CHAPTER THREE

MULTI-EGO-CENTRED COMMUNITIES

MAXIMILIEN DANISCH, JEAN-LOUP GUILLAUME AND BENEDICTE LE GRAND

Introduction

In social networks, *communities* are groups of users who share common features or have similar interests; studying this community structure has thus many applications for advertising or for market research. Given a set of users, the most common way of identifying communities consists in classifying them into classes which may be predefined or not. This is what traditional classification and clustering approaches do respectively.

In the context of real-world graphs, community detection generally aims at finding a partition of nodes, i.e. disjoint communities where each node belongs to exactly one community. However, in social networks it is hard to conceive that a user belongs to only one group, indeed, he/she clearly belongs simultaneously to a family, a group of colleagues, and various groups of friends. Overlapping communities should, therefore, be allowed in order to take this critical remark into account. However, computing all 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 complexity of the algorithm, the interpretation of obtained results may be very difficult.

An interesting compromise consists in focusing on the groups related to one specific node, referred to as *ego-centred* communities. We suggest adopting a novel approach based on proximity between nodes rather than on a cost function approach, as commonly seen in the literature. The use of cost functions may lead to a local minimum and imply hidden scale parameters. Despite promising initial results, ego-centred community detection is still a difficult problem because a single node can still belong to numerous groups. Therefore, we suggest focusing on specific communities and taking into account the context by identifying the communities of a set of nodes, called *multi-ego-centred communities*. Indeed, as we show in this chapter, a small set of nodes is generally sufficient to define a unique community, which is generally not the case with one single node.

We have worked so far (Danisch et al. 2012, Danisch et al. 2013) on small synthetic networks and small real-world networks, but also on a very large Wikipedia dataset containing more than 2 million labelled pages and 40 million links (Palla et al. 2008). This chapter details four recent contributions to the state of the art:

1. A new proximity measure between nodes based on opinion dynamics, which we call the *carryover opinion*. This proximity measure is parameter-free, takes into account the whole graph (rather than only a local view) and is very fast to compute: the algorithm is in $\mathcal{O}(te)$ where e is the number of edges and t is relatively small. Calculating the proximity between one given node and all other nodes takes only a few seconds for the whole Wikipedia dataset.
2. The possibility of characterising a node with regard to its *ego-centred community structure*, i.e. of stating whether it is in the centre of a community or between several, thanks to the carryover opinion and its time-efficient computation.
3. The new concept of *multi-ego-centred communities*: communities related to a set of nodes, which extends the already established concept of ego-centred communities.
4. An *algorithm* that unfolds all ego-centred communities of a given node through unfolding multi-ego-centred communities on the node of interest and some other carefully selected nodes.

This chapter is organised as follows. After this introductory section, the second section describes the state of the art of community detection algorithms and node proximity measures for community detection. The third section presents a new proximity measure, called carryover opinion, and its application for the detection of ego-centred communities. The fourth section describes the way the carryover opinion can be used to unfold multi-ego-centred communities and this approach is validated on real graphs. The fifth section details the algorithm that unfolds all ego-centred communities of a given node. Finally, the last section of this chapter concludes and presents perspectives for future work.

State of the Art

Community Detection

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 which are more connected to one another than to the nodes of other groups. The notion of community is also related to information propagation: information propagates faster within a community than through different communities.

As stated in the introduction, even though most community structures are made of overlapping communities, most initiatives for community detection in very large graphs (i.e. dozens of thousands of nodes) are limited to the identification of disjoint communities. A common way to extract such disjoint communities consists in maximising a quality function, a popular one being modularity (Girvan and Newman 2002). Even though maximising 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 et al. 2005), where a metric based on random walks maps nodes into points in a Euclidean space, and thus transforms the problem of community detection into a clustering task. The Infomap method (Rosvall and Bergstrom 2008), borrows techniques from data compression; and, finally, (Morarescu and Girard 2011), use opinion dynamics, as we do to compute ego-centred communities.

However, algorithms adapted to overlapping community structures do exist. The most popular one is the k-clique percolation (Palla et al. 2005), which considers a community as a set of cliques of size where each clique overlaps another one by $k-1$ nodes. Another interesting approach consists in partitioning links instead of nodes, which results in an overlapping node community structure (Ahn et al. 2010). This can be done by applying the techniques established for disjoint communities to the line-graph of the considered graph (Evans and Lambiotte 2009). Another technique uses the non-determinism of algorithms to obtain overlapping communities (Wang and Fleury 2011).

Another trend in the literature related to community structures focuses on one node. In addition to being a good compromise between the realism of overlapping communities and the feasibility of disjoint communities, this third approach has emerged because real networks, such as the Internet, Facebook or the web, are huge and dynamic. In this context, it is hard to find out the complete structure of the network, while it is still possible to discover the structure around the neighbourhood of one specific node. In the literature, algorithms dealing with this problem design and optimise a fitness function. Most of the time it is a function of the number of internal and external edges (Clauset 2005, Luo et al. 2008, Bagrow 2008, Chen et al. 2009, Ngonmang et al. 2012). In (Friggeri et al. 2011), the fitness function, called Cohesion, compares the triangles made of three nodes within a community to triangles with only two nodes in the community.

However, in addition to suffering from local minimum problems, these functions often have a hidden scale parameter. For instance, Cohesion, which depends on the density of triangles, decreases in $\mathcal{O}(s^3)$ (where s is the number of selected nodes) in 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 neighbours. In that case, since complex networks are not locally sparse, the density of triangles decreases slower and the function is less biased in favour of small egommunities.

Another interesting algorithm based on fitness function is detailed in (Sozio and Gionis 2010). The algorithm starts with all nodes in the community and removes some of them by greedily maximising the minimum degree of the sub-graph induced by the remaining nodes in the community. Even though the algorithm is greedy, it is proved to reach a global optimum, however, while the other algorithms are biased towards small communities, this one favours very big communities. Due to the local minimum problems and since an unbiased cost function (with regard to scale) remains very hard to define, we suggest using a proximity-based approach. The principle of our method can be split into three consecutive steps:

1. Calculate the proximity between the node of interest and all other nodes.
2. Rank nodes in decreasing proximity order.
3. Find irregularities in the decrease, if they exist, as they can give information about the community structure.

Node Proximity Measure

Even though using a node proximity measure (or metric) is novel for the study of ego-centred communities, proximity measures have already been used for disjoint community detection. For instance (Pons and Latapy 2005) developed a metric based on random walks to map nodes into points in a Euclidean space. They thus transformed the problem of community detection into a clustering task. They then used an agglomerative clustering algorithm to obtain a partition of nodes.

In our context, various existing node proximity measures or metrics may be used. However, they all have one of the three following drawbacks: (i) they are too restrictive; (ii) they need an *a priori* parameter; (iii) they are too slow to be computed for very large graphs. A selection of commonly-used proximity measures or metrics is presented in the following:

1. Number of hops between nodes. This metric is not selective enough (drawback (i)) since the number of distinct integer values is small with regard to the size of the graph.
2. 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 2005). This metric depends on t (drawback (ii)) and moreover it favours high degree nodes.
3. Jaccard similarity coefficient. For 2 nodes a and b it is given by
$$\mathcal{J}(a,b) = \frac{|N_a \cap N_b|}{|N_a \cup N_b|},$$
where N_a (resp. N_b) is the set of the neighbours of a (resp. b). However, two nodes which do not have any common neighbour have a proximity equal to zero. This is too restrictive for our problem (drawback (i)).
4. Personalised page-rank (Page et al. 1999), 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 t iterations, X_0 is the zero vector except for the picked node which is set to one, T is the transition matrix: $T_{ij} = \frac{l_{ij}}{d_j}$, where l_{ij} is the weight of the link between nodes i and j, and d_j is the degree of node j. $\alpha \in]0,1[$ is a parameter which controls the depth of network exploration. The problem of personalised page-rank is that the result depends significantly on α (drawback (ii)) and gives an advantage to high-degree nodes.
5. Hitting time (resp. commuting time). This metric is the expected number of steps that a random walker would take to go from a source node to a target node (resp. to go to a target node and come back to the source).

   With the node of interest as a target and all nodes set alternatively as sources, all hitting times can be calculated with a fix-point algorithm as detailed in (Norris 1997). However, for very large graphs the fixed-point method converges too slowly. Each iteration takes $\mathcal{O}(e)$ (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 greater than n (number of nodes). Thus, this proximity suffers from drawback (iii).

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

A New Node Proximity Measure for Ego-centred Communities

Carryover Opinion Metric

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

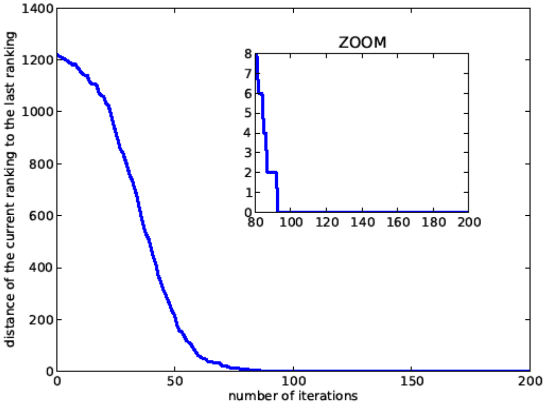
Given a node of interest, the framework consists in first setting the opinion of this node to one and the opinion of all other nodes to zero. Then, at each time step, the opinion of every node is averaged with the opinion of one of its neighbours. The opinion of the node of interest is then reset to one. Thus, its opinion does not change throughout the process and remains equal to one.

As such, this process might seem useless because it converges to an opinion of one for every node. However, the speed of convergence is interesting. Indeed, nodes that are *closer* to the starting node will converge faster to the opinion of that node. Our idea is to measure that speed to characterise to what extent nodes are similar to the node of interest. The higher the speed, the more similar the node. Two conjectures are needed to carry on:

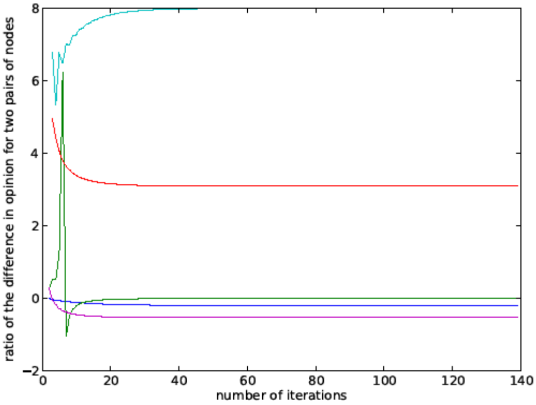
Conjecture 1: after a sufficient number of iterations, the ranking of nodes according to their opinion no longer changes.

Conjecture 2: after a sufficient number of iterations, the difference between the opinion of two nodes decreases proportionally to the difference between the opinions of any other two nodes.[1]

[1] Even though conjecture 2 implies conjecture 1, we think it is clearer to dissociate the two.



(3-1a)



(3-1b)

Fig. 3-1. Figure 3-1a validates conjecture 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 no longer changes. The distance between the rankings is the number of misclassified nodes. Figure 3-1b validates conjecture 2 by plotting the ratio of the difference of two randomly chosen pairs of nodes. The experiment has been conducted five times. As we can see on the corresponding five curves, after only 40 iterations the ratio is quite constant, thus the differences in the opinion of a given pair of nodes is proportional to the opinion of any other pair.

These conjectures simply state that, given four nodes a, b, c and d with opinion O_a^t, O_b^t, O_c^t and O_d^t respectively at iteration t, 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 only on nodes a, b, c and d. These conjectures have been tested on various benchmarks and real-world networks with conclusive results. Figure 3-1 shows the results of the experiment carried out on the symmetrised polblogs network (Adamic and Glance 2005), a network of blogs and hyperlinks consisting of 1,222 nodes and 16,717 edges.

It is thus possible to rescale the opinion at each iteration such that the lowest opinion is zero. The highest value 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 fixed point. We call this value after convergence the *carryover opinion*, because, even though the simple opinion process detailed above converges towards one for all nodes, this 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 being labelled i, each iteration thus consists of three steps:

1. Averaging: $X_t = MX_{t-1}$
2. Rescaling: $X_t = \frac{X_t - \min(X_t)}{1 - \min(X_t)}$
3. Resetting: $X_t^i = 1$

where:

- 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 to the zero vector, except for the node of interest, i, with value one.
- M is the averaging matrix, i.e. the transposed of the transition matrix: $M_{ij} = \frac{l_{ij}}{d_i}$, where l_{ij} is the weight of the link between the nodes i and j, and d_i is the degree of node i.

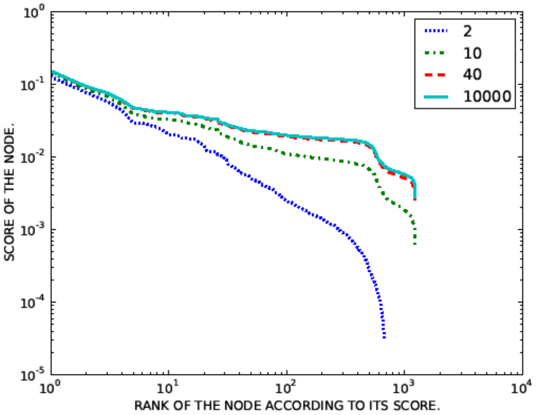


Fig. 3-2. Experiment showing the convergence towards the carryover opinion. The experiment was carried out on the (Newman 2006) polblogs network for which we randomly selected a node. The plot shows the score of each node as a function of its score ranking itself for 2, 10, 40 and 10,000 iterations. Even though the order of nodes changes slightly during the first 100 iterations, as proved in Figure 3-1a, the changes are negligible after 40 iterations.

We tested the algorithm on the polblogs network (see Figure 3-2). 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 labelled as liberal and 443 labelled 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. The randomly picked node is pointed out by an arrow. We then coloured nodes according to their scores following a logarithmic scale (see Figure 3-3). As we can see, colours are consistent with labels. The randomly picked node is actually a liberal blog and most liberal blogs are dark while conservative blogs remain white. When nodes are ranked in decreasing 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.

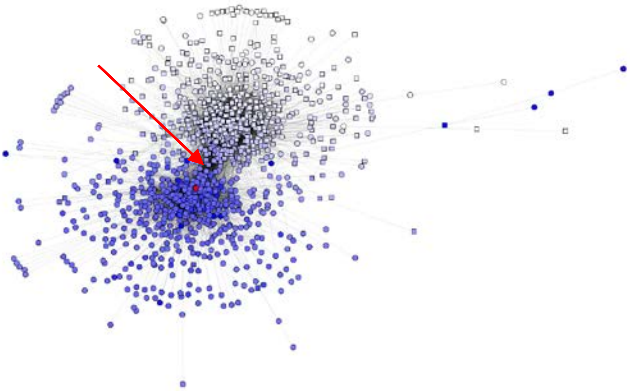
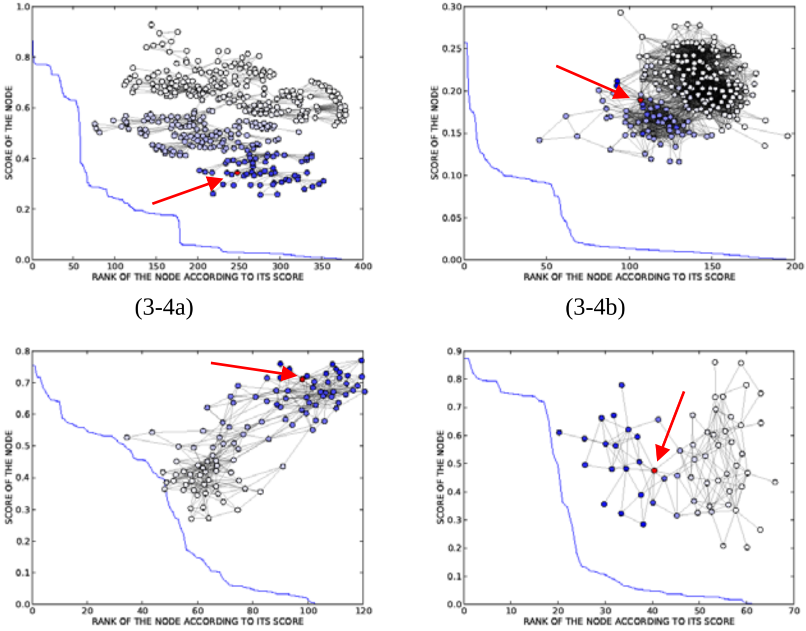


Fig. 3-3. Representation of the polblogs graph with a spring layout (Fruchterman and Reingold 1991).

We applied this technique to smaller networks in order to visualise them more easily. Interesting results have been obtained, as shown in Figure 3-4a which represents 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 graph where three different levels of colour emerge. The succession of plateaus and decreases (on Figures 3-4b, 3-4c and 3-4d) for three other networks also shows how useful the carryover opinion can be in unfolding ego-centred communities.

As we can see in Figure 3-5a, 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 Erdős-Rényi graphs. Each graph contains one hundred nodes with a link probability of 0.3. Two nodes from different Erdős-Rényi graphs have a probability of 0.05 of being linked. The value obtained for the first neighbours of the picked node somewhat dominates the artificially generated community structure, in fact, the neighbours of the picked node have a high score even if they are in different Erdős-Rényi 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 node those two nodes already constitute a community. Actually, the minimal value for a first neighbour with degree d is $1/d$, which makes sense if all other neighbours of this first neighbour are *far away* from the picked node, then this first neighbour is still $1/d$ part of the community(ies) of the picked node.



(3-4a) (3-4b)

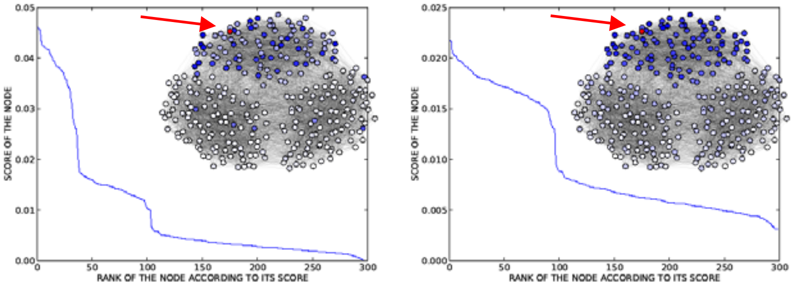
(3-4c) (3-4d)

Fig. 3-4. Results for four small visualisable networks. On the drawing of the networks, arrows point to the selected nodes, while the higher the score, the darker 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. Figure 3-4a: co-authorship network of 379 nodes and 914 edges (Newman 2006). Figure 3-4b: co-appearance network of jazz musicians of 198 nodes and 5,484 edges (Gleiser and Danon 2003). Figure 3-4c: citation network of political books of 105 nodes and 441 vertices (Krebs). Figure 3-4d: social network of dolphins of 62 nodes and 159 edges (Lusseau et al. 2003).

This effect (due to the communities of two nodes) can, however, be easily eliminated, as shown in Figure 3-5b, 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 neighbours. This affects only the first neighbours and it is the same as applying the transformation:

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

where S is the carryover opinion of a first neighbour.



(3-5a) (3-5b)

Fig. 3-5. Figure 3-5a shows the results for three Erdős-Rényi graphs (100,0.3), while nodes in different Erdős-Rényi graphs are linked with probability 0.05. Figure 3-5b 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 neighbours, i.e. a final averaging step is performed without the picked node. The higher the score, the darker the node.

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 3-5a, the distance effect sometimes dominates the redundancy effect. We argue that this is because the carryover opinion considers a pair of linked nodes as a community. The question to answer is how, or if, this affects the results for the nodes at distance two or more. To investigate this, we compared the decrease of the carryover opinion as a function of the distance for the Wikipedia network (choosing the page “boxing”) and an Erdős-Rényi graph of the same average degree. As shown in Figure 3-6, while on the Erdős-Rényi graph the decrease is exponential, on the Wikipedia network only the neighbours 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 suggest solves this problem.

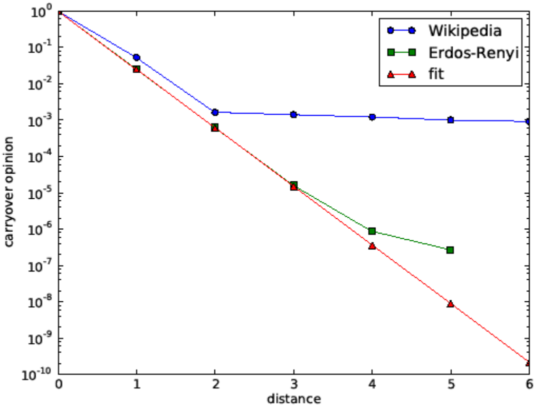
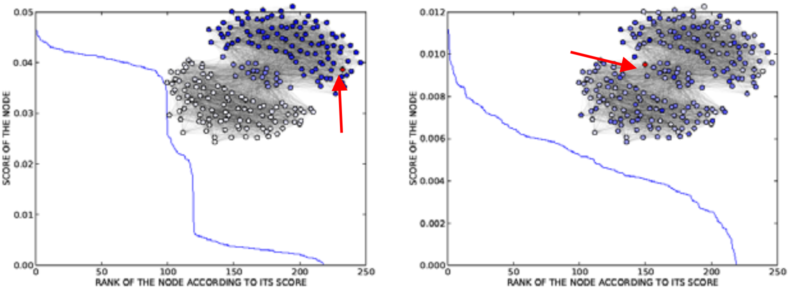


Fig. 3-6. These plots show the average carryover opinion for nodes at a given distance from the node of interest as a function of the distance. Wikipedia is for the Wikipedia network containing n=2,070,367 nodes and e=42,336,614 edges. Erdős-Rényi is for an Erdős-Rényi graph containing this same number of edges and nodes. Fit represents the curve $\frac{1}{degree^{distance}}$ where $degree$ is set to the average degree of the previous graph, i.e. $degree = \frac{2e}{n} = 40$.



(3-7a) (3-7b)

Fig. 3-7. Results given by the carryover opinion with the correcting step for two overlapping Erdős-Rényi graphs of 110 nodes with an edge probability of 0.3 overlapping on 20 nodes. The higher the score, the darker the node. As we can see in Figure 3-7a, when the picked node is at the centre of a community the plateau-decreases structure is clear, while it is not when the node is peripheral (Figure 3-7b).

Such an ideal structure of plateaus and strong decreases (as seen in Figures 3-4 and 3-5) does not always appear. The shape of the curve depends on two things:

1. The position of the picked node, i.e. central in a community or peripheral and thus within several communities. As shown in Figure 3-7, when the node is central the plateaus are clear while when the node is peripheral, no plateau emerges.
2. The structure of the community itself, i.e. whether that community is well defined or not, as we can see in Figure 3-8.

Ego-centred Communities: Results on Large Graphs

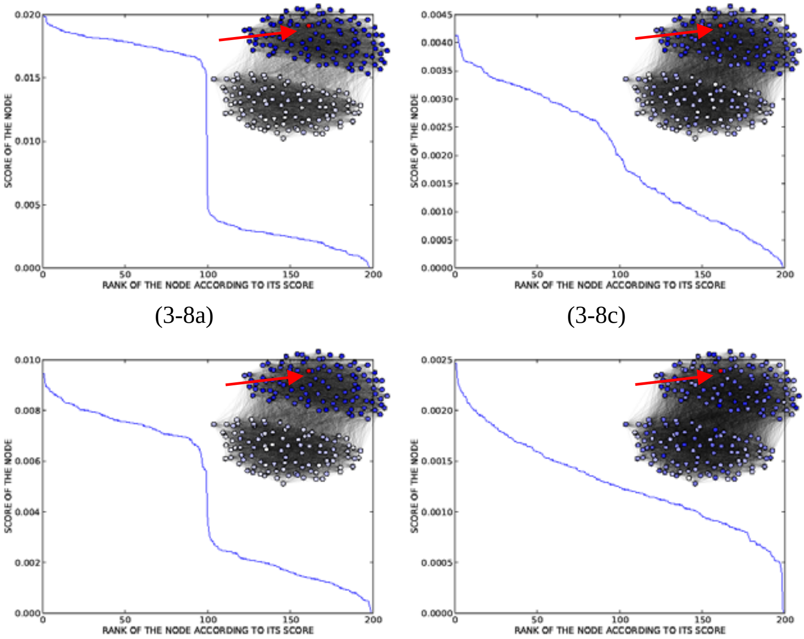
The technique presented above does not require any *a priori* input parameter (other than the graph) and is very time-efficient. It can thus be used in very large graphs to find "the community" or "the communities" of a node if there is one. However, as already discussed, a node often belongs to numerous communities and such a succession of plateaus and decreases is only occasionally observed.

Given randomly chosen nodes from the Wikipedia network, Figure 3-9a (resp. 3-9b) shows four 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 illustrate the four major trends one can obtain: sharp transition, smooth transition, deformed power-law and perfect power-law.

These four very different types of curves reflect the very different structural properties of the nodes. Let us first notice that the correcting step does not significantly modify the curves, the bias due to communities of two nodes is thus minimal here. This may actually mean that there are only few weak ties (i.e. links between very different communities) in the Wikipedia network. Let us explain these four behaviours by analysing the curves and the ranking of pages without the correcting step:

1. The "sharp transition" curve corresponds to the "Cotton Township, Switzerland County, Indiana" page. The first six nodes constitute a plateau. These nodes correspond to the page "Switzerland County, Indiana" and the five other townships of Switzerland County. Then, we withstand a decrease on the next seven 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 has a total of 1,005 townships). The next decrease after about 1,000 nodes is composed of nodes related to townships and Indiana and also a little about Illinois, while the following plateau

after 1,000 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 distantly related contexts, passing through Indiana related topics to Ohio townships, Michigan townships, other states' townships, US related topics...

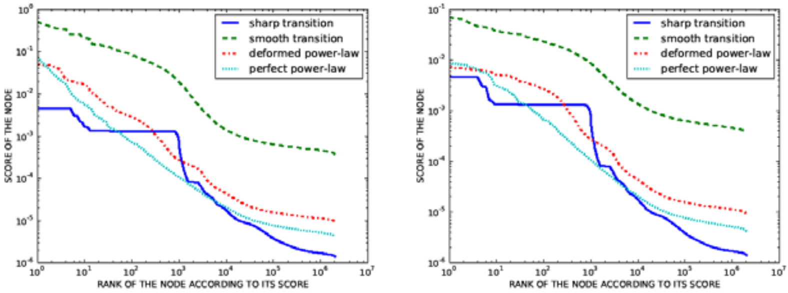


(3-8a) (3-8c)

(3-8b) (3-8d)

Fig. 3-8. Results given by the carryover opinion with the correcting step for two Erdős-Rényi graphs (100, 0.5). In Figure 3-8a (resp. 3-8b, 3-8c, 3-8d) two nodes in different Erdős-Rényi graphs are linked with probability 0.1 (resp. 0.2, 0.3, 0.4).

2. The "smooth transition" curve is obtained for the "Mafia" page. This node can characterise a community by itself. The first thousand pages are Mafiosi names or topics related to organised crime. However, this community is more fuzzily defined than the communities of "Cotton Township, Switzerland County, Indiana".



(3-9a) (3-9b)

Fig. 3-9. Plots of the carryover opinion of all nodes as a function of their ranking for four randomly picked nodes in the Wikipedia network (Figure 3-9a), and the same plots with the correcting step (Figure 3-9b). 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.

3. The "deformed power-law" curve results from the "Mi-Hyun Kim" page. This page is mainly linked to pages about golf and Korea. The first thousand pages are related to one or both topics, and 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 both topics/communities (Korea and golf).
4. The "perfect power-law" curve is obtained for the "JNCO" page, which is a clothing brand. The plot is a perfect power-law which 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 "Los Angeles", "Jeans", "Hip-hop", "J.C. Penney", "Graffiti", "Kangaroo", "Boxing" and "Nu Metal" pages, from which hardly any context can emerge.

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

1. Well-defined communities, such as Switzerland County's or Indiana's communities.
2. Fuzzily defined communities, such as the Mafia's community.

Moreover, these communities can be multi-scale: Switzerland County 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 that can, by themselves, define a community such as "Cotton Township, Switzerland County, Indiana" or "Mafia".
2. Nodes that are in the middle of very few communities, such as "Mi-Hyun Kim".
3. Nodes that are in the middle of a large number of communities, such as "JNCO".

For a given node, these features can all be deduced from the shape of the curve representing their carryover opinion as a function of the ranking.

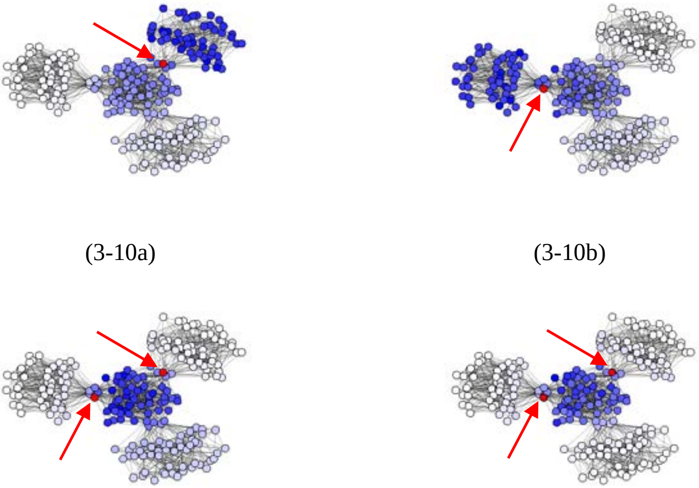
A New Vision of Communities

Multi-Ego-centred 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 believe that a well-chosen small set of nodes could define a single community.

The question is: how may the communities shared by a set of nodes be unfolded? We suggest using the previously established proximity measure. The idea is that a node belonging to both a community of $node_1$ AND a community of $node_2$ has to be somewhat similar to $node_1$ AND to $node_2$. The following example in Figure 3-10 shows how to proceed:

1. For all nodes, evaluate the proximity to node1 and to node2.
2. The proximity to the set $\{node_1, node_2\}$ is then given by the minimum, or by the geometric mean of the similarities to $node_1$ and the similarities to $node_2$. This quantity measures to what extent a node is close to $node_1$ AND $node_2$.

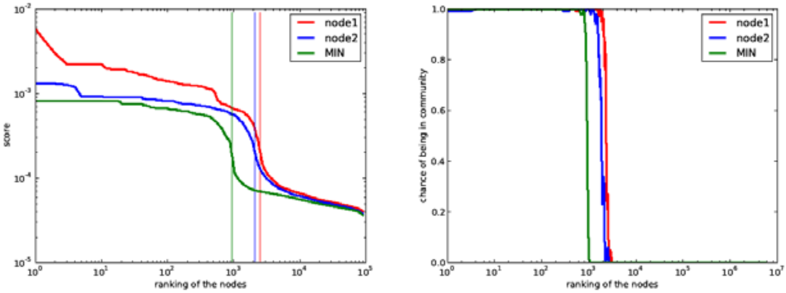


(3-10a) (3-10b)

(3-10c) (3-10d)

Fig. 3-10. Results for four overlapping Erdős-Rényi graphs of 50 nodes and an edge probability of 0.2 overlapping on five nodes. The darker a node, the higher its score. Arrows point to selected nodes. Figure 3-10c (resp. figure 3-10d) gives the (rescaled) minimum (resp. geometric mean) of the scores in the experiments presented in Figures 3-10a and 3-10b. The community shared by both red nodes is emerging.

The method is easily generalisable to a set of more than two nodes. To validate the technique presented here, we extensively tested it and obtained good results on various homemade visualisable networks and also on the LF benchmark for overlapping communities (Lancichinetti and Fortunato 2009). We present here the results for a particular trial on the benchmark. We built a network of 100,000 nodes with 10,000 nodes belonging to three communities and the others belonging to only one community. We used a mixing parameter of 0.2 and kept default values of power-law coefficients for the degrees distribution and communities' sizes distribution. We picked two nodes, each belonging to three communities and each sharing one community in common. The results are presented in Figure 3-11. As we can see, the unions of the three communities for both nodes is identified almost perfectly, as is the community shared by both nodes. Indeed, the Jaccard coefficient between the real communities and the one unfolded by the framework is always greater than 0.9.



(3-11a) (3-11b)

Fig. 3-11. Figure 3-11a shows the carryover opinion of all nodes as a function of their ranking for the two nodes having three communities while sharing one (node1 and node2). It also shows the minimum of these two scores for all nodes as a function of the ranking (MIN). The highest slope of each curve is identified by a vertical bar. Figure 3-11b shows the proportion of nodes (on a sliding window containing 100 nodes) in one of the three communities, as well as the proportion of nodes in the shared community, as a function of the same rankings. We can see that the highest slopes correspond to the transition: “in the community/out of the community”.

Multi-Ego-centred Communities: Results on Large Graphs

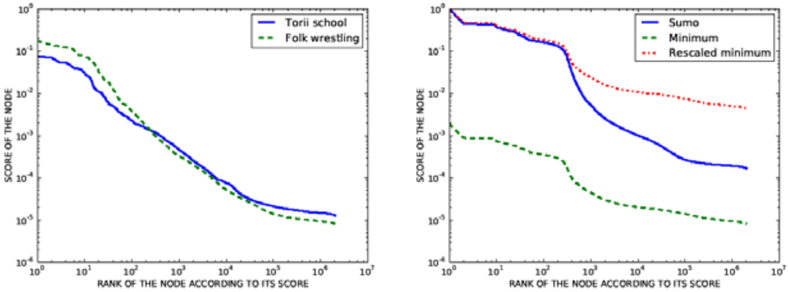
We applied the framework described above to the Wikipedia network using the minimum proximity of the picked nodes. Figure 3-12a shows the results for two nodes: “Folk wrestling” and “Torii school”. One is dedicated to the various types of traditional wrestling around the world, while the other one is dedicated to a traditional Japanese art school. Both curves are slightly deformed power-laws and do not show any community.

Figure 3-12b shows the results 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 one.

The two curves have exactly the same structure: a plateau followed by a decrease at about the 350th node. “Folk wrestling” and “Torii school” are related to “Sumo” in a transversal way. Keeping the minimum of the scores for these two pages shows how nodes are related to “Folk wrestling” and “Torii school” which actually correspond to “Sumo”. Comparing the 350 first nodes of each experiment 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”.



(3-12a) (3-12b)

Fig. 3-12. Figure 3-12a shows the results for two nodes, “Folk wrestling” and “Torii school”: two power-laws. Figure 3-12b 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 one.

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

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

How to Find All Ego-centred Communities of a Given Node

In this section we propose an approach to find all ego-centred communities of a given node, by finding multi-ego-centred communities of the node of interest and some other candidates. We show the results of our method when applied to a real large graph, that is, the whole Wikipedia network containing more than 2 million labelled pages and 40 million edge hyperlinks (Palla et al. 2008).

Framework

Given a specific node u, we measure the proximity[2] of all nodes in the graph to u and then try to find irregularities in the decrease of these proximity values, as explained in the previous sections. Such irregularities can reflect the presence of one or more communities. However, this routine often leads to a power-law with no plateau and from which no scale can be extracted; this happens when lots of communities of various sizes overlap, which is often the case. To cope with this problem, we use the notion of *multi-ego-centred community* (in particular, a *bi-ego-centred community*), i.e. centred on a set of nodes instead of a single node. We thus need to intelligently pick another node, v, evaluate the proximity of all nodes in the graph to v, and then, for each node in the graph, compute the minimum of the score obtained from u and the score obtained from v. This minimum evaluates to what extent a node is similar to u AND v. Note that doing this sometimes leads to the identification of a community that does not contain u and/or v, however, since we are interested only in communities containing u, we use v as an artifact and keep a community only if it contains u, regardless of v. The framework consists in doing this for enough candidate nodes v in order to obtain all communities of u. We will now detail the steps of the framework.

Choice of Candidates for v

First, the carryover opinion of node u has to be computed, providing the value of each node's proximity to u. The carryover curve is obtained by sorting the obtained values and plotting them as a function of their ranking.

[2] Even though other proximity measures can be used, we use the carryover opinion.

If the outcome is a power-law, there is no relevant scale and u certainly belongs to several communities of various sizes.

The goal is then to pick a node v such that v and u share exactly one community. This is very unlikely if v is very dissimilar from u. Computing the minimum of the scores obtained from u's and v's carryover opinion will lead to very small values. Indeed, if the two nodes share no community, at least one of the scores will be very low. Conversely, if v is extremely similar to u then the two nodes will share many communities. The carryover opinion values obtained from u and v will be roughly the same and doing the minimum will not give more information. No single community can be isolated in this case.

Thus, v must be similar enough to u, but not too similar. Its score in u's carryover should be neither too high nor too low. A low and high proximity threshold can be manually tuned to select all nodes at the right distance in order to quicken the execution.

It is quite likely that many of these nodes at the right distance will lead to the identification of the same community. Therefore, not all of them need to be candidates; a random selection can be performed if the running time of the algorithm matters. More precise selection strategies will be discussed in the future work section.

Identification of the Ego-centred Community of u and v

In order to identify the potential community centred on both u and v, we must compute the minimum of the carryover values obtained from u and from v for each node, w, of the graph. The minimum value of both scores is used to measure the belonging of w to the community of (u and v). We sort these minimum values and plot the minimum carryover curve. Once again, an irregularity in the decrease, i.e. a plateau followed by a strong decrease, indicates that all nodes before the decrease constitute a community of (u and v).

The automated detection of this plateau/strong decrease pattern can be done by searching for the maximum slope and keeping the outcome if the slope is larger than a given threshold. This threshold should be manually tuned. If there are several sharp decreases, we currently only detect the sharpest. This could be improved in the future.

If a plateau/strong decrease pattern is detected, several situations may then occur:

- u and v are before the decrease: a community of both nodes has been identified.

- u is before the decrease and v is after: v helped to identify a specific community of u even if v does not belong to it.
- u is after the decrease and v is before: a community of v has been identified but we are only interested in communities of u so the community is not kept.
- u and v are after the decrease: a community has been identified but again the community is not kept. This can happen, for instance, if there is a small community at the intersection of u's community and v's community.

As such, this method is not very efficient if u is a very high degree node and is connected to a very large number of communities. In that case, u's carryover will be high for every node in the graph. Calculating the minimum with the scores obtained from a less popular node (with lower scores) will simply result in the values obtained with this second node. A rescaling before doing the minimum can fix the problem. Indeed, as the lowest values obtained by running the carryover opinion result in a plateau, rescaling (in logarithmic scale) the values such that these plateaus are at the same level solves this problem.

Cleaning the Output and Labelling the Communities

The output of the two previous steps is a set of communities (where each node is scored), since each candidate node v can yield a community. These communities need to be post-processed, since many of them are very similar.

We propose computing the Jaccard similarity[3] (or any other similarity measure between sets) between every two pairs of communities to identify redundancies. If the similarity value is very high, we only keep the intersection of both. For each node in this new (intersection) community, the score is the sum of the scores in the original communities.

An additional optional cleaning step can enhance the results: if a community is dissimilar to all other communities, we remove it. Indeed, a "good" community should appear for several candidate nodes. We observed that such communities come from the detection of a plateau/decrease structure that does not exist in reality (this may happen if the threshold is too low). Note, however, that if u is in or around a large community, we have a high probability of unfolding it, and this probability increases with

[3] For two sets A and B, the Jaccard similarity is $Jac(A,B) = \frac{|A \cap B|}{|A \cup B|}$.

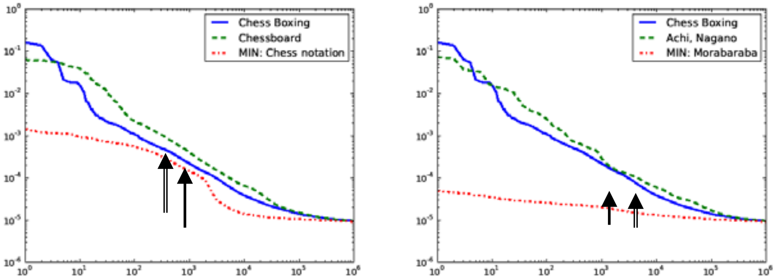
the size of the community. If very large communities exist, the algorithm can have some difficulty in unfolding other small communities. We will come back to that problem in the future work section.

Finally, we label each remaining community with the label of its best ranked node, i.e. the node whose score is the highest. If two communities have the same label we suggest keeping both (they can be different scales of the same community).

This algorithm finally returns a set of distinct, labelled communities. We will now show some results obtained on a real network.

Results and Validation

In this section we will show the results obtained when node u is the Wikipedia page entitled “Chess Boxing”[4]. This page exhibits good results which are easily interpretable and can be validated by hand.



(3-13a) (3-13b)

Fig. 3-13. Each figure shows the curves corresponding to a trial. The y axis represents the scores and the x axis represents the ranking of the nodes according to their scores. The first (resp. second) curve is the carryover opinion run from the “Chess Boxing” node (resp. a candidate for v, the legend shows the label of the candidate), while the third curve shows the minimum, the label of the first ranked node is in the legend. The first trial is successful, while the second is not (no plateau/decrease structure). The double arrow points to the “Chess Boxing” node, while the simple arrow indicates the sharpest slope.

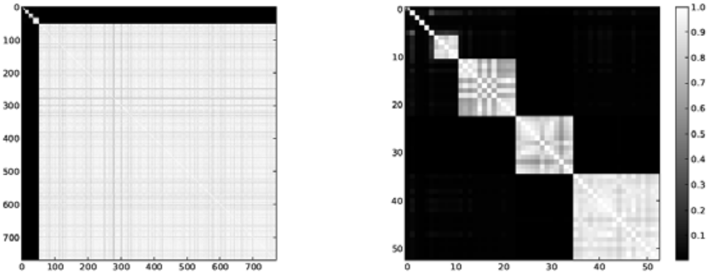
For the “Chess Boxing” node (u), the algorithm iterated over 3,000 nodes (v) chosen at random from the nodes between the 100th and the

[4] Chess boxing is a sport mixing chess and boxing in alternated rounds.

10,000th best ranked nodes, leading to 770 groups of nodes. Figure 3-13 shows a successful trial leading to the identification of a group and an unsuccessful trial.

Figures 3-14a shows the Jaccard similarity matrix of the 770 unfolded communities before cleaning. The columns and lines of the matrix have been rearranged so that columns corresponding to similar groups are close to each other. We see that there are 716 communities very similar to one another, while not similar to other communities (note the big white square in the bottom right corner).

When zooming in on the rest of the matrix (Figure 3-14b) we see four smaller groups of communities and six groups containing only a single community. These are actually mistakes produced by the plateau/decrease detection part of the algorithm and these groups are automatically deleted during the cleaning step.



(3-14a) (3-14b)

Fig. 3-14. Rearranged Jaccard similarity matrix. Figure 3-14b shows a magnification of the top left corner of the matrix.

This decomposition into five main groups (one large and four small) is easily obtained by intersecting similar groups (for this we used a Jaccard similarity threshold of 0.7). The labels and sizes of the five groups are “Enki Bilal” (35 nodes), “Uuno Turhapuro” (26 nodes), “Da Mystery of Chessboxin'” (254 nodes), “Gloria” (55 nodes) and “Queen's Gambit” (1.619 nodes). As we can see, the algorithm identifies groups with very different sizes (from 26 nodes to 1.619 nodes on this example) which is a positive feature since other approaches are quite often limited to small communities.

Some labels are intriguing. However, by checking their meanings on Wikipedia online, all of them can be justified very easily:

1. Enki Bilal is a French cartoonist. Wikipedia indicates that “Bilal wrote [...] *Froid Équateur* [...] acknowledged by the inventor of

chess boxing, Iepe Rubingh as the inspiration for the sport”. The nodes in this group are mostly composed of *Froid Équateur*’s other cartoons.

2. *Uuno Turhapuro* is a Finnish movie. It is also acknowledged as the inspiration for the sport, with a scene “where the hero plays blindfold chess against one person using a hands-free telephone headset while boxing another person”. The nodes in this group are mostly other cartoon characters or actors in the movie or are strongly related to Finnish movies.
3. “Da Mystery of Chessboxin'” is a song by American rap band Wu-Tang Clan. The nodes in the community are related to the band and rap music, which is also relevant.
4. “Gloria” is a page of disambiguation linking to many pages containing Gloria in their title. The current Wikipedia page for “Chess Boxing” contains the sentence: “On April 21, 2006, 400 spectators paid to watch two chess boxing matches in the Gloria Theatre, Cologne”. However, there is no hyperlink to the page “Gloria Theatre, Cologne” which is a stub. Looking at the Wikipedia records, we found that a link for the page Gloria was added to the page “Chess Boxing” on 3 May 2006 and then removed on 31 January 2008. Due to the central nature of the “Gloria” page within the Gloria community, “Chess Boxing” was part of the Gloria community between these two dates, i.e. when the dataset was compiled.
5. Finally, “Queen's Gambit” is a famous chess opening move. This is consistent with the content of the community as it is composed of chess related pages. “Queens' Gambit” is very specific to chess and thus characterises this community very well.

Surprisingly, the algorithm did not find any community related to boxing. However, the Wikipedia page “Chess Boxing” explains that most chess boxers come from a chess background and learn boxing afterward. They might be important within the community of chess, but less so within the boxing community. Therefore, this could explain why the “Chess Boxing” node lies within the community of chess, but is at the limit of the boxing community.

Comparison to Another Approach

As stated in the related work section, there are other methods for finding ego-centred communities, all of them based on the optimisation of a quality function. We have compared our results to the approach proposed by (Ngonmang et al. 2012) which, we believe, is the most advanced approach since it corrects many of the drawbacks of previous methods.

Quality function techniques, due to the non-convexity of the optimisation problem, often lead to small communities, while our approach does not suffer from this drawback. We can indeed check this on the previous example for which the approach of (Ngonmang et al. 2012) finds only two small communities:

1. The first one contains seven nodes: Comic book, Enki Bilal, Cartoonist, La Foire aux immortels, La Femme Piège, Froid équateur and Chess boxing. This community is strikingly similar to our community labelled “Enki Bilal” and is very relevant.
2. The second one contains five nodes: Germany, Netherlands, 1991, International Arctic Science Committee and Chess boxing. This second community is not similar to any of the communities we found and we could not find its meaning.

Conclusion and Perspectives

While studying the global overlapping structure of a real-world network is too complex, studying its community structure as a partition is too restrictive. The local overlapping structure around a node (ego-centred community structure) is a good compromise between simplicity and realism. Trying to unfold ego-centred communities by optimising a quality function often leads to poor results because the optimisation landscape is highly non-convex and the optimisation often ends up in local minima. In this chapter, we have suggested looking for irregularities in the decrease of a proximity measure to avoid this problem. We have suggested a new proximity measure called the carryover opinion. It has good properties for this application: it is fast to compute, not restrictive and parameter-free. Note, however, that our framework may be used with other proximity measures.

This proximity shows how likely it is for two nodes to share at least one community. It also allows us to see whether a node characterises a community by itself (a plateau/decrease structure), is in the middle of a few communities (wavy power-law) or is in the middle of many communities (quasi-perfect power-law). In large graphs, the decrease of the carryover opinion often follows a scale-free law because a node often belongs to many overlapping communities, fuzzily defined and of different sizes. In this case, no scale can be extracted from the measure and this first approach is limited.

To cope with this limitation we introduced the concept of *multi-ego-centred communities*. While a node often belongs to many communities, a well-chosen small set of nodes can characterise a single community. Following this idea, we introduced an algorithm which, given a node, finds all communities centred on that node. Contrary to other existing algorithms, ours avoids local minima, finds communities of various sizes and densities, and also allows labelling of the obtained communities. This algorithm is time efficient and works with very large graphs. We validated the results on toy graphs, benchmarks and a real-world very large graph extracted from Wikipedia.

Still, some features of the algorithm can be improved. For instance, the detection of irregularities only returns the sharpest decrease. It would be good to find all relevant irregularities, which would provide multi-scale communities.

Furthermore, the algorithm currently only looks for bi-centred communities, but some communities might appear only when centred on three or more nodes. It would be interesting to incorporate this feature. However, it will increase the running time of the algorithm, especially because of unsuccessful trials. More advanced selection of candidates thus needs to be developed. We could, for instance, add the following criterion: if a candidate is chosen for v, nodes very similar to this candidate might be neglected since they would probably lead to the same result. The speed of the algorithm is indeed a very important feature and is central to making it practical for the study of evolving communities.

The algorithm can have some difficulty in finding very small communities if there exist very large ones around the node of interest. This might be the reason why, when applied on a globally popular node such as “Biology” or “Europe” in the Wikipedia network, the algorithm only returns one very big community, while we expect the communities of various sub-fields of biology or European country related topics. Two directions should be investigated to improve this: re-launching the algorithm again on the sub-graph induced by the nodes of the large community, or removing the nodes belonging to the big community from the graph and running the algorithm again.

In this book chapter we have mainly focused on a single application of the concept of multi-ego-centred communities, that is, finding all ego-centred communities of a node by unfolding its multi-ego-centred communities using well-chosen candidates. At least two other straightforward applications of multi-ego-centred communities are currently under investigation: (i) unfolding all nodes of a community given only some of its members and (ii) unfolding all (overlapping) communities of a network by unfolding multi-ego-centred communities of many small sets of nodes. In the long term, this notion of multi-ego-centred community could also help the study of communities in evolving networks. Finally, the definition of weighted-multi-ego-centred communities (potentially with negative weights) may also enhance this technique.

Acknowledgments

The authors are supported in part by the French National Research Agency contract DynGraph ANR-10-JCJC-0202. The authors would also like to thank Daniel Bernardès, Sergey Kirgizov, and Lionel Tabourier for helpful discussions.

Bibliography

Adamic, L.A. and Glance, N. “The political blogosphere and the 2004 US election: divided they blog”. Proceedings of the 3rd international workshop on Link discovery. 36-43. 2005.

Ahn, Y.-Y. and Bagrow, J.P. and Lehmann, S. “Link communities reveal multiscale complexity in networks”. Nature, 466, 7307, 761-764, 2010.

Bagrow, J.P. “Evaluating local community methods in networks”. Journal of Statistical Mechanics: Theory and Experiment. 2008, 05, P05001, 2008.

Blondel, V.D. and Guillaume, J.-L. and Lambiotte, R. and Lefebvre, E. “Fast unfolding of communities in large networks”. Journal of Statistical Mechanics: Theory and Experiment, 10, P10008, 2008.

Chen, J. and Zaiane, O. and Goebel, R. “Local community identification in social networks”. Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in. 237-242, 2009.

Clauset, A. “Finding local community structure in networks”, Physical Review E, 72, 2, 026132, 2005.

Danisch, M. and Guillaume, J.-L. and Le Grand, B. “Towards multi-ego-centered communities: a node similarity approach”. Int. J. of Web Based Communities. 2012.

Danisch, M. and Guillaume, J.-L. and Le Grand, B. “Unfolding ego-centered community structures with a similarity approach”. CompleNet 2013, 145-153, Berlin.

Evans, T.S. and Lambiotte, R. “Line graphs, link partitions, and overlapping communities”. Physical Review E, 80, 1, 016105, 2009.

Fortunato, S. “Community detection in graphs”. Physics Reports. 486, 3 75-174, 2010.

Fruchterman, T.M.J. and Reingold, E.M. Graph drawing by force-directed placement. Software: Practice and experience. 21, 11, 112--1164, 1991.

Girvan, M. and Newman, M.E.J. “Community structure in social and biological networks”. Proceedings of the National Academy of Sciences. 99, 12, 7821-7826, 2002.

Gleiser, P. and Danon, L. Community structure in jazz. arXiv preprint cond-mat/0307434. 2003.

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

Lancichinetti, A. and Fortunato, S. “Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities”. Physical Review E, 80, 1, 016118, 2009.

Luo, F. and Wang, J.Z. and Promislow, E. “Exploring local community structures in large networks”. Web Intelligence and Agent Systems. 6, 4, 387-400, 2008.

Lusseau, D. and Schneider, K. and Boisseau, O.J. and Haase, P. and Slooten, E. and Dawson, S.M. “Finding community structure in networks using the eigenvectors of matrices”. Behavioral Ecology and Sociobiology. 34, 4, 396-405, 2003.

Moraresc, I.-C. and Girard, A. “Opinion dynamics with decaying confidence: application to community detection in graphs”. Automatic Control, IEEE Transactions on. 56, 8, 1862-1873, 2011.

Newman, M.E.J. “Finding community structure in networks using the eigenvectors of matrices”. Physical review E 74, 3, 036104, 2006.

Ngonmang, B. and Tchuente, M. and Viennet, E. “Local community identification in social networks”. Parallel Processing Letters. 22, 01, 2012.

Norris, J. R. “Markov chains”. 17, 1997.

Page, L. and Brin, S. and Motwani, R. and Winograd, T. “The PageRank citation ranking: bringing order to the web”. 1999.

Palla, G. and Derényi, I. and Farkas, I. and Vicsek, T. “Uncovering the overlapping community structure of complex networks in nature and society”. Nature, 435, 7043, 814-818, 2005.

Palla, G. and Farkas, I.J. and Pollner, P. and Derényi, I. and Vicsek, T. "Fundamental statistical features and self-similar properties of tagged networks". New Journal of Physics, 10, 12, 123026, 2008.

Pons, P. and Latapy, M. "Computing communities in large networks using random walks". Computer and Information Sciences-ISCIS 2005, 284-293, 2005.

Rosvall, M. and Bergstrom, C.T. "Maps of random walks on complex networks reveal community structure". Proceedings of the National Academy of Sciences. 105, 4, 1118-1123, 2008.

Sozio, M. and Gionis, A. "The community-search problem and how to plan a successful cocktail party". Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. 939-948, 2010.

Wang, Q. and Fleury, E. "Uncovering Overlapping Community Structure". Complex Networks, 176-186, 2011.