Parallel Processing Letters


# DIFFUSION CASCADES: SPREADING PHENOMENA IN BLOG NETWORK COMMUNITIES

ABDELHAMID SALAH BRAHIM
BÉNÉDICTE LE GRAND
MATTHIEU LATAPY
*LIP6 – CNRS and UPMC (Université Pierre et Marie Curie)*
*4 place Jussieu 75252 Paris cedex 05, France.*
*email: First-name.Last-name@lip6.fr*

Received September 2011
Revised November 2011
Communicated by Guest Editors

ABSTRACT

A diffusion cascade occurs when information spreads from one node to the rest of the network through a succession of diffusion events. So far diffusion phenomena have been mostly considered at a macroscopic scale i.e. by studying all nodes of the network. We give a complementary way to analyse network interactions by considering the problem at different scales. To that purpose, we use the *community structure* of the network to characterize diffusion between nodes (and between communities) and to identify interactions behaviour patterns.

*Keywords*: Cascade, diffusion phenomena, blog network, communities

## 1. Introduction and related work

Diffusion phenomena occur in a network when an action, information or idea becomes adopted due to the influence of neighbors in the network [5]. The results of studies on diffusion are used in many applications. For example, viral marketing exploits existing social networks and encourages customers to share product information with their friends [8]. The "cascade" diffusion model allows to investigate which individual dynamics lead to global spreading phenomena. Cascades have been theoretically analyzed in random graphs using a threshold model [14]. However, only few empirical studies of the topological patterns of cascades have been done [2, 6]. Previous work aimed at characterizing cascade topological properties. In this paper, we give a complementary way to analyse network interactions by considering the problem at different scales. To that purpose, we use the *community structure* of the network [11].

Indeed, it has been observed that nodes with common features tend to interact preferentially with each other [4, 13]. These groups of nodes are called *communi-*

*ties*. Among the various definitions of "community" which exist in the literature, we use the following one: *"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 levels (groups of nodes). In addition to topological properties, we investigate community information to understand diffusion cascades. We apply this approach to a French blog network which has a topical community structure obtained manually be professional blog analysts. In a first step, we aim at characterizing cascades using several topological, temporal and community metrics. Indeed, in addition to classical graph metrics we define a *community distance* measure to determine if a link relates nodes from *close* or *distant* communities. Subsequently, we investigate how cascades spread through communities, and show that the community of the cascade origin has an impact on cascade properties and especially on average community distance for which diffusion behaviours differ noticeably.
The paper is structured as follows: In Section 2, we introduce the community structure formalism and define the *community distance*. We also explain how cascades are computed and present first statistics. In Section 3, we study at a macroscopic level the different properties which characterize a cascade and investigate their impact on each other. Finally, in Section 4 the impact of the community of the cascade origin is observed at the node level (i.e. at microscopic level).

## 2. Definitions and cascades computation

### 2.1. *Data corpus*

A blog is composed of a set of posts written at a given time with a body text and references to other pages on the web (for example pictures, videos and websites). The text may contain references to previous posts, from the same blog (auto-citation) or from another blog, by quoting the corresponding $URL_s$, which are called *citation links*. Citation links are very important as they represent a diffusion of information in the network. 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 $Pb$. In terms of information spreading, we can say that $Pa$ has 'adopted' $Pb$'s content or that $Pb$'s content has been spread towards $Pa$.

The corpus analyzed in this paper was obtained by daily crawls of 10.309 blogs during five months (from February $1^{st}$ to July $1^{st}$ 2010). 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 active blogs which provide rich information for activity and dynamics study. The dataset is composed of 10.309 blogs 848.026 posts and 1.079.195 citation links.

We first 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.

### 2.2. *Hierarchical community structure*

The methodology we propose to characterize diffusion cascades requires a hierarchical community structure of the blog network. This structure may be obtained in two different ways. First, by executing an automatic community detection algorithm. Second, by classifying *manually* each blog into hierarchical classes. 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 *automatic* classification [1].

In this case the classification into *communities* has been done manually by professional blog analysts according to blogs topics. The hierarchical community structure we consider for this dataset comprises 5 levels: level 0 corresponding to a single community (with all blogs), level 1 with 3 communities called continents (*Leisure*, *Individuality*, *Society*), level 2 with 16 regions, level 3 with 96 territories and finally level 4 with the 10.309 individual blogs. For instance, the blog *http://www.sailr.com* belongs to the *leisure continent*, the *sport region* and the *sailing territory*.

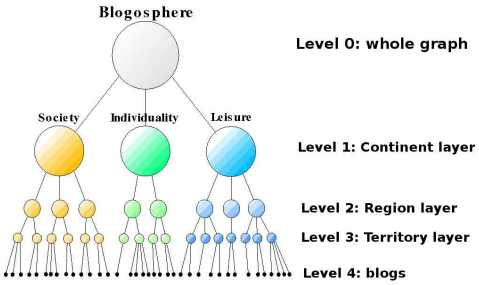


Fig. 1. Blog network community structure

In the following, we explain the formalism we will use in the paper. Let a graph $G = (V, E)$, with $V$ a set of nodes and $E$ a set of edges. Our methodology requires a community structure such that each node of $V$ belongs to exactly one community at each level of the tree (more general hierarchical community structures will be considered in the future to allow overlapping communities).

**Definition 2.1. Hierarchical Community Structure**
Given a community partition $P = \{C_1, C_2, ..., C_l\}$ of $V$, a sub-partition $P' = \{C'_1, C'_2, ..., C'_m\}$ 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 $P_k \sqsubseteq P_{k-1} ... \sqsubseteq P_2 \sqsubseteq P_1 \sqsubseteq P_0$ with $P_0 = V$, i.e. $P_0$ contains only one community which

is the whole set of nodes and $P_k = \{\{v\}, v \in V\}$, i.e. $P_k$ contains $n$ communities containing only one node.

In order to characterise whether links relate nodes from *close* or *distant* communities, we introduce the notion of *community distance* $d(u, v)$ between two nodes $u$ and $v$.

**Definition 2.2. 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 define the community distance of the spreading link $(u, v)$ as:

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

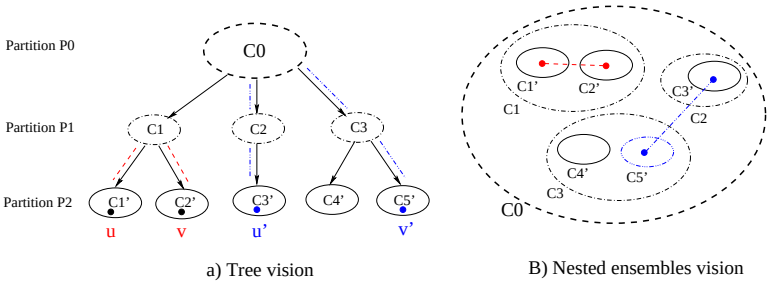


Fig. 2. Community distance example

If we consider nodes $u, v, u', v'$ in Figure 2.(a) (with $u \in C'_1$, $v \in C'_2$, $u' \in C'_3$ and $v' \in C'_5$), $d(u, v) = 1$ and $d(u', v') = 2$. The link between $u'$ and $v'$ therefore connects two nodes from more distant communities than the link between $u$ and $v$. Figure 2.(b) provides another representation of the same community structure.

### 2.3. *Cascades computation*

Cascades are subgraphs of the post network, where nodes correspond to posts and edges to citation links. In order to compute post cascades, we start by posts which do not cite any other post i.e. with no outgoing link; each of them represents the beginning of a cascade called "origin". Consider such a post; if it is cited by one or several posts, the process carries on recursively: posts which have cited this citing post are looked for and so on. Each post can belong to several cascades (e.g. post $F$ in Figure 3), represented as Directed Acyclic Graphs. In Figure 3, the cascades origins are $A$, $B$ and $C$, respectively. Information is therefore spread from the origin to the leaves (posts with no incoming link).

We focus on information diffusion among different blogs, therefore links between two posts from a same blog were removed. Indeed, self-citations can make cascades

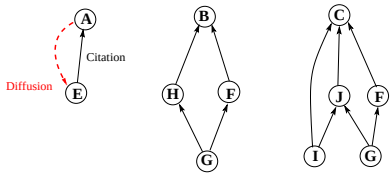


Fig. 3. Cascade samples.

longer but do not represent a diffusion process from one blog to another. This hypothesis has an impact on cascades sizes however ignoring blog self-citations removes some biase and leads to more relevant cascades. The total number of cascades is $10,659$.

## 3. Macroscopic analysis: cascades features

### 3.1. *Cascade shapes*

This section aims at studying cascades shapes, in order to know what types of cascades appear frequently in the blog network. Do they look like trees, stars or chains? We computed all cascades and we used an isomorphism algorithm to determine whether a cascade was identical to another.

*Isomorphic cascades*

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

We have used the VF2 isomorphism algorithm based on a depth-first strategy [3]. 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. There are in total $10,659$ cascades and 641 isomorphic shapes. The most common post network cascade shapes are given in Table 1, where the red post is the cascade origin. 65% of cascades are composed of two nodes; the second most frequent shape represents 10% with three nodes. We may observe that cascades tend to be *stars* (e.g. cascade 19) rather that *chains* (e.g. cascade 30). This is more obvious if we compare the shape frequency of cascades which contain the same number of nodes. For example, if we consider cascades 2 and 30 which contain 3 nodes and 19 and 4 which contain 5 nodes the *star* shapes are more frequent.

Now we focus on the largest cascades represented in Figure 4. In terms of frequency, all those shapes appear only once as they are very complex. Unlike most frequent shapes described in Table 1, large cascades seem to have *tree-like* shapes. When looking at cascades topology we may notice that they are very complex and we may distinguish nodes with a more important role in the cascade spreading phenomena. This raises the following questions: which properties have an impact on cascade topological characteristics? What makes a cascade longer or bigger? Those questions are addressed in the next section.

Table 1. cascade shapes ordered by frequency.

| ID | shape | # nodes | # links | frequency |
|---|---|---|---|---|
| 1 | | 2 | 1 | 6992 |
| 2 | | 3 | 2 | 1173 |
| 81 | | 4 | 3 | 397 |
| 30 | | 3 | 2 | 370 |
| 19 | | 5 | 4 | 182 |
| 29 | | 4 | 3 | 134 |
| 88 | | 6 | 5 | 83 |
| 4 | | 5 | 4 | 56 |
| 101 | | 3 | 3 | 52 |
| 702 | | 4 | 3 | 46 |
| 418 | | 7 | 6 | 33 |
| 107 | | 5 | 4 | 30 |
| 333 | | 4 | 3 | 30 |
| 122 | | 5 | 4 | 29 |

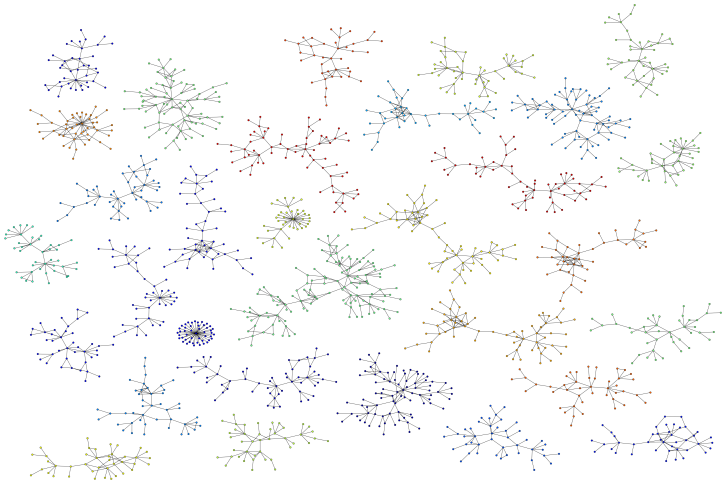

Fig. 4. Largest cascades. All cascades have a number of node >= 40.

### 3.2. *Cascade topological, temporal and community properties*

To understand cascades we need to characterise them precisely with regard to all their features detailed below (see Table 2). In this section we consider properties

Table 2. Cascade properties.

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

at the cascade scale rather than at the node scale. We classify cascade properties into three categories: topological, temporal, and community-related.

The topological features regroup classical graph measures. The temporal features considered in this paper are the total duration of the cascade ($T$) and the timestamp of cascade origin ($T_s$). A first step consists in considering each property and studying its distribution using cumulative density distributions (also called PDF probability density functions).

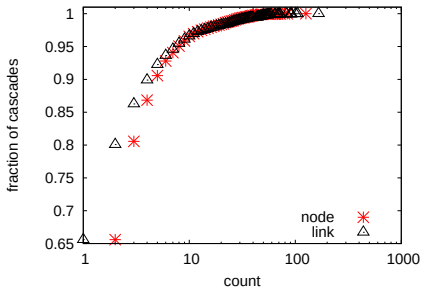


Fig. 5. Cumulative distribution of the number of nodes (in red) and links (in green).

In Figure 5, the cumulative distribution of the number of nodes ($N_n$) and links ($N_l$) per cascade shows a similar heterogeneous distribution (in Figure 5, a point $(x, y)$ with $x = 10$ and $y = 95\%$ means that there are 95% cascades which contain less than 10 nodes. Moreover, 65% of cascades (6992) are composed of only two nodes, as most post are only cited 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. Indeed, the numbers of edges and nodes increase similarly which suggests that the average degree in the cascade remains constant as the cascade grows.

In addition, we analyse the cascade depths (i.e. their number of levels) noted $N_{lvl}$. This depth corresponds to the length of the maximum path between the cascade

origin and the "leaves". Figure 6 represents the $N_{lvl}$ cumulative density function. A point $(x, y)$ with $x = 4$ and $y = 98\%$ means that 98% of cascades contain at most 4 levels. We also observe that almost 84% of cascades have $N_{lvl} = 1$.

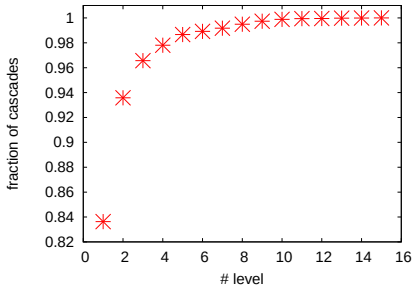


Fig. 6. $N_{lvl}$ cumulative density function

In the following, we study the distribution of cascade density (noted $\delta$). 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 $\delta = \frac{m}{n.(n-1)}$. Density is a fraction that goes from a minimum of 0 if no edge is present (which is not possible in our case because a cascade has at least one edge) to 1 if all edges are present.
Here we consider only cascades with $N_n > 2$. Figure 7 shows the density distribution; we may observe that it is heterogeneous and that cascade density is mostly comprised between 0.4 and 0.7. These density values may be considered as high, but this result is not surprising because most cascades are small. In Figure 7, we show the correlation between cascade density and size, and we observe that density is inversely proportional to size.

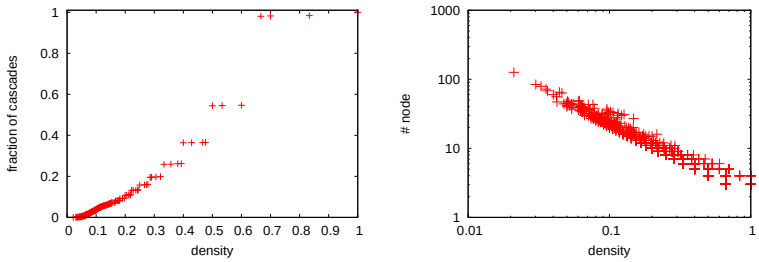


Fig. 7. Cumulative distribution of cascade density, left: $\delta$ cumulative density function, right: density VS number of node.

Another typical feature of real world networks is the tendency of nodes of a given degree to be connected with other nodes of similar degree. This property may be measured by *degree assortativity* noted $r$ [9, 12]. 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 if $r = -1$ the network is completely disassortative. Newman [9] 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 8 represents the cumulative distribution of cascade degree assortativity (only cascades with $N_n > 2$ are considered). We observe that 95% of cascades have a negative degree assortativity. It means that cascades in the blog network tend to have a disassortative behaviour. In addition, the degree assortativity measure gives an indication of cascade shapes [10]. Indeed a disassortative graph has high-degree nodes which tend to connect low-degree ones, therefore creating a star-like structure.

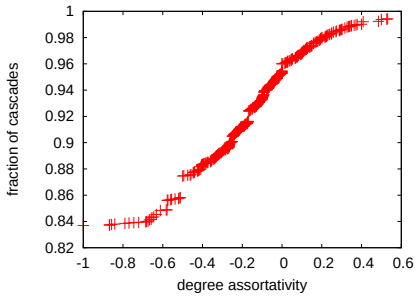


Fig. 8. Cumulative distribution of cascade degree assortativity.

In addition to topological properties, we used the community distance defined in Section 2.2 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 calculate the community distance for each link. Afterwards, we calculate 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 in Figure 9 shows that the average community distance has an heterogeneous distribution.

### 3.3. *Correlation between cascade features*

In the previous section the different cascades properties have been studied independently. Now we go further to determine how the topological, temporal and community features may impact one another. We first study the impact of the average community distance on the duration of the cascade. The intuition is that if a cascade goes through links which have a high community distance the cascade duration may be longer.

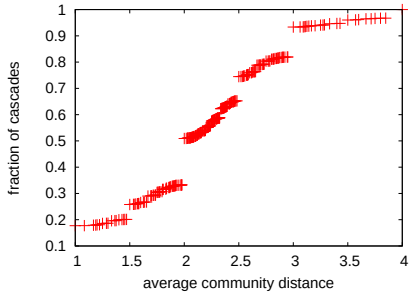


Fig. 9. Cumulative distribution of average community distance.

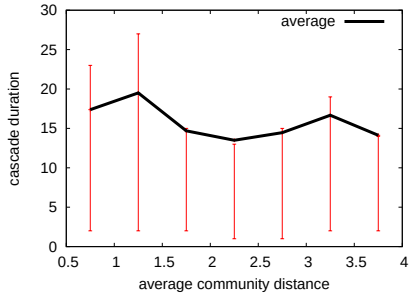


Fig. 10. Impact of community distance on cascade duration.

In figure 10 we represent the average community distance in relation to cascade duration. For each 0.5 $A_{cm}$ interval we calculate the 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 rather small or on the contrary rather high. As an example, the average difference between cascades with $A_{cm}$ comprised between 1 and 1.5, and between 2 and 2.5 is about 5 days. The interpretation is that cascades have in average a longer duration when citations are made between topically close posts (i.e. the same territory) or on the contrary between semantically distant posts.

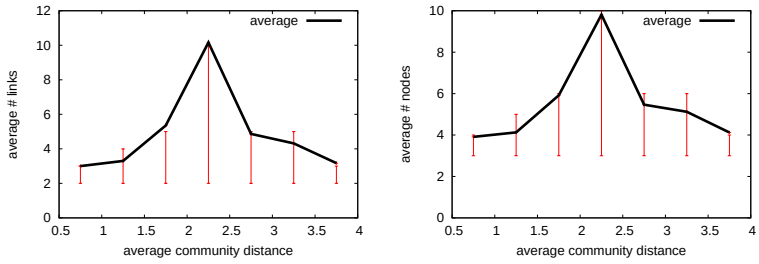


Fig. 11. Impact of community distance on cascade size.

After studying the impact of communities on cascade duration, we now observe their impact on cascade size (see Figure 11). This size (in terms of number of nodes and links) tends to be maximum for cascades with an average community distance comprised between 2 and 2.5. Citation therefore have to occur mostly between blogs from the same region to ensure a large diffusion. However, range of community distance values which maximizes cascade size corresponds in figure 10 to shortest cascades (in therms of duration). A conclusion is that community distance values may not simultaneously maximize cascade size and duration.

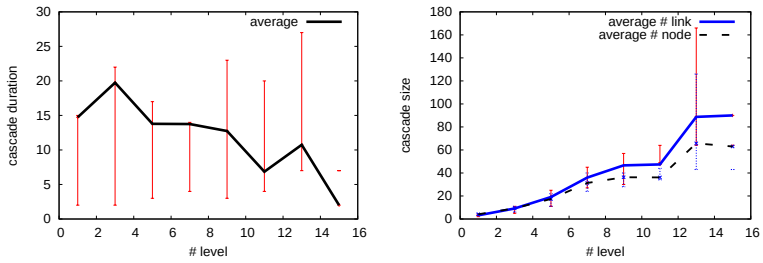


Fig. 12. Number of level and size cascade impact. Left:$N_{lvl}$ VS $T$, Right: Cascade size VS $T$.

Next, we investigate whether the number of levels $N_{lvl}$, which is a topological metric, is correlated to the cascade duration. One may think that the higher the number of levels, the longer the cascade. The result is shown on the left plot of Figure 12 where the cascade duration increases for values of $N_{lvl}$ comprised between 2 and 3 with an average of 5 days. After that, the cascade duration decreases. This means that cascades with a longer duration contain in average a small number of levels (2 and 3) rather than a high number as one may expect. We can also conclude that spreading speed is higher when the cascade has more levels (and tends to have a *chain* shape). On the other hand, the cascade size has a different impact (see right plot in Figure 12 where the average cascade duration increases proportionally to the size for both $N_l$ and $N_n$).

## 4. Microscopic analysis: impact of individual nodes on cascades

In the previous section we have studied cascades properties at a macroscopic level. Now, we go deeper and investigate how a given node impacts the rest of the cascade. The originality of this work is that we include a community aspect, at different layers. As explained earlier, the community structure has been built manually by professional blog analysts according to blog topics. It is composed of three hierarchical levels: *continent*, *region* and *territory* (from the most general to the most specific). In this section, we study the impact of individual nodes' communities - origin and intermediate nodes - on cascades.

### 4.1. *Impact of cascade origin*

We compare cascade properties depending on the community they start to spread from. In this section we focus on *Continent* layer (which corresponds to the level 1 in the hierarchical community structure), with the three communities: *Leisure*, *Individuality* and *Society*.

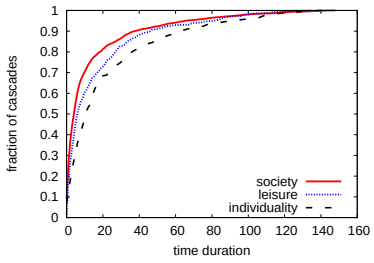


Fig. 13. Impact of community origin on cascade time duration.

We first study cascade duration, by considering the cumulative density functions (Figure 13). We may see that cascades which start from *Society* community have a shorter duration than *Leisure* cascades and that nodes from *Individuality* community tend to initiate cascades of longer duration.

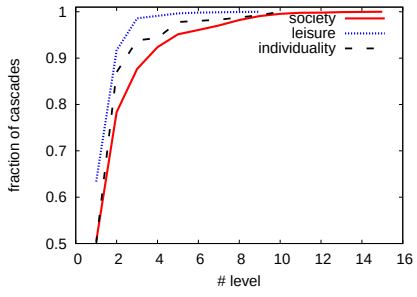


Fig. 14. Impact of community source on cascade levels.

With regard to the number of levels (see Figure 14), 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 higher numbers of levels and nodes. The same analysis may be done for communities of Region and Territory levels.

Next, we study the average community distance of each cascade link. The first observation is that 70% of cascades which start in *Leisure* continent have a commu-

nity distance $A_{cm} = 1$. The *Leisure* cascades tend to be smaller than those starting from other continents and also have a small community distance. On the other hand, *Society* cascades have a higher community distance and are larger.

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

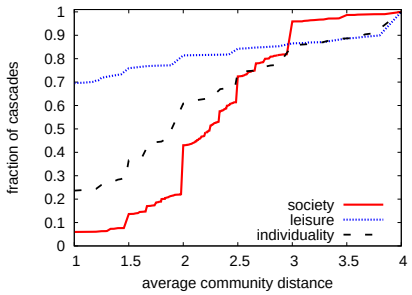


Fig. 15. Impact of community origin on average community distance

### 4.2. *Impact of intermediate blogs*

Table 3. Table of symbols.

| Notation | Description |
|---|---|
| $T_u$ | Time when 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 in the cascade 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 $u$ |

Now, we investigate the impact of all nodes on cascades (see notation in Table 3) and not only the cascade origin.

We give an example in Figure 16: 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 $lvl\_max_u = 2$. These three

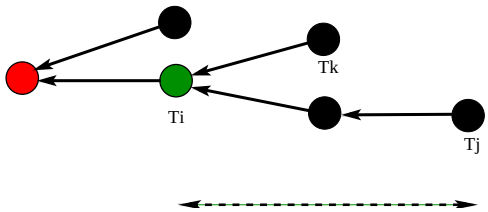


Fig. 16. An example of node impact.

metrics regroup temporal and topological properties.

We study each property for three communities in each level (*continent*, *region* and *territory*).

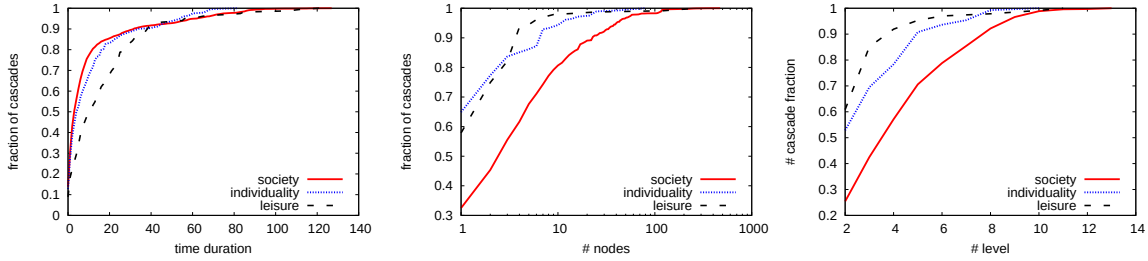


Fig. 17. Impact of node community at Continent level. Left: $T\_max_u$, center: $N\_node\_max_u$, right: $lvl\_max_u$.

We start by the individual impact at *Continent* level (Figure 17). With regard to the time delay $T\_max$ induced by a post (see left plot in Figure 17), *Society* and *Individuality* posts have approximatively 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 cascade number of nodes (note that x-axis is at log scale) and levels (center and right plots in Figure 17). The conclusion is that when a diffusion goes through *Society* community, cascades tend to be larger but not longer in time.

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*. With regard to cascade duration, the impact of the three communities is almost similar. However, if we consider the number of nodes and levels (center and right plots in Figure 18), *Agora* and *Politics* posts have a very similar distribution with, a more important impact than posts from *Technology* community. *Agora* and *Politics* posts induce very similar diffusion processes.

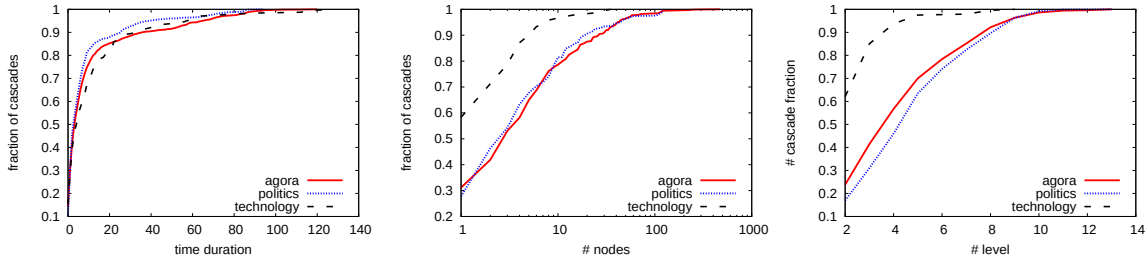


Fig. 18. Impact of node community at Region level. Left: $T\_max_u$, center: $N\_node\_max_u$, right: $lvl\_max_u$.

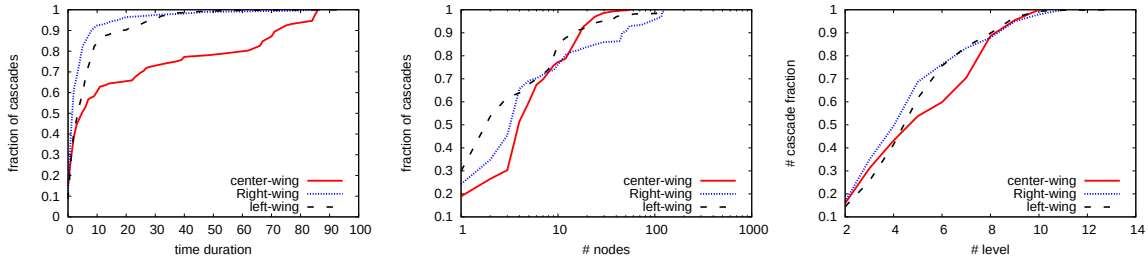


Fig. 19. Impact of node community at Territory level. Left: $T\_max_u$, center: $N\_node\_max_u$, right: $lvl\_max_u$.

In order to illustrate our methodology 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, which means that cascades spread during a longer period. However, we do not observe a significant difference for cascades sizes. We note that the number of nodes does not exceed 110 for each community.

## 5. Conclusion

We have proposed a new approach for the empirical study of diffusion cascades. We gave a definition of what we considered as a cascade in particular we did not consider citations within the same blog in cascade computation. We have observed that cascade shape frequencies were very similar to those of previous work done on American blog network [7], which suggests that cascade shape in blog networks is a common property. Moreover, the topological properties distribution analysis has shown a heterogeneous behaviour; in particular, we observed that cascade density was inversely proportional to cascade size and that cascades tended to be disassortative.

Second, in addition to topological properties, we have investigated community information to understand how cascades spread through communities. Therefore, we have used a topical community structure with three hierarchical levels which allows

analysis at different scales. The community of cascade origin has shown an impact on cascade properties and especially on average community distance for which diffusion behaviours differ noticeably.

Finally, we also considered the microscopic scale: we showed that at the node level, the community of cacade origin has various impact on other cascade features. At Continent level, *Society* nodes have a significant impact on cascade size and number of levels but not on their duration. Communities related to politics have a similar impact on cascade sizes however *Center-wing* blogs increase cascade duration. One perspective of this work is to consider the impact of other nodes properties, for example *Betweenness centrality* or degree in the blog network.

## References

[1] A. S. Brahim, B. L. Grand, L. Tabourier, and M. Latapy. Citations among blogs in a hierarchy of communities: Method and case study. *Journal of Computational Science*, 2(3):247 – 252, 2011.

[2] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi. Characterizing social cascades in flickr. In *Proceedings of the first workshop on Online social networks*, WOSN '08, pages 13–18, New York, NY, USA, 2008. ACM.

[3] 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.

[4] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.

[5] M. Granovetter. Threshold Models of Collective Behavior. *The American Journal of Sociology*, 83(6):1420–1443, 1978.

[6] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. 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.

[7] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *Proceedings of 7th SIAM International Conference on Data Mining*, 2007.

[8] A. L. Montgomery. Applying quantitative marketing techniques to the internet. *Interfaces*, pages 90–108, 2001.

[9] M. E. J. Newman. Assortative Mixing in Networks. *Physical Review Letters*, 89(20):208701+, Oct. 2002.

[10] M. E. J. Newman. *Networks: an introduction*. Oxford University Press, Tavistock, London, 2010.

[11] M. E. J. Newman, A. L. Barabási, and D. J. Watts, editors. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

[12] J. Park and M. E. J. Newman. The origin of degree correlations in the internet and other networks. *PHYS.REV.E*, 68:026112, 2003.

[13] S. Wasserman and K. Faust. *Social network analysis*. Cambridge University Press, 1994.

[14] D. J. Watts and P. S. Dodds. Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research*, 34(4):441–458, Dec. 2007.