# Towards multi-ego-centered communities: a node similarity approach

## Maximilien Danisch*

ComplexNetworks.fr
Laboratoire d'Informatique de Paris 6.
Université Pierre et Marie Curie
4 Place Jussieu, 75005 Paris, France
E-mail: maximilien.danisch@gmail.com
*Corresponding author

## Jean-Loup Guillaume

ComplexNetworks.fr
Laboratoire d'Informatique de Paris 6.
Université Pierre et Marie Curie
4 Place Jussieu, 75005 Paris, France

## Bénédicte Le Grand

ComplexNetworks.fr
Laboratoire d'Informatique de Paris 6.
Université Pierre et Marie Curie
4 Place Jussieu, 75005 Paris, France

**Abstract:** The community structure of a graph is defined in various ways in the literature: (i) Partition, where nodes can belong to only one community. This vision is unrealistic and may lead to poor results because most nodes belong to several communities in real-world networks. (ii) Overlapping community structure, which is the most natural view, but is often very difficult to identify in practice due to the complex structure of real-world networks, and the huge number of such possible communities. (iii) Ego-centered community which focuses on individual nodes' communities and seems to be a good compromise.

In this paper we investigate the third vision; we propose a new similarity measure between nodes based on opinion dynamics to unfold ego-centered communities. We call it the *carryover opinion*. In addition to be parameter-free, the carryover opinion can be calculated in a very time-efficient way and can thus be used in very large graphs.

We also go further in the idea of ego-centered communities by introducing the new concept of *multi-ego-centered communities*, i.e., focusing on the communities of a set of nodes rather than of a single node. A key idea is that, although one node generally belongs to numerous communities, a small set of appropriate nodes can fully characterize a single community.

**Keywords:** ego-centered communities, multi-ego-centered communities, carryover opinion, local communities, overlapping

communities, metric on nodes, node similarity, node proximity, opinion dynamics.

**Reference** to this paper should be made as follows: M. Danisch, J.-L. Guillaume and B. Le Grand (xxxx) 'Towards multi-ego-centered communities: a node similarity approach', *Int. J. of Web Based Communities*, Vol. x, No. x, pp.xxx–xxx.

**Biographical notes:** Maximilien Danisch is a Ph.D student under the supervision of Jean-Loup Guillaume and Bénédicte Le Grand at Université Pierre et Marie Curie (Paris 6) since September 2011. In 2010, he obtained a master's degree in physics from ENS Cachan after an internship on granular matter at the City College of New York under the supervision of Hernan A. Makse. In 2010-2011 he spent one year working on machine learning problems as an intern at Columbia University under the supervision of Tony Jebara.

Jean-Loup Guillaume is an Associate Professor at Université Pierre et Marie Curie (Paris 6) since 2007. He obtained his Ph.D in computer science from University Denis Diderot (Paris 7) in 2004 after which he spent two years and a half in post-doc position in France Telecom and at Université Catholique de Louvain in Belgium. His research interests are centered on complex networks. He works on many aspects of these networks: analysis, modeling and algorithmics. He is the co-author of the article presenting the most efficient community detection algorithm available today. In particular he has created the software (available on the web page http://findcommunities.googlepages.com/). Jean-Loup Guillaume is also reviewer for high quality journals and conferences, in particular for articles about complex networks and community detection.

Bénédicte Le Grand is an Associate Professor at Université Pierre et Marie Curie (Paris 6) since 2002. She will be a full Professor at Université Paris 1 Panthéon Sorbonne from September 2012. Her research activities deal with information retrieval and navigation in large complex networks such as the Internet, the Web or social networks. She has published her work in international journals and conferences, and she contributed to several books. She also works as an expert for the European Commission, in the context of projects evaluation.

## 1 Introduction

In social networks, communities are groups of users who share common features or have similar interests; studying the community structure has thus many applications for advertising as well as market research. Given a set of users, the most common way of identifying communities consists in classifying them in identified or unknown classes; this is what classical classification and clustering approaches do, e.g., k-means or others.

In terms of graphs, community detection generally aims at finding a partition of nodes, which means that each node belongs to one and only one community. However, if we consider social networks, where edges may represent friendship between users, it is hard to conceive that a user belongs to only one group: he clearly belongs to numerous groups, e.g., his family, colleages, various groups of friends. In order to be consistent with this, overlapping communities should be allowed. However, computing all these overlapping groups in a network leads to numerous problems, in particular the number of potential groups in a network is $2^n$, where $n$ is the number of nodes: in addition to the time and space of the algorithm, the interpretation of obtained results may be very difficult.

An interesting compromise is to focus on the groups related to one node. This type of communities is referred to as *ego-centered communities*. For this task, we suggest to adopt a novel approach based on similarity between nodes instead of a cost function approach, as commonly seen in the literature, which suffers from local minimum and hidden scale parameter.

Even though we obtain interesting results, in some cases, ego-centered community detection is still a difficult problem because a single node can still belong to numerous groups (up to $2^{n-1}$); we therefore suggest to take into account the context by identifying the communities of a set of nodes, called *multi-ego-centered communities*. In particular, we show that a small set of nodes is generally sufficient to define a unique community, which is generally not the case with one single node.

In addition to results obtained on small synthetic networks and small real-world networks, we also worked on a very large network which is a wikipedia dataset containing more than 2 million labeled pages and 40 million links.

This article goes beyond the state of the art through the three following contributions:

1. A new similarity measure between nodes based on opinion dynamics, which we call *the carryover opinion*. This similarity measure is parameter-free, takes into account the whole graph and not only a local view and is very fast to compute: the algorithm is in $O(te)$, where e is the number of edges and t is relatively small. (Calculating the similarity of one given node to all other nodes takes only few seconds for the whole wikipedia dataset of more than 2 million nodes).

2. The possibility of characterizing a node in terms of its ego-centered community structure, i.e., stating whether it is in the center of a community or more peripheral in between several communities, thanks to the carryover opinion and the time-efficiency of its computation.

3. The new concept of multi-ego-centered communities: communities related to a set of nodes, which extends the already established concept of ego-centered communities.

The first section being the introduction, the following of this article is organized in four sections: the second section is a state of the art of community detection algorithms and node similarity measures for community detection. The third section presents a new similarity measure, called carryover opinion, and its

applications for the detection of ego-centered communities. The fourth section shows how to use the carryover opinion to unfold multi-ego-centered communities with some validations on real graphs. Finally, the last section concludes and presents the perspectives for future works.

## 2 State of the art

### *2.1 Community detection*

It has been found that most complex networks exhibit a community structure, Girvan and Newman (2002). However, the concept of *community* itself is not well-defined. A common fuzzy definition is: a group of nodes more connected to one-another than to the nodes of the other groups. The idea of a community is also related to information propagation: information will propagate faster within a community than through different communities. In most practical cases, communities are simply the output of an algorithm, without a more accurate definition.

As detailed in the introduction, even though the most realistic way of seeing the community structure is to consider overlapping communities, most initiatives in community detection applicable to very large graphs (i.e., dozens thousand nodes) are limited to the identification of a partition of nodes. A common way to unfold the community structure seen as a partition consists in (keeping in mind the fuzzy definition) maximizing a quality function, a popular one being modularity, Girvan and Newman (2002). Even though maximizing this quality function is NP-hard, a good local minimum can be found very efficiently using the Louvain method, Blondel et al. (2008). Other approaches also exist, such as Pons and Latapy (2006), where a metric based on random walks maps nodes into points in a Euclidean space, and thus transforms the problem of community detection into the one of clustering; the infomap method, Rosvall and Bergstrom (2008), using techniques from data compression; or Morarescu and Girard (2011), using opinion dynamics, which is similar to the approach we will follow for ego-centered communities.

There however exist algorithms to cope with the problem of overlapping community structure. The most popular is the k-clique percolation, Palla et al. (2005), where a community is seen as a set of cliques of size k where each clique overlaps, at least, another one by k-1 nodes, where k is a parameter controlling the size of the cliques. Another interesting approach consists in partitioning the links instead of the nodes, which results in an overlapping community structure on nodes, Ahn et al. (2010). This can be done by applying the techniques established for communities seen as partition to the line-graph of the considered graph, Evans and Lambiotte (2009) and Evans and Lambiotte (2010). Another technique uses the non-determinism of algorithms for community seen as partition to obtain overlapping communities, Wang and Fleury (2010).

Another trend in the literature related to the community structure focuses on one node. In addition of being a good compromise between the realism of overlapping communities and the feasibility of communities seen as a partition, this third way seems to have emerged because real networks, such as Internet, Facebook or the Web are huge and dynamic; this makes it hard to know the

complete structure of the network, while it is still possible to know the structure around the neighborhood of one node. In the literature the algorithm dealing with this problem consists in designing and optimizing a fitness function. Most of the time it is a function of the number of internal and external edges, Clauset (2005); Bagrow (2008); Chen (2009); Ngonmang et al. (2012). Another work based this fitness function on triangles, Friggeri et al. (2011): the function, called Cohesion, compares the triangles made of three nodes within a community to triangles with only two nodes in the community and thus pointing out.

However, in addition to suffer from local minimum problems, these functions often have a hidden scale parameter. For instance Cohesion, incorporating a density of triangle term, decreases in $O(s^3)$ (where $s$ is the number of selected nodes) on sparse graphs and thus leads to very small communities. This cost function is actually used to find *egommunities*, i.e., communities related to a node taking into account only its neighbors. In that case, since complex networks are not locally sparse, the density of triangle decreases slower and the function is less biased in favor of small size egommunities.

Because of the local minimum problems and since an unbiased cost function (with regard to scale) remains very hard to define, we suggest to use a similarity approach. The principle of our method can be split into three consecutive steps:

1. Calculate the similarity between the node of interest and all other nodes.
2. Rank nodes in decreasing similarity order, with regard to the node of interest.
3. Find irregularities in the decrease, if they exist, that can be due to the community structure.

### *2.2 Node similarity measure*

Even though using a similarity measure (or metric) on nodes approach is novel for the study of ego-centered communities, similarity measures have already been used for community detection seen as partition. For instance Pons and Latapy (2006) developped a metric based on random walks to map nodes into points in a Euclidean space. They thus transformed the problem of community detection into the one of clustering. They then used an agglomerative clustering algorithm to obtain a partition of nodes.

For our problem, various existing similarity measures or metrics on nodes may be used. However they all have one of the three following drawbacks: (i) they are too restrictive, or (ii) they need an a priori parameter, or (iii) they are too slow to be computed for huge graphs. A selection of commonly-used similarity measures or metrics is presented in the following:

- Distance between nodes. This metric is too restrictive since it takes integer values which are small in front of the size of the graph. It falls in category (i).
- Probability for a random walker who started to walk from the picked node to be on a given node after $t$ iterations, Pons and Latapy (2006). This metric depends on the parameter $t$ and belongs to category (ii). Moreover it gives an advantage to high degree nodes.

- Jaccard similarity coefficient. For 2 nodes $a$ and $b$ it is given by

$$J(a,b) = \frac{|N_a \cap N_b|}{|N_a \cup N_b|}$$

where $N_a$ (resp. $N_b$) is the set of the neighbors of $a$ (resp. $b$). However, with this similarity two nodes that do not share any neighbor have a similarity equal to zero. This is too restrictive for our problem and falls in category (i).

- Personalized page-rank, Page et al. (1998), which is given by the following fix-point algorithm:

$$X_{t+1} = (1-\alpha)TX_t + \alpha X_0$$

where $X_t$ is the vector of the scores after n iterations, $X_0$ is initialized with the vector of all zeros except for the picked node which is set to one, T is the transition matrix: $T_{kl} = \frac{l_{kl}}{d_l}$, where $l_{kl}$ is the weight of the link between the nodes $k$ and $l$, and $d_l$ is the degree of node $l$. $\alpha \in ]0,1[$ is a parameter which controls the depth of network exploration. The problem is that the result highly depends on $\alpha$ and gives an advantage to the nodes with a high degree. This similarity falls in category (ii).

- Hitting time (resp. commuting time) could be a solution. It is, for a source node and a target node, the expected number of steps that a random walker would take to go (resp. to go and come back) from the source to the target.

  For the node of interest set as a target, all hitting times (i.e. for all nodes set alternatively as a source) can be calculated with a fix-point algorithm as detailed in Norris (1997). However for very large graphs the fixed-point method is too slow to converge. Each iteration takes $O(e)$ ($e$, number of edges) and the number of iterations is about the maximum of the expected number of steps for all source nodes, which can be bigger than ($n$, number of nodes). Thus this similarity falls in category (iii).

To our knowledge there is no similarity measure without at least one of the three identified drawbacks.

## 3 A new node similarity measure for ego-centered communities

### *3.1 Carryover opinion metric*

In this section, we define a similarity measure based on opinion dynamics, which takes into account all the depth of the graph, is parameter free and is fast to compute.

Given a node of interest, the framework consists in first setting the opinion of this node to one and the opinion of all other nodes to zero. Then, at each time step, the opinion of every node is averaged with the one of this neighbors. The opinion of the node of interest is then reset to one. Its opinion thus does not change all along the process and remains equal to one (which means that the similarity between the node of interest and itself is one).

Such as, this process is useless because it converges to an opinion of one for every node, however we have the feeling that nodes *closer* to the starting node will converge faster. The idea is to obtain a measure of that speed to characterise how nodes are similar to the node of interest: the higher the speed the more similar is the node.

Two conjectures are needed to carry on :

**Conjecture 3.1:** *After a number of iterations sufficiently large, the ranking of the nodes according to their opinion is not changing.*

**Conjecture 3.2:** *After a number of iterations sufficiently large, the difference between the opinion of two nodes decreases proportionally to the difference between the opinion of any other two nodes.*[a]

The conjectures simply states that given four nodes $a$, $b$, $c$ and $d$ with opinion at iteration $t$ noted $O_a^t$, $O_b^t$, $O_c^t$ and $O_d^t$ respectively. We have:

$$\lim_{t\to\infty} \frac{O_a^t - O_b^t}{O_c^t - O_d^t} = C_{a,b,c,d}$$

where $C_{a,b,c,d}$ is a constant depending on the nodes $a$, $b$, $c$ and $d$.

These conjectures have been tested on various benchmarks and real-world networks with conclusive results. We show the results on figure 1, where the experiment is carried out on the symmetrized polblogs network, Adamic and Glance (2005), a network of blogs and hyperlinks consisting in 1222 nodes and 16717 edges. As we can see, after a few iterations, the ranking of nodes according to their opinion is not changing, while the difference between opinions becomes proportional.

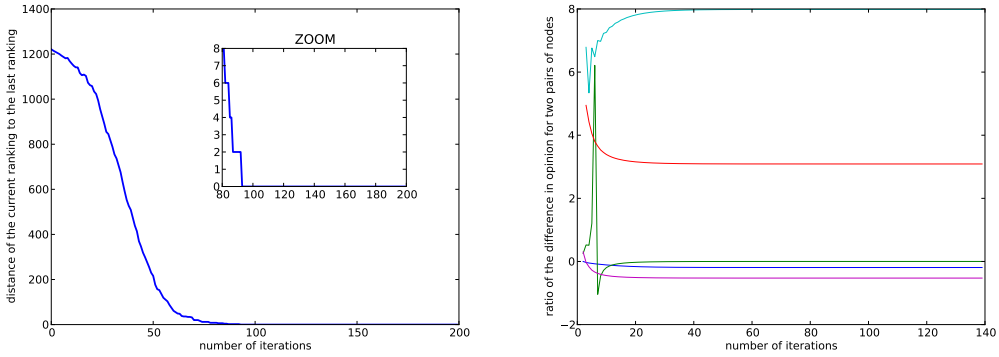
It is thus possible to rescale the opinion at each iteration such that the lowest opinion is zero. The highest is always one, which is the opinion of the node of interest. Scores between one and zero are thus obtained for each node at each iteration and the process converges towards a fix point. We call this value after convergence the carryover opinion, because even though the simple opinion process detailed above converges towards one for every nodes, the rescaling allows us to capture the proximity of nodes to the node of interest, which is carried over the whole process.

The node of interest beeing labeled $i$, each iteration thus consists in three steps:

$$\begin{aligned} X_t &= MX_{t-1} \quad \ldots\ldots\ldots\ldots\ldots \text{AVERAGING} \\ X_t &= \frac{X_t - min(X_t)}{1 - min(X_t)} \quad \ldots\ldots\ldots \text{RESCALING} \\ X_t^i &= 1 \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{RESETTING} \end{aligned} \tag{1}$$

where,

[a] Even though conjecture 3.2 implies conjecture 3.1, we think it is clearer to dissociate the two.




(a) (b)

**Figure 1**: Experiments validating conjectures 3.1 and 3.2. The experiments are carried out on the symmetrized polblogs network Adamic and Glance (2005), a network of 1222 nodes and 16717 edges. Figure 1a validates conjecture 3.1 by comparing the ranking of nodes according to their opinions to the ranking according to the last opinions obtained (for 200 iterations). As we can see, after only 95 iterations the ranking is not changing. The distance between the ranking we used is simply the number of mis-classed nodes. Figure 1b validates conjecture 3.2 by plotting the ratio of the difference of two randomly chosen pairs of nodes. The experiment has been made 5 times, there is therefore five curves. As we can see, after only 40 iterations the ratio is quite constant, thus the differences in the opinion of a pair of nodes is proportional to the one of any other pair.

- $X_t$ is the score vector after $t$ iterations and the component $j$ of the vector $X_t$ is noted $X_t^j$.
- $X_0$ is set the null vector, except for the node of interest, $i$, with value one.
- M is the averaging matrix, i.e., the transposed of the transition matrix : $M_{kl} = \frac{l_{kl}}{d_k}$, where $l_{kl}$ is the weight of the link between the nodes $k$ and $l$, and $d_k$ is the degree of node $k$.

We tested the algorithm on the polblogs network, see figure 2. After the convergence, which is nearly obtained after 40 iterations, the decrease in loglog scale is composed of two plateaus separated by a significant decrease in score values. This decrease appears around the 600th node. Actually the dataset contains 759 political blogs labeled as liberal and 443 labeled as conservative. In order to determine whether the nodes of the first plateau correspond to the picked node's community, we plotted the graph using the spring layout of Fruchterman and Reingold (1991), using a circle (resp. square) shape for liberal (resp. conservative) blogs. We then colored the nodes in blue according to their scores following a logarithmic scale, except the randomly picked node which is colored in red, see in figure 3. As we can see, the colors are consistent with labels: the randomly picked node was actually a liberal blog and most liberal blogs are colored in blue while the conservative blogs remain white. When nodes are ranked in decreasing

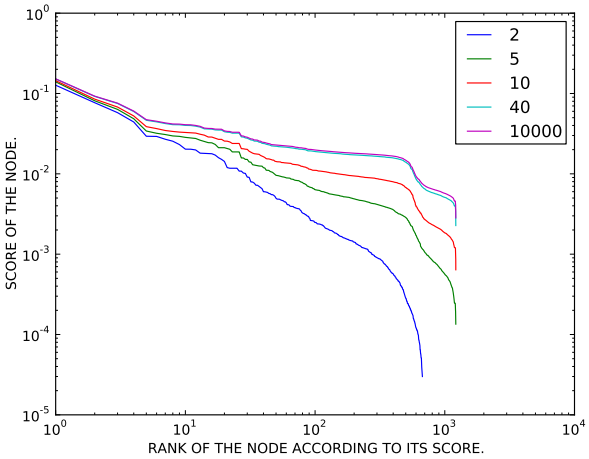



**Figure 2**: Experiment showing the convergence towards the carryover opinion. The experiment is carried out on the Newman (2006) polblogs network for which we randomly selected a node. The plot shows the score of each node as a function of its score ranking itself for 2, 5, 10, 40 and 10000 observed figure 1b. Even though the order of nodes slightly changes during the first hundred iterations, as proved on the figure 1a, the changes are negligible after 40 iterations.

order according to the carryover opinion: 561 liberal nodes are among the 600 first ranked nodes, i.e., 93.5% of the 600 first ranked nodes are liberal; 617 liberal nodes are among the 759 first ranked nodes, i.e., 81, 4% of the 759 first ranked nodes are liberal.

We applied this technique to smaller networks, therefore easier to visualize. Interesting results were obtained, as shown on figure 4: Figure 4a shows the carryover opinion of nodes as a function of their carryover opinion ranking for a co-authorship network, Newman (2006). The curve exhibits two major drops: the first one around the 50th node (the first 50 nodes therefore constitute the closest community of the picked node) and another one around the 180th (the first 180 nodes thus correspond to a larger community of the picked node, i.e., a community at a lower resolution). The corresponding nodes can be seen on the drawing where three different levels of color emerge. The succession of plateaus and decreases (on figures 4b, 4c and 4d) for three other networks also shows how useful the carryover opinion is to unfold ego-centered communities.

As we can see on figure 5a, results obtained with the carryover opinion are not always the expected ones: this experiment has been carried out on a synthetic network consisting of three Erdos-Renyi graphs of hundred nodes with a link probability of 0.3, while nodes belonging to different Erdos-Renyi graphs have a probability of 0.05 to be linked. The value obtained for the first neighbors of the picked node somewhat dominates the community structure artificially generated, in fact the neighbors of the picked node have a high score even if they are in different Erdos-Renyi graphs. However one can argue that we are looking for the community(ies) of one node and, in that sense, if a node is linked to the picked

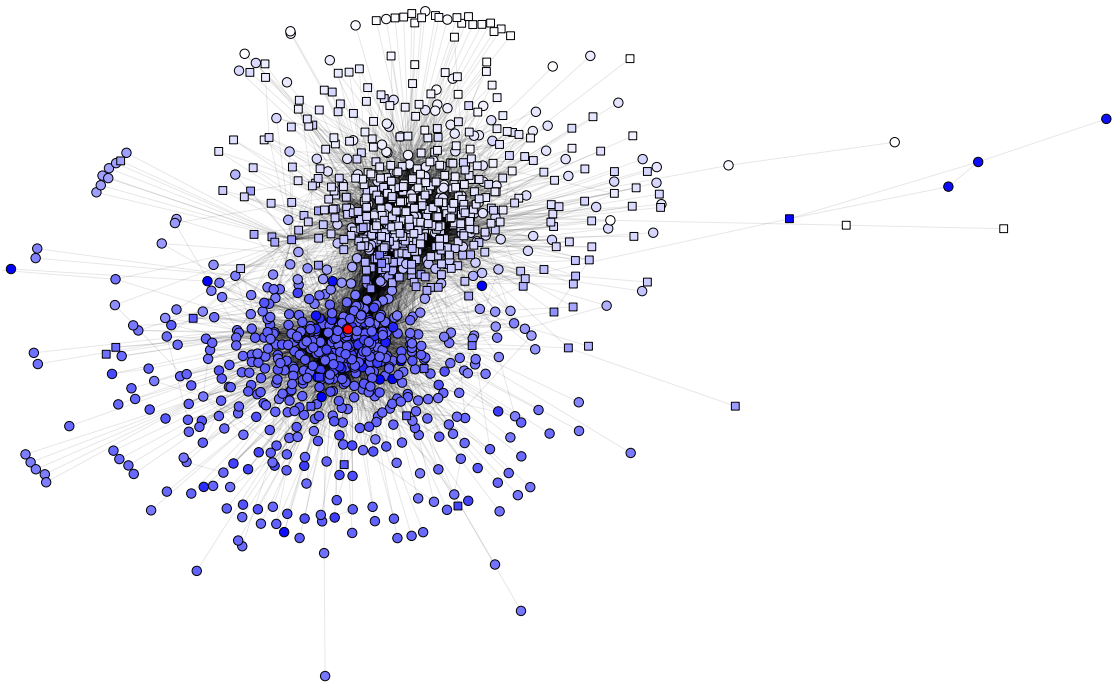


**Figure 3**: Drawing of the polblogs graph following the spring layout of Fruchterman and Reingold (1991). The circles represent liberal blogs, while squares represent conservative blogs. The picked node is in red, while the higher the carryover opinion of a node, the more intense its blue color, following a logarithmic scale.

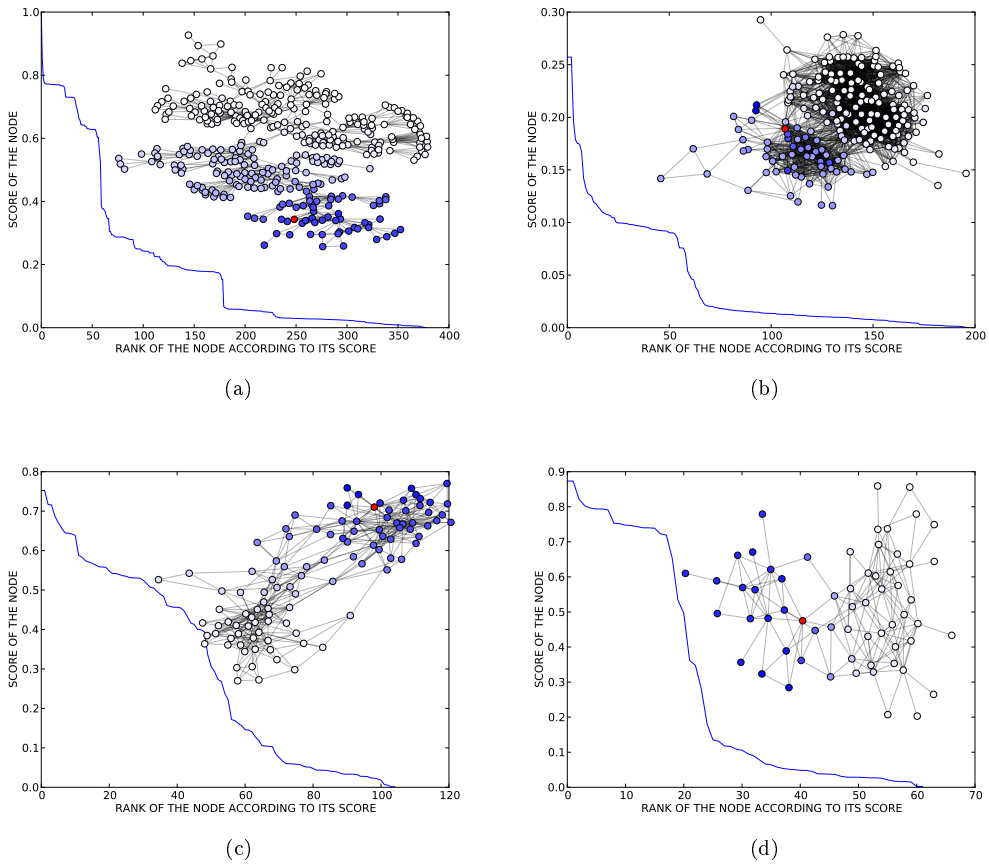
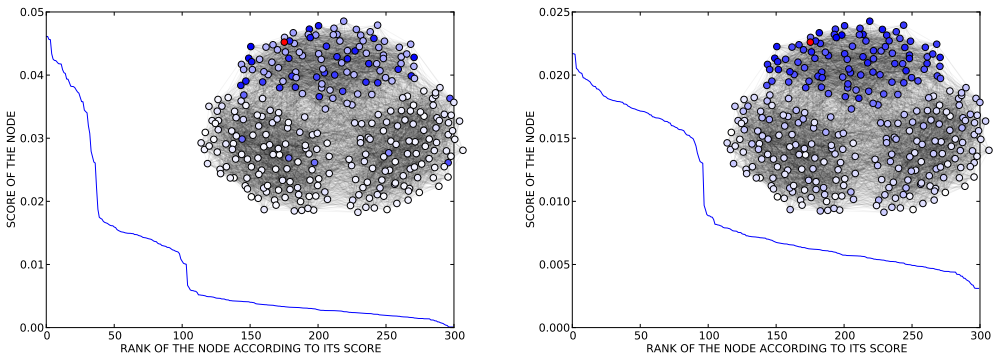



**Figure 4**: Result for four small visualisable networks. On the drawing of the networks, the picked node is in red. For the other nodes, the higher the score the more bluish the node. The graphs are plotted using the graphviz layout. On small graphs a simple linear scale for the plot of the carryover opinion can be used. 4a is for a co-authorship network of 379 nodes and 914 edges, Newman (2006). 4b is for a co-appearance network of jazz musicians of 198 nodes and 5484 edges, Gleiser and Danon (2003). 4c is for a citation network of political books of 105 nodes and 441 vertices, Krebs. 4d is for a social network of dolphins of 62 nodes and 159 edges, Lusseau et al. (2003).




(a) (b)

**Figure 5**: 5a shows the result for 3 Erdos-Renyi graphs (100,0.3), while nodes in different erdos-Renyi graphs are linked with probability 0.05. Figure 5b, shows the same result, but with an additional step: the picked node is removed and the value for each node is set to the average value of its neighbors, i.e., a final averaging step is performed without the picked node.

node those two nodes already constitute a community. Actually the minimal value for a first neighbor with degree $d$ is $\frac{1}{d}$, which makes sense: if all of the other neighbors of this first neightbor are *faraway* from the picked node, then this first neighbor is still $\frac{1}{d}$ part of the comunity(ies) of the picked node.

This effect (due to the communities of two nodes) can however be easily eliminated, as shown on figure 5b, by adding an additional step after the convergence of the carryover opinion: the picked node is removed from the graph and the value for each node is set to the average value of its neightbors. This affects only the first neighbors and it is the same as applying the transformation:

$$S = (S - \frac{1}{d})\frac{d}{d-1},$$

where S is the carryover opinion of a first neighbor.

We also can see that there are two effects that result in the final value of the carryover opinion: (i) 'a distance effect' and (ii) 'a redundancy effect' due to the community structure. As shown in figure 5a, the distance effect is sometimes dominating the redundancy effect. We argued that this is because the carryover opinion sees a pair of linked nodes as already a community. The question is to know how (if) this will affect the result for the nodes at distance two or more. To investigate this, we compare the decrease of the carryover opinion as a function of the distance for the wikipedia network (choosing the page 'boxing') and an Erdos-Renyi graph of the same average degree. As shown in figure 6, while on the Erdos-Renyi graph the decrease is exponential, on the wikipedia network only the neighbors of the picked node are affected. This means that there is no correlation between the distance and the value of the carryover opinion for nodes at distance two or more from the picked node. Thus this effect is only due to the fact that two linked nodes are considered as a community and the correcting step we suggested

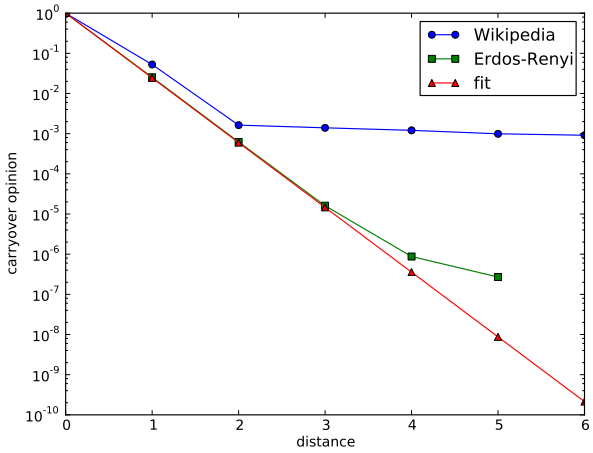



**Figure 6**: These plots show the average carryover opinion for nodes at a given distance from the node of interest as a function of the distance: Wikipedia is for the wikipedia network containing $n = 2,070,367$ nodes and $e = 42,336,614$ edges. Erdos-Renyi is for an Erdos-Renyi graph containing this same number of edges and nodes. Fit represents the curve $\frac{1}{degree^{distance}}$ where the degree is set to the average degree of the previous graph, i.e., $degree = \frac{2e}{n} = 40$

is efficient to eliminate this effect.

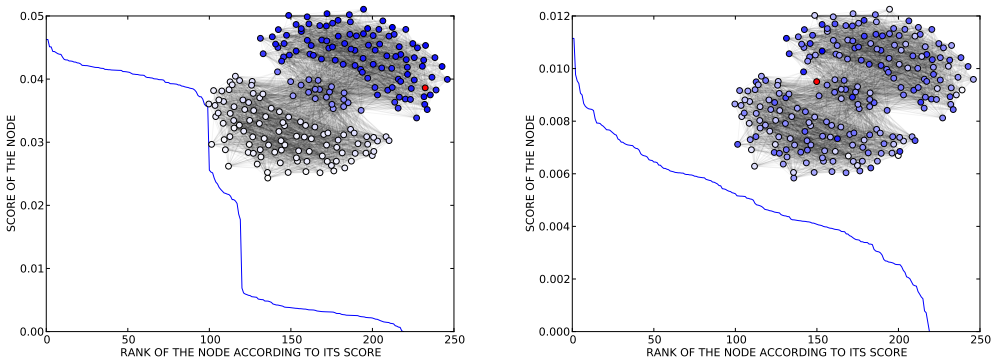
Such an ideal structure of plateaus and strong decreases (as seen on figures 4 and figures 5) does not always apear. In fact it depends on two things: (i) The position of the picked node, i.e., central in a community or peripheral and thus within several communities. As shown on Figure 7, when the node is central the plateaus are clear while when the node is peripheral, no plateau is emerging. (ii) The structure of the community itself, i.e., if the community is well defined or not, as we can see on figure 8.

### *3.2 Ego-centered communities: Results on large graphs*

The technique presented above does not need any a priori input parameter other than the graph and is very time-efficient. It can thus be used in huge graphs to find 'the community' or 'the communities' of a node if there is one, looking for various rates in the decrease. However, as already discussed, a node often belongs to numerous communities and such a succession of plateaus and decreases is only occasionally observed.

Given randomly chosen nodes from the wikipedia network, figure 9a (resp. 9b) shows the plots of the carryover opinion (resp. with the additional correcting step) for all nodes as a function of their ranking. The four types of curves show the four major trends one can obtain:

- sharp transition,




(a) (b)

**Figure 7**: Results given by the carryover opinion with the correcting step for two overlapping Erdos-Renyi graphs of 110 nodes with an edge probability of 0.3 overlapping on 20 nodes. As we can see on figure 7a when the picked node is at the center of a community the plateaus-decreases structure is clear, while it can be unclear when the node is peripheral, figure 7b.

- smooth transition,
- deformed power law,
- perfect power law.

These four very different types of curves reflect very different structural properties of the nodes. Let us first notice that the correcting step is not modifying much the curves, the bias due to communities of two nodes is thus minimal here. This may actually mean that there is only a little amount of weak ties (i.e., links between very different communities) in the wikipedia network.

Let us explain these four behaviors through analyzing the curves and the ranking of pages without the correcting step:

- The 'sharp transition' curve corresponds to the 'Cotton Township, Switzerland County, Indiana' page. As we can see the first 6 nodes constitute a plateau. These nodes correspond to the page 'Switzerland County, Indiana' and the 5 other townships of The Switzerland County. Then we withstand a decrease on the next 7 nodes which are tightly related to 'Township, Switzerland County' and 'Indiana'. The next 970 nodes constituting the second plateau all correspond to other townships in Indiana (with no exception, Indiana counting 1005 townships). The next decrease on about 1000 nodes is composed by nodes related to townships and Indiana and also a little about Illinois, while the following plateau on about 1000 additional nodes is composed of the pages of the townships of Illinois (with a few exceptions). The wavy decrease towards the final plateau smoothly transits towards far away related contexts, passing through Indiana related topics,

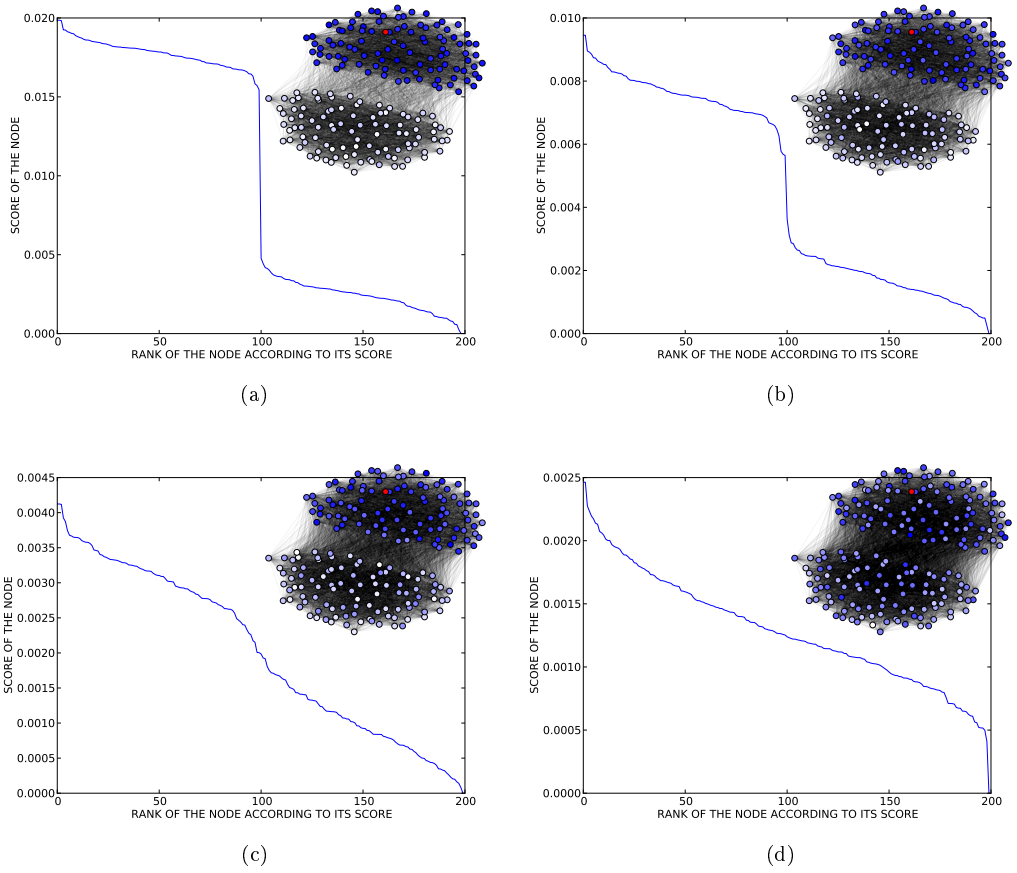
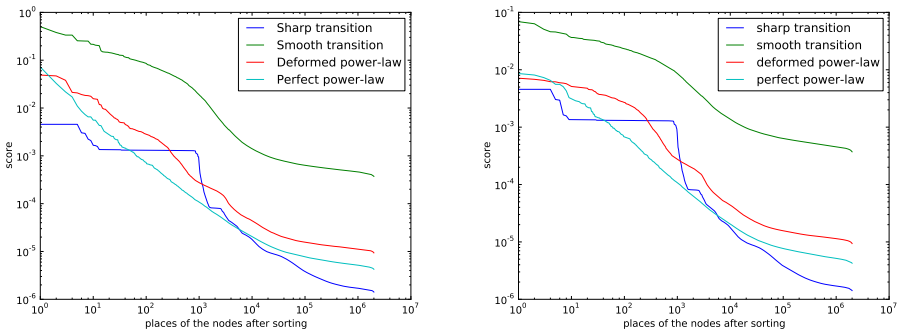



**Figure 8**: Results given by the carryover opinion with the correcting step for two Erdos-Renyi graphs (100,0.5). In Figure 8a (resp. 8b, 8c, 8d) two nodes in different Erdos-Renyi graphs are linked with probability 0.1 (resp. 0.2, 0.3, 0.4).

Ohio's townships, Michigan's townships, other states townships, US related topics...

- The 'Smooth transition' curve is obtained for the page 'Mafia'. This node can characterize a community by itself: the first thousands pages are mafiosi names or organized crime related topics. However the community is more fuzzily defined than the ones for 'Cotton Township, Switzerland County, Indiana'.
- The 'Deformed power-law' curve is for 'Mi-Hyun Kim' page. The page is mainly linked to pages about Golf and Korea topics. The first thousand pages are related to one or two of these topics, we obtain a superposition of the score of these topics, which leads to this wavy power law; this behaviour is even clearer after applying the correcting step: we can then see two waves corresponding to a mixture of the two topics/communities (Korea and Golf).




(a) (b)

**Figure 9**: Plots of the carryover opinion of all nodes as a function of their ranking for four randomly picked nodes in the wikipedia network (left), and the same plots but after adding the correcting step (right). Sharp transition corresponds to the 'Cotton Township, Switzerland County, Indiana' node. Smooth transition corresponds to the 'Mafia' node. Deformed power law corresponds to the 'Mi-Hyun Kim' node. Perfect power law corresponds to the 'JNCO' node.

- The 'perfect power-law' curve is for the 'JNCO' page, which is a clothing brand. As we can see the plot is a perfect power law that finishes with a low plateau. No community structure emerges from this plot; this is because the page is indeed linked to many different nodes that are part of various communities of different sizes fuzzily overlapping: 'JNCO' is linked to the pages 'Los Angeles', 'Jeans', 'Hip-hop', 'J.C. Penney', 'Graffiti', 'Kangaroo', 'Boxing', 'Nu Metal', from which hardly any context can emerge.

Concerning communities, we found that, in the same network, there seems to be two types of communities and we may characterize them as:

1. well-defined communities, like the one of Switzerland country or Indiana.
2. fuzzily defined communities, like the one of mafia.

Also, these communities can be multiscale: Switzerland country is a sub-community of Indiana.
Concerning nodes, we found that, in the same network there are mainly three types of nodes (regarding communities):

1. Nodes which can, by themselves, define a community like 'Cotton Township, Switzerland County, Indiana' or 'mafia'.
2. Nodes which are in the middle of very few communities, like 'Mi-Hyun Kim'.
3. Nodes which are in a middle of a large number of communities, like 'JNCO'.

For a given node, the properties can all be deduced from the shape of the curve: carryover opinion as a function of the ranking according to it.

## 4 A new vision of communities

### *4.1 Multi-ego-centered communities*

It appears that, on the wikipedia network, most nodes have a -carryover opinion VS ranking- curve whose behaviour is between deformed power-law and perfect power-law. Thus, in this network, nodes seem to belong to many communities; however, we have the intuition that a well chosen set of few nodes could define a single community.

The question is: how may the communities shared by a set of nodes be unfolded? We suggest to use the previously established similarity. The idea is that a node belonging to a community of node1 AND to a community of node2 has to be similar to node1 AND to node2. The following example in figure 10 shows how to proceed:

1. Evaluate for all nodes the similarity to node1 and to node2.
2. The similarity to the set {node1,node2} is then given by the minimum, or by the geometric mean of the similarities to node1 and the similarities to node2. This quantity measures to what extent a node is near from node1 AND node2.[b]

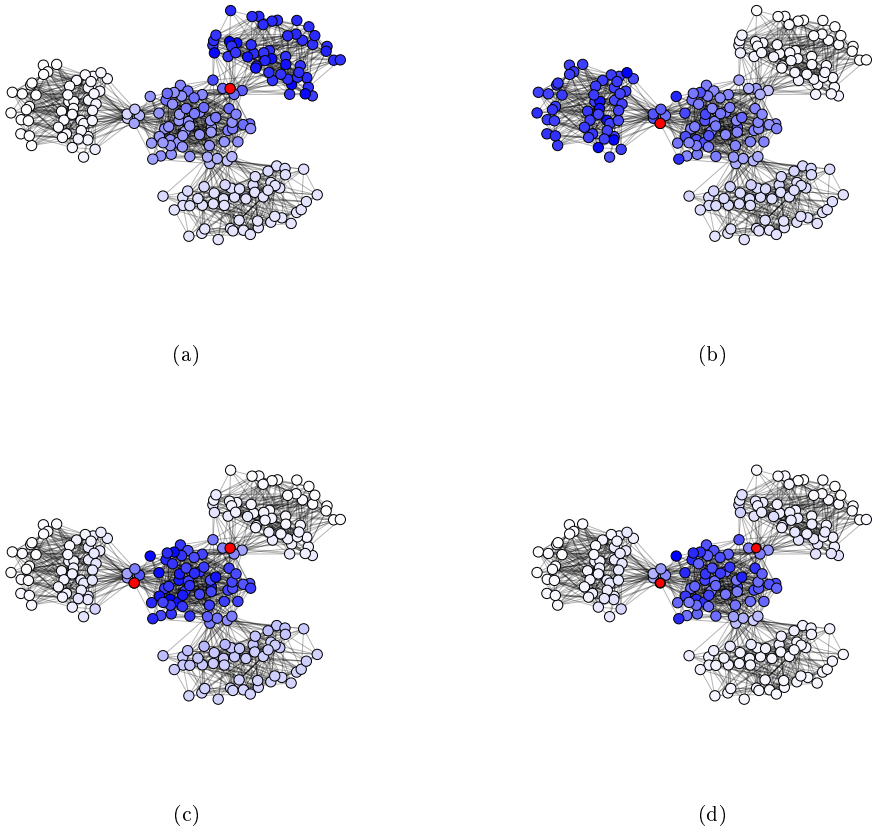
The method is easily generalisable to a set of more than two nodes.

### *4.2 Multi-ego-centered communities: results on large graphs*

We applied the framework described above to the wikipedia network using the minimum similarity of the picked nodes. Figure 11a shows the results for two nodes : 'Folk wrestling' and 'Torii school'. One is dedicated to the various types of traditional wrestling around the world, while the other one is dedicated to a traditional Japanese art school. Both curves are slightly deformed power-laws and do not uncover any community.
Figure 11b shows the result for sumo along with the minimum of the scores for the pages 'Folk wrestling' and 'Torii school' and the same rescaled minimum, such that it starts at 1.
As we can see the two curves have exactly the same structure: a plateau followed by a decrease at about the 350th node. 'Folk wrestling' and 'Torii school' where related to 'Sumo' in a transversal way. Doing the minimum of the scores for these two pages gives us a score of how nodes are related to 'Folk wrestling' and Torii school' which actually correspond to 'Sumo'. Comparing the 350 first nodes of each experiments gives that:

- 14 nodes are in the first 350 nodes of 'Sumo' and 'Torii school',
- 12 nodes are in the first 350 nodes of 'Sumo' and 'Folk wrestling',
- 337 nodes are in the first 350 nodes of 'Sumo' and the minimum of 'Folk wrestling' and 'Torii school'.

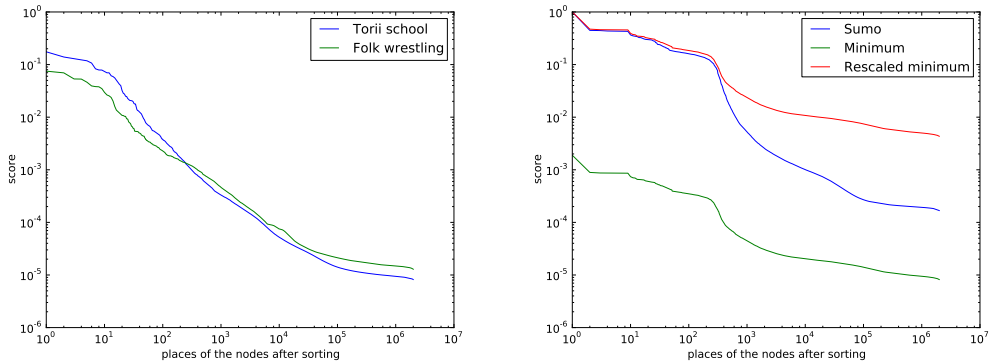
[b] Doing the arithmetic mean of the similarity or their maximum is not relevant for our problem, since this would unfold nodes that are part of a community of node1 OR node2.



(a) (b)

(c) (d)

**Figure 10**: Result for 4 overlapping Erdos-Renyi graph of 50 nodes and an edge probability of 0.2 overlapping on 5 nodes. The picked nodes are in red, the darker blue a node, the higher its score. Figure 10c (resp. figure 10d) gives the (rescaled) minimum (resp. geometric mean) of the scores on the experiments presented on figures 10a and 10b. The community shared by both red nodes is emerging.




(a) (b)

**Figure 11**: Figure 11a shows the results for two nodes: 'Folk wrestling' and 'Torii school'. Figure 11b shows the result for 'Sumo' along with the minimum of the scores for the pages 'Folk wrestling' and 'Torii school' and the same rescaled minimum, such that it starts at 1.

Also, the node having the highest score when doing the minimum of the carryover opinion for 'Folk wrestling' and 'Torii school' is actually 'Sumo'. In that case we found a set of pages which define a community already defined by a single node (the ego-centered community of 'Sumo'), but we believe that it is also possible to find multi-ego-centered communities which are not ego-centered.

It seems that using the minimum of both values could be more effective, however doing the geometric mean can allow to weight the set (possibly weighting some nodes negatively) to better investigate the overlapping. Also, using the minimum may be less stable in large graphs, since a single node added to the initial set could highly change the result (for instance if a node that has nothing to do with the rest of the set is added). Conversely adding a node very similar to a node already present in the set would not change the result. However, in our experiments, we obtained better results doing the minimum.

## 5 Conclusion and future works

We presented a new similarity measure between nodes of a graph that we call the carryover opinion. Its calculation can be performed very efficiently and does not require any parameter influencing the result. This new similarity can be used to unfold ego-centered communities, even though in very large graphs a deformed power-law decrease is often obtained because nodes generally belong to numerous fuzzily overlapping communities. Nevertheless this similarity shows how likely it is for two nodes to share at least one community. It also allows to see whether the node characterizes a community by itself (succession of plateaus and decreases), is in the middle of a few communities (wavy power-law) or in a middle of many communities (quasi-perfect power-law).

We also introduced a new vision of communities: multi-ego-centered communities. In this problem, we consider a set of nodes and we look for the communities shared by all nodes in the set. We showed that a very small set of nodes, e.g., 2, is often enough to characterize a single community using the previously established similarity measure.

Future works should deal with the formalization of multi-ego-centered communities and further validation in practical cases. Moreover, multi-ego-centered communities link community detection and recommendation systems. Another perspective may therefore consist in going this way by working on weighted-multi-ego-centered communities, where the initial set of nodes is weighted (possibly some nodes weighted negatively). We could thus look for nodes belonging to the communities of positively weighted nodes privileging highly weighted nodes, while not belonging to the communities of negatively weighted nodes. This could help investigate further overlapping communities by studying the structure of overlaps.

## Acknowledgements.

This work is supported in part by the French National Research Agency contract DynGraph ANR-10-JCJC-0202 and by the DiRe project, funded by the city of Paris Emergence program. The authors would also like to thank Daniel Bernardès, Sergey Kirgizov, Amélie Medem and Lionel Tabourier for helpful discussions and proofreadings.

## References

M. Girvan and M. E. J. Newman. 'Community structure in social and biological networks'. PNAS June 11, 2002, *Biometrika*, vol. 99 no. 12, pp. 7821-7826.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre. 'Fast unfolding of communities in large networks'. J. Stat. Mech. (2008).

Pascal Pons and Matthieu Latapy. 'Computing communities in large networks using random walks'. Journal of Graph Algorithms and Applications (JGAA) Vol. 10, no. 2, pp. 191-218, 2006

Martin Rosvall and Carl T. Bergstrom. 'Maps of random walks on complex networks reveal community structure' 1118-1123, PNAS, January 29, 2008, vol. 105. no. 4.

Irinel Constantin Morarescu, Antoine Girard. 'Opinion Dynamics with Decaying Confidence: Application to Community Detection in Graphs'. IEEE (2011).

Palla, G., I. Derenyi, I. Farkas and T. Vicsek. 'Uncovering the overlapping community structure of complex networks in nature and society'. Nature 2005.

Yong-Yeol Ahn, James P. Bagrow and Sune Lehmann. 'Link communities reveal multiscale complexity in networks'. Nature. Vol 466, 5 August 2010, doi:10.1038/nature09182.

T.S. Evans and R. Lambiotte. 'Line Graphs, Link Partitions and Overlapping Communities'. Phys.Rev.E 80 (2009) 016105, DOI: 10.1103/PhysRevE.80.016105.

T.S. Evans and R. Lambiotte. 'Line Graphs of Weighted Networks for Overlapping Communities'. Eur. Phys. J. B 77 (2010) 265-272.

Qinna Wang and Eric Fleury. 'Uncovering Overlapping Community Structure'. CompleNet 2010, Oct 2010, Rio de Janeiro, Belize.

Aaron Clauset. 'Finding local community structure in networks'. PHYSICAL REVIEW E 72, 026132, 2005.

James P. Bagrow. 'Evaluating local community methods in networks'. J. Stat. Mech. (2008).

Jiyang Chen, Osmar R. Zaiane and Randy Goebel. 'Community Identification in Social Networks'. Local 2009 Advances in Social Network Analysis and Mining.

Blaise Ngonmang, Maurice Tchuente, and Emmanuel Viennet. 'Local communities identification in social networks'. Parallel Processing Letters, 22(1), March 2012.

Adrien Friggeri, Guillaume Chelius, Eric Fleury. 'Triangles to Capture Social Cohesion'. IEEE (2011).

L. Page, S. Brin, R. Motwani, and T. Winograd. 'The pagerank citation ranking: Bringing order to the web'. Technical report, Stanford Digital Library Technologies Project, 1998.

J. R. Norris. 'Markov chains'. Cambrige university press 1997.

MEJ Newman. 'Finding community structure in networks using the eigenvectors of matrices'. Physical Review E, 2006, APS.

Lada A. Adamic and Natalie Glance. 'The political blogosphere and the 2004 US Election'. In Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem (2005).

Fruchterman, Thomas M. J.; Reingold, Edward M. (1991). 'Graph Drawing by Force-Directed Placement'. Software, Practice and Experience (Wiley) 21 (11): 1129-1164. doi:10.1002/spe.4380211102.

P. Gleiser and L. Danon. Adv. Complex Syst.6, 565 (2003).

D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, 'The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations'. Behavioral Ecology and Sociobiology 54, 396-405 (2003).

V. Krebs, unpublished. http://www.orgnet.com/

## Bibliography

S. Fortunato. Phys. Rep. 486, 75-174, 2010. 'Community detection in graphs'.

K. Thiel and M. R. Berthold. 2010 IEEE International Conference on Data Mining. 'Node similarities from Spreading Activation'.