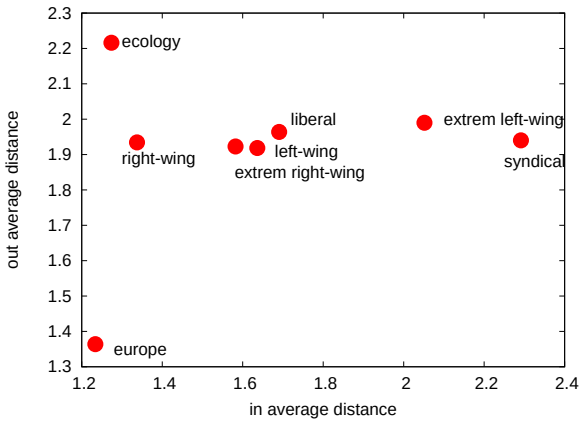


Figure 8: Average in and out links distance correlation in political communities

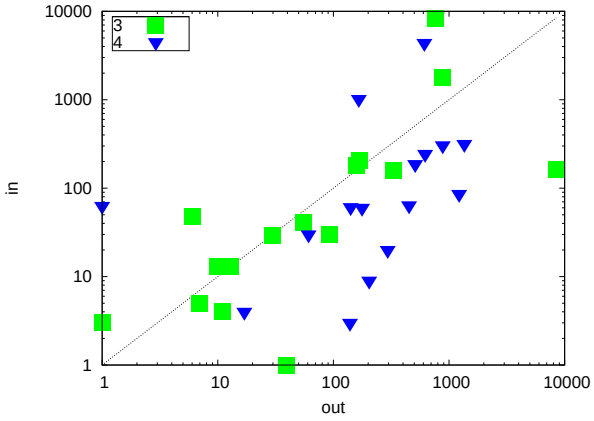


Figure 9: in and out links distance correlation at region layer (to help readability the two axes are in log scale). Each point corresponds to a region R and a distance d and has coordinates $in_d(R)$ and $out_d(R)$. All points with $d = 1, 2$ would be on the diagonal so we do not display them. Squares correspond to $d = 3$; triangles to $d = 4$.

are much lower as triangles are not on the diagonal. In addition, 13 regions out of 16 are below the diagonal and only two are above, indicating that most regions tend to cite far blogs much more than they are cited by them. This also means that the majority of distance 4 links are made from regions below the diagonal to the two above. Those two regions belong to *society* continent and are *agora* and *technology*. We can qualify them as *popular* communities as they attract citations from distant blogs.

4 Related work

Many complex network studies have investigated different aspects related to structure and dynamics [3] and in this context understanding links creation is a key issue. In most previous works, the study of the impact of different factors on interaction processes has been studied at node scale. In particular, diffusion process such as rumor spreading [12], diffusion of innovation [17] or e-mail communication process [10] has been studied with regards to scale free networks properties [16, 21]. With the emergence of online social networks more attention has been given to individual human behaviour to understand global interaction phenomena, for example in [1, 19] the authors investigate how activity bursts are related to human dynamics. The study of citation links to understand information spreading phenomena in blog networks is a very active field [9, 6, 8, 11, 4].

Our approach is original with regard to those as it studies complex networks properties (in this case blogs citation behaviour) with regards to a community structure, i.e. at various community levels (and not at the node level). The authors of [6] study blogs interactions with regard to their semantics. The community structure we use is also based on blogs topics, but we consider a hierarchy of communities and not a single partition of nodes. Moreover, our methodology is generic as it may be used on any type of community structure (i.e. built according to any criterion).

The methodology we have presented requires a hierarchical community structure, but the creation of this structure (if it does not already exist as it is the case with our blog network) is not in the scope of our work. Many automatic community detection algorithms exist, which identify groups of nodes which have similar properties in general based on node similarity [7, 14]. [2] proposes a very efficient algorithm to compute a hierarchical community structure in very large graphs. Another approach proposed by [20] to detect hierarchical communities consists in creating link communities instead of nodes communities.

5 Conclusion and perspectives

We have proposed a generic methodology to analyse interaction behaviour in complex networks, with regard to a hierarchical community structure defined over their nodes. This approach mainly relies on two measures: *homophily* and *community distance*. The former evaluates in an unbiased way the tendency of nodes to interact within their own community. We have compared homophily with the *modularity* quality function and have shown their complementarity. Links community distance captures whether nodes of a network interaction with nodes from *close* or *distant* communities. We have applied this approach to a citation network of French blogs captured during four months, manually classified according to their topics. Citation links have been studied at various scales, which has given new insight on blogs topical communities. Finally, we have proposed a synthetic map based on an average value of community distances and have illustrated it at the region and territory levels.

One perspective of this work is to make a similar study using automatic community detection algorithm. Having several types of hierarchical community structures is a good opportunity to compare the information we get regarding citation links. In a second step, we want to investigate other networks and determine statistical metrics to better classify and understand hierarchical communities.

Acknowledgement

This research work is conducted and funded by Ville de Paris DiRe project and the European Commission through the EULER project (Grant No.258307) part of the Future Internet Research and Experimentation (FIRE) objective of the Seventh Framework Programme (FP7).

References

[1] Albert-Laszlo Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, May 2005.

[2] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech*, 10008, 2008.

[3] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.*, 81(2):591–646, May 2009.

[4] M. Cha, J.A.N. Pérez, and H. Haddadi. Flash Floods and Ripples: The Spread of Media Content through the Blogosphere. 2009.

[5] Aaron Clauset, M. E. J. Newman, , and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, pages 1– 6, 2004.

[6] Jean-Philippe Cointet and Camille Roth. Socio-semantic dynamics in a blog network. In *SocialCom 09 Intl Conf Social Computing*, pages 114–121. IEEE, 2009.

[7] Santo Fortunato. Community detection in graphs. *Physics Reports*, 2009.

[8] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *WWW '04*, pages 491–501. ACM, 2004.

[9] Jure Leskovec, Mary Mcglohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Cascading behavior in large blog graphs. In *In SDM*, 2007.

[10] R. Dean Malmgren, Daniel B. Stouffer, Adilson E. Motter, and Luís A. Amaral. A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences of the United States of America*, (47):18153–18158, November 2008.

[11] M. McGlohon, J. Leskovec, C. Faloutsos, M. Hurst, and N. Glance. *Finding patterns in blog shapes and blog evolution.* icwsm, 2007.

[12] Yamir Moreno, Maziar Nekovee, and Amalio F. Pacheco. Dynamics of rumor spreading in complex networks. *Phys. Rev. E*, 69(6):066130, Jun 2004.

[13] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, Sep 2003.

[14] M E J Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, (2):26113, 2004.

[15] Mark E. J. Newman, Albert L. Barabási, and Duncan J. Watts, editors. *The Structure and Dynamics of Networks.* Princeton University Press, 2006.

[16] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(14):3200–3203, Apr 2001.

[17] E. Rogers. *Diffusion of innovations.* Free Press., 1995.

[18] Abdelhamid Salah Brahim, Benedicte Le Grand, and Matthieu Latapy. Some Insight on Dynamics of Posts and Citations in Different Blog Communities. *IEEE ICC 2010 workshop "SocNets"*, 2010.

[19] Alexei Vázquez, João Gama Oliveira, Zoltán Dezsö, Kwang-Il Goh, Imre Kondor, and Albert-László Barabási. Modeling bursts and heavy tails in human dynamics. *Phys. Rev. E*, 73(3):036127, Mar 2006.

[20] Sune Lehmann Yong-Yeol Ahn, James P. Bagrow. Link communities reveal multiscale complexity in networks, 2010.

[21] Damián H. Zanette. Critical behavior of propagation on small-world networks. *Phys. Rev. E*, 64(5):050901, Oct 2001.