

**THESE DE DOCTORAT DE
L'UNIVERSITE PIERRE ET MARIE CURIE**

Spécialité

SYSTEMES INFORMATIQUES

Présentée par

M. SALAH BRAHIM ABDELHAMID

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE PIERRE ET MARIE CURIE

**Diffusion d'information et structure en communautés
dans un réseau de blogs**

Soutenance prévue le : le 8 Décembre 2011

Jury :

Isabelle CHRISMENT	Rapporteur	Professeur, Henri Poincaré University, Nancy 1
Marie-Aude AUFAURE	Rapporteur	Professeur, Ecole Centrale Paris
Marcelo DIAS DE AMORIM	Examineur	Directeur de recherche CNRS, UPMC
Christophe PRIEUR	Examineur	Maître de Conférences, Paris-Diderot
Cécile BOTHOREL	Examineur	Maître de Conférences, Telecom Bretagne
Matthieu LATAPY	Co-directeur de thèse	Directeur de recherche (DR) CNRS, UPMC
Bénédicte LE GRAND	Co-directeur de thèse	Maître de Conférences (HDR), UPMC



**DOCTOR OF SCIENCE THESIS
PIERRE AND MARIE CURIE UNIVERSITY**

Specialization

COMPUTER SCIENCE

presented by

M. SALAH BRAHIM ABDELHAMID

for obtaining the degree of

**DOCTOR OF SCIENCE FROM PIERRE AND MARIE CURIE
UNIVERSITY**

**Information diffusion and community structure
in a blog network**

Defence: December, 8th 2011

Commettiee :

Isabelle CHRISMENT	Reviewer	Professor, Henri Poincaré University, Nancy 1
Marie-Aude AUFAURE	Reviewer	Professor, Ecole Centrale Paris
Marcelo DIAS DE AMORIM	Examiner	Researcher (DR) CNRS, UPMC
Christophe PRIEUR	Examiner	Associate Professor, Paris-Diderot
Cécile BOTHOREL	Examiner	Associate Professor, Telecom Bretagne
Matthieu LATAPY	Co-supervisor	Researcher (CR) CNRS, UPMC
Bénédicte LE GRAND	Co-supervisor	Associate Professor (HDR), UPMC

Remerciements

Une thèse est le fruit d'un travail collectif. C'est la raison pour laquelle je tiens à remercier ici toutes les personnes m'ayant aidée tout au long de mon parcours.

Tout d'abord, je remercie vivement mon encadrante Bénédicte LE GRAND et mon directeur de thèse Matthieu LATAPY pour m'avoir en premier lieu, donnés la chance de réaliser ma thèse dans les conditions les plus favorables, ensuite pour m'avoir guidé, conseillé, soutenu et pour avoir été disponible tout au long de ma thèse.

J'aimerais aussi exprimer toute ma gratitude à Isabelle CHRISMENT et Marie-Aude AUFAURE pour avoir accepté d'être rapporteurs de cette thèse. Je remercie aussi vivement Marcelo DIAS DE AMORIM, Cécile BOTHORE et Christophe PRIEUR qui m'ont fait l'honneur de faire partie de mon jury de thèse.

Je tiens à remercier également tous les membres de l'équipe "Complex Networks" du Lip6, en particulier Jean-loup Guillaume et Clémence Magnien avec qui j'ai beaucoup appris. Je remercie également tous les thésards et les post-doctorants qui m'ont accompagné pendant ces années et grâce à qui j'ai travaillé dans une bonne ambiance. Merci à eux pour leur gentillesse et les nombreuses discussions partagés. Merci aussi à tous les membres du personnel du laboratoire, qui m'ont facilité les tâches administratives, en particulier Véronique Varenne pour sa gentillesse et sa précieuse aide.

Merci à toutes mes amies pour leur soutien et leur écoute, je citerais particulièrement Oussama, Lamia, Assia, Nadjib, Thomas, Nadjet, Anissa et tous les autres avec qui j'ai partagé des bons moments au Lip6. Je remercie mes amis de long date, les deux Hocine, Hosni, toute la bonde de Clichy et plus spécialement Youcef.

Enfin, je tiens à exprimer ma profonde affection et mes plus chaleureux remerciements à ma famille, en particulier mes chers parents Saliha et Kaddour qui m'ont toujours soutenue et qui m'ont fait confiance. Leur amour est pour beaucoup dans ce travail, je leur dédie spécialement cette thèse. Merci à mon frère Arslan et ma soeur Fatima et à Maha qui m'ont toujours encouragé. Je ne pourrais pas oublier ma grand-mère et toutes mes tantes, oncles, cousins et cousines et plus spécialement mon oncle Ali et Adel.

Pour terminer, je voudrais remercier plus généralement, toute personne qui a contribué de près ou de loin à l'aboutissement de ma thèse.

Table of contents

1	Introduction	7
2	Context and Objectives	13
2.1	Introduction	14
2.2	Complex networks: definitions and research area	14
2.2.1	Measurement and metrology	15
2.2.2	Analysis	15
2.2.2.1	Basic definitions	16
2.2.2.2	Common properties of complex networks	18
2.2.3	Modeling	18
2.2.4	Algorithmics	18
2.3	Social networks and interaction links	19
2.4	Diffusion phenomena	21
2.4.1	Influence patterns	21
2.4.2	Information cascades	22
2.5	Communities	24
3	Blog network analysis: global approach	27
3.1	Introduction	28
3.2	Blog network description	28
3.2.1	Blogs	28
3.2.2	Data description	31
3.2.3	Blog and post network	33
3.2.4	Data Cleaning	35
3.3	Activity analysis	36
3.3.1	Blogosphere activity	36
3.3.2	Blog and post networks characterization	37
3.4	Post popularity and community structure	41

3.4.1	Dynamics of post and citation arrivals	41
3.4.2	Biases	42
3.4.3	Post popularity	43
3.4.4	Two types of citation dynamics	44
3.4.5	Post popularity and impact of community structure	45
3.5	Conclusion	47
4	Citation link study: community oriented approach	49
4.1	Introduction	50
4.2	Framework	51
4.2.1	Hierarchical community structure	51
4.2.2	Homophily	52
4.2.3	Community distance	54
4.3	Case study: topical community structure	55
4.3.1	Dataset and community structure description	56
4.3.2	Citation links homophily	56
4.3.3	Citation links community distance	60
4.3.3.1	Community profiling based on links distance	60
4.3.3.2	Community mapping based on community distance	63
4.4	Case study: Automatic community structure	66
4.4.1	Louvain community detection algorithm and community cores	67
4.4.2	Community structure	68
4.4.3	Results	68
4.5	Summary and perspectives	72
5	Network cascades	75
5.1	Introduction	76
5.2	Cascades definition and computation	76
5.2.1	Data corpus	76
5.2.2	Cascade computation	77
5.3	Macroscopic analysis: cascade structure and community impact	78
5.3.1	Cascades shapes	78
5.3.2	Cascades topological, temporal and community properties	81
5.3.3	Temporal and topological cascades properties correlation	85
5.4	Microscopic analysis: impact of individual nodes on cascades	88
5.4.1	Impact of cascade origin	88
5.4.2	Impact of intermediate blogs	89
5.5	Conclusion	93
6	Conclusion and Perspectives	97

List of figures	101
List of tables	105
References	107

Introduction

We can cite many examples of complex networks in the real world: in computer science we find the internet (networks of interconnected routers, autonomous systems or computers), the web (a set of web pages related by hyper-links), overlay networks (eg. Peer-to-peer networks) and content sharing networks (email or file exchanges). We can also cite many examples of complex networks in other fields: in sociology and human science (all types of social networks), biology (the brain network or protein interaction networks), linguistics (for example co-occurrence and relations between words) and many others. These networks are generally modelled as mathematical graphs, where nodes are the elements of the network and links represent nodes interactions.

Complex networks are also called *interaction networks*, as nodes are characterized by their own features but also by their relationships with other nodes, as interactions occur between networks' elements. For example, in a social network such as *Facebook*, users interact and build relations with one another (for example a friendship relation). These interactions have not been fully understood yet. Therefore, data analysis appears as a necessary step towards a better understanding of phenomena occurring in real-world complex networks. Understanding how *interaction networks* are structured and which events may occur within them, are indeed key questions of the field. As most of these networks share non-trivial statistical properties [BA99b, Lat07, WS98a], it is relevant to consider them as a coherent group.

The *analysis of interaction networks* aims at characterizing their structure and the evolution of their properties over time. The objective is to capture the key features of the

graph. This topic has led to an important stream of studies [BGLLf08, AJB00].

Interactions may be observed at different levels, microscopic at the node level or macroscopic at the graph level. More precisely, interactions within the network induce the creation of *links* between nodes. Through links, members may create new relations, exchange information or influence each other. For example, if one of my friends shares an information I am interested in, I may spread this information: I may share this information with my friends (my neighbours in the network). This phenomenon is called *diffusion* or *spreading* process, and is at the core of this thesis. Studying diffusion has a high interest and also many applications. For example, viral marketing aims at using existing social networks and encouraging customers to share product information with their friends [Mon01]. Information dissemination is also an important feature where the challenge is to determine the best way for an information to be widely adopted [KKT03, HPV]. In the opposite, we want to be aware of any new information that appears in the network with a minimum cost (monitoring only a relevant portion of the network) [LKG⁺07]. In addition to those applications, it is essential to better understand those phenomena to be able, in a second step, to produce more realistic diffusion models.

Investigating how links appear and characterizing this behaviour is crucial to better understand diffusion phenomena over those networks. Indeed, links may be a way to spread information, an opinion or to propagate an influence. Diffusion phenomena may be studied in different ways. First, by studying how a node propagates its content towards its neighbours. Second, by investigating how a node gets an information from its neighbours (how it adopts it). Third, by considering a diffusion *cascade* where an information spreads from one node to the rest of the network through a succession of diffusion events.

Until now diffusion phenomena have been mostly considered at a macroscopic scale i.e. by studying all nodes of the network as a whole. We give a complementary way to analyse the network interaction by considering the problem at different scales. To that purpose, we use the *community structure* of the network [NBW06]. It has been observed that nodes with common features tend to interact preferentially with each other [WF94, GN02]. These groups of nodes, called *communities*, are also central in this thesis. Many definitions of "community" have been proposed in literature. We adopt the following definition: "*A community is a set of nodes with common features or interests*". The community structure enables an analysis at different scales: local (individual nodes), global (whole network) and intermediate scales (groups of nodes).

The approach adopted in this work is empirical. Indeed, I analyse a real network collected with a *measuring* procedure. The benefit of the empirical approach is that the

results are based on real observations. However it requires a high attention to validate the observations and results. The collected data may indeed contain many biases for different reasons and dealing with them is itself a challenging problem.

The goal of my thesis is to empirically study diffusion phenomena in complex networks, at various scales provided by their community structures.

I conduct this study on a *French blog network*¹. A blog is a web site constituted of a set of posts which are short publications, usually dedicated to a specific subject and written by the author of the blog. This work has been achieved in the context of an ANR project named *Webfluence* that aimed to study a blog network from different aspects, sociological as well as graph approaches.

Contributions I present here an overview of the contributions of my thesis summarized in three points:

- First, I have studied post and blog "popularity" within the blogosphere. The aim is to capture how the influence of a post (or a group of posts) evolves over time. I have been able to identify several patterns with regards to the community structure which showed that observing interactions at community scale was relevant.
- Second, I have studied the tendency of a node to be linked to nodes from its own community. I have defined two metrics named *homophily* and *community distance* to measure it. I have been able to distinguish different behaviour patterns and also to produce synthetic maps to classify communities.
- Finally, I have investigated how an opinion starting from one post influences and spreads towards other blogs. This succession of *influence* is represented as a graph and is called a *cascade*. I have studied cascades through three angles: topological properties (e.g. size), temporal properties (e.g. cascade duration) and community properties.

This manuscript is organized as follows. In chapter 2, I introduce the context of this work and propose a state of the art regarding methods and results used for the analysis of diffusion phenomena and community structure analysis. I present especially works related to social networks. Chapter 3 starts with a description of the blog network I used for this work. Then I present a methodology for analysing post popularity and detecting a community pattern behaviour. In Chapter 4 I introduce two metrics to evaluate the impact of the community structure on diffusion phenomena: links *community homophily*

1. See description in Section 3.2

and *distance*. These metrics allow to classify communities according to their citation behaviour. Chapter 5 investigates diffusion cascades and the influence of individual and community behaviors. Finally, Chapter 6 presents my conclusions and some key directions for further work.

Context and Objectives

Contents

2.1	Introduction	14
2.2	Complex networks: definitions and research area	14
2.2.1	Measurement and metrology	15
2.2.2	Analysis	15
2.2.3	Modeling	18
2.2.4	Algorithmics	18
2.3	Social networks and interaction links	19
2.4	Diffusion phenomena	21
2.4.1	Influence patterns	21
2.4.2	Information cascades	22
2.5	Communities	24

2.1 Introduction

Complex networks study is a recent field consisting in considering networks collected from a real context also called *real-world networks*. We can cite many examples such as computer, social, biological, or linguistic networks, internet maps, web graphs, data exchanges and co-authoring networks, protein interactions or word occurrence networks. These networks may be modeled with graphs where networks' elements are *nodes* (*vertices* in graph theory) and are related by *links* (*edges* in graph theory).

These graphs are studied through graph theory and algorithmic approaches, which are traditionally used to study algorithmic properties of random graphs (i.e. graphs defined mathematically). Many designations have been proposed to represent this type of graphs. In this manuscript, I will use equivalently *interaction networks* and *complex networks*.

Understanding how interaction networks are structured, how they evolve and grow, and what phenomena impact them are the central questions in this research field [FN93, Lat07]. Earlier studies have shown fact that most complex networks, even from different context share many characteristics [BA99a].

In this chapter I present a state-of-the-art of notions and methods used for network analysis. I start with basic definitions that will be used in the manuscript. The state-of-the-art is organized in three parts: interaction network characterization (static and dynamic), community structure in complex networks and finally link diffusion and spreading phenomena.

2.2 Complex networks: definitions and research area

Many real world networks share some non-trivial properties. Moreover, many questions and problems raised by these networks are general and transversal. Therefore, it is possible to work on problems independently from the case of study, and try to apply these results to other real world cases.

These problems can be divided into four principal research axes to group the main questions of the field [Lat07]. I give a brief description of each one below with more emphasis on analysis as I mostly contributed to that research axis.

2.2.1 Measurement and metrology

In general, interaction networks are not directly available. The *measurement* is the operation which consists in collecting information about the nodes and links of the considered object. Measurement methods naturally depend on the studied network. In most cases it is impossible to measure the whole network. Indeed, the measurement procedure is limited, and a trade-off must be found between measurement frequency, duration and the quantity of collected information. As an example, we may cite the case of internet [LMO08,DRFC05] and Peer-to-peer networks measurements [ALM09b,ALM09a].

Metrology consists mainly in the study of biases introduced by the measurement procedure. As we cannot capture the totality of nodes and links, metrology analyses the gap between the measured sample and the real graph and the effects the various results that can be obtained with this sample. Many works have been done in the field and especially in the context of the internet [ACKM09,GL05]. One bias due to measurement is when a property of a node is studied over time. As nodes arrivals are not synchronised, the observation duration may be different from a node to another and the conclusion could therefore be biased [Mag10]. In Section 3.3 I show in details how such a bias may be removed, based on a methodology used in another context [SR06].

2.2.2 Analysis

One may consider the analysis part as the most crucial for understanding real world networks and phenomena occurring within them. When we perform an analysis, the underlying social network can be considered at two different levels of resolution: one in which the network may be seen as a heterogeneous population of individuals and is observed in an aggregate way, and another which considers the structure of the network and studies how individuals are influenced by their neighbours. Indeed, a first step in real world network analysis is to describe the main graph features and properties. Therefore, we use statistical and structural notions in order to synthesize relevant graph characteristics. In addition, one may be interested in analyzing the network at a microscopic scale, for example at node scale for a static analysis (e.g. with the degree or betweenness centrality) or a dynamic analysis (for example link arrival over time) [BHKL06].

Networks from different contexts continually evolve over time. This dynamics impacts the network topology, as nodes and edges are added and deleted. This evolution can be observed at two scales:

1. The evolution of macroscopic network properties, like diameter and network density (through a series of network snapshots) [LKF07, CSN07].
2. The network evolution at the level of individual edges and nodes (for example nodes popularity). Studying individual evolution is important as microscopic mechanisms induce some properties observed at the macroscopic level [CMAG08, HPV].

2.2.2.1 Basic definitions

In this Section I present basic definitions and notations used throughout this manuscript.

A graph $G = (V, E)$ is defined by a set V of *nodes* (*vertices* in graph theory) and a set $E \subseteq V \times V$ of *links* (*edges* in graph theory). We denote by $N(u) = \{v \in V, (u, v) \in E\}$ the neighbourhood of a node u in G i.e. all nodes which are connected to u by a link of the graph G . The number of nodes in $N(u)$ is the *degree* of u : $d(u) = |N(u)|$. In *undirected* graphs, there is no distinction between links (u, v) and (v, u) .

A *directed graph*, also called a *digraph*, is a network in which each edge has a direction, from one vertex to another. In a directed network each vertex has two degrees. The *in-degree* is the number on incoming links and the *out-degree* is the number of outgoing links.

The basic statistics describing such a graph are its size $n = |V|$ and its number of links $m = |E|$. Its average degree is $k = \frac{1}{n} \sum_{u \in V} d(u) = \frac{2m}{n}$; its density is $\delta = \frac{2m}{n \cdot (n-1)}$, i.e. the number of links divided by the number of possible links between all pairs of nodes.

Moreover, a value can be assigned to every link to represent a property (for example a cost to follow the link). In this case this is a *weighted graph*. For example, in the internet, the bandwidth of links may be modeled by weight.

Degree distribution

The degree distribution of a graph indicates for each integer k , the number P_k of nodes with a degree equal to k : $P_k = |\{u \in V : d(u) = k\}|$.

Degree distributions play a key role in graph analysis. In particular *homogeneous* and *heterogeneous* ones may be distinguished.

In homogeneous distributions (such as normal, Gaussian and Poissonian distributions) the degrees of nodes are very close to the average degree. This means that the average degree gives an important information as it indicates the expected behavior of nodes.

Heterogeneous distributions (such as Zipf and power-law distributions) are such that there are several orders of magnitude between degrees, and most nodes have a degree very different from the average degree. Then, the average value gives little information: it is very different from the degree of most nodes, and randomly chosen nodes may have very

different degrees.

Many works have demonstrated that real world characteristics follow power laws. For Example, in the Internet most routers have a very low degree, while a few routers have extremely high degrees [FFF99]. Power-law distributions attempt to fit this degree distribution. We can cite many other examples as citation graphs [Red98], web network [BKM⁺00, KKR⁺99], online social networks [CZF04], Internet AS (Autonomous Systems) graph [FFF99].

Power law distributions have been used to characterize many network properties. In [Mit03, New06, CSN07] one can find a detailed mathematical analysis of network properties that usually follow power law distributions.

Clustering coefficient

The clustering coefficient of a node u is defined as the fraction of existing connections among its nearest neighbours divided by the total number of possible connections. It is a measure of the local density of nodes and corresponds to the probability that two nearest neighbours of u are connected with each other. In other words, it is the probability that any two neighbors of any node are linked together.

It is defined for any node u of degree greater or equal to 2:

$$cc_{\bullet}(u) = \frac{|\{(v, w) \in E \text{ s.t. } v, w \in N(u)\}|}{\frac{d(u) \cdot (d(u)-1)}{2}}.$$

The clustering coefficient of the graph itself is the average of this value for all nodes:

$$cc_{\bullet}(G) = \frac{\sum_{u \in V, d(u) \geq 2} cc_{\bullet}(u)}{|\{u \in V, d(u) \geq 2\}|}.$$

A second notion of clustering coefficient (sometimes called *transitivity ratio*) applies directly to the whole graph G :

$$cc_{\vee}(G) = \frac{3N_{\Delta}}{N_{\vee}}$$

where N_{Δ} denotes the number of triangles, *i.e.* sets of three nodes with three links in G , and N_{\vee} denotes the number of connected triples, *i.e.* sets of three nodes with at least two links in G . This notion of clustering is slightly different from the previous one since it gives the probability, when one chooses two links with one common extremity, that the two other extremities are linked together.

2.2.2.2 Common properties of complex networks

Earlier studies have shown that most complex networks have non-trivial properties in common [WS98b]. This suggests that even if we study some specific complex networks, some results may be valid for complex networks in general. The common properties of complex networks [Lat07] are:

- **Low diameter:** the average distance between nodes is low, i.e. there generally exists a short path between every pair of nodes. This property leads to the notion of *small world* [Mil67](see Section 2.3) or also "six degree of separation" according to which two nodes are separated by no more than six nodes.
- **Very low density:** this is equivalent to a very small average degree comparing to n ; in other words, when two nodes are randomly chosen, the probability that they are linked is very small.
- **High clustering coefficient:** there are significantly more chances that two nodes are linked if they have a neighbour in common than two nodes chosen randomly.
- **Heterogeneous degree distribution:** it is generally approximated by a power law, $p_k \sim k^{-\alpha}$, with α between 2 and 3 in general.

These properties are now considered as fundamental in the complex networks domain and are often completed by others applying to specific networks. For example, [RB02, PGF02] show that real-world networks are resilient to random node attacks, i.e., connectivity is not highly impacted even if many random nodes are removed.

2.2.3 Modeling

In order to conduct simulations, it is essential to capture the observed properties in practice through complex networks models. A first class of models aims at generating synthetic graphs with specific properties. For example the degree distribution which follows a power-law to reflect real-world observation as in the model proposed in [?]. A second class of models focuses on diffusion phenomena rather than graph structure. A large amount of work on information diffusion or influence has been done in different contexts. For example, *epidemiology* studies diseases or viruses spreading [Bai75, AM92].

2.2.4 Algorithmics

Because of their large size, working on real networks naturally leads to algorithmic problems. A first class of algorithms deals with classical problems in graph theory (as graph diameter or shortest path) but are no longer applicable in the context of complex networks. This is principally due to the scale of such objects. The second class concerns

new algorithmic problems that appeared with the emergence of real interaction networks. For example:

- Community detection: a decomposition into groups such that the number on intra-group links is maximized. Therefore, a local density measure is used, such as clustering coefficient or modularity [NG04].
- Information spreading: e.g. find the best way to maximize a spreading flow while minimizing the cost of the operation [LKG⁺07].

2.3 Social networks and interaction links

A social network is in general a representation of a set of relations between some individuals. It is modeled with a graph where nodes are people and edges are relations between them, such as a friendship. It is traditionally studied by sociologists who analyze the connections between individuals or collective behaviours and social structures. In sociology we refer to vertices or people as *actors* and edges as *ties*. Sociologists usually obtain their data from interviews with the analyzed people. This data, although very detailed, can be difficult to obtain. The process of interviewing people is often long and costly and so the obtained datasets are rather small, with several hundreds of analyzed relations in the best cases.

For most people, social networks nowadays refer to online social networking as *Facebook* or *MySpace*. This term started to be used at the beginning of the twentieth century by Georg Simmel. In the 1930s, Jacob L. Moreno was a pioneer in collecting and analyzing social interactions in small groups, especially classrooms and work groups. In 1954, John A. Barnes [Bar54] started using the term *social network* for patterns of ties. Many other works have contributed to the development of the field as Elisabeth Bott's on kinship [Bot57], Sigfried Nadel and Harrison White on social structure [Nad57].

One of the most famous social network experiment is certainly small-world experiment done by Stanley Milgram in the 1960s [mil69,Mil67]. Milgram was interested in quantifying the distance between two actors in social networks. It corresponds to the minimum number of edges that must be traversed to travel from one vertex to the other through the network. He found that the average length of completed paths was only 5.9. This result is the origin of the idea of the "six degrees of separation".

An edge in a social network may be defined in many different ways. A particular definition will depend on what questions we want to answer. An edge can represent a friendship or professional relation, communication pattern or money exchange.

Recently, different aspects related to the dynamics of real world networks have been investigated, in particular key measures like link creation and deletion [SBF⁺08]. In the context of social networks, such metrics are closely related to interaction patterns, as they shed light on typical human behaviours, for example bursting activity patterns [?].

Online social networks — and blogs in particular — provide interesting grounds for such studies: they may be described globally, giving insights on overall macroscopic trends [BGH⁺10]. Moreover they give a lot of information, enabling rich descriptions of datasets, like posting and commenting profiles [MT10].

A recurrent topic of interest in this context is the detection of information bridges among groups, like political parties [AG05a]. One possible way consists in tracking similar contents throughout the dataset as in [GGLNT04]. On the other hand, citation links are supposed to play the role of communication pathways, so they may be the elementary bricks of diffusion processes, and may thus be compared to models of such phenomena [BGH⁺10,PSV01]. This hypothesis has been studied in several contexts, including photo-sharing platforms [CPH09] or blog networks [LMF⁺07]. It is then acknowledged that their configuration gives insights on the mechanisms of influence among actors [CR09]. Finally, studying their evolution gives clues to understand dynamical trends such as the evolution of popularity [SBLGL10].

One class of networks is *citation networks*. A classical example is author / co-author citation network [BHKL06]. It corresponds to a network of citations between academic papers. The network is constructed with vertices as papers and there is a directed edge from paper *A* to paper *B* if *A* cites *B* [LKF05] as I will explain in Section 3.2.1. The important difference between citation networks and others (for example the world wide web) is that a citation network is *acyclic*: there is no closed loop of directed edges. The reason is that if a paper wants to cite another paper, this one must already have been written. In blog networks, we may find closed loops for many reasons, like local time error, human mistake or even collecting data bias. In this case a preliminary correction data step is required as described in Section 3.4.2.

With the growth of the *internet*, many new interaction networks have emerged. These networks are called *information networks* and the most studied object is the *World Wide Web*. Many of these networks have a social aspect, for example networks of e-mail communication, network of social networking websites such as *facebook* [fac] or *twitter* [twi] and networks of blogs and online journals. In this thesis I will mainly focus on a blog network described in Section 3.2. The blog network is also a citation network and has similar characteristics. The evolution of the web and the appearance of large online social networks gives an opportunity to observe diffusion phenomena in new contexts. In these networks

information spreads across social links. Content, under the form of ideas, products, and messages, spreads across social connections.

2.4 Diffusion phenomena

The two main families of diffusion models from the literature are *threshold* [Gra78, GLM01] and *cascade* models [AA05, CR09]. In threshold models, one node is 'infected' if a given number (or proportion) of its neighbors are infected. In this context, it can be seen as an opinion adoption model. In cascade models, each infected node spreads information towards a given proportion of its neighbors.

2.4.1 Influence patterns

Influence has been studied in many different contexts like sociology, communication, marketing, and political science [KP05, Rog62]. The notion of influence has an importance in different contexts, for example in understanding how businesses operate, how a fashion spreads [Gla02] and how people vote [EK]. Studying influence patterns can help us better understand why certain trends or innovations are adopted faster than others and how we could help advertisers and marketers design more effective campaigns. However it is difficult to get readily available quantification, and essential components like human choices and the ways our societies function cannot be reproduced within the confines of the lab.

A more recent view has investigated the role of influentials. It shows that the key factors impacting influence are the interpersonal relationships between ordinary users and the preparedness of people to adopt [WD07]; this approach has been used for marketing strategies such as collaborative filtering. In this thesis, in addition to studying the role of influential nodes, I want to identify roles and influences of group of communities on information propagation.

Online communities have become a significant way to receive new information, and influence in such communities needs to be explored.

Identifying influence patterns may be seen differently depending on the context. In this thesis I study a blog network (described in Section 3.2) with an innovative approach. More precisely, I measure the influence of a blog according to the way it spreads information and to the way it is cited by other blogs.

The authors of [AG05b] have studied discussion topics and citation patterns of political bloggers for the 2004 U.S. Presidential election. They have focused on political blogs and

have shown that blog behaviors could be different according to the type or topic of the blog.

The diffusion of topics over time has also been addressed in [GGLNT04]. This study concluded with the existence of two types of topics: *chatter* topics, with a stable popularity over time, in opposition to *spiker* topics with varying popularity values. In Chapter 3 I also investigate the impact of the topic of a blog on its dynamics.

2.4.2 Information cascades

Interactions between groups of individuals in real networks have a fundamental role in information spreading. Cascades are also known as *fads*. Many works observed how an idea or action was suddenly widely spread through the network. In blog networks, a story or piece of information may be widely cited by the blogger community and is eventually also referred to by the mass media.

A cascade is due to an activation of nodes (individuals) in a graph. A node is activated if it is influenced by the state of its neighbors. A related formalism is a graph where a directed edge (i, j, t) indicates that node i influenced node j at time t .

In many contexts, information cascades are spreading phenomena which can result in a wide adoption [BHW92]. Cascades have been studied for many years by sociologists to understand diffusion of innovation [Rog62]; more recent works have investigated cascades for the purpose of selecting trendsetters for viral marketing [DR01], finding inoculation targets in epidemiology [NFB02], and explaining trends in blogosphere [KNRT05].

In this thesis I investigate information spreading based on cascade extraction. The statistics computed are based on Leskovec et al. methodology [LMF⁺07] which uses citations between blogs to extract cascades rather than content analysis methods, which produce less realistic cascading behaviors [GGLNT04]. However, I have proposed original approaches to define blog classes based on topological and community features of the network (see Section 3.4).

Figure 2.1 illustrates a cascade sample which corresponds to a Directed Acyclic Graph. In a cascade the node which starts the spreading is called *cascade initiator*¹. Its role may be critical for the cascade propagation. Therefore, I have investigated (see Section 5.4) how the community of the first node of the cascade impacts the cascade properties.

1. We consider only one cascade initiators but a cascade with multiple initiator can be considered in some contexts.

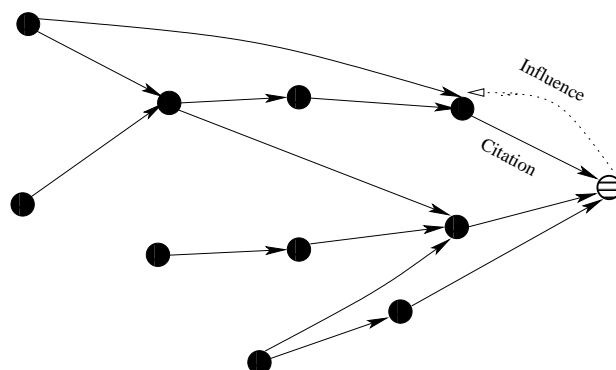


Figure 2.1: Cascade sample

Social cascades have been studied through popular photos and a social network collected from Flickr: [CMAG08] investigated how information disseminates through social links in online social networks and showed that social cascades were an important factor in the dissemination of content.

Previous research on the diffusion of information and influence over networks has been done in the context of epidemiology and the spreading of diseases [Gra78, GLM01]. There are many models of influence spreading in social networks. I will not detail epidemiological models because they are not our purpose. We have to note that for example a model of product diffusion predicts the number of people who will adopt an innovation over time. It does not explicitly account for the structure of the social network but it assumes that the rate of adoption is a function of people who have adopted. In my work I will not only consider how wide a diffusion is but also topological, temporal and community characteristics.

Most works have focused on the study of the *topological patterns* of the underlying contact networks and their influence on the properties of spreading phenomena in social networks such diffusion of information, innovations, computer viruses, opinions [BBV08, DNMKKL09, IKMW07]. However, most of these studies of dynamical phenomena on social networks do not consider the community structure of the network.

2.5 Communities

Community research field is strategic to understand and analyze complex networks. The first type of research on communities is related to community detection algorithms. Many automatic community detection algorithms exist, in general based on node similarity, which identify groups of nodes which have similar properties [For09,NG04]. [BGLL08] proposes a very efficient algorithm to compute a hierarchical community structure in very large graphs. Another approach proposed by [YYA10] to detect hierarchical communities consists in creating link communities instead of nodes communities. All these algorithms try to maximise an objective function (for example modularity [NG04]) and have been tested on synthetic and real graphs. Community detection is not our purpose in this thesis.

The second type of research on communities is their analysis. Unfortunately only few works tried to understand the community structure of real-world complex networks. During the 2004 U.S. Presidential election, the authors of [AG05b] studied discussion topics and citation patterns of political bloggers. They showed that blog behaviors could be different according to blog types or topical communities. For example news blogs do not behave as personal diaries. In [BHKL06] authors have proposed a framework for comparing the different kinds of communication dynamics within different communities of social communication networks.

In this chapter, I have presented the methods and the notions mostly used for analysing real-world networks in general and social networks in particular. I have presented how links have been studied in different contexts and for various purposes. As we have seen, most of these works are based on topological graph properties, but the community structure of the network is rarely considered to analyse spreading. In this thesis, I address topological and temporal blog network characteristics and show community impact with different determined patterns.

Blog network analysis: global approach

Contents

3.1	Introduction	28
3.2	Blog network description	28
3.2.1	Blogs	28
3.2.2	Data description	31
3.2.3	Blog and post network	33
3.2.4	Data Cleaning	35
3.3	Activity analysis	36
3.3.1	Blogosphere activity	36
3.3.2	Blog and post networks characterization	37
3.4	Post popularity and community structure	41
3.4.1	Dynamics of post and citation arrivals	41
3.4.2	Biases	42
3.4.3	Post popularity	43
3.4.4	Two types of citation dynamics	44
3.4.5	Post popularity and impact of community structure	45
3.5	Conclusion	47

3.1 Introduction

In this chapter I explore new approaches and methods to characterize blogs dynamics. In particular, the evolution of post popularity over time is studied, as well as behaviour patterns. I aim at going beyond traditional approaches by defining classes of dynamic behaviors based on topological features of the post network, and by investigating the impact of topical communities on blog dynamics.

The chapter is organized as follows. After a description of a blog network and of the dataset I used for this study in Section 3.2, Section 3.3 presents the results of traditional blog and post analysis. My contributions related to the characterization of post and citation dynamics according to their community structure are described in Section 3.4, before concluding and presenting perspectives of this work.

The work presented in this chapter was published in [SBLGL10].

3.2 Blog network description

Over the last years, the activity and popularity of online networks in general and blogs network in particular, have been increasing continually. It has become an important medium of communication and information on the Web for a large population due to an easy use and an intuitive interaction. In the research domain, this virtual network is extremely popular due to its timely nature and the observation of the creation and spreading of information and opinions at a large scale. The work presented in this chapter was published in [SBLGL10].

3.2.1 Blogs

A blog is a sort of website. Blogs are usually maintained by an individual who regularly publishes information to describe events or to comment multimedia resources such as images or video. A publication is called a *post* and is commonly displayed in reverse-chronological order. Two features of posts are important in this thesis. First, each post has an unique hyper-link address so other posts can access it and potentially refer to it. The blog webpage displays the last published post. Second, the time of post publication is known: each post has a timestamp which I use for temporal analysis.

Most blogs are interactive: visitors may write a comment and send messages to the author; this interactivity distinguishes blogs from static websites. Many blogs provide comments or news on a particular subject; others function as more personal online diaries. A typical blog contains links to other blogs, text, images, and links and other media related

to its topic. Blogs are generally textual, however some blogs focus on art (art blog), videos (video blogging), and audio (podcasting).

Types of blogs

There are many different types of blogs. One may classify blogs according to their content but also to the way this content is written and presented.

Many blogs focus on a particular subject, eg. politics, education, fashion, house, travel, project, music, quizzing, family, mom blog (blog discussions especially about home and family), legal issues and many others.

Most blogs are personal some others operate in a collaborative way. Personal bloggers usually take pride in their blog posts, even if their blog is read by only few people. Political blogs may have many contributors with the same goals, who discuss their political orientation and give their opinion on daily people preoccupations. This is also the case in blogs dedicated to news (considered as a mass media). Journalists and specialists of various fields (sociology, art, sport, music) maintain their blogs and their publications are an extension of the journalistic work. Therefore, they are more professional and have a different way of writing and presenting their content, and making citations to other blogs (which has a direct impact on the analysis I make in this thesis).

Blog structure

As I said above, a blog is constituted by a set of posts written at a determined time (*date* in Figure 3.1) with a text body with references to other pages on the web (for example pictures, videos and websites). In the text we may find a reference to a previous post, from the same blog (auto-citation) or from another blog, by quoting the corresponding URL, which is called a *citation link* (see Figure 3.1). Citation links are very important in my thesis as they represent the citation interaction between posts (and consequently blogs).

In addition to citation link and publication date the blog is composed of a *blog-roll* which is a static set of blogs which the current blog is referring to as blogs with the same interest. The *blog-roll* is associated to the whole blog and not to individual posts and is not my focus here.

Each blog and post is identified by its hyper-link address. 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*).

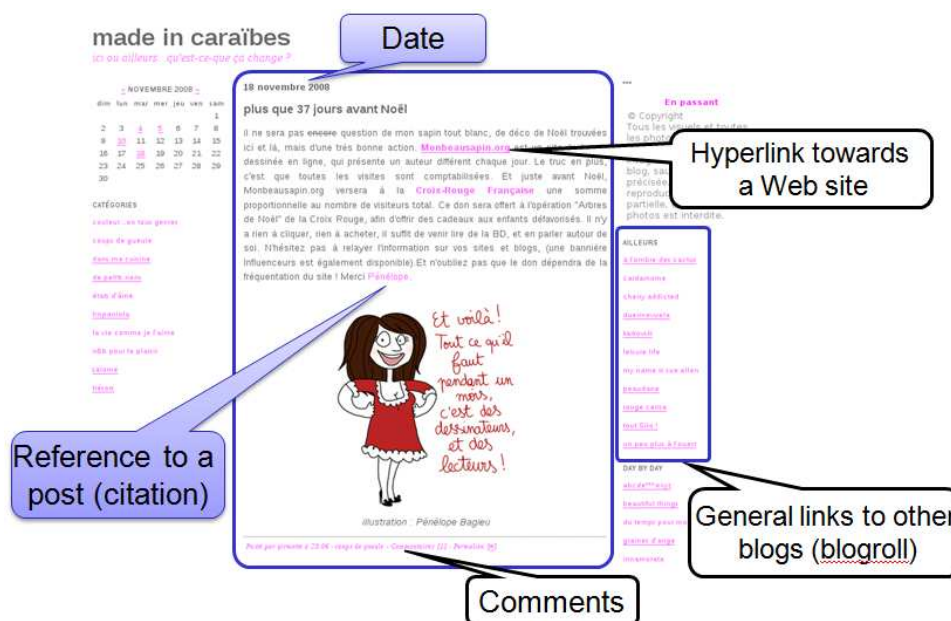


Figure 3.1: Blog sample

In terms of information spreading, we can say that P_a has 'adopted' P_b 's content or that P_b 's content has been spread towards P_a . The citation relationships between posts may be extended to the blogs they belong to: we may then say that blog A cites blog B and consider that A has adopted B 's information (or that B 's information has been spread towards A).

Citation links should not be mistaken with *comments*. If someone comments an existing post, this contribution is not a post (as it does not start a new discussion)¹.

One may be interested in identifying important nodes in the network. This problematic is essential in many real-networks and in particular for the web or blog networks. Most popular blogs have many links pointing to them and few going out. On the contrary, some blogs have many outgoing links. In the first case we call the blog an *authority* or *popular* blog; in the second one it is a *hub* (see Figure 3.2.) In reality all blogs have a double nature because best hubs are also authorities [Kleinberg1999]. Looking for authority blogs may be a way to optimize topic research. The idea is that a page cited by a good hub is more authoritative than one cited by a very bad hub.

Previous works have studied the process of post and link creation and shown that their activity was bursty; results show that the network dynamics follows a power law.

1. Comments have not been studied in my work, but this is one of my perspectives.

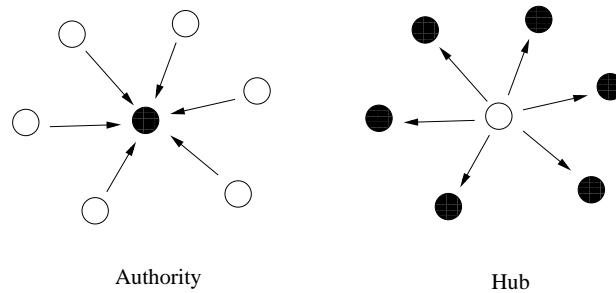


Figure 3.2: Authority and hub

This burstiness results from human behavior [WCP⁺02, VOD⁺06] and may be observed in different contexts. As said earlier, power law distribution is indeed observed in many social network properties [BA99b].

3.2.2 Data description

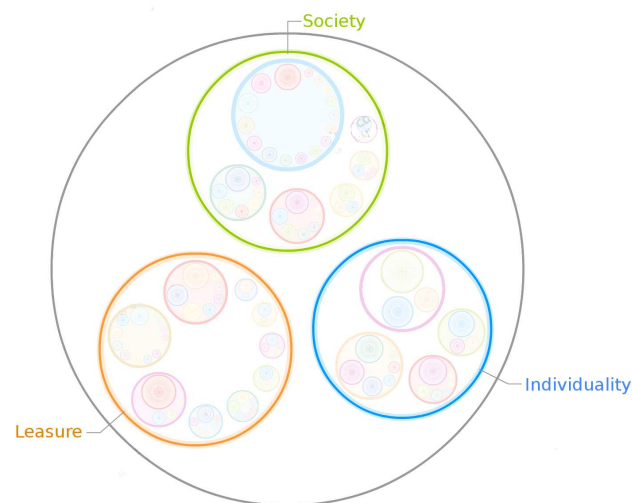


Figure 3.3: Three continents

The corpus analyzed in this section was obtained by daily crawls of 6344 blogs (1,230,692 posts) during 4 months from November 1st, 2008 to March 1st, 2009. These blogs have been chosen according to their popularity and activity in the French-speaking blogosphere. They have been selected by a company specialized in blog and opinion analysis

(<http://linkfluence.net>) as being the most active and productive blogs which provide rich information for activity and dynamics study. Blogs provide a description of their publications by RSS (Rich Site Summary); RSS represents a key feature to capture temporal features as all posts published by the 6344 blogs are recorded as they are published. It allows a fast and automatic information update.

Topical community structure

Table 3.1: Most active blogs for each continent

Continent	Blog
Society	http://www.lepoint.fr http://www.engadgetmobile.com http://www.macgeneration.com http://www.liberation.fr http://www.lefigaro.fr
leisure	http://www.sailr.com http://www.autoblog.com http://www.autobloggreen.com http://www.jeuxcherche.com http://www.gamekult.com
Individuality	http://www.beaute-test.com http://www.plurielles.fr http://www.sweetange.fr http://www.designspotter.com http://philippe-watreLOT.blogspot.com

Blogs have been classified manually into communities by blogs analysts according to their topics; we therefore call this classification *semantic* or *topical*.

The existence of this community structure is a major advantage as it allows us to study the impact of topical communities on network topology and on post dynamics (see Section 3.3). This manual classification provides three abstraction levels: Continent, Region and Territory (from the most general to the most specific). In this Chapter we focus only on the top layer (Continent layer). The whole hierarchical community structure will be addressed in the next Chapter.

The three continents are: *Society*, *Individuality* and *Leisure*. Figure 3.3 is a representation of the continents where the size of the circles corresponds to the number of blogs

within each community. It also shows the sub communities in each continent. In Table 3.1 I have cited the five most active blogs for each continent. *Society* continent include for example blog related to news (as <http://www.lepoint.fr>), politics and society topics. *Leisure* is composed, among other, of sport blogs, videogame blogs, cars blogs. *Individuality* continent regroups more personal blogs where people share their stories, experience or their centers of interest.

Blogs have been selected according to their activity. However, in order to represent the 3 continents fairly, approximatively the same number of blogs has been chosen in the 3 corresponding communities, as shown in Table 3.2².

In chapter 3 I will give a more detailed description and analysis of this manual community structure. For now, I want to show the impact of the community structure on citation behaviour and try to distinguish different pattern classes.

Table 3.2: Activity by continent

Continent	# of blogs	# of posts	# of links
Society	2245	564535	736324
Leisure	2045	469896	429748
Individuality	2054	150927	326228

3.2.3 Blog and post network

The post network is a graph where nodes are posts and edges are citations between posts (see Figure 3.4.a). The analysis at post network layer gives a macroscopic vision of interactions within the blogosphere. For example, it allows to observe the *impact* of a post in terms of number of citations which corresponds to the *impact* of one topic (for example election results or virus danger) during a specific time. The post network is a Directed Acyclic Graph (DAG) because as we mentioned previously we cannot cite a post which has not been published yet. The edges are not weighted and have a timestamp corresponding to the time of citation. All outgoing links of a given post have the same timestamp. On the other hand, incoming links may have different timestamps as they come from different posts.

The blog network is a graph representing blogs connected by inter-blog citations. Nodes of the graph are blogs and edges, corresponding to citation links, are directed and weighted

2. In this table, all filtering steps have been applied except removing links referring to resources.

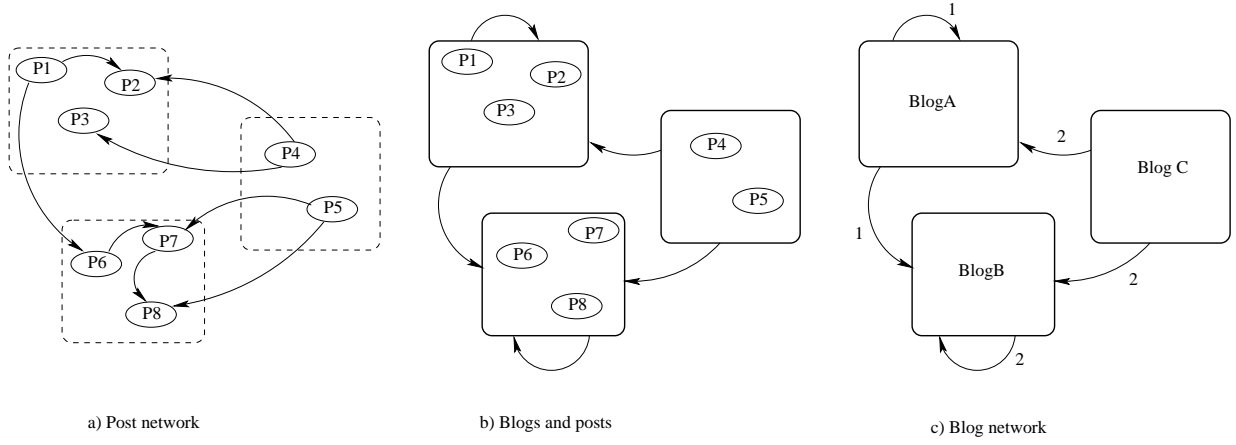


Figure 3.4: Blog and post network

(see Figure 3.4.c). In terms of structure, blog A is linked to blog B with a weight n if there are n posts from blog A which have cited posts from blog B . The blog network is obtained by regrouping the nodes at post network layer. The citation timestamp is not determinant in blog network analysis because we aggregate all posts activities of the blog.

Now I explain how the data are represented and how I build the blog network. Starting from the crawled pages for each post P (the following information was extracted for each post P which belongs to a blog B and published at time T). The first line should be read as follows: at time T , the post P from the blog B cites the post P_i from the blog B_i (created at time T_i). A post P_i cited by P is published at time T_i (4_{th} column) induces that $T_i \leq T$ as a post cannot refer to a post which has not been published yet.

T	P	B	T_1	P_1	B_1
T	P	B	T_2	P_2	B_2
		
		
T	P	B	T_k	P_k	B_m

If a post P makes multiple citations to the same post P_i , only one citation is considered as it corresponds to citations at the same time and links between posts are not weighted. From this representation, the post P has made k citations toward k posts and the blog B has made k citations toward m blogs. It is obvious that $k \geq m$ as several posts cited by P can belong to the same blog (B_1 to B_m).

Considering this set of citations, the post network is created as follows:

- Add node p
- For i between 1 and k , add the node P_i if it does not exist i.e. if this is the first time this post appears in the data.
- Add k directed edges between P and P_i for i between 1 and k .

The blog network layer is updated as follows:

- If blog B does not exist, add it as a node in the blog network.
- For all $B_i \in \{B_1, B_2, \dots, B_m\}$, add the node B_i if it does not exist.
- For all $B_i \in \{B_1, B_2, \dots, B_m\}$:
 - If no edge exists between B and B_i , add a directed and weighted edge between B and B_i ; the weight of the edge is the number of citation links between B and B_i created by P .
 - If an edge already exists, increment the weight of the edge by the number of new citations made from B to B_i .

3.2.4 Data Cleaning

A preliminary step consists in cleaning the dataset, in order to remove errors and ensure that data represent the actual blogs dynamics.

The data consist of posts and citation links. Each link connects a post to another post at a given time (indicated by a timestamp). An example of error in the initial dataset is when a post cites a more recent post (i.e. a post which has a higher timestamp); this would mean that it refers to a post which has not been created yet. This may happen if the corresponding blogs are hosted on servers in different time zones; this may also be due to a human manipulation. Such 'impossible' links are filtered as the information they provide is erroneous (at least with regard to the temporal information).

In order to study spreading cascades between different blogs, self-citation links (i.e. citations of posts within the same blog) have been removed, as they do not provide any information about diffusion towards other blogs.

Out-links can point to a post inside the dataset (if the post has been published by one of the 6344 blogs of the corpus), or outside the dataset if the post refers to a post belonging to another blog, a resource (picture, video...) or any web page. During the cleaning process those links were removed for two reasons: first, resources like pictures cannot 'cite' any post and therefore cannot contribute to the citation dynamics. Second, no temporal information is available about blogs outside the dataset, so they are unusable for our study of the blogosphere dynamics.

When building the blog graph, blogs with no in-link and with out-links only towards blogs outside the initial dataset were removed as these blogs are not connected to any blog from the initial dataset. After this filtering process, the blogs network contains 4907 nodes (blogs) and 28,258 edges (citation links among blogs).

In this section, we have described the blog corpus. In the following section, traditional blog and post analysis is performed in order to check data consistency with existing work on the subject and study the dynamic evolution of blogs and posts over time.

3.3 Activity analysis

In this section I present a first blog and post network characterisation, starting by a temporal study of the blogosphere activity. Then, I analyse blog network topology and blogs degree with a traditional approach (used in similar studies [KNRT05, AZAL04]) based on the blog network community structure.

3.3.1 Blogosphere activity

I first study the number of posts created per day in order to know the daily activity. One goal is to see if the measuring procedure collects the real blogs activity. Moreover, it is useful to detect periodic behaviors.

Figure 3.5a shows the number of posts published daily during 4 months. Instead of starting from a low value and increasing gradually, the number of posts is high from the beginning, which is due to the crawling technique: the activity of all blogs of the corpus is monitored by linkfluence³ (as opposed to a 'snow ball' approach starting from a single blog and discovering new blogs through citation links of this blog).

Another observation is that the number of posts increases suddenly after the 40th day. This is also due to the crawling method, as a new set of blogs was introduced in the corpus at that time. To avoid any bias due to measurement the first 40 days have therefore not been taken into account. We can also observe an activity decrease between days 50 and 63. This corresponds to the period of Christmas holidays, during the two last weeks of December, which indicates that bloggers are less active during holidays. Figure 3.5b displays the number of posts created for each day of the week from Saturday to Friday. A weekly periodicity may be observed on this figure, as week-end days show a lower production activity. This effect has also been observed in previous work [LMF⁺07].

3. linkfluence is a SME specialized in blog analysis, <http://fr.linkfluence.net/>.

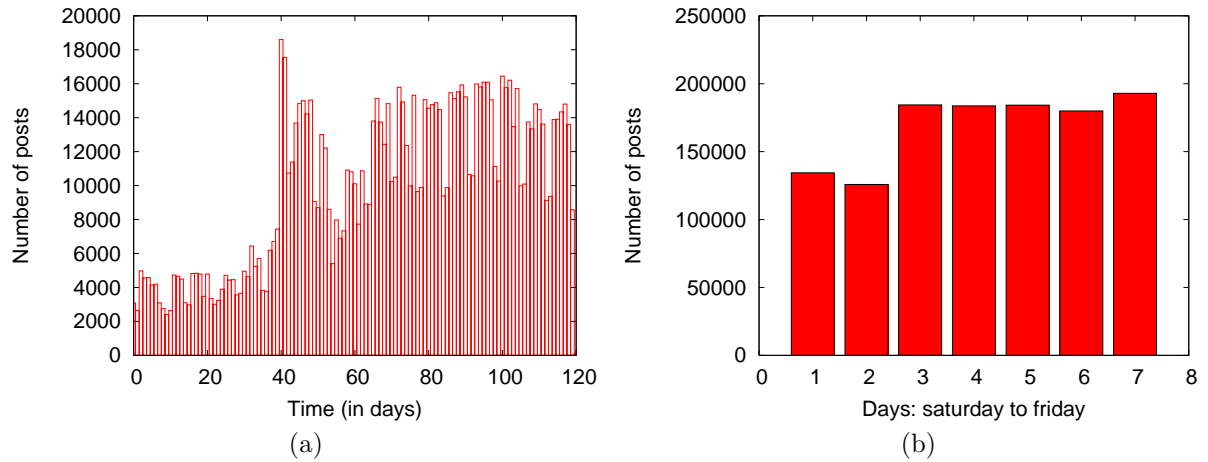


Figure 3.5: Number of posts a) per day during 4 months, b) for each day of the week

3.3.2 Blog and post networks characterization

In this section I use a classical methodology to characterize real-network topological properties. All analysis have been done on 80 days of corpus data (as I have removed the first 40 days as mentioned previously). First, I use statistical distributions, and investigate the blog degree, distinguishing between incoming (in-degree) and outgoing links (in-degree). I also use the community structure to observe its impact on degree distributions.

Figure 3.9 represents the in-degree and out-degree distribution and cumulative distribution. We may observe that in-degree and out-degree follow similar distributions (Figures 3.6a, 3.6c). They are very heterogeneous and well fitted by power laws with exponents 1.4 and 1.1. The two exponents are lower than what we may find in literature for other real-world networks (between 2 and 3). This statical distribution can be difficult to interpret as dots may superpose. Therefore I complement it by a cumulative distribution where it is easy to identify major values as well as extreme behaviors.

If we observe the cumulative distribution of in-degree in Figures 3.6b we see that almost 90% of blogs have a in-degree and out-degree smaller than 100. However, some blogs have a degree greater than 30000. The same distribution is observed for out-degree (Figure 3.6d).

To go further I have plotted the in and out-degree with regards to the blog *continent*. The aim is to observe the impact of topical communities on blog degree. To be able to compare the degree distributions I have plotted cumulative density distributions (also called PDF probability density functions). A dot with $x = 10$ and $y = 0.5$ means that 50%

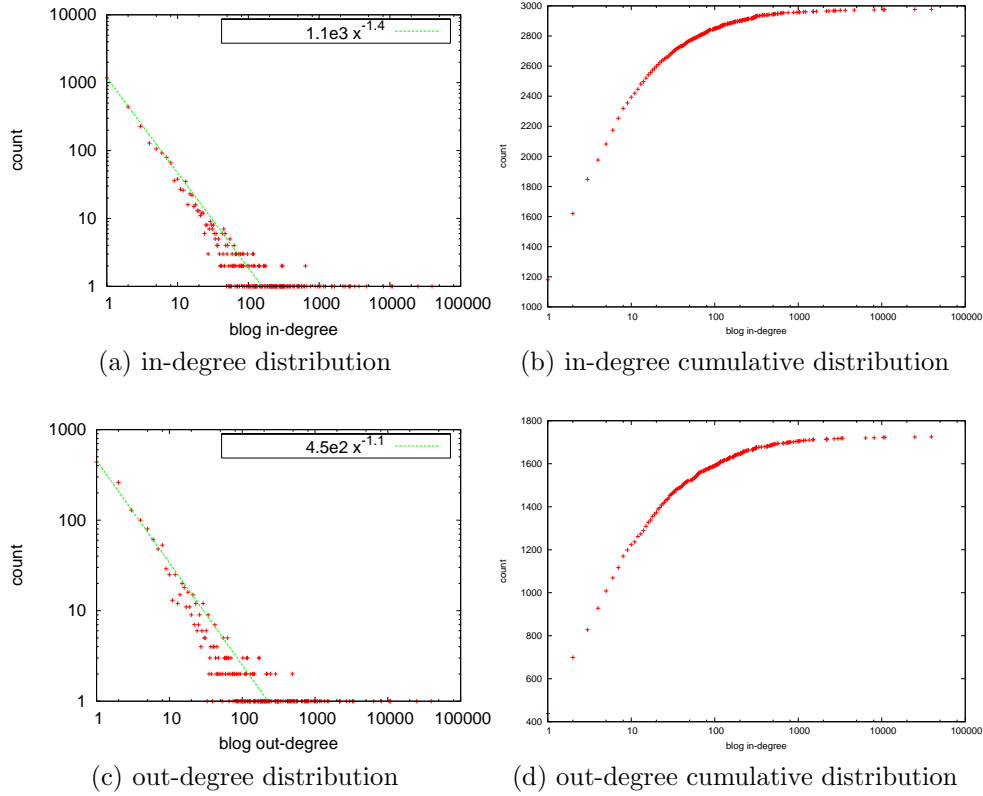


Figure 3.6: a) c) In-degree and out-degree distributions and b) d) cumulative distribution in the blogs network.

of blogs a degree less or equal to 10.

Figure 3.7a shows in-degree cumulative density distribution. We observe that the distribution of blogs from *Society* continent is lower than for *Individuality* and *Leisure* continents while *Individuality* and *Leisure* continents have a close distribution. This indicates that in-degree in *Society* continent blogs is higher than those of blogs from *Individuality* and *Leisure* continents. The distinction between *Leisure* and *Individuality* distributions is more visible for out-degree distribution. *Individuality* blogs tend to have a smaller out-degree than *Leisure* blogs.

One other important characterization with regard to blogs degree is the correlation between in and out-degrees. Figure 3.8 shows that these degrees are correlated for high degree values (approximately when the degree is higher than 180). The interpretation is that if the blog is active enough than it will get an attention (i.e. a number of in-links) proportional to his activity. However, as the number of dots with low degree values is high

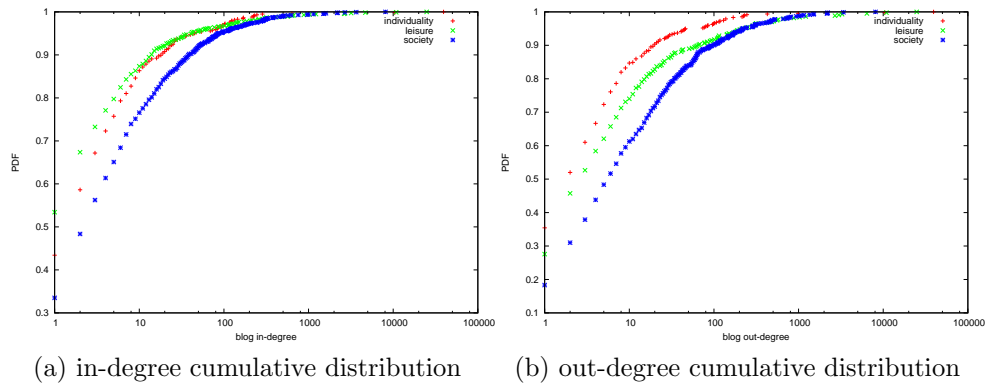


Figure 3.7: in-degree and out-degree cumulative density distribution per continent in the blog network

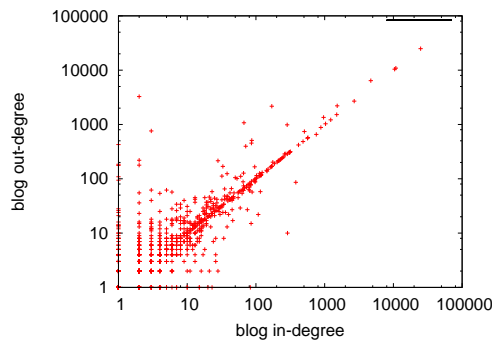


Figure 3.8: log in-out correlation

the correlation coefficient between in and out degrees is only equal to 0.36. The small value of this coefficient contradicts the intuition that the popularity of a blog (indicated by its in-links) is related to its activity (its out-links). However, this result is in line with outcomes of previous studies [LMF⁺07] which observed an even smaller correlation coefficient.

In addition of blog degree, we are interested in the number of links (citations) between two blogs. As mentioned previously, I do not take in count auto-citation links (links from a post to another post from the same blog) because I want to characterize how two distinct blogs may cooperate with each other. Figure 3.9a shows the cumulative distribution of inter-blog links number. In addition of being a heterogeneous distribution we also observe on that plot that 55% (4700/8500 blogs) of blog to blog interactions correspond to only

one link. The majority (88%) of intra-blog link numbers do not exceed 10 however we may find some blogs which interact more than 20000 times. In complementarity of the number of intra-blog links I have been interested in the number of days during which two blogs have interacted with each other at least once. We observe that 64% of blog pairs interact only one day and 94% less than 20 days. Those results also illustrate that most blogs have only very few interactions with other blogs whereas some blogs have a high activity and during a long time.

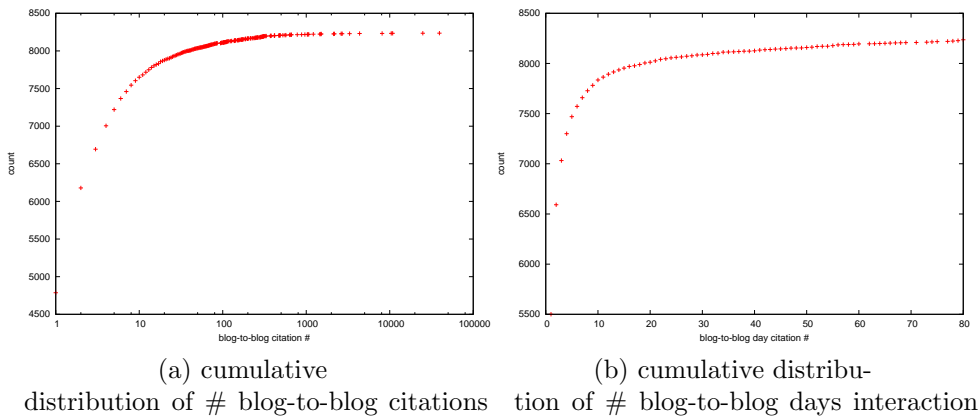


Figure 3.9: Inter-blog activity

The activity of a blog can be measured through the number of posts it sends (publishing activity) or through the number of citations by others blogs (in-linking activity which reflects the influence of the blog). I have calculated the distribution of the number of citations per blog. In the blog graph this represents the sum of weights of out-going edges. It is close to a power-law distribution with exponent 1.2. The total number of citations is 666,191 which represents an average of 135 citations per blog during 4 months. The number of posts per blog is also close to a power law with exponent 1.25. The distribution of edges weights in the blog network, which corresponds to the number of blog-to-blog links also follows a power-law distribution with exponent 2.27. The post network has similar characteristics. I have found that distributions of posts in- and out-degrees are close to power laws with exponents 2.6 and 3.1 respectively. All these results are consistent with the state of the art [LMF⁺07].

3.4 Post popularity and community structure

In this section I explore new approaches and methods to characterize post and citation dynamics in different blog communities. The post popularity measures the impact of posts over time. More precisely I study the evolution of the number of incoming links starting from the day of publication. I also pay attention to biases that occur when studying a property over time (in this case post popularity). I address how such a bias is introduced and how it may be removed for better results. Here I present a methodology which goes beyond traditional approaches by defining classes of dynamic behaviors based on topological features of the post network, and by investigating the impact of topical communities on post popularity dynamics.

3.4.1 Dynamics of post and citation arrivals

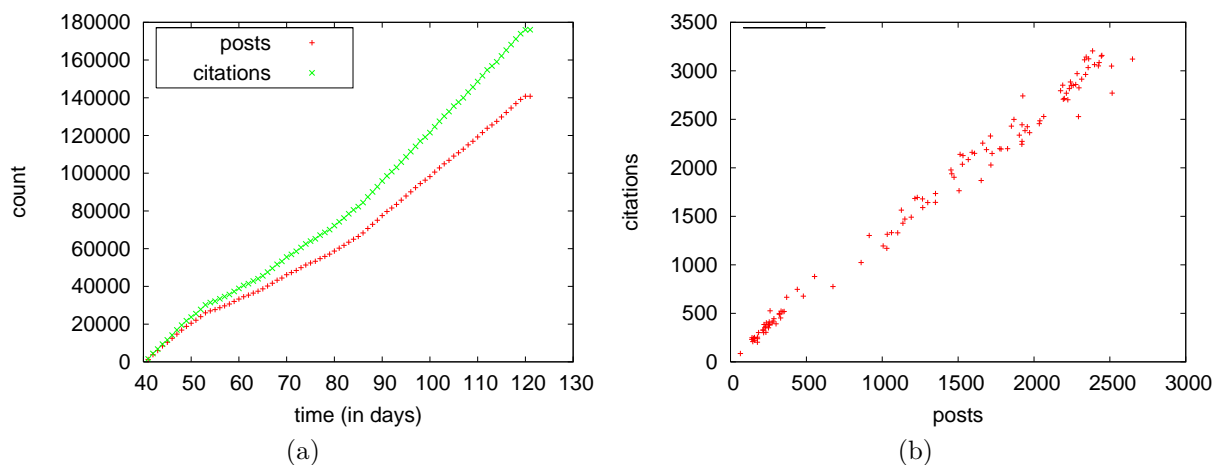


Figure 3.10: a) Evolution of the number of posts and citations. b) Citation and post correlation.

I investigated how the overall number of citations and posts evolved over time. Figure 3.10a presents this evolution during the 80 days of measurement (the first 40 days have not been taken into account as explained in Section 3.3). Figure 3.10a shows that the numbers of posts and citations grow in similar ways. Figure 3.10b presents the correlation between the number of citations and posts per day. The daily ratio between total number of citations and post remains approximatively constant during the 80 days, with an average value of 1.28. This high correlation of post and link numbers per day means that the ratio

between the total number of citations per post is constant even when blogosphere topology changes.

3.4.2 Biases

Trying to study the dynamics of a complex network is usually biased since only a partial view of the studied object may be observed [BM].

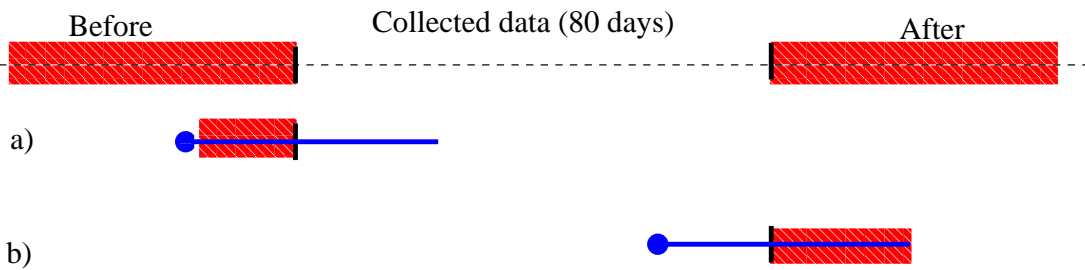


Figure 3.11: Biases causes

A bias is induced by missing information, typically all links that appear before or after the capture. This type of bias is observed when we study a property over time and the nodes arrive in the system in a desynchronized way. Such biases have been studied in Peer-to-Peer networks and other contexts [SR06, BM]. In this case I consider citation post arrivals. Each post has a date of publication, afterwards, it may be cited by more recent posts and possibly after the end of the collected data. If a post is published before the start of the trace (see Figure 3.11.b) the studied property will be underestimated. Border effect also appears when links arrive after the end of the measurement (see Figure 3.11.a). This creates a bias toward posts appearing at the end the capture as they cannot be cited during as many days as earlier posts. The result is that we overestimate the observed property at the end of the trace.

The solution consists in observing the popularity of all posts during the same period of time (see Figure 3.12). I have divided the data trace into equal period of 40 days. Each post created during the first 40 days (effective period in Figure 3.12) of the capture has been monitored during 40 days after its publication. All other data are not taken into account. This correction method was applied for all the rest of chapter studies and results.

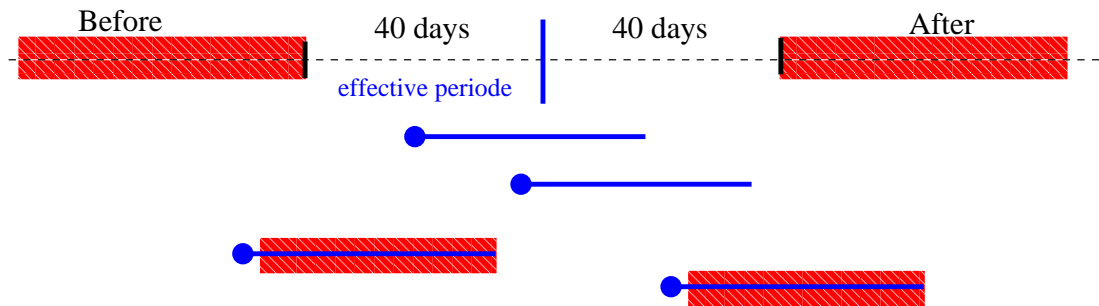


Figure 3.12: Biases correction

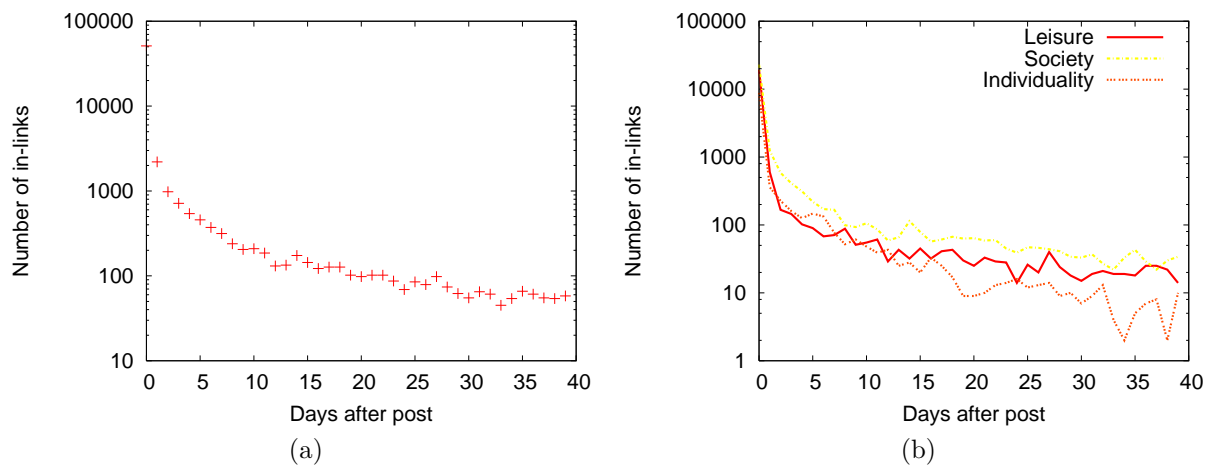


Figure 3.13: Evolution of post popularity a) overall evolution b) according to post community.

3.4.3 Post popularity

Now I investigate the evolution of post popularity evolution after publication day, in terms of number of links each post gathers from its first day of appearance (i.e. its posting day) until 40 days later, according to the methodology to correct bias described in Section 3.3. In Figure 3.13a we observe that most citations are made within the first 24 hours which is in accordance with previous studies [AA05]. Results differ from [LMF⁺07] where popularity decreases significantly at the end of the measurement, but this is due to the bias I mentioned earlier. Here, popularity is divided by half after the first 24 hours but later on it decreases at a much slower pace.

Figure 3.13b represents the evolution of posts popularity according to their community

(Leisure, Society, Individuality). We observe that popularity of posts from Individuality community decreases more than the popularity of posts from Leisure and Society after 13 days after post creation. Moreover, although the number of citations of posts from Society community is higher than the number of citations of posts from Society, both plots tend to converge after 27 days after post creation. As a conclusion, we may say that post popularity evolves quite similarly in the 3 studied communities.

3.4.4 Two types of citation dynamics

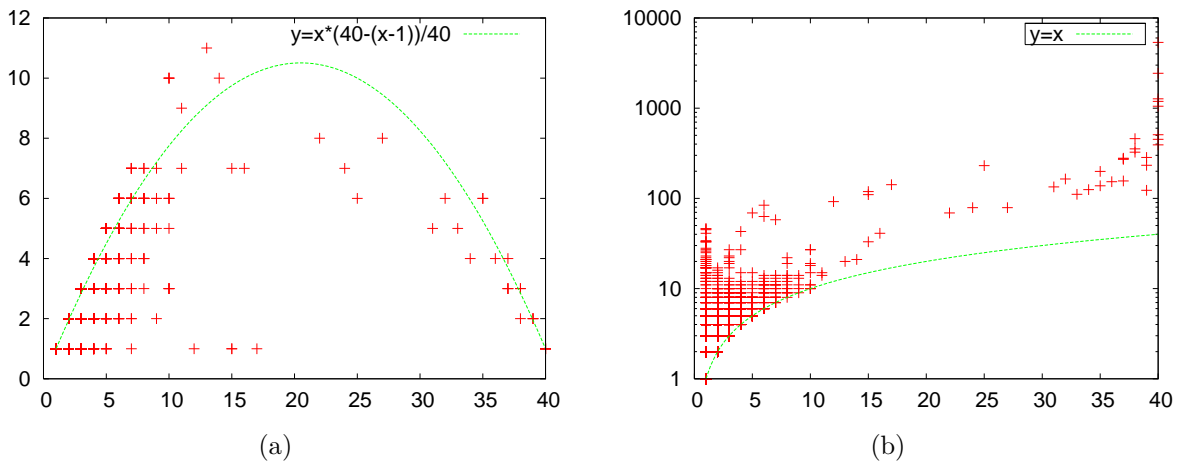


Figure 3.14: Post classification - for both plots, the value on the x-axis (denoted by X) represents the number of days during which the posts have been cited. a) The value on the y-axis (denoted Y) corresponds to the number of distinct citation periods (e.g. for a post cited on days 1, 2, 3, 5, 7, 8, $X=6$ as it is cited on 6 days, and $Y=3$ as these citations occurred during 3 periods of time: $\{1, 2, 3\}$, $\{5\}$ and $\{7, 8\}$). b) The value on the y-axis represents the sum of all citations of a post for the corresponding number of days. So for example a dot with coordinates $(3; 34)$ represents a post which has been cited on 3 distinct days and 34 times in total.

An interesting approach to study posting behavior is to classify posts according to incoming citations. Thanks to the bias correction method, all incoming citations for each post are known during 40 days after its creation. Figures 3.14a and 3.14b explore this information.

On Figure 3.14a, an additional curve is displayed, corresponding to a stochastic behavior. It corresponds to the probability to be cited the day x and not the day before

$(x - 1)$, multiplied by the number of days x if the x days are chosen at random among the 40 observation days. We obtain the plot $y = x * (40 - (x - 1))/40$. We can observe that the dots are very close to this stochastic behaviour. Apart from posts which are cited only a few days (i.e. for which X is low), the citation of posts is neither oscillating (as Y values are not higher than those of the stochastic curve) nor continuous (as there are very few dots with low Y values when X is high).

One limit of Figure 3.14a is that the number of citations of posts on each day is not visible; Figure 3.14b completes the view; We observe that most posts are cited only few days. However a group of posts can be identified which are cited more than 30 days; they correspond to a specific class of behavior. Moreover, posts which are cited on many days (i.e. with high X values) are cited more than once a day, as Y values tend to be higher than the minimum bound corresponding to the curve $y=X$. We may therefore conclude that the higher the number of citation days, the higher the frequency of citations per day.

Figure 3.14b indicates a statistically significant classification of posts into two groups:

- posts which are cited on at most 10 distinct days; they represent 99% of all posts.

The number of posts cited only one day represents 89% of posts.

- posts cited on more than 10 days.

Going further, we study in the next section how posts from the three continents (Leisure, Society and Individuality) are distributed among these two types of popularity.

3.4.5 Post popularity and impact of community structure

Let us first focus on posts which are cited only one day. Figure 3.15a shows the distribution of the total number of in-links for all posts cited only one day and for each community at continent level (*Society*, *Individuality* and *leisure*). The reasons for this restriction to posts cited only one day are the following: first, the number of these posts represents 89% of the total number of posts in the dataset; second, taking into account all posts of the first class would be confusing e.g. a post cited 10 times on a single day would appear like a post cited once on 10 distinct days. The distributions of the posts of each class are heterogeneous. However, the behaviors within each community differ: the *Individuality* continent has the lowest total number of citations (corresponding to the surface below its curve) which indicates that personal blogs have a smaller popularity than posts from other blogs. The popularity of posts from the *Leisure* community is much higher but these posts are not cited many times as the maximum x value is 9 (i.e. posts cited 9 times on the only day they have been cited). On the other hand we observe that all posts with $x > 10$ belong to the *Society* continent. Those posts have a high popularity and the corresponding topics

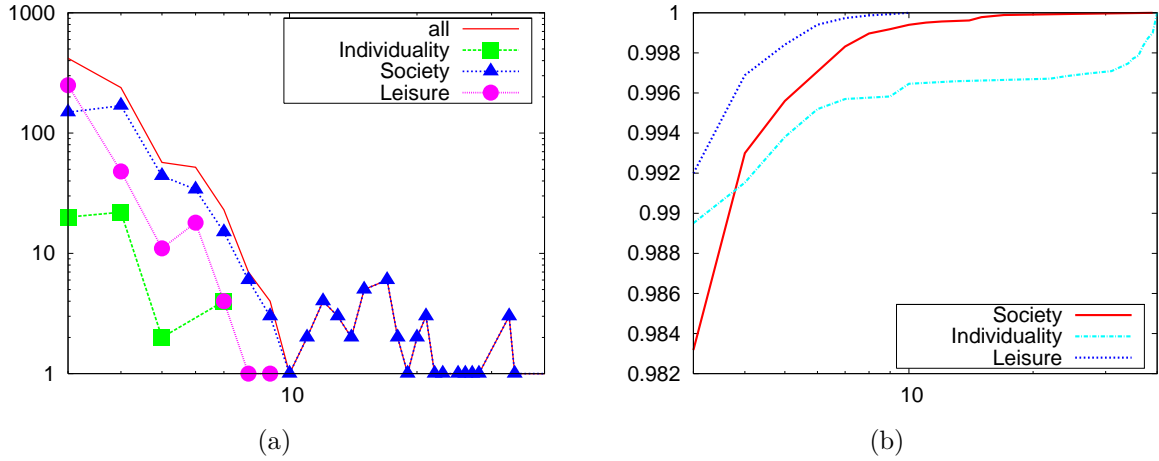


Figure 3.15: a) Distribution of citations of posts cited only one day. A dot with coordinates (x, y) means that there are y posts which have been cited x times on their single citation day (as $X=1$). b) Cumulative number of citations per total number of citation days. This plot represents the cumulative probability density function (CDF) of the number of citations per number of days for X greater than 3.

can be considered as *spikers* because they are cited a lot only one day. After studying these posts in more detail, we found out that they belonged mainly to blogs related to news sites, thus confirming their interpretation as spikers.

We have focused so far on posts cited only one day. We now study the citation dynamics of posts cited on at least 3 different days, illustrated on Figure 4.8a. The curve representing the popularity of the Individuality community increases more slowly than the 2 other curves for low X values and increases significantly for the maximum values of X . This rapid increase at the end of the measurement is a border effect due to the fact that each post is monitored during only 40 days. Therefore the maximum value (1) has to be reached at the end of the plot, and it is not possible to know whether posts cited 40 days have actually been cited 40 days, or 41 days (or more). An isolated group of posts cited on more than 30 days may be identified in this community; they belong to blogs with an important forum-like activity. These blogs may be considered as 'different' from traditional blogs and this representation is therefore interesting to detect outliers. On the other hand, the two other communities (Leisure and Society) show a different behavior in terms of popularity evolution. In particular, no specific increase can be observed at the end of the measurement on these curves as the maximum value is reached around $X = 10$ and $X = 20$ for Leisure and Society communities respectively. The conclusion is that posts of these communities

are not cited on more than 10 (resp. 20) days after their publication and that these posts therefore do not need to be monitored more than 10 days (resp. 20 days).

3.5 Conclusion

I have proposed a new approach and methods to characterize post and citation dynamics in different blog communities. In particular, the evolution of post popularity over time has been studied. I have gone beyond traditional approaches by defining classes of post popularity evolution from topological features of the post network, and by investigating the impact of topical communities on citation dynamics. I have proposed a new representation of posts' popularity using daily incoming citations, in order to classify posts. Moreover, I showed that distinct patterns related to both the duration and the frequency of citations could be observed in the various communities. By comparing only posts with a small citation duration (for example only 1 day) I have been able to isolate posts with a high popularity (spikes as the citation duration is very small) and have shown that this type of pattern is present in one continent and not the others.

In Section 3.3 I have presented a methodology to correct biases towards most recent posts and blogs in the dataset. I have observed that post popularity decreased slowly but remained significant even after 40 days. An interestingly high correlation was found between the total numbers of posts and links per day, which means that the ratio between total number of citations and posts remain constant even when the blogosphere topology evolve over time. In particular, the topology of the post network has allowed us to identify two main classes of post popularity. A deeper study of each class has shown that blogs related to spiker topics could be detected easily as well as blogs with forum-like activity. Finally, the impact of topical communities on the evolution of post popularity over time and on information diffusion cascades has been investigated. In particular, it showed that the measurement's duration may be reduced for specific blog communities. The study of the impact of community information on post and citation dynamics showed that dynamic patterns were related to social behaviours.

One conclusion of this first work is that the study of complex system properties (in this case the characterisation of post citation activity) through the community structure point of view is a promising approach. This approach goes beyond traditional methods and provides a complementary interpretation to traditional methods of network dynamics. This community-based approach is developed in the following chapter.

Citation link study: community oriented approach

Contents

4.1	Introduction	50
4.2	Framework	51
4.2.1	Hierarchical community structure	51
4.2.2	Homophily	52
4.2.3	Community distance	54
4.3	Case study: topical community structure	55
4.3.1	Dataset and community structure description	56
4.3.2	Citation links homophily	56
4.3.3	Citation links community distance	60
4.4	Case study: Automatic community structure	66
4.4.1	Louvain community detection algorithm and community cores	67
4.4.2	Community structure	68
4.4.3	Results	68
4.5	Summary and perspectives	72

4.1 Introduction

Understanding interaction patterns in real-world networks is an important topic with both fundamental and practical implications [LMF⁺07]. 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* [NBW06] tend to interact preferentially with each other, but limited knowledge is available on this topic for real-world data [CR09].

In Chapter 3, I have shown the impact of topical communities on citation behaviour [SBLGL10]. I have focused mainly on post popularity and have observed that a community approach can be efficient to classify and detect specific patterns with regard to this property.

In this chapter I go further and propose a generic methodology in Section 4.2 to study interaction links in complex networks with regard to their community structure. To do so, I define 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. I also study variations between incoming and outgoing links.

My approach is original as it studies citation patterns with regards to a predefined hierarchical community structure. Community detection is often an automatized task, using algorithms which rely on the structural properties of the network (for a review see [For09]). In addition to topical communities, I have implemented my methodology with synthetic community structures using Louvain algorithm [BGLL08] which proposes an efficient algorithm to compute a hierarchical community structure in very large graphs.

I apply this methodology to the same blog network as before [SBLGL10] in Section 3.2. This approach allows to study interactions with regard to the community structure and conversely to characterize communities according to link homophily and distances. The citation behavior of the studied blogs is analyzed with regard to two different community structure:

- in Section 4.3, the topical community structure provided by blog analysts is considered;
- in Section 4.4, an automatic community detection algorithm is used on the dataset to identify a partition of blogs according to citation links in the blog network, thus showing that this methodology may be applied in a general case even if no community

structure is given.

The work presented in this chapter was published in [BGTL11].

4.2 Framework

The methodology I propose consists in studying interaction links in a network with regard to its community structure. The construction of this structure can be obtained in different ways; I have applied the proposed methodology to the same blog network studied in Chapter 2 because it presents a high interest and has many advantages. However, it is important to note that these methodology and metrics can be applied to any other interaction network. Therefore, I explain in Section 4.2.1 how a hierarchical community structure may be obtained for any complex network.

This section is structured as follows: I first introduce definitions related to the hierarchical community structure. I then define in Section 4.2.2 two metrics to evaluate whether interaction links relate nodes from a same community (at all levels of the hierarchical structure). I explain how these metrics are complementary to *modularity* which is used traditionally to evaluate partitions quality [NBW06]. In Section 4.2.3 I introduce the notion of *community distance* to evaluate whether interaction links between nodes relate “close” or “distant” communities.

4.2.1 Hierarchical community structure

Let a graph $G = (V, E)$, with V a set of nodes and E a set of edges. The methodology I propose requires a community structure such that each node of V belongs to exactly one community at each level of the tree¹. 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. Case of studies will be given in Section 4.3 and 4.4 with semantic and topological community structures respectively.

Definition 1 *Hierarchical Community Structure*

Given a community partition $P = \{C_1, C_2, \dots, C_l\}$ of V , a sub-partition $P' = \{C'_1, C'_2, \dots, C'_m\}$ 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 $P_k \sqsubseteq P_{k-1} \dots \sqsubseteq P_2 \sqsubseteq P_1 \sqsubseteq P_0$ with $P_0 = V$, i.e. P_0 contains only one community which is

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

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.

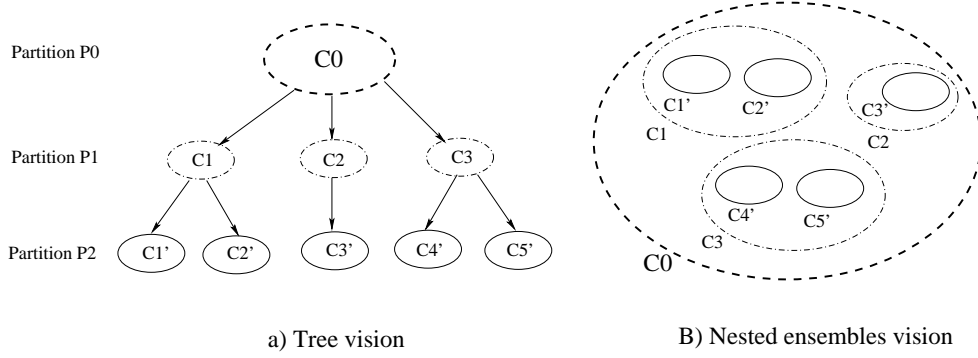


Figure 4.1: Hierarchical community structure example

In figure 4.1, I give two example representations of a same community structure, a tree and nested partition vision. In this example, the community structure has 3 partitions P_0 : the whole network, P_1 : 3 communities, P_2 : fives communities .

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. $v \in C$.

4.2.2 Homophily

The approach I present requires an interaction network and a hierarchical community structure (or a community tree), formally defined in Section 4.2.1. The first step of my methodology consists 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 j^{th} level of the community tree, with $j > i$.

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

Note that, in this definition, the value of $\Delta_j(C)$ may be biased by the number of links in communities at the 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 $\Delta_j(C) \div \psi_j(C)$, 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 j^{th} 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 j^{th} level of the hierarchy, independently of the number of edges in these communities. The *modularity* function [New03, CNM04] 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. However, the metrics Δ and ψ I propose do not have the same goal: they measure the proportion of internal links with regards to a random distribution.

For example, $\Delta_1(G)$ (resp. $\Delta_2(G)$, $\Delta_3(G)$) measures the fraction of citation links in G between two nodes from the same *continent* (resp. *region*, *territory*) in the whole graph.

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

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

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 4.3.2.

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

4.2.3 Community distance

In the previous section I have defined the probability for a link to relate blogs from the same community. Now I want to characterise the interaction behaviour more precisely, in particular for links which do relate nodes from a same community: do these nodes belong to *close* or *distant* communities? What is the proportion of close versus distant links? In order to answer these questions, I introduce the notion of *community distance* $d(u, v)$ between two nodes u and v .

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 k^{th} level to characterise interaction links between nodes.

Note: If the community tree is a balanced tree then the definition of community distance can be simplified as the distance to the closest common ancestor in the community tree. Given a couple of communities $u, v \in P_h$, there exists a minimal integer t such that there is a community C in P_t with $u \subset C$ and $v \subset C$.

$$d(u, v) = h - t$$

Figure 4.2 shows an example of two community distances measurements. In 4.2.a we measure the community distances links made between u, v in red ($u \in C1'$ and $v \in C2'$) and u', v' in blue ($u' \in C3'$ and $v' \in C5'$). In this case, $d(u, v) = 1$ and $d(u', v') = 2$ which indicates that the link between u' and v' connects two communities more distant than the

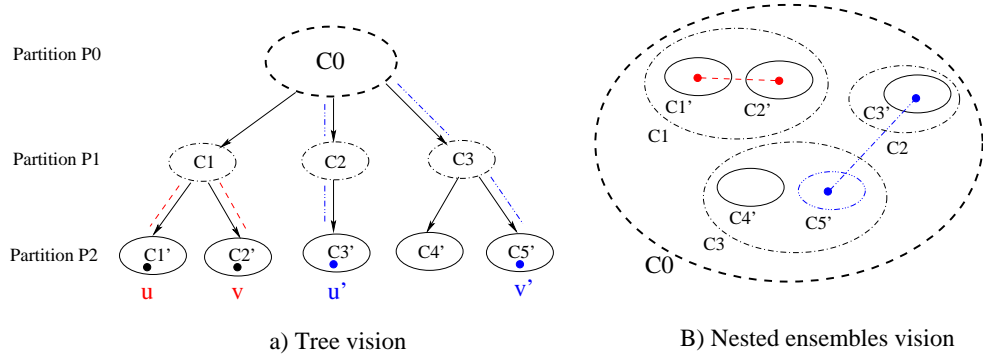


Figure 4.2: Community distance example

link between u and v . In figure 4.2.b) we can observe in a simple way the two links and how they connect distant communities.

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

I 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 4.3.3).

4.3 Case study: topical community structure

In this section, I use the formalism introduced in Section 4.2 to analyze the blog dataset introduced in Section 3.2.

4.3.1 Dataset and community structure description

The dataset was obtained by daily crawls of 6007 blogs in the French-speaking blogosphere (1,074,315 posts) during 4 months from November 1st, 2008 to March 1st, 2009. The blog network dataset we used for the experimentation is the same dataset described in Chapter 3. In the previous Chapter I have focused on post popularity and therefore have worked at post network layer. In this study, I focus on blog interactions. 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 A to B .

To get a hierarchical community structure of the blog network two possibilities are available. First, perform a community detection algorithm and get an *automatic* classification. Second, classify *manually* each blog into hierarchical classes based on topical blogs' knowledge. In this study I use a manual community classification. Such a classification is generally hard to obtain due to the large size of datasets, but is very interesting as it is validated manually, unlike in *automatic* classification.

In this case classification into *communities* has been built manually by professional blog analysts according to blogs topics (<http://linkfluence.net>). This topical classification is organized in three hierarchical levels: *continent*, *region* and *territory* (from the most general to the most specific, see Figure 4.3). For instance, the blog <http://www.sailr.com> belongs to the *leisure continent*, the *sport region* and the *sailing territory*. The hierarchical community structure I 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 4.2, I therefore consider the directed graph $G = (V, E)$ where V is the set of blogs and E is the set of citation links.

4.3.2 Citation links homophily

Table 4.1: Probability to link blogs from the same community at various levels.

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

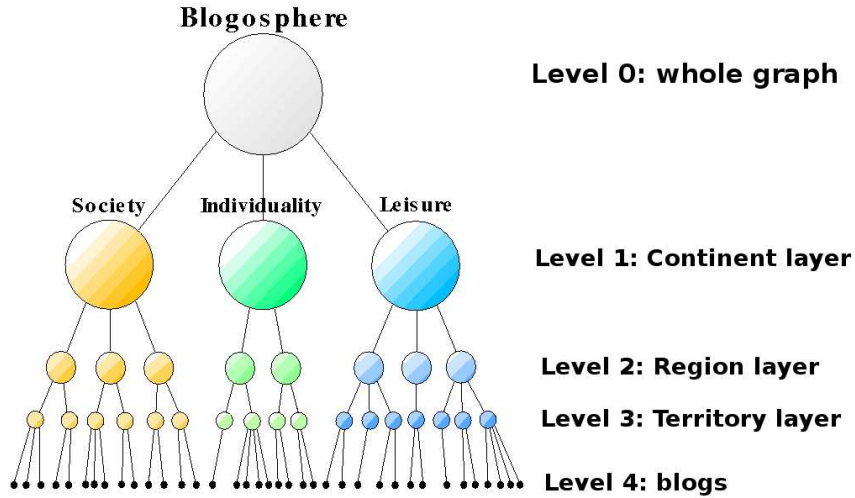


Figure 4.3: Blog network community structure

First, I 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 homophily².

The results presented in Table 4.1 show that $\Delta_1(G) = 89\%$ which means that 89% 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 $\Delta_1(C) \div \psi_1(C)$, 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 $\Delta_3(C) \div \psi_3(C)$ indicates that links at territory level are significantly more homophile than expected in a random case.

Table 4.2: Probability to link blogs from the same region in each continent and associated modularity

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

After considering the whole graph, I now focus on links within each continent at the re-

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

gion layer, i.e. the tendency of blogs from the same continent to cite blogs within the same region (Table 4.2). We may see that homophily values are very high. In particular, *Individuality* and *leisure* continents have Δ_2 values greater than 98%. However, $\psi_2(\textit{Individuality})$ is also very high (97%) which means that the high value of $\Delta_2(\textit{Individuality})$ is more expectable than the value of $\Delta_2(\textit{Leisure})$.

Table 4.3: Probability to link blogs from the same *Territory* for each *Region*

Region	# of link	$\Delta_3(Ci)$	$\psi_3(Ci)$	$\Delta_3(Ci) \div \psi_3(Ci)$	$Q(Ci)$
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 random cases. However, the high value of $\Delta_2(\textit{society}) \div \psi_2(\textit{society})$ 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.

I now study citation links homophily within selected regions at the territory level, i.e. the tendency for blogs from a same region to cite blogs from the same territory (Table 4.3). Let us notice the low values of modularity: this indeed not suprising as the classification into continents, regions and territories is based on blog topics and not on their interaction links.

We may also observe in Table 4.3 a very low homophily probability with regards to random values, for example, $\Delta_3(\textit{automobile}) \div \psi_3(\textit{automobile}) = 0.041$. 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.

Conversely, the homophily probability $\Delta_3(\textit{sports})$ is high = 0.95 (with also a high $\Delta_3 \div \psi_3$ value): in this case, the division into territories is consistent with the citation behaviour.

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.

Figure 4.4a shows that $\Delta_2(C) \div \psi_2(C)$ and modularity $Q_2(C)$ are correlated for almost all regions as most of them are close to the diagonal: both homophily and modularity consider the density of links within a community; $\Delta_j(C) \div \psi_j(C)$ could therefore be considered as a kind of unbiased modularity (with regards to the number of links in each community). More precisely, I illustrate the differences between both functions in Figure 4.4b which plots the values of $\Delta_2(C)$, $\psi_2(C)$ and $Q_2(C)$ for all regions, sorted by increasing modularity value $Q_2(C)$. We may observe that regions with very close modularity values may have very different $\Delta_j(C)$ and $\psi_j(C)$ or $\Delta_j(C) \div \psi_j(C)$ values (eg. regions 5, 6, 9 corresponding respectively to *agora*, *cooking* and *notebook*). This confirms that in order to study homophily we have to consider at the same time the value of $\Delta_j(C)$ and $\Delta_j(C) \div \psi_j(C)$ and not only one of them.

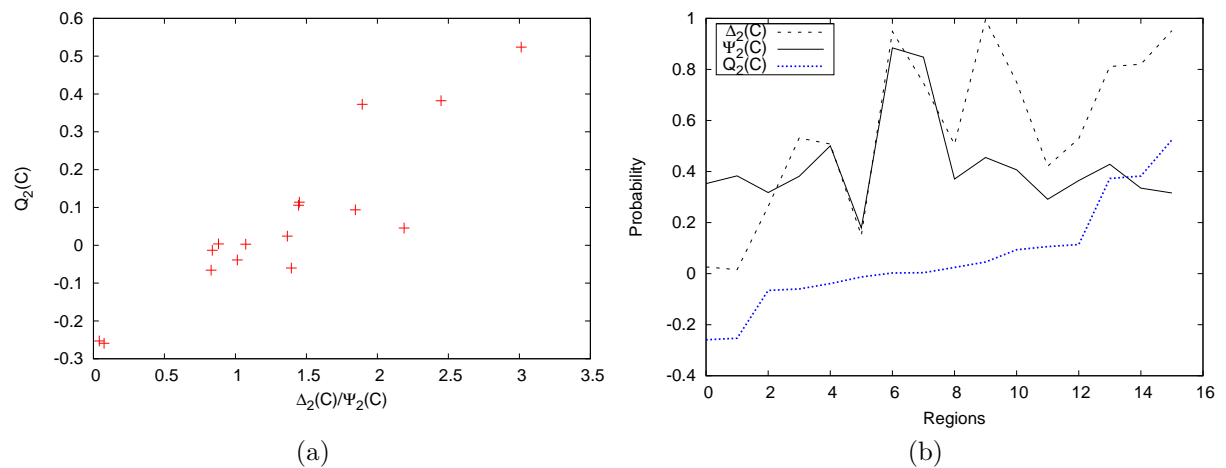


Figure 4.4: Delta vs modularity at region layer

4.3.3 Citation links community distance

Table 4.4: Distribution of community distances in G

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

Let us now study the citation behaviour more precisely by considering the distribution of community distances for all links of G (given in Table 4.6). 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*.

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.

4.3.3.1 Community profiling based on links distance

In the previous section, we observed that homophily probability was very low for some communities and that most citations were between blogs from the same *region*. In this section, we go further by first studying for each *region* its number (resp. fraction) of links at each distances 1, 2, 3 and 4, see Figure 4.5a (resp. 4.5b).

We would like to compare the citation behaviour independently from the size of the communities, but as 78% of links come from the agora region, we focus on *fractions* of links numbers (Figure 4.5b) which are easier to read. Different patterns appear but there is generally one (or two) main distances. Communities with very similar profiles also appear. For example, *automobile* and *video-games* or *sport* and *cooking*. More precisely, *sport* and *cooking* have a majority of their links at distance 1 and most others at distance 4.

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

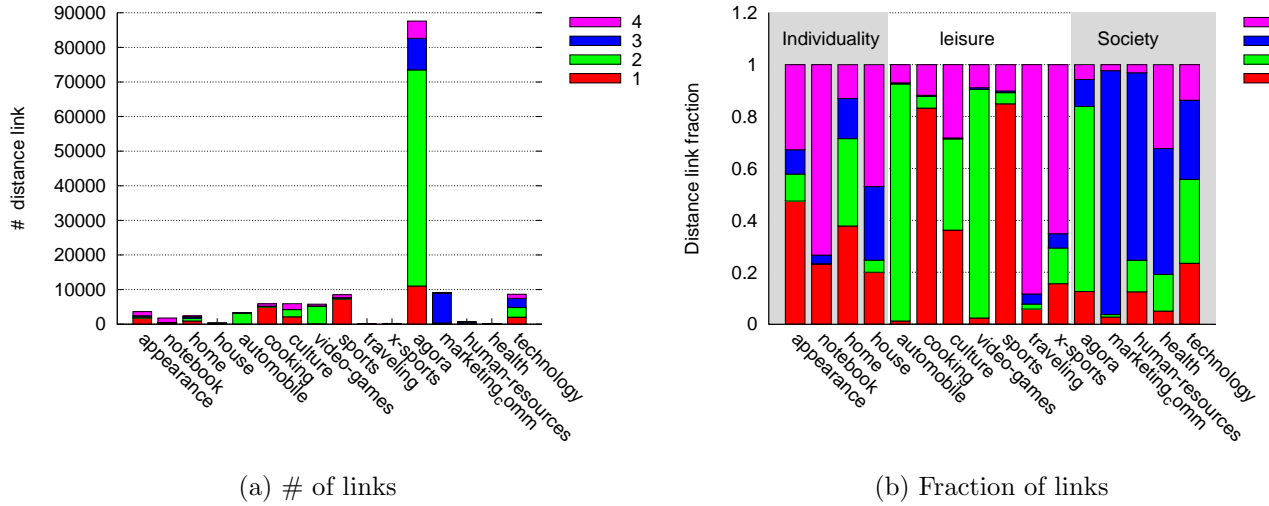


Figure 4.5: Number of link distances by region

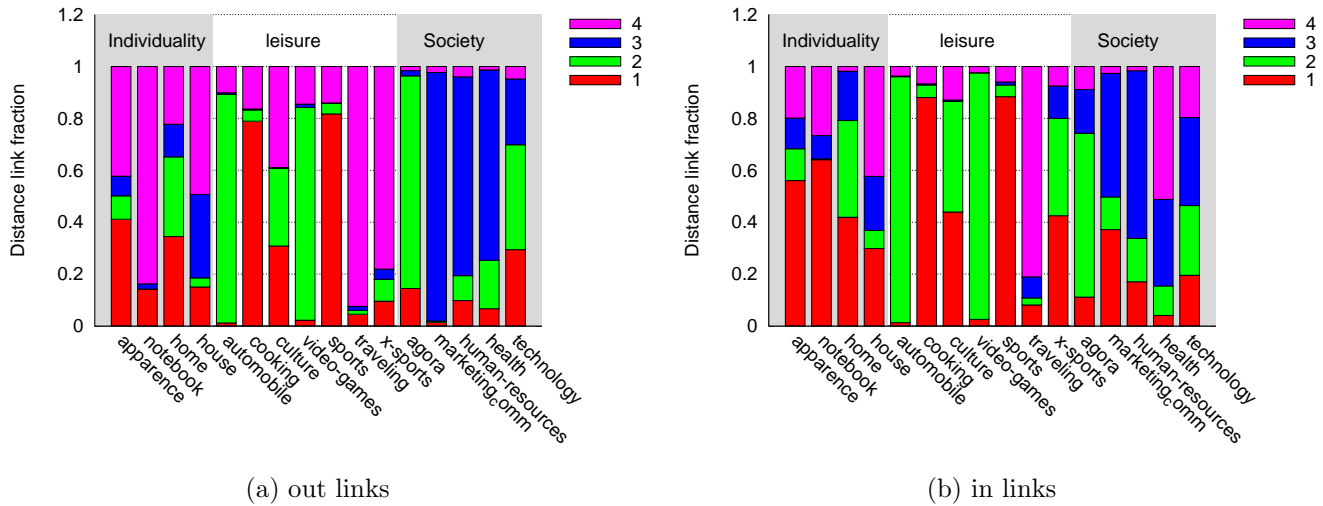


Figure 4.6: Fraction of in and out link distances by region

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 4.6 we characterise blogs from each region according to their fraction of out links $out_{\kappa}(C)$ (Figure 4.6a) at each distance (resp. in links $in_{\kappa}(C)$ in Figure 4.6b): blogs from *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*³ 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.

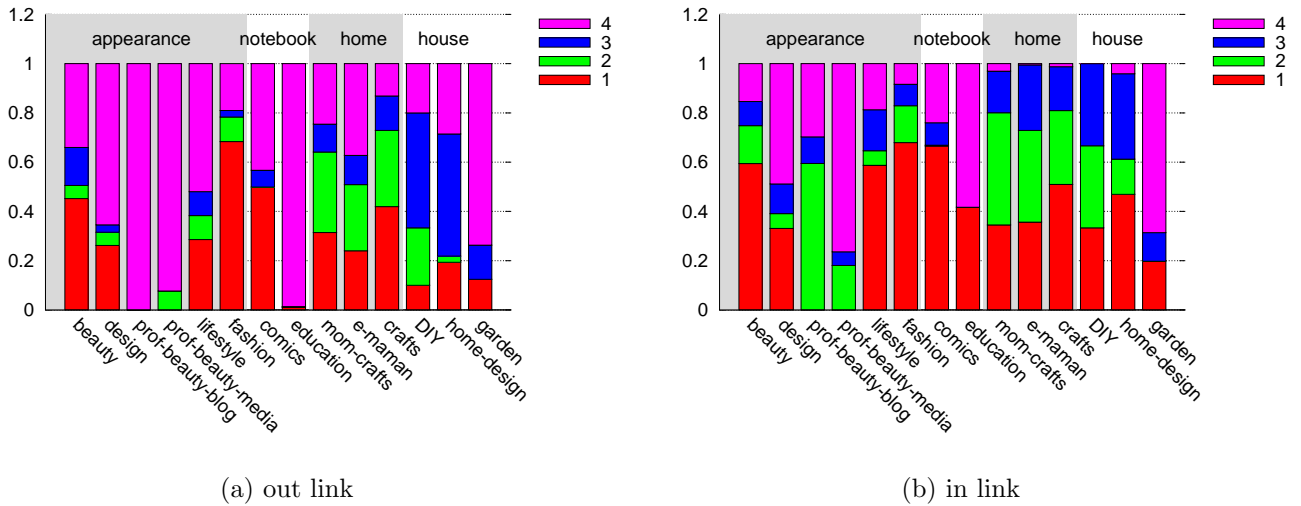


Figure 4.7: Fraction of in and out link distances by territory in individuality continent.

One may deepen these observations by considering the territory layer. Figure 4.7 rep-

3. X-sport: extreme sports

resents 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 4.7). 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 4.6). 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. There are 6 territories in this class which belong to *home* and *house* regions (the 6 last territories in Figure 4.7). 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.

4.3.3.2 Community mapping based on community distance

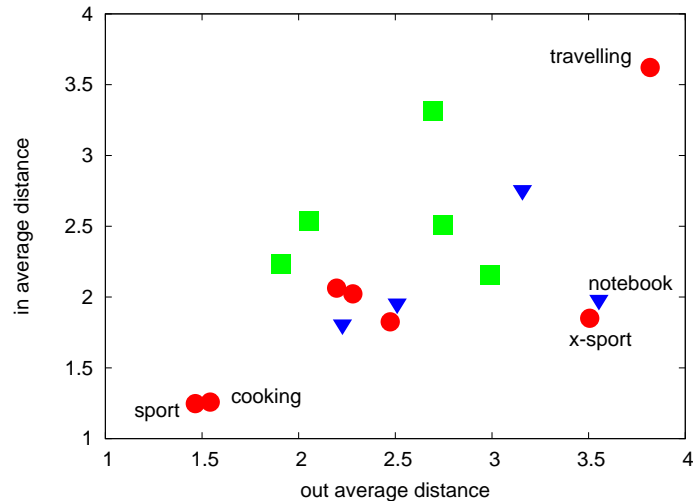


Figure 4.8: Average in and out links distance correlation at region layer. Green square=society, red circle=leisure, blue triangle=individuality.

So far, I have analyzed the fraction of link distances in each region. In order to provide an overall picture, I have used average distances of in and out links as the coordinates of each region on a 2D map (Figure 4.8). We may note that the patterns found in Figure 4.6

are confirmed here despite the use of an average value. For example *x-sport* and *notebook* are clearly grouped together. *Self-centered* communities also appear with low values of out-distances, 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).

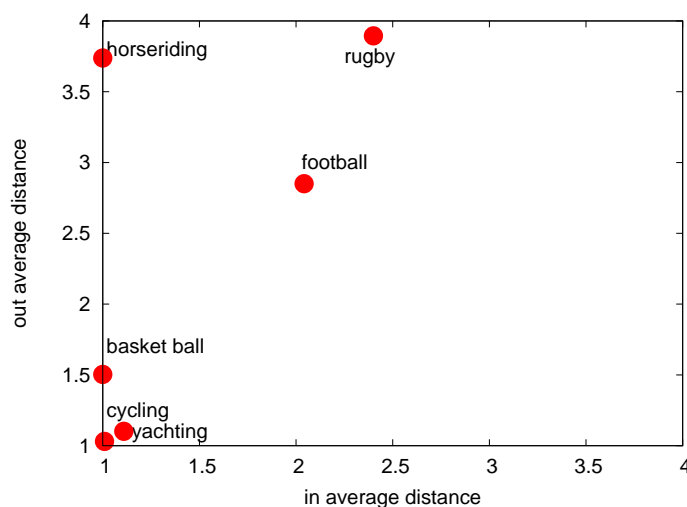


Figure 4.9: Average in and out links distance correlation in *sport* community

Now I focus on more specific communities at territory level. I first consider territories within *sport* region (Figure 4.10). In this example I only study *sports* territories which deal with one sport in particular and not blogs related to sport news. First we may 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 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 4.10). First we may 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.

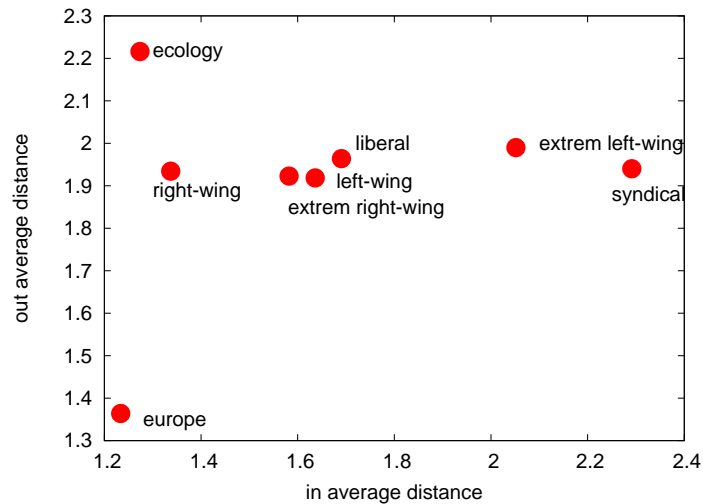


Figure 4.10: Average in and out links distance correlation in political communities

Until now I have studied community citation profiles in terms of proportion of links and average distance rather than number of links. I complete the study in Figure 4.11 where I 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 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

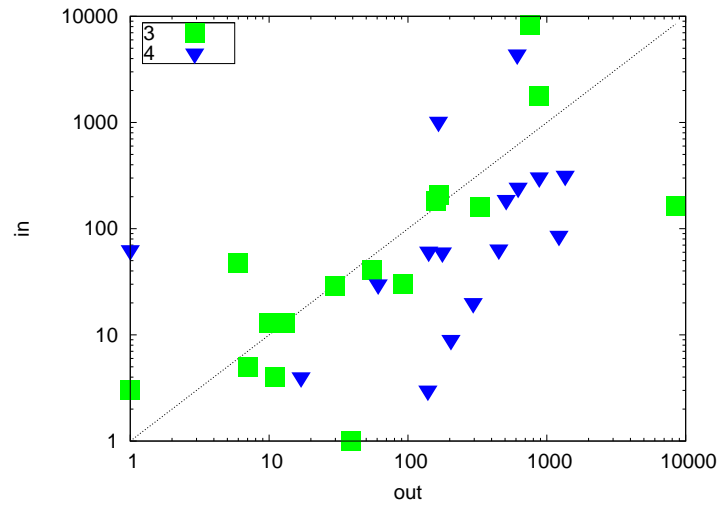


Figure 4.11: in and out links distance correlation at region layer (to help readability the two axes are in log scale). Each point corresponds to a region R and a distance d and has coordinates $in_d(R)$ and $out_d(R)$. All points with $d = 1, 2$ would be on the diagonal so we do not display them. Squares correspond to $d = 3$; triangles to $d = 4$.

regions belong to *society* continent and are *agora* and *technology*. We may qualify them as *popular* communities as they attract citations from distant blogs.

4.4 Case study: Automatic community structure

In Section 4.3 I have applied the proposed community-based methodology to the topical community structure obtained manually. In this section I apply it to the same blog network (to study citation behavior) but this time with regard to an automatic community structure built according to topological features (i.e. citation links). The first goal is to compare the results obtained for the automatic and manual cases. The second objective aims at generalizing the proposed methodology to other types of community structures.

As mentioned previously, most graphs from different contexts can be decomposed in sub-graphs of communities. In community detection algorithm field the most common definition of communities is a set of sub-graphs with nodes densely connected and the nodes belonging to different communities being sparsely connected. There are several algorithms in literature as mentioned in section 2.5. One problem in most community detection algorithms is that they are not deterministic which means that two executions on

the same graph may produce two distinct results. For this study I have used an extension of Louvain community detection algorithm [BGLL08] (a description of the algorithm is given in the next section). One benefit of this algorithm is that it calculates community partitions which are stable over time, called *communities cores*. In the previous Section I have studied link citation properties with regards to a topical community structure. Here I consider this automatic community detection algorithm which uses topological information to detect density connected sub-graphs.

4.4.1 Louvain community detection algorithm and community cores

Community detection algorithms aim at partitioning the network into communities of densely connected nodes. Many algorithms have been proposed to find good partitions in a reasonably fast way. Community detection algorithms can be classified into three categories: divisive algorithms, which detect inter-community links and remove them from the network, agglomerative algorithms, which merge similar nodes/communities recursively and optimization methods, based on the maximisation of given a function. The quality of the partitions resulting from these methods is often measured with modularity. Louvain algorithm is considered as an agglomerative algorithm.

The algorithm is divided in two phases repeated iteratively. In this case we assume that we have a weighted network of N nodes. First, the algorithm assigns a different community to each node of the network; in this initial partition there are thus as many communities as there are nodes. Then, for each node i it considers the neighbors j of i and evaluates the gain of modularity that would take place by removing i from its community and by placing it in the community of j . The node i is then placed in the community for which this gain is maximum only if this gain is positive. For each algorithm execution we obtain a partition.

The detection of core communities consists in detecting the most stable communities [QF10] appearing in several executions of the algorithm⁴. First, the Louvain algorithm is applied n times (in our case we fix $n = 100$). Second, for each pair of nodes (i, j) we calculate the number of executions (frequency noted $P_{(i,j)}$) for which i and j appear in the same community. After, a threshold that is applied to $P_{(i,j)}$ values. For example if the threshold is equal to 1 it means that we consider only communities in which the nodes have appeared in the same community n times. The variation of the threshold produces a hierarchical community structure.

4. The algorithm is non-deterministic which means that two executions on the same graph may give two different results.

4.4.2 Community structure

I have run the Louvain community detection algorithm and communities cores and have obtained a hierarchy of three levels (see Figure 4.12). The Louvain algorithm has been applied 100 times. A matrix P is built, containing for each pair of nodes i, j the number of times i and j have been in the same community out of n executions of the algorithm. This matrix may correspond to a graph where each pair of nodes (i, j) is connected, with a weight equal to P_{ij} . After, we variate the threshold from 0% to 100% to obtain the hierarchy. A threshold of 10% means we delete all edges with weight inferior to 10%. Then if threshold= 0%, there is only one community containing all nodes. A threshold equal to 100% means two nodes i, j are in the same community only if $P_{ij} = 100$. The obtained hierarchical structure has 3 three communities at level 1 and six at level 2. However the resulting hierarchy had initially more levels and communities that where removed. Indeed, I have kept only significant communities and ignored very small communities (less than 50 blogs).

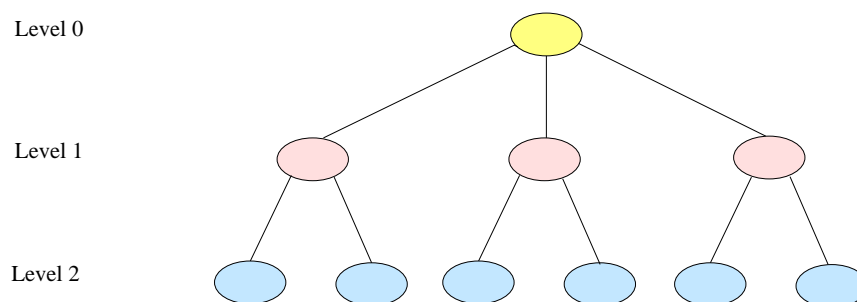


Figure 4.12: Automatic community structure

4.4.3 Results

Homophily

First, I measure the Δ_j probabilities over the whole graph $G = (V, E)$ in order to evaluate the impact of the level j in the automatic hierarchical community on homophily. The results show that $\Delta_1(G) = 99\%$ which means that 99% of blogs cite blogs from the same community at level 1. At level 2 the homophily remains very high with $\Delta_2(G) = 98\%$. The homophily for the whole graph is higher in this case than in topical communities (89% at continent level, 74% at region level and 21% at territory level). If we compare with topological communities (see Section 4.3.2) we observe that the homophily decreases

significantly which is not the case in automatic hierarchical community. My hypothesis is that in the topical communities case the more we try to sub-divide a community topic into subtopics, the higher the probability that blogs in subtopics interact. This is not the case in automatic communities where the goal is to minimize inter-community interaction.

Table 4.5: Probability to link blogs from the same community

Community ID	$\Delta_2(C_i)$
0	0.99
1	0.96
6	0.97

After considering the whole graph, I now focus on links at level 1 which contains three communities (Table 4.5). The homophily values remain very high for all sub-communities of these three communities.

Community distance

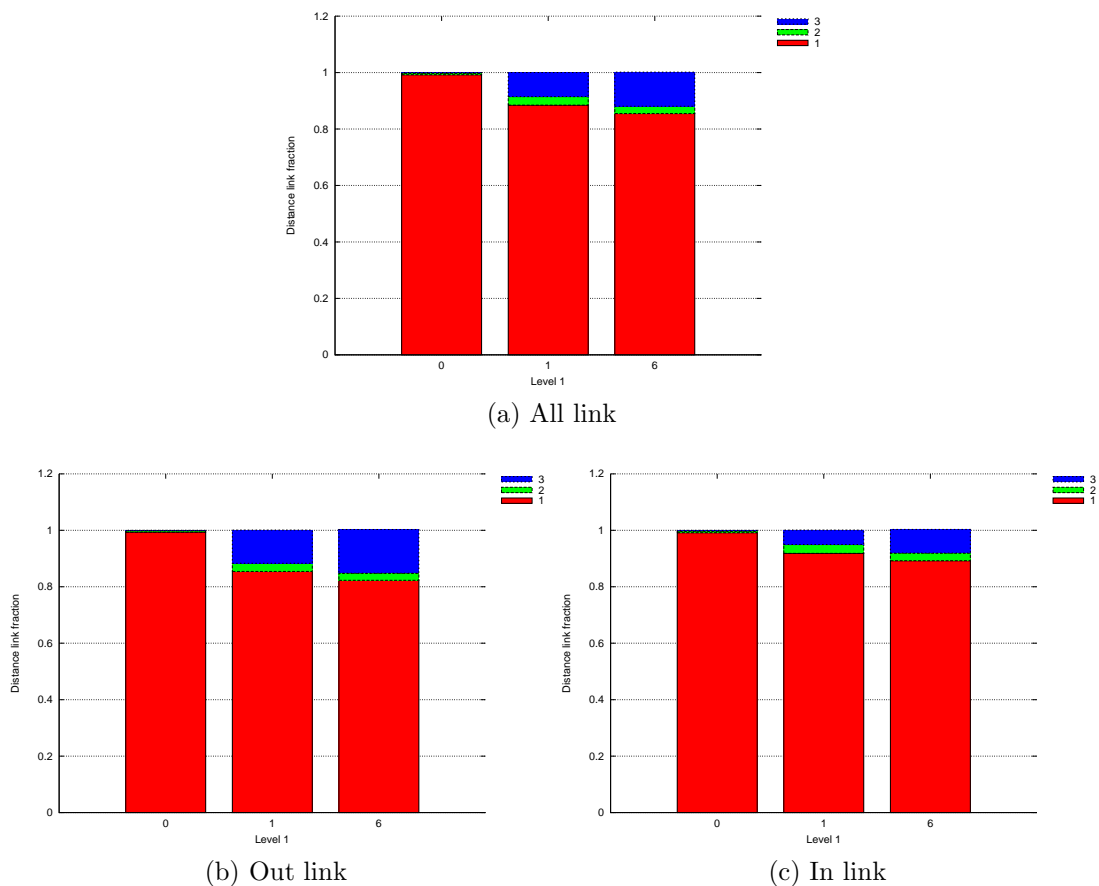
Table 4.6: Distribution of community distances in G

Distance k	# links	% of links
1	56607	98%
2	329	0.5%
3	267	0.4%

The community distance is calculated based on the hierarchical community described in Section 4.4.2. As we have three levels 0, 1 and 2 we obtain three distance values. A distance 1 means that the two blogs are from the same community at the level 1 and 2 and on the opposite a distance 3 means that the two blogs do not belong to the same community for any level. The distribution of community distances in the whole graph G shows that 98% of links have a community distance of 1. This is not a surprising result: the community detection algorithm seeks to maximize modularity, which reduces community distance.

Now, we calculate for each *community* its fraction of links at each distance $d = 1, 2$ and 3, for each community level (Figure 4.13 and 4.14). At level 1 we do not observe significant

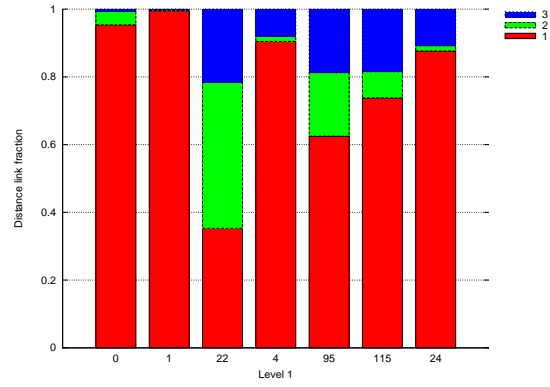
Figure 4.13: Fraction of in and out link distances community at level 1



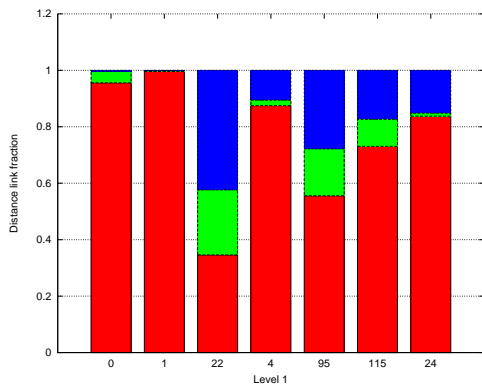
pattern differentiation nor any difference between incoming and outgoing behaviors. On the other hand, at level 2 we can observe some communities with a significant number of links with a community distance greater than 1. Those links may be considered as non intuitive and can be helpful to detect outliers. However, the majority distance is 1 for most of them.

In my opinion, one interesting feature of using automatic community structure is to study non topological graphs. One can imagine to build a network where nodes are blogs, and a link is made between two blogs if they share a semantic content. The automatic community structure can be a way to study such a network at different scales and give interpretations on how the semantic content is diffused through topological communities.

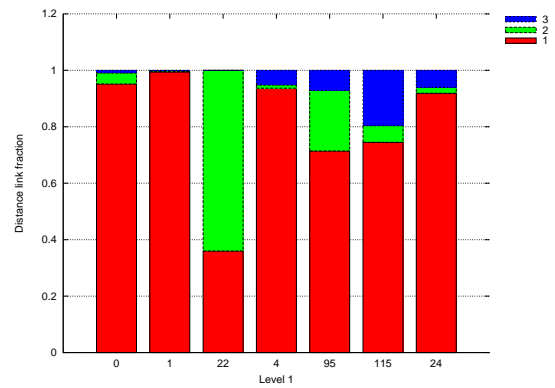
Figure 4.14: Fraction of in and out link distances community at level 2



(a) All link



(b) Out link



(c) In link

The goal in this Section was to show that automatically detected communities could also be used for the proposed community-based methodology.

4.5 Summary and perspectives

I 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. I have compared homophily with the *modularity* quality function and have shown their complementarity. Links community distance captures whether nodes of a network interact with nodes from *close* or *distant* communities. I 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. I have also studied the case of automatic hierarchical communities and have compared the two studies. Finally, I 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 study other community structures, identified by other automatic community detection algorithms. I will pay specific attention to overlapping communities. The methodology I proposed may also be applied in this case. Having several types of hierarchical community structures is a good opportunity to compare the information we get regarding citation links. In a second step, I want to investigate other networks and determine statistical metrics to better classify and understand hierarchical communities. In the following chapter, I will study diffusion cascades, i.e. successions of citations of a given post.

Network cascades

Contents

5.1	Introduction	76
5.2	Cascades definition and computation	76
5.2.1	Data corpus	76
5.2.2	Cascade computation	77
5.3	Macroscopic analysis: cascade structure and community impact	78
5.3.1	Cascades shapes	78
5.3.2	Cascades topological, temporal and community properties	81
5.3.3	Temporal and topological cascades properties correlation	85
5.4	Microscopic analysis: impact of individual nodes on cascades	88
5.4.1	Impact of cascade origin	88
5.4.2	Impact of intermediate blogs	89
5.5	Conclusion	93

5.1 Introduction

Diffusion phenomena happen when an action, information or idea becomes adopted due to the influence of neighbors in the network [Gra78]. As we have seen in Section 2.4 there are many models of influence spreading in social networks. Many diffusion models predict the number of people who will adopt (an opinion or innovation) over time and do not explicitly account for topological patterns. On the contrary, the *cascade* point of view allows to investigate which individual dynamics lead to global spreading phenomena. My goal in this chapter is, in a first step, to characterize cascades using several metrics and in a second step, to determine how community properties impact cascade characteristics.

Cascades have been theoretically analyzed in random graphs using a threshold model [WD07]. However, only few empirical analyses of the topological patterns of cascades have been done [LSK06,CMAG08]. Those works aim at characterizing cascade topological properties separately and demonstrate that they follow a power-low distribution. In this work I propose to go beyond known studies; in addition to topological properties, I investigate community information to understand how cascades spread through communities. I apply this approach to the topical community structure described in the previous chapter. This community structure has three hierarchical levels, which allows an analysis at different scales.

A typical question I address in this chapter is: what is the impact of the community from which the cascade starts on cascade size, shape or duration?

The chapter is structured as follows. First, I explain how cascades are computed and present first statistics. Second, I investigate cascade shapes to determine which cascade topologies are more frequent. Then, I study at a macroscopic level the different properties which characterize a cascade and investigate their impact on each other. Finally, at the node level (i.e. microscopic level) the impact of the community of the cascade origin is observed with differences for each metric.

The work presented in this chapter is currently under submission.

5.2 Cascades definition and computation

5.2.1 Data corpus

The corpus analyzed in this section was obtained with the same crawling methods as the dataset used previously but on a different period of time. The advantages of this new dataset is that it contains a higher number of blogs and was measured during a longer period. In addition, having multiple versions of the same object, on different periods of

time, is helpful to validate the various datasets and obtained results. This dataset covers a period of five months, from February 1st to July 1st 2010. The dataset is composed of 10,309 blogs, 848,026 posts and 1,079,195 citation links.

I proceeded to a dataset cleaning which consisted in:

- removing links that point to posts outside the dataset or to other resources on the web as pictures or web-pages.
- removing all posts with incorrect time stamps (i.e. out of the measuring period).
- removing auto-citation links (links between two posts from the same blog).
- removing links which cite a future post (mistake).

The resulting post network is composed of 20,885 posts and 25,837 links; it was used to extract cascades as explained in the next section.

5.2.2 Cascade computation

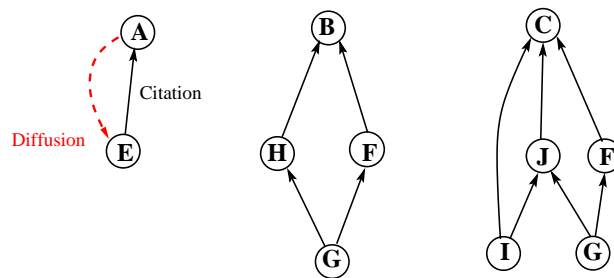


Figure 5.1: Cascade samples.

Cascades are subgraphs of the post network, where each node corresponds to a post and edges to citation links. In order to compute post cascades, I started by posts which do not cite any other post; each of them represents the beginning of a cascade called "origin". Consider one such post; if it is cited by one or several posts, the process carries on recursively: posts which have cited this citing post (or posts) are looked for and so on. Each post can belong to several cascades, represented as Directed Acyclic Graphs (e.g. post *F* in Figure 5.1). On Figure 5.1, the cascades origins are *A*, *B* and *C*. Information is therefore spread from the origin to the leaves (posts with no incoming link).

Propagation flows represent influence between blogs through their posts. I focus on information diffusion among different blogs, therefore links between two posts from a same blog were removed. Indeed, self-citations can make cascades longer but do not represent a diffusion process from one blog to another. This hypothesis has an impact on cascades

sizes however ignoring blogs self-citations removes some biases and leads to more relevant cascades.

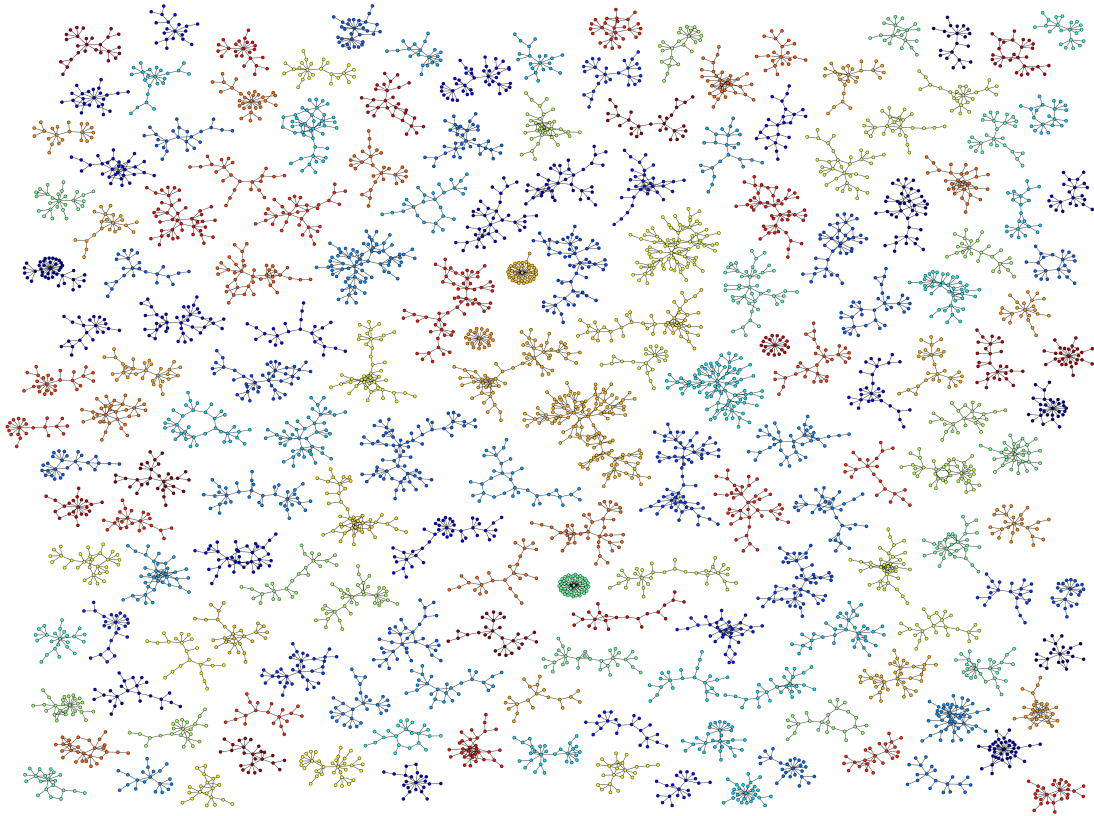


Figure 5.2: Extracted Cascade samples.

The total number of cascades is 10,659. The most common cascade is the trivial one composed of two posts and represents 65% of the whole. Figure 5.2 show samples which have more than 10 nodes, representing 1,5% of all cascades.

5.3 Macroscopic analysis: cascade structure and community impact

5.3.1 Cascades shapes

This section aims at studying cascade shapes, in order to know what types of cascades appear frequently in the blog network. Do they look like trees, stars or chains? I explored

the structure of cascades by performing the following procedure. After computing all cascades, I used an isomorphism algorithm to determine whether a cascade was identical to another.

Isomorphic cascades

Two graphs G and G' are isomorphic if there is a one-to-one mapping from the nodes of G to the nodes of G' that preserves the adjacency of nodes.

I have used VF2 algorithm based on a depth-first strategy [CFSV01]. There is no known polynomial time algorithm for graph isomorphism, however, the computational time in this case is reasonable because we deal with small graphs.

The most used algorithms are Ullmann, Schmidt and Druffel (SD), VF, VF2, and Nauty. Brute force graph isomorphism results in a depth-first search tree. Ullmann reduces the search space through backtracking [Ull76]. SD is another backtracking algorithm; however, it relies on information contained in the distance matrix representation of the graphs [Ull76]. VF is based on a depth-first strategy with a set of rules to prune the search tree [Ull76]. VF2 uses the same concept but stores the information in more efficient data structures [CFSV01]. Nauty is based on a set of transformations that reduce the graphs to a conical form on which isomorphism tests are faster [McK81]. These algorithms, while exponential, strive to be efficient in practice.

To reduce complexity, some algorithms give an approximate, heuristic solution. They use an efficiently computable signature to compare graph isomorphisms.

Interpretation

There are a total of 10,659 cascades and 641 isomorphic shapes. The most common post network cascades shapes are given in Table 5.1, where the red post is the cascade origin.

65% of cascades are composed of two nodes (6,992/10,659). The second most frequent shape represents 10% with three nodes. If we compare all shapes we may observe that cascades tend to be *stars* (e.g. cascade 19) rather than *chains* (e.g. cascade 30). It is more obvious if we compare the shape frequency of cascades with the same size. For example, if we consider cascades 2, 30 which contain 3 nodes and 19, 4 which contain 5 nodes the *star* shapes are more frequent. If we compare the shape frequencies with previous work done on another blog network [LMF⁺07] we observe that the ranking order is almost the same.

Now, I compare the cascade shapes in the three continents (three last columns in Table 5.1), in order to see if some shapes are more frequent in a continent. Therefore, for each

continent I have computed the fraction of each shape over all cascades which start in this continent. I observe that for the three continents, the shapes frequency distributions are very close. This means that at continent layer the cascade origin has not a significant impact on cascade shape. However, we may find a more important correlation between cascade origin and shape between communities at a lower level in the community structure.















ID	shape	# nodes	# links	frequency	Society	Individuality	leisure
1		2	1	6992	71.6%	79.5%	72.3%
2		3	2	1173	13.2%	8.5%	11.4%
81		4	3	397	4.3%	2.2%	4.2%
30		3	2	370	3.5%	4.1%	5.5%
19		5	4	182	2.0%	1.1%	1.2%
29		4	3	134	1.4%	1.2%	1.6%
88		6	5	83	1.4%	0.5%	0.4%
4		5	4	56	1.0%	0.2%	0.2%
101		3	3	52	0.8%	0.7%	0.6%
702		4	3	46	0.4%	0.3%	0.5%
418		7	6	33	0.3%	0.2%	0.5%
107		5	4	30	0.4%	0.1%	0.1%
333		4	3	30	0.2%	0.4%	0.5%
122		5	4	29	0.3%	0.4%	0%

Table 5.1: cascade shapes ordered by frequency

Now I focus on the largest cascades represented in Figure 5.3¹. In terms of frequency, all those shapes appear only once as they are very complex. Unlike most frequent shapes described in Table 5.1, large cascades seem to have *tree-like* shapes.

When looking at cascade topologies we may note that they are very complex and we may distinguish nodes with a more important role in the cascade spreading phenomena.

1. In this visualization links are not directed for readability.

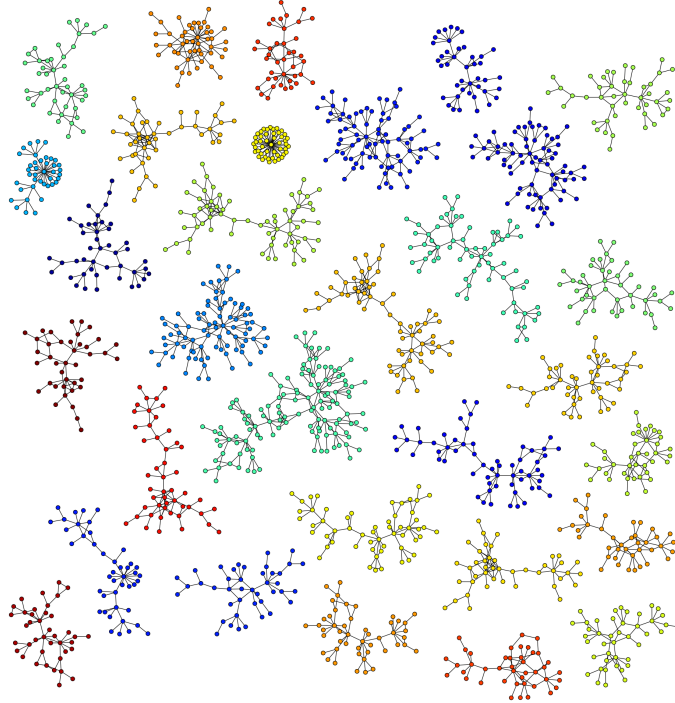


Figure 5.3: Largest cascades in our dataset. All cascades has at least 30 nodes.

This raises the following questions: what makes a cascade have certain characteristics? What makes a cascade longer or bigger? Those issues are addressed in the next section.

5.3.2 Cascades topological, temporal and community properties

To understand cascades characteristics and understand how they are formed we need to characterise them precisely with regard to all their features detailed below (see Table 5.2). We classify cascade properties in three categories: topological, temporal, and community-related. In this section I consider properties at the cascade scale rather than at the node scale. The topological characteristics regroup classical graph measures. Temporal features are measured by the total duration of the cascade (T) and the timestamp of cascade origin (T_s).

A first step is to consider each property and study its distribution using cumulative density distributions (also called PDF probability density function).

Table 5.2: Cascade properties

Notation	Description
N_n	Number of nodes
N_l	Number of links
N_{lvl}	Number of levels
r	Degree assortativity
δ	Cascade density
T_s	Timestamp of the cascade start
T	Total duration
A_{cm}	Average link community distance

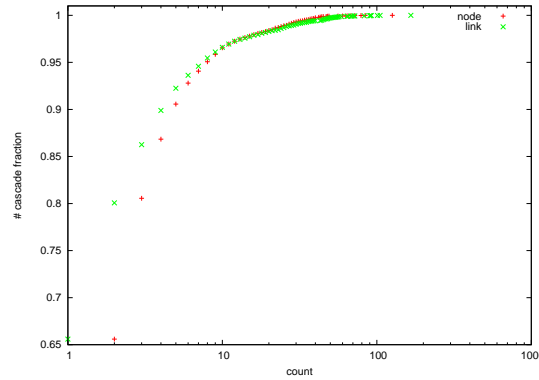


Figure 5.4: Cumulative distribution of the number of nodes (in red) and links (in green).

The number of nodes and links indicate the size of the cascade. The cumulative distribution of the number of nodes (N_n) and links (N_l) per cascade shows a similar heterogeneous distribution (Figure 5.4). The total number of cascades is 10659. A point (x, y) with $x = 10$, $y = 95\%$ on the red plot means that there are 95% cascades which have a number of nodes inferior to 10. We observe that 65% of cascades are composed of only two nodes (6992 cascades). In fact most posts are not cited more than once. In addition, 5% of cascades have sizes (N_n and N_l) greater than 10. There are also very large cascades with over 100 nodes. The same results are observed for the number of links. In fact the numbers of edges and nodes increase similarly which suggests that the average degree in the cascade remains constant as the cascade grows. This also suggests that they are mostly tree-like.

In addition to the size of cascades, I analyse their depth (i.e. their number of levels)

noted N_{lvl} . It corresponds to the length of the maximum path between the cascade origin and a leaf. The algorithm is described in Algorithm 1. This algorithm is recursive and covers the cascade from origin to the leaves. Figure 5.5 represents the N_{lvl} cumulative density function. A point (x, y) with $x = 4$, $y = 98\%$ means that 98% cascades have at most 4 levels. We observe that almost 84% of cascades have $N_{lvl} = 1$, but some have up to 16 levels.

Algorithm 1: Level number computing

Input: Cascade (DAG graph) with the *root* node r .

Output: Integer N_{lvl}

Predecessors(node i): set of nodes which cite i

$N_{lvl} \leftarrow 1$

Rec(*node* i , *level* l):

for $u \in \text{Predecessors}(i)$ **do**

$N_{lvl} \leftarrow \text{maximum}(l, N_{lvl})$
 Rec(u , $l + 1$)

Rec(r , 0)

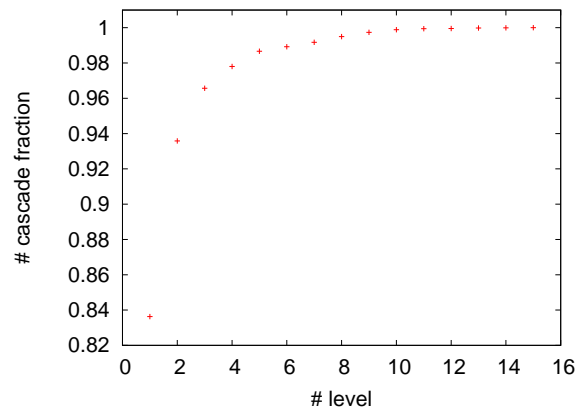


Figure 5.5: N_{lvl} cumulative density function

Next, I study the distribution of cascade density (noted δ). The density of a graph is the number of links divided by the number of possible links between all pairs of nodes. Given a cascade with $N_n = n$ and $N_l = m$, the density is $\delta = \frac{m}{n \cdot (n-1)}$. Note that this density formula is for directed graphs. Density is a fraction that goes from a minimum of

0 if no edge is present² to 1 if all edges are present³.

Here we consider only cascades with $N_n > 2$. Figure 5.6a shows the density distribution; we may observe that it is also heterogeneous and that most cascades have a density comprised between 0.4 and 0.7. The density values may be considered as high, but this result is not surprising because most cascades have a small size. The question one may ask is if the size of a cascade impacts the density: in Figure 5.6b we show the correlation between density and size of cascades. We observe that the density is inversely proportional to the size of the cascade.

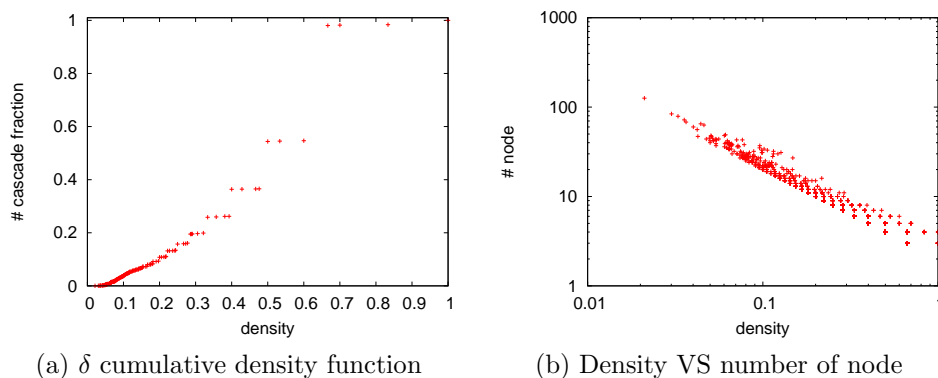


Figure 5.6: Cumulative distribution of cascade density.

Another typical feature of real world networks is the tendency of nodes of a certain degree to be connected with other nodes with similar degree. We measure it by *degree assortativity* noted r [New02, PN03]. Positive values of r indicate a correlation between nodes of similar degree, while negative values indicate relationships between nodes of different degrees. r lies between -1 and 1 . When $r = 1$, the network is said to have perfect assortative mixing patterns, while at $r = -1$ the network is completely dissociative. Newman [New02] has compared many networks and noted that biological and technological networks show disassortative behaviour while social networks are assortative. The reasons for such results are not completely understood.

Figure 5.7 represents the cumulative distribution of cascade degree assortativity (only cascades with $N_n > 2$ are considered). We observe that 0.95% of cascades have a negative degree assortativity. It means that cascades in the blog network tend to have a disas-

2. which is not possible in our case because a cascade has at least one edge
 3. This is also impossible because there can be no cycle in a cascade.

sortative behaviour. In addition, the degree assortativity measure gives an indication of cascade shape [New10]. A disassortative graph has high-degree nodes which tend to connect low-degree ones creating a star-like structure. In an assortative graph, high-degree nodes are linked together, surrounded by a less dense periphery of nodes with lower degree.

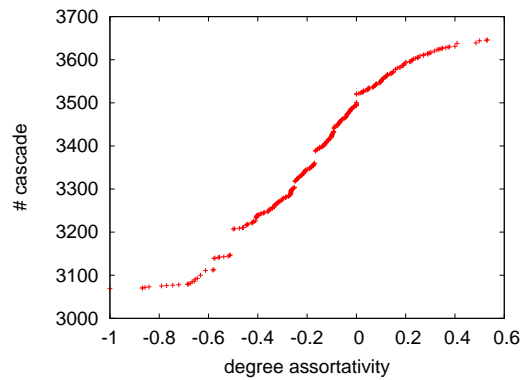


Figure 5.7: Cumulative distribution of cascade degree assortativity.

In addition to topological properties, I used the community distance defined in Section 4.2.3 and the topical community structure to measure the tendency of cascades to relate *close* or *distant* communities. This is captured with a value noted A_{cm} measured as follows. First, we compute the community distance for each link. Afterwards, we compute for each cascade the average community distance of its links. We then investigate the impact of the A_{cm} measure on other cascade properties. The cumulative distribution on Figure 5.8 shows that the average community distance has a heterogeneous distribution with majority between an average of 2 and 3.

5.3.3 Temporal and topological cascades properties correlation

In the previous section the different cascades properties have been studied independently. Now I go further to determine how topological and temporal features may impact each other. I first study the impact of the average community distance on the duration of the cascade. The intuition is that if a cascade crosses over links which have a high community distance the cascade duration may be longer.

In Figure 5.9 I represent the average community distance with the 25th and 75th percentile in relation to cascade time duration. For each 0.5 A_{cm} interval I compute the

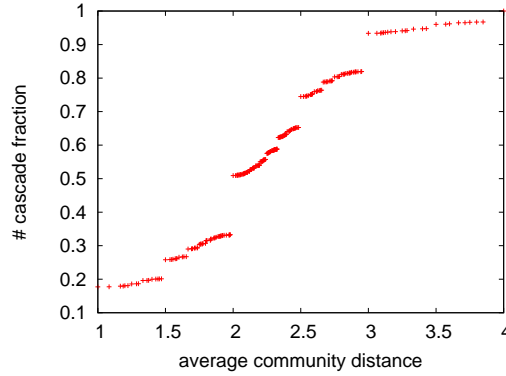


Figure 5.8: Cumulative distribution of average community distance.

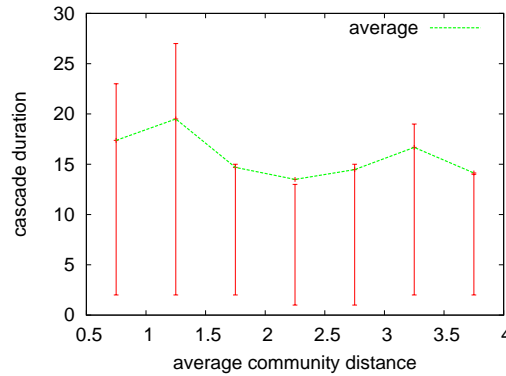
(a) δ cumulative density function

Figure 5.9: Impact of community distance on cascade duration.

average duration of all cascades in the interval. We may observe two peaks (in $[1, 1.5]$ and $[3, 3.5]$ intervals). The cascade duration seems to be higher when the average community distance is very small or high. As an example, the average difference between cascades with A_{cm} in $[1, 1.5]$ and A_{cm} in $[2, 2.5]$ is about 5 days.

After studying the impact of communities on cascade duration, we now observe the impact on cascade size (see Figure 5.10). The cascade size (in terms of number of nodes and links) tends to be maximum for cascades with an average community distance between 2 and 2.5.

Next, we investigate whether the number of levels N_{lvl} which is a topological metric, is correlated to cascade duration. One may have the intuition that the higher the number of levels, the longer the cascade. The result is shown in Figure 5.11a where the cascade

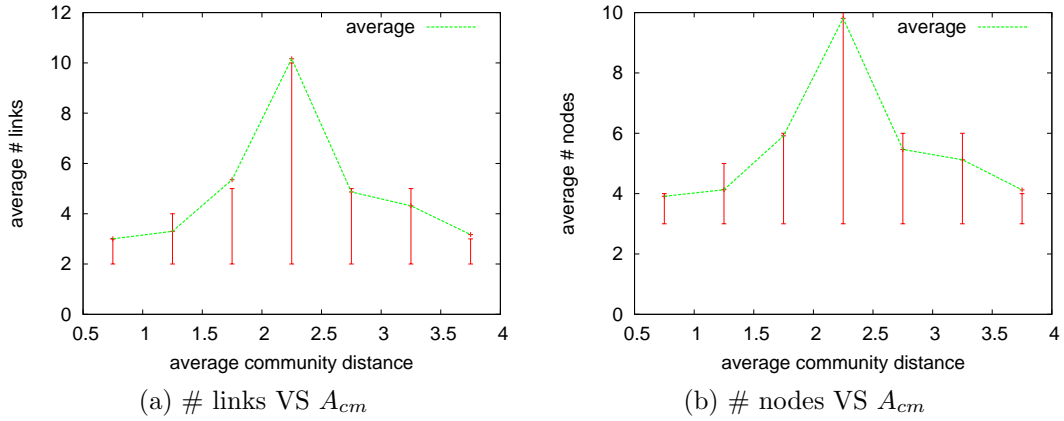


Figure 5.10: Impact of community distance on cascade size.

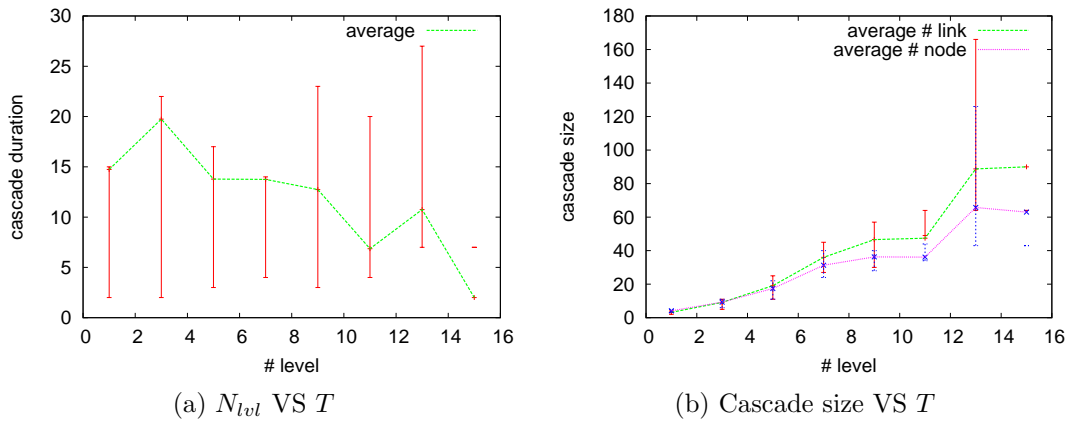


Figure 5.11: Number of level and size cascade impact.

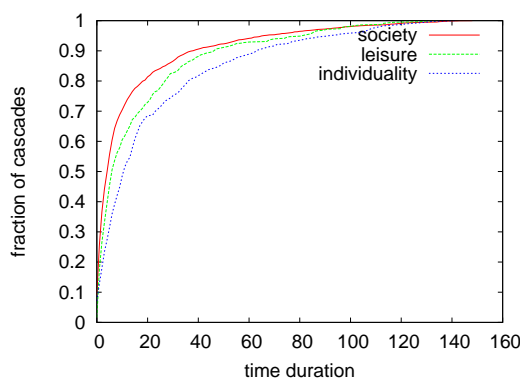
duration increases for values of N_{lvl} between 2 and 3 with an average of 5 days. After that, the duration is decreasing. So, the cascades with a longer duration are in average cascades with a small number of levels (2 and 3) rather than a high number as one may expect. We also observe that spreading speed is higher when the cascade has more levels and so tends to have a *chain* shape. On the other hand, the cascade size has a different impact (see Figure 5.11b where the average cascade duration increases proportionally to the size for both N_l and N_n). This means that the higher the number of nodes in the cascade, the longer the cascade duration.

5.4 Microscopic analysis: impact of individual nodes on cascades

In the previous section I have studied cascades properties at macroscopic level. Now, I go deeper and investigate in what way a node impacts the rest of the cascade. The originality of this work is that I include a community aspect, at different layers. As explained earlier, the community structure has been built manually by professional blog analysts according to blogs topics. It is composed of three hierarchical levels: *continent*, *region* and *territory* (from the most general to the most specific). First, I study the impact of continent of the node which started the cascade. Second, for each node cascade node I study how communities affect the cascades in order to differentiate their impact on diffusion.

5.4.1 Impact of cascade origin

I first study the differences between cascades depending on which community they start to spread from. In this section I focus on *Continent* layer (which corresponds to the level 1 in the hierarchical community structure), with the three communities: *Leisure*, *Individuality* and *Society*.



(a) Cumulative density function

Figure 5.12: Impact of community origin on cascade time duration.

The first metric is cascade duration represented by the cumulative density functions (Figure 5.12). I use the cumulative density function in order to compare the various community impacts. We may see that cascades which start from *Society* community have a shorter duration than *Leisure* cascades and that in the opposite *Individuality* community tends to initiate cascades with a longer duration.

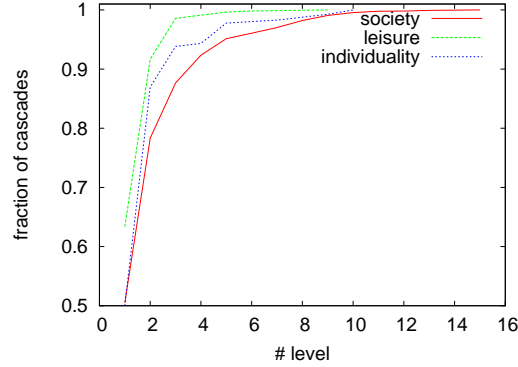


Figure 5.13: Impact of community source on cascade levels.

For the number of levels, we observe the opposite (see Figure 5.13). The *Society* community produces in proportion longer cascades (with a higher N_{lvl}). The same distribution has been observed for cascade sizes (the figure is not presented here). In summary, *Society* community induces shorter cascades in time duration, with a longer number of levels and a higher number of nodes. The same analysis may be done for the other communities.

Next, we study the average community distance for each community. The first observation is that the average community distance of cascades which start in *leisure* continent have at 70% a community distance $A_{cm} = 1$. The *leisure* cascades tend to be smaller than those starting from other continents and also have a small community distance which means that citations come from very close communities. On the other hand, *Society* cascades have a higher community distance and larger cascades.

This gives an indication on the correlation between topological and community properties. We may suppose that the links community distance impacts the diffusion flow. Indeed, observing the link community distance can help understand resulting cascade characteristics.

5.4.2 Impact of intermediate blogs

Now, we investigate the impact of not only the cascade origin but of all nodes of the cascade (see notation in Figure 5.3). We illustrate an example in Figure 5.15. The post i (in green) is published at T_i and belongs to con_i , reg_i and ter_i communities at the continent, region and territory levels respectively. $T_max_i = \max(T_j, T_k)$ represents the impact of the post on the total duration of the cascade. $N_node_max_i = 3$ and $lvl_max_u = 2$. The

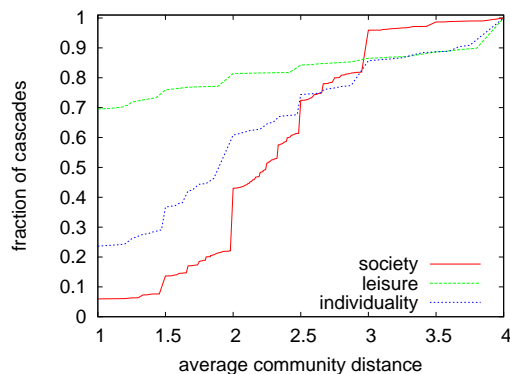


Figure 5.14: Impact of community of origin on average community distance.

Table 5.3: Table of symbols

Notation	Description
T_u	Time at which post u was published
con_u	Post u Continent
reg_u	Post u Region
ter_u	Post u Territory
lvl_u	Post u level in the cascade
lvl_max_u	Number of levels after the post u
T_max_u	Time delay between u and the last published post
$N_node_max_u$	Number of nodes after the post u

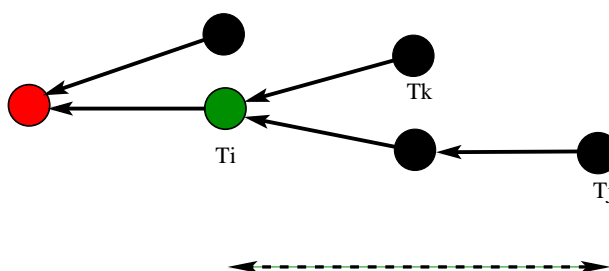


Figure 5.15: An example of node impact.

three metrics regroup temporal and topological properties.

We study each property for three communities in each level *continent*, *region* and *ter-*

ritory.

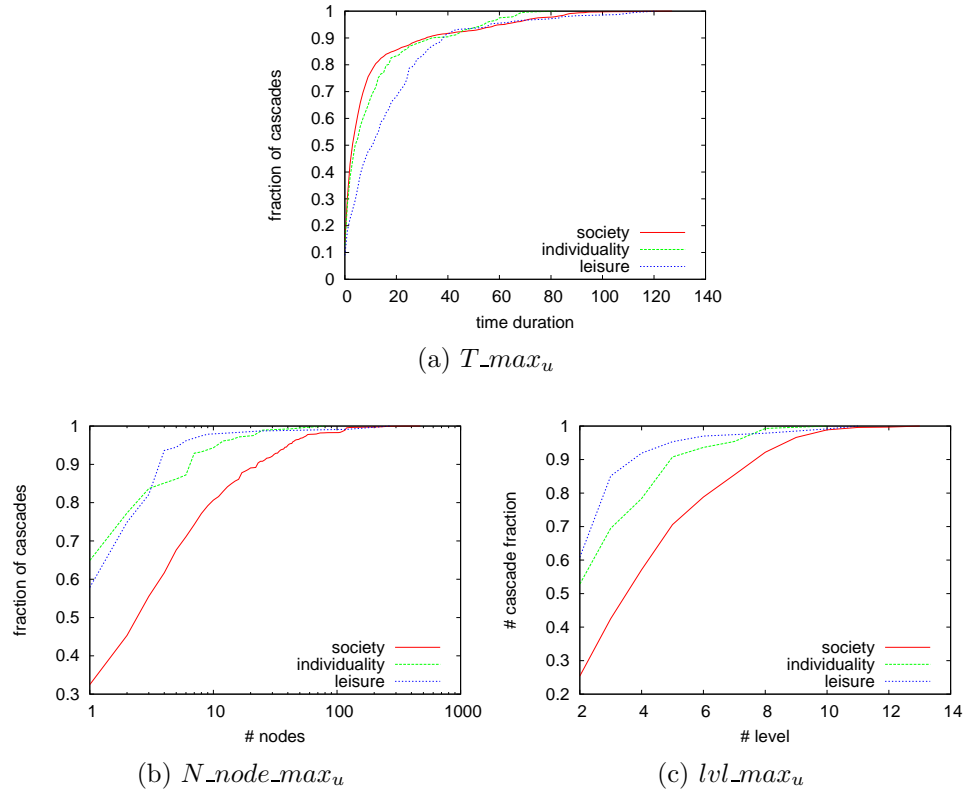


Figure 5.16: Impact of node community at Continent level.

We start by the individual impact at *Continent* level (Figure 5.16). With regard to the time delay T_{max} which a post causes (see Figure 5.16a), *Society* and *Individuality* posts have approximately the same impact while *Leisure* posts have a more important impact on cascade duration. On the other hand, posts belonging to *Society* community have an important impact on cascades in terms of number of nodes (note that x-axis is at log scale) and levels (Figure 5.16c and 5.16b). The conclusion is that when a diffusion passes through *Society* community, cascades tend to be larger but not with a higher time duration.

At *Region* level we consider three communities, *Agora* (which regroups mass media and opinion actuality), *Politics* and *Technology* blogs. The three communities belong to *Society Continent*. For cascade duration, the impact of the three communities is almost similar. However, with regard to the number of nodes and levels, *Agora* and *Politics* have a very similar distribution with a more important impact than posts from *Technology* community.

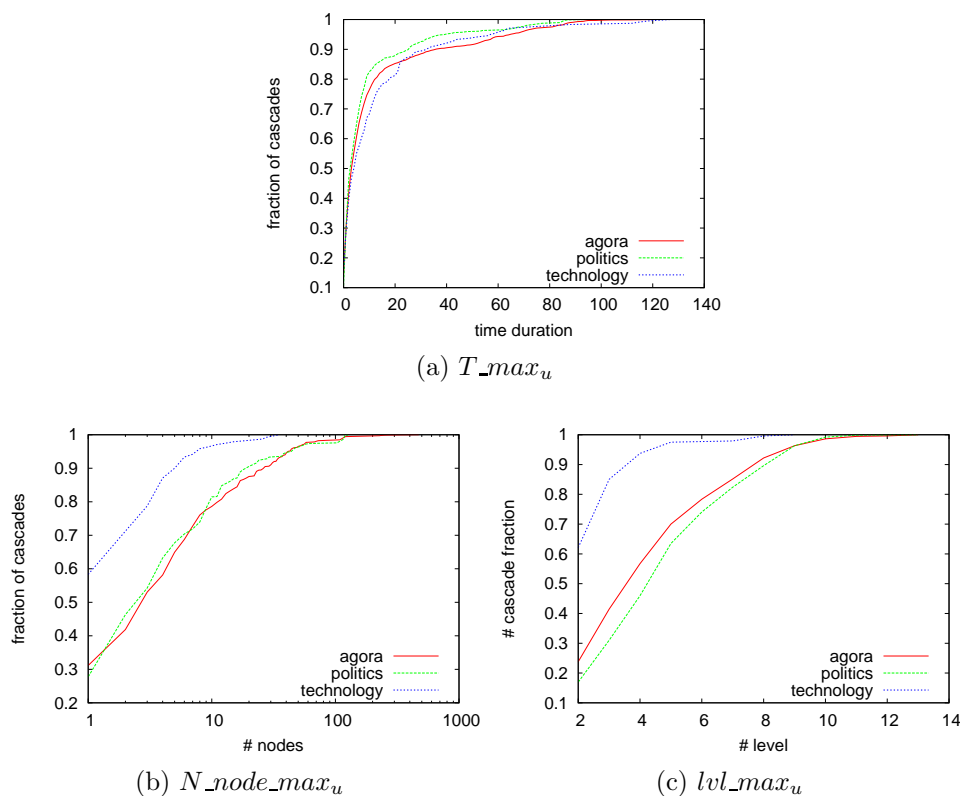


Figure 5.17: Impact of node community at Region level.

Agora and *Politics* posts provoke almost the same diffusion processes and therefore have a similar diffusion pattern in terms of diffusion size.

At *Territory* layer we consider three communities from *Politics* community: *Left-wing*, *Right-wing* and *Center-wing*. We observe that *Center-wing* has a more significant impact on cascade duration. It means that the cascades spread for a longer period. However, we do not observe a significant difference for cascades size. We note that the number of nodes does not exceed 110 for the three communities.

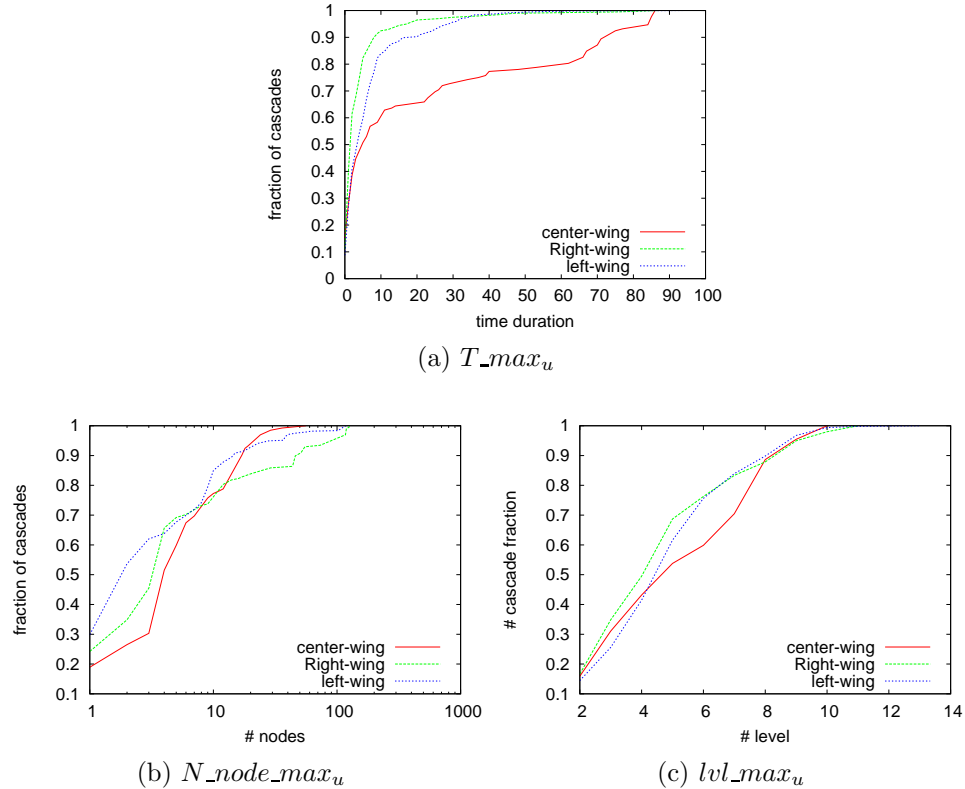


Figure 5.18: Impact of nodes community at Territory level.

5.5 Conclusion

In summary, I have proposed a new approach for the empirical study of diffusion cascades. I gave a definition of what I considered as a cascade; in particular I did not consider citations within the same blog in cascade computation. I have observed that cascades shape frequencies were very similar to previous work done on American blog network [LMF⁺07], which suggests that cascades shape in blog networks is a common property. The topological properties distribution analysis has shown a heterogeneous behaviour in general. In particular, I observed that cascade density was inversely proportional to cascade size and that cascades tended to be disassortative.

Second, in addition to topological properties, I have investigated community information to understand how cascades spread through communities. Therefore, I have used a topical community structure with three hierarchical levels which allows analysis at different scales. The community of cascade origin has an impact on cascade properties and especially

on average community distance for which the diffusion behaviours differ noticeably.

Finally, I did not consider only macroscopic scale: in Section 5.4, I showed that at node scale, community origin has a different impact on the rest of cascades features. At Continent level, *Society* nodes have a more significant impact on cascades size and number of levels but not on their duration. Political Communities related to politics have a similar impact on cascades size however *Center-wing* blogs increase the cascade duration. One perspective is to consider the impact of other nodes properties, for example nodes *Betweenness centrality* or node degree in the blog network.

Conclusion and Perspectives

In this thesis I have studied real-world diffusion phenomena in a blog network through the analysis of citation links. Understanding node interactions is a crucial to investigate diffusion phenomena. The approaches classically used for studying diffusion within a real-world network ignore the community structure. Indeed, they consider essentially node to node interactions while in this thesis in addition to those approaches I characterized diffusion behavior with regard to the network community structure.

I have shown that it was relevant to consider the community structure to better characterize and classify diffusion patterns at different scales.

I have applied this empirical methodology to a *French blog network* which have many advantages as it was observed during several months, and a community structure is defined over it.

In a first step, I have proposed a new approach and method to characterize post and citation dynamics in different blog communities. I have been able to identify classes of post popularity evolution from topological features of the network, and have investigated the impact of topical communities on citation dynamics.

I have shown that distinct patterns related to both the duration and the frequency of citations could be observed in the various communities. I have identified communities with a high popularity among posts with small citation duration.

The second problem I have addressed consists in characterizing the tendency of nodes to interact within their own community. I have proposed a methodology to analyse interaction behavior, with regard to the hierarchical community structure defined over the nodes. This

approach mainly relies on two measures: *homophily* and *community distance*.

Community distance measures whether nodes of a network interact with nodes from *close* or *distant* communities. Citation links have been studied at various scales, in the case of manual and automatic hierarchical communities and I have compared the two studies.

I have proposed a synthetic map based on an average value of community distances for incoming and outgoing citations and have illustrated it at the *region* and *territory* levels. Communities with apparently close topics (e.g. political wings) have a different behaviour and a different impact on other communities.

Finally, I have considered diffusion cascades. In a first step, I have characterized cascades using several topological, community and temporal metrics and in a second step, I have studied how community properties impact cascade characteristics. I have used a topical community structure with three hierarchical levels which allows an analysis at different granularities. The community of the cascade origin has an impact on cascade properties and especially on the average community distance for which diffusion behaviours differ significantly.

At a microscopic level, I have observed a real impact of intermediate nodes on cascade dynamics. I have shown that communities have a role on diffusion phenomena and I have been able to observe and quantify it.

The work conducted in this thesis opens several perspectives.

One perspective of this work is to study other community structures, identified by other automatic community detection algorithms. I would pay specific attention to overlapping communities. The methodology I proposed may also be applied in this case. Having several types of hierarchical community structures is a good opportunity to compare the information we get regarding citation links. The study of automatic community structures may allow us to detect "unexpected" links which connect distant communities and understand when such events occur in the network.

Another relevant investigation is to consider other interactions within the blogosphere. Indeed, the blog network is composed of multiple interaction networks. In addition to blogs and post networks, there are also blogroll and comment networks. The blog activity may be seen as the result of the sum of all networks which have different dynamics. The aim is to have a more complete view of interactions which compose the blogs dynamics. Two problems may be addressed: first how can we aggregate all information and study it as one single network? Second, what is the impact of each network on the others? In particular,

how comments could lead to new interaction in post networks?

I have studied in this thesis diffusion phenomena at different scales of a community structure. One perspective is to consider other network properties, for example characterize nodes activity by considering the local clustering coefficient, centrality measure or time activity in the system. The approach based on community structure will highlight nodes (or groups of nodes) with common characteristics and will help determine different patterns.

Finally, a key perspective is to design models for diffusion phenomena which take into account not only basic topological features (for example the size of the diffusion) but also community information. For example such models could generate a diffusion with a determined community distance distribution.

List of figures

2.1	Cascade sample	23
3.1	Blog sample	30
3.2	Authority and hub	31
3.3	Three continents	31
3.4	Blog and post network	34
3.5	Number of posts a) per day during 4 months, b) for each day of the week .	37
3.6	a) c)In-degree and out-degree distributions and b) d) cumulative distribution in the blogs network.	38
3.7	in-degree and out-degree cumulative density distribution per continent in the blog network	39
3.8	blog in-out correlation	39
3.9	Inter-blog activity	40
3.10	a) Evolution of the number of posts and citations. b) Citation and post correlation.	41
3.11	Biases causes	42
3.12	Biases correction	43
3.13	Evolution of post popularity a) overall evolution b) according to post com- munity.	43

3.14	Post classification - for both plots, the value on the x-axis (denoted by X) represents the number of days during which the posts have been cited. a) The value on the y-axis (denoted Y) corresponds to the number of distinct citation periods (e.g. for a post cited on days 1, 2, 3, 5, 7, 8, $X=6$ as it is cited on 6 days, and $Y=3$ as these citations occurred during 3 periods of time: $\{1, 2, 3\}$, $\{5\}$ and $\{7, 8\}$). b) The value on the y-axis represents the sum of all citations of a post for the corresponding number of days. So for example a dot with coordinates $(3; 34)$ represents a post which has been cited on 3 distinct days and 34 times in total.	44
3.15	a) Distribution of citations of posts cited only one day. A dot with coordinates (x, y) means that there are y posts which have been cited x times on their single citation day (as $X=1$). b) Cumulative number of citations per total number of citation days. This plot represents the cumulative probability density function (CDF) of the number of citations per number of days for X greater than 3.	46
4.1	Hierarchical community structure example	52
4.2	Community distance example	55
4.3	Blog network community structure	57
4.4	Delta vs modularity at region layer	59
4.5	Number of link distances by region	61
4.6	Fraction of in and out link distances by region	61
4.7	Fraction of in and out link distances by territory in individuality continent.	62
4.8	Average in and out links distance correlation at region layer. Green square=society, red circle=leisure, blue triangle=individuality.	63
4.9	Average in and out links distance correlation in <i>sport</i> community	64
4.10	Average in and out links distance correlation in political communities	65
4.11	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 $d = 4$	66

4.12	Automatic community structure	68
4.13	Fraction of in and out link distances community at level 1	70
4.14	Fraction of in and out link distances community at level 2	71
5.1	Cascade samples.	77
5.2	Extracted Cascade samples.	78
5.3	Largest cascades in our dataset. All cascades has at least 30 nodes.	81
5.4	Cumulative distribution of the number of nodes (in red) and links (in green).	82
5.5	N_{lvl} cumulative density function	83
5.6	Cumulative distribution of cascade density.	84
5.7	Cumulative distribution of cascade degree assortativity.	85
5.8	Cumulative distribution of average community distance.	86
5.9	Impact of community distance on cascade duration.	86
5.10	Impact of community distance on cascade size.	87
5.11	Number of level and size cascade impact.	87
5.12	Impact of community origin on cascade time duration.	88
5.13	Impact of community source on cascade levels.	89
5.14	Impact of community of origin on average community distance.	90
5.15	An example of node impact.	90
5.16	Impact of node community at Continent level.	91
5.17	Impact of node community at Region level.	92
5.18	Impact of nodes community at Territory level.	93

List of tables

3.1	Most active blogs for each continent	32
3.2	Activity by continent	33
4.1	Probability to link blogs from the same community at various levels.	56
4.2	Probability to link blogs from the same region in each continent and associated modularity	57
4.3	Probability to link blogs from the same <i>Territory</i> for each <i>Region</i>	58
4.4	Distribution of community distances in G	60
4.5	Probability to link blogs from the same community	69
4.6	Distribution of community distances in G	69
5.1	cascade shapes ordered by frequency	80
5.2	Cascade properties	82
5.3	Table of symbols	90

References

- [AA05] Eytan Adar and Lada A. Adamic. Tracking information epidemics in blogspace. In *Web Intelligence*, Compiegne, France, September 2005.
- [ACKM09] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling. *Journal of the ACM*, 56(4):1–28, 2009.
- [AG05a] L.A. Adamic and N. Glance. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.
- [AG05b] Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.
- [AJB00] R Albert, H Jeong, and A-L Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–82, 2000.
- [ALM09a] Frédéric Aidouni, Matthieu Latapy, and Clémence Magnien. Ten weeks in the life of an edonkey server. In *Proceedings of HotP2P'09*, 2009.
- [ALM09b] Oussama Allali, Matthieu Latapy, and Clémence Magnien. Measurement of edonkey activity with distributed honeypots. *CoRR*, abs/0904.3215, 2009. informal publication.
- [AM92] R. M. Anderson and R. M. May. *Infectious Diseases of Humans Dynamics and Control*. Oxford University Press, 1992.
- [AZAL04] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In *WWW2004*, 2004.

-
- [BA99a] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science (New York, N.Y.)*, 286(5439):509–512, October 1999.
- [BA99b] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.
- [Bai75] Norman Bailey. *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London, 1975.
- [Bar54] J. A. Barnes. Class and Committees in a Norwegian Island Parish. *Human Relations*, 7(1):39–58, February 1954.
- [BBV08] Alain Barrat, Marc Barthlemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, New York, NY, USA, 2008.
- [BGH⁺10] D. Balcan, B. Gonçalves, H. Hu, J.J. Ramasco, V. Colizza, and A. Vespignani. Modeling the spatial spread of infectious diseases: The GLObal Epidemic and Mobility computational model. *Journal of Computational Science*, 1:132–145, 2010.
- [BGLL08] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, jul 2008.
- [BGLLf08] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment*, page P10008, 2008.
- [BGTL11] Abdelhamid Salah Brahim, Bénédicte Le Grand, Lionel Tabourier, and Matthieu Latapy. Citations among blogs in a hierarchy of communities: Method and case study. *Journal of Computational Science*, 2(3):247 – 252, 2011.
- [BHKL06] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 44–54, New York, NY, USA, 2006. ACM.

-
- [BHW92] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *The Journal of Political Economy*, 100(5):992–1026, 1992.
- [BKM⁺00] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web: experiments and models. In *Proceedings of the Ninth International World-Wide Web Conference (WWW9, Amsterdam, May 15 - 19, 2000 - Best Paper)*. Foretec Seminars, Inc. (of CD-ROM), Reston, VA, 2000.
- [BM] Lamia Benamara and Clémence Magnien. Estimating properties in dynamic systems: the case of churn in p2p networks.
- [Bot57] Elizabeth Bott. *Family and Social Network*. Tavistock, London, 1957.
- [CFSV01] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento. An improved algorithm for matching large graphs. In *In: 3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition, Cuen*, pages 149–159, 2001.
- [CMAG08] Meeyoung Cha, Alan Mislove, Ben Adams, and Krishna P. Gummadi. Characterizing social cascades in flickr. In *Proceedings of the first workshop on Online social networks, WOSP '08*, pages 13–18, New York, NY, USA, 2008. ACM.
- [CNM04] Aaron Clauset, M. E. J. Newman, , and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):1– 6, 2004.
- [CPH09] M. Cha, J.A.N. Pérez, and H. Haddadi. Flash Floods and Ripples: The Spread of Media Content through the Blogosphere. In *Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media*, 2009.
- [CR09] 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.
- [CSN07] Aaron Clauset, Cosma R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Reviews*, June 2007.

- [CZF04] Deepayan Chakrabarti, Yiping Zhan, and Christos Faloutsos. R-mat: A recursive model for graph mining. In *SDM*, 2004.
- [DNMKKL09] Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. How opinions are received by online communities: A case study on Amazon.com helpfulness votes. In *Proceedings of WWW*, pages 141–150, 2009.
- [DR01] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 57–66, New York, NY, USA, 2001. ACM.
- [DRFC05] Benoit Donnet, Philippe Raoult, Timur Friedman, and Mark Crovella. Efficient algorithms for large-scale topology discovery. In *Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, SIGMETRICS '05, pages 327–338, New York, NY, USA, 2005. ACM.
- [EK] J Berry E Keller. *One American in ten tells the other nine how to vote, where to eat, and what to buy. They are the influentials.*
- [fac] facebook: <http://www.facebook.com/>.
- [FFF99] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, SIGCOMM '99, pages 251–262, New York, NY, USA, 1999. ACM.
- [FN93] Jonathan Frenzen and Kent Nakamoto. Structure, Cooperation, and the Flow of Market Information. *Journal of Consumer Research*, 20:360–null, 1993.
- [For09] Santo Fortunato. Community detection in graphs. *Physics Reports*, 2009.
- [GGLNT04] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *WWW '04*, pages 491–501. ACM, 2004.

-
- [GL05] J. L. Guillaume and M. Latapy. Complex Network Metrology. *Complex Systems*, 2005.
- [Gla02] Malcolm Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. Back Bay Books, January 2002.
- [GLM01] J. Goldenberg, B. Libai, and E. Muller. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters*, pages 211–223, August 2001.
- [GN02] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.
- [Gra78] Mark Granovetter. Threshold Models of Collective Behavior. *The American Journal of Sociology*, 83(6):1420–1443, 1978.
- [HPV] Shawndra Hill, Foster Provost, and Chris Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, page 2006.
- [IKMW07] Nicole Immorlica, Jon Kleinberg, Mohammad Mahdian, and Tom Wexler. The role of compatibility in the diffusion of technologies through social networks. In *Proceedings of the 8th ACM conference on Electronic commerce, EC '07*, pages 75–83, New York, NY, USA, 2007. ACM.
- [KKR⁺99] Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S. Tomkins. The Web as a Graph: Measurements, Models, and Methods. pages 1+. 1999.
- [KKT03] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, New York, NY, USA, 2003. ACM.
- [KNRT05] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the bursty evolution of blogspace. *World Wide Web*, 8:159–178, June 2005.
- [KP05] E. Katz and F. Paul. *Personal Influence: The Part Played by People in the Flow of Mass Communications*. Transaction Publishers, 2005.

-
- [Lat07] Matthieu Latapy. Grands graphes de terrain – mesure et métrologie, analyse, modélisation, algorithmique. *Mémoire d’habilitation à diriger les recherches*, UPMC, 2007. <http://www-rp.lip6.fr/~latapy/HDR/>.
- [LKF05] Jurij Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In *KDD*, pages 177–187, 2005.
- [LKF07] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution : Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (ACM TKDD)*, 1(1), 2007.
- [LKG⁺07] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD ’07*, pages 420–429, New York, NY, USA, 2007. ACM.
- [LMF⁺07] Jure Leskovec, Mary Mcglohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Cascading behavior in large blog graphs. In *Proceedings of 7th SIAM International Conference on Data Mining*, 2007.
- [LMO08] Matthieu Latapy, Clémence Magnien, and Frédéric Ouédraogo. A radar for the internet. In *ICDM Workshops’08*, pages 901–908, 2008.
- [LSK06] Jure Leskovec, Ajit Singh, and Jon Kleinberg. Patterns of Influence in a Recommendation Network. In *Advances in Knowledge Discovery and Data Mining*, volume 3918 of *Lecture Notes in Computer Science*, chapter 44, pages 380–389. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2006.
- [Mag10] Clémence Magnien. Intégrer mesure, métrologie et analyse pour l’étude des graphes de terrain dynamiques. *Mémoire d’habilitation à diriger les recherches*, UPMC, 2010.
- [McK81] Brendan D. McKay. Practical graph isomorphism, 1981.
- [Mil67] S. Milgram. The small world problem. 1967.
- [mil69] An Experimental Study of the Small World Problem. *Sociometry*, 32(4):425–443, 1969.

-
- [Mit03] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2), 2003.
- [Mon01] Alan L. Montgomery. Applying quantitative marketing techniques to the internet. *Interfaces*, pages 90–108, 2001.
- [MT10] M. Mitrović and B. Tadić. Bloggers behavior and emergent communities in Blog space. *The European Physical Journal B-Condensed Matter and Complex Systems*, 73(2):293–301, 2010.
- [Nad57] SF Nadel. *The Theory of Social Structure*. Cohen and West, London, 1957.
- [NBW06] Mark E. J. Newman, Albert L. Barabási, and Duncan J. Watts, editors. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.
- [New02] M. E. J. Newman. Assortative Mixing in Networks. *Physical Review Letters*, 89(20):208701+, October 2002.
- [New03] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066133, Sep 2003.
- [New06] M. E. J. Newman. Power laws, Pareto distributions and Zipf’s law. *Contemporary Physics*, 46(5):323–351, May 2006.
- [New10] Mark E. J. Newman. *Networks: an introduction*. Oxford University Press, Tavistock, London, 2010.
- [NFB02] M. E. J. Newman, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101+, 2002.
- [NG04] M E J Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, (2):26113, 2004.
- [PGF02] Christopher R. Palmer, Phillip B. Gibbons, and Christos Faloutsos. Anf: A fast and scalable tool for data mining in massive graphs. In *INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING*, pages 81–90. ACM, 2002.
- [PN03] Juyong Park and M. E. J. Newman. The origin of degree correlations in the internet and other networks. *PHYS.REV.E*, 68:026112, 2003.

-
- [PSV01] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200–3203, 2001.
- [QF10] Wang Qinna and Eric Fleury. Uncovering Overlapping Community Structure. In *Workshop on Complex Networks*, Rio, Brésil, October 2010.
- [RB02] Albert Reka and Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, June 2002.
- [Red98] S. Redner. How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B - Condensed Matter and Complex Systems*, 4(2):131–134, August 1998.
- [Rog62] E.M. Rogers. *Diffusion of Innovations*. The Free Press, New York, 1962.
- [SBF⁺08] A. Scherrer, P. Borgnat, E. Fleury, J.L. Guillaume, and C. Robardet. Description and simulation of dynamic mobility networks. *Computer Networks*, 52(15):2842–2858, 2008.
- [SBLGL10] Abdelhamid Salah Brahim, Benedicte Le Grand, and Matthieu Latapy. Some Insight on Dynamics of Posts and Citations in Different Blog Communities. In *Proceeding of the First International Workshop on Social Networks*. IEEE, 2010.
- [SR06] Daniel Stutzbach and Reza Rejaie. Understanding churn in peer-to-peer networks. In *IMC '06: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 189–202, New York, NY, USA, 2006. ACM.
- [twi] twitter: <http://twitter.com/>.
- [Ull76] J. R. Ullmann. An algorithm for subgraph isomorphism. *J. ACM*, 23:31–42, January 1976.
- [VOD⁺06] Alexei Vazquez, Joao Gama Oliveira, Zoltan Dezso, Kwang-Il Goh, Imre Kondor, and Albert-Laszlo Barabasi. Modeling bursts and heavy tails in human dynamics, 2006.
- [WCP⁺02] Mengzhi Wang, Ngai Hang Chan, Spiros Papadimitriou, Christos Faloutsos, and Tara Madhyastha. Data mining meets performance evaluation: Fast algorithms for modeling bursty traffic. *Data Engineering, International Conference on*, 0:0507, 2002.

-
- [WD07] Duncan J. Watts and Peter S. Dodds. Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research*, 34(4):441–458, December 2007.
- [WF94] S. Wasserman and K. Faust. *Social network analysis*. Cambridge University Press, 1994.
- [WS98a] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684), jun 1998.
- [WS98b] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.
- [YYA10] Sune Lehmann Yong-Yeol Ahn, James P. Bagrow. Link communities reveal multiscale complexity in networks. *Nature*, 2010.

Abstract

Real-world networks are also called *interaction networks* as nodes have relationships with one another and different interactions occur between them. Understanding how *interaction networks* are structured and which events may occur within them, are indeed key questions of the field. Interactions within the network induce the creation of *links* between nodes. Through links, members may create new relations, exchange information or influence each other. This phenomenon is called *diffusion* process.

Until now diffusion phenomena have been mostly considered at a macroscopic scale i.e. by studying all nodes of the network as a whole. We give a complementary way to analyse the network interaction by considering the problem at different scales. To that purpose, we use the *community structure* of the network.

We propose in this thesis an analysis of diffusion phenomena in a *French blog network* with regard to its community structure. Methodologies we introduce are general and may be used in other contexts of real-world networks. We have studied diffusion phenomena in different ways. First, by studying how a node propagates its content towards its neighbours. Second, by investigating how a node gets an information from its neighbours. Third, by considering a diffusion *cascade* where an information spreads from one node to the rest of the network through a succession of diffusion events. Considering the *community structure* of the network in link analysis is relevant as we have been able to characterize diffusion between nodes (and also communities) and to identify interaction behaviour patterns.

Key Words:

complex networks, interaction links, community structure, homophily, blog network, statistic analysis, cascades

Diffusion d'information dans les réseaux de blog et structure en communautés

Résumé.

On peut modéliser de nombreux objets issus du monde réel par des graphes. Ces objets sont issus de contextes très différents (ex. réseaux informatiques, sociaux ou biologiques), cependant ils se ressemblent au sens de certaines propriétés statistiques. On les désigne sous le terme général de *graphes de terrain* (*complex networks* en anglais) ou grands graphes d'interaction.

L'*analyse* des graphes de terrain est probablement le plus grand champ de recherche du domaine et l'étude des phénomènes de diffusion constitue un des axes importants dans la compréhension de ces objets. Beaucoup de précédentes études ont été menées sur la diffusion avec une approche théorique mais avec l'apparition de données issues du monde réel de plus en plus riches, une approche empirique de l'analyse de ces réseaux est apparue comme une nécessité.

La diffusion peut être de différentes natures : diffusion d'information, d'idées ou d'opinion. Cette diffusion est vue dans la plupart des travaux comme le résultat de l'interaction entre les éléments du réseau (i.e. les nœuds du graphe). En complément de cette vision, nous considérons dans cette thèse que la diffusion, en plus de se produire entre les nœuds, est aussi le résultat de l'interaction entre des groupes de nœuds, appelés *communautés*, qui ont des propriétés en commun. On dit que le réseau possède une *structure en communautés*. Cette approche ouvre de nouvelles perspectives pour la compréhension et la caractérisation des graphes de terrain.

L'objectif de cette thèse est d'étudier les phénomènes de diffusion de manière empirique non seulement à l'échelle des nœuds mais à différents niveaux de la structure en communautés. À l'aide d'une approche statistique, nous proposons un ensemble de méthodes et de métriques pour aborder la diffusion sous un nouvel angle et aller plus loin dans la caractérisation de ces phénomènes. Nous nous proposons d'étudier les liens de diffusion au sein d'un réseau de blogs francophones. Nous montrons en premier lieu l'impact des communautés sur la popularité des blogs et distinguons des classes de comportement. Cela nous conduit à investiguer les interactions entre les communautés. Pour ce faire, nous définissons deux mesures : la *distance communautaire* et l'*Homophilie*. En dernier lieu, nous étudions la diffusion de proche en proche dans le graphe, caractérisée par des *cascades de diffusion*. Nous montrons que notre approche permet de détecter et d'interpréter les différents comportements de diffusion et de faire le lien entre les propriétés topologiques, temporelles et communautaires.

Mots clés. graphes de terrain, communautés, diffusion, réseau de blogs, analyse statistique, cascade de diffusion.
