Journal of Computational Science 2 (2011) 247–252

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

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

ARTICLE INFO

Article history:
Received 22 November 2010
Received in revised form 7 May 2011
Accepted 28 May 2011
Available online 6 June 2011

Keywords:
Complex network
Community structure
Citation links
Blogs

ABSTRACT

In this paper we propose a generic methodology to study the correlation between nodes interactions in complex networks and their organization into groups called *communities*. We illustrate it on citations in a blog network. We first define a *homophily* probability evaluating the tendency of blogs to cite blogs from the same community. We then introduce the notion of *community distance* to capture whether 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 us to study citation links at various abstraction levels, and conversely, to characterize communities with regard to their citation behaviour.

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* [12] 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 [14]. 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 two 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 [14] in Section 3. This approach allows us to study interactions with regard to the 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 explain in Section 2.1 how a hierarchical community structure may be obtained for any complex network. In this section, we first introduce definitions related to the hierarchical community structure we used in the paper. We then define in Section 2.2 two metrics to evaluate whether interaction links relate nodes from a same community (at all levels of the hierarchical structure). We explain how these metrics are complementary to *modularity* which is used traditionally to evaluate partitions quality [12]. In Section 2.3 we introduce the notion of *community distance* to evaluate whether interaction links between nodes 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 community structure such that each node of V belongs to exactly one community at each level of the tree.[1] Communities may be based on nodes features, e.g. groups of web pages dealing with similar topics, or on topological information, e.g. hyperlinks between these pages.

Definition 1 *(Hierarchical community structure).* Given a community partition $P=\{C_1, C_2, \ldots, C_l\}$ of V, a sub-partition $P' =$

[☆] This work is supported in part by the ANR Webfluence and Ville de Paris DiRe projects.

* Corresponding author.
E-mail address: First-name.Last-name@lip6.fr (A. Salah Brahim).

[1] More general hierarchical community structures will be considered in the future to allow overlapping communities.

1877-7503/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jocs.2011.05.010

$\{C'_1, C'_2, \ldots, C'_m\}$ of P is a partition of V such that $\forall C'_i \in P'$, $\exists C_j \in P$ such that $C'_i \subseteq C_j$. This is denoted $P' \sqsubseteq P$.

A hierarchical community structure of G is defined as a series of partitions $P_k \sqsubseteq P_{k-1} \ldots \sqsubseteq P_2 \sqsubseteq P_1 \sqsubseteq P_0$ 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 at distance j of C in the community tree, i.e. $D_j(C) = \{C' \in P_{i+j},\ C' \subseteq C\}$, with $(i+j) < k+1$. Note that j is a relative distance with regard to the current level.

Definition 2 *(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 an interaction network and a hierarchical community structure (or a community tree), formally defined in Section 2.1. The first step of our methodology consist in evaluating, 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 3 *(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 proportion of edges of E' that connect two nodes from the same community at the jth level of the community tree, with $j > i$.

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

Note that, in this definition, the value of $\Delta_j(C)$ may be biased by the number of links in communities at the jth 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 $\Delta_j(C) \div \psi_j(C)$, 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 jth level of the hierarchy:

$$\psi_j(C) = \frac{\sum_{C' \in D_{j-i}(C)} |E'| \cdot (|E'| - 1)}{|E| \cdot (|E| - 1)}$$

High values of $\Delta_j(C) \div \psi_j(C)$ indicate a high homophily, i.e. a significant fraction of links between nodes from a same community at the jth level of the hierarchy, independently of the number of edges in these communities. The *modularity* function [11,5] has been defined to evaluate the quality of a partition; a high value of modularity means that there is a high density of links within communities of the partition and a low density of links between distinct communities. 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 therefore compare the value of $\Delta_j(C) \div \psi_j(C)$ with the value of modularity $Q_j(C)$:

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

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.

We will study interaction links homophily by first comparing $\Delta_j(C)$ value for the various communities at different levels of the community tree. We will then use $\psi_j(C)$ values two identify the most relevant $\Delta_j(C)$ values when these values are close for several communities.

2.3. Community distance

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

Definition 4 *(Community distance).* Given a couple of communities $u \in P_i$ and $v \in P_j$, there exists a minimal integer t such that there is a community C in P_t with $u \subset C$ and $v \subset C$. We then define the community distance of the spreading link (u, v) as:

$$d(u, v) = \frac{(i - t) + (j - t)}{2}$$

This distance will be used for communities at the kth level to characterize interaction links between nodes. 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:

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

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 blogs interactions (see Section 3.3).

3. Application to a real-world case

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

3.1. Dataset

The dataset we used for our experiment was obtained by daily crawls of 6007 blogs in the French-speaking blogosphere (1,074,315 posts) during 4 months in 2009.

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*