Egomunities, Exploring Socially Cohesive Person-based Communities

Adrien Friggeri, Guillaume Chelius, Eric Fleury

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Egomunities
Exploring Socially Cohesive Person-based Communities

Adrien Friggeri — Guillaume Chelius — Eric Fleury

N° 7535 — version 2

initial version 11 February 2011 — revised version 18 February 2011

ISSN 0249-6399 ISRN INRIA/RR--7535--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE - RHÔNE-ALPES

Egomunities
Exploring Socially Cohesive Person-based Communities

Adrien Friggeri , Guillaume Chelius , Eric Fleury

Thème : Réseaux et télécommunications
Équipe-Projet DNET

Rapport de recherche n° 7535 — version 2 — initial version 11 February 2011 — revised version 18 February 2011 — 19 pages

Abstract:

The last years, there has been a great interest in detecting overlapping communities in complex networks, which is understood as dense groups of nodes featuring a low outbound density. To date, most methods used to uncover overlapping communities stem from the field of disjoint community detection by attempting to decompose the whole network into several possibly overlapping groups of nodes. In this article we take an orthogonal approach by introducing a novel point of view to the problem of overlapping communities, namely the concept of egomunities, which are subjective communities centered around a given node, more precisely inside its neighborhood. In order to construct those egomunities, we propose a general metric on graphs, the cohesion, inspired by sociological considerations. The cohesion quantifies the community-ness of one given set of nodes, based on the notions of weak ties and triangles – triplets of pairwise connected nodes, instead of the classical view using only edge density. A set of nodes has a high cohesion if it features a high density of triangles and intersects few triangles with the rest of the network. We build upon the cohesion to construct a heuristic algorithm which uncovers egomunities of a given node by attempting to maximize their cohesion. We illustrate the pertinence of our method with a detailed description of one person's egomunities among Facebook friends and promising results from an ongoing large scale Facebook experiment. We finally conclude by describing promising applications of egomunities such as information inference and interest recommendations, and present a possible extension to cohesion in the case of weighted networks.

Key-words: social networks, complex networks, real-world graphs, community detection, overlapping communities, data mining, modelisation

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Egomunautés
Exploration de communautés socialement cohésives et personne-centrées

Résumé :

Ces dernières années, l'intérêt pour la détection de communautés recouvrante dans les réseaux réels s'est intensifié. Celles ci sont des groupes de nœuds possédant une forte densité interne et présentant une densité faible vers le reste du réseau. À ce jour, la majorité des méthodes utilisées pour calculer de telles communautés héritent de celles développées dans le domaine de la détection de communautés disjointes, ou bien en étendant le concept de modularité à un contexte recouvrant, ou bien en essayant de décomposer le réseau entier en plusieurs sous ensembles éventuellement recouvrant. Dans ce rapport, nous abordons la question de manière orthogonale en introduisant une mesure, la cohésion, reposant sur des considérations sociologiques. La cohésion permet de quantifier l'aspect communautaire d'un ensemble de nœuds à partir des notions de triangles – triplets de nœuds interconnectés – et de liens faibles, au lieu de la vision classique utilisant des arêtes. En substance, nous introduisons une caractérisation numérique des communautés: des ensembles de nœuds possédant une cohésion élevée. Nous présentons ensuite une nouvelle approche au problème des communautés recouvrantes en introduisant le concept d'ego-munauté: des communautés subjectives centrées sur un nœud donné, précisément incluses dans son voisinage. Nous utilisons la cohésion pour élaborer un algorithme heuristique construisant les ego-munautés d'un nœud en tentant de maximiser leur cohésion. Finalement, nous présentons des résultats préliminaires, sous la forme d'une description détaillé des ego-munautés d'amis Facebook d'une personne. Nous concluons en décrivant des applications prometteuses des ego-munautés, par exemple l'inférence d'information sur le sujet ou la recommandation de centre d'intérêt, et présentons une extension possible à la cohésion dans le cas de réseaux pondérés.

Mots-clés : réseaux sociaux, réseaux complexes, graphes réels, détection de communautés, communautés recouvrantes, data mining, modélisation

1 Introduction

Although community detection has drawn tremendous amount of attention across the sciences in the past decades, no formal consensus has been reached on the very nature of what qualifies a community as such. In addition to the contributions of sociology, several propositions have also emerged from the physics and computer science communities [2, 4]. Despite the lack of globally accepted analytical definition, all authors concur on the intuitive notion that a community is a relatively dense group of nodes which somehow features less links to the rest of the network. Unfortunately, this agreement does not extend to the specific formal meanings of *dense* and *less links*.

However, the past few years have witnessed a paradigm shift as the idea of defining the nature of communities was progressively left aside. It has become apparent, and widely accepted that it suffices to compare several sets of communities and choose the *best* obtained division – relative to a given metric – in order to detect communities.

The metric the most used to that effect is Newman's Q-modularity [9], which compares the density of links inside a given community to what would be expected if edges where distributed randomly across the network (null model). This method has proven to give sensible results on several networks and gained traction in the *communities* community. Since maximizing the Q-modularity on general graphs is an established NP-hard problem, several heuristics have been proposed [1, 5, 12].

Most approaches were mainly focused on partitioning a network, leading to non overlapping communities (each node belonging to one and only one group). In the recent years, there has been a growing interest in the study of overlapping communities; a distribution of the nodes across different groups which reflect more precisely what one might expect intuitively, namely that a given node might belong to different communities – for example, in a social network, an individual might simultaneously belong to a family, a friends group and coworkers groups.

Due to the historical evolution of the field, to this day, most methods used to detect overlapping communities are inspired by, or adapted from, existing counterparts for disjoint community detection. If some of those methods take a literal approach to the issue and are built upon extensions to the modularity [8, 11], others have taken another path, such as clique percolation [10]. However, we assess that all those methods aim at finding all communities in a network.

In this article, we propose to take a step back and focus on a specific type of a user-centric communities, which we call *egomunities*. Those egomunities are overlapping communities contained in the neighborhood of one given node. In order to detect those egomunities, we introduce a graph metric, the *cohesion*, upon which we construct a heuristic algorithm. Drawing inspiration from well established sociological results, the *cohesion* is based on the notions of weak ties and triangles – triplets of pairwise connected node – instead of the classical view that uses edges to rate the *communityness* of one given set of nodes. Preliminary yet promising results from a large scale ongoing Facebook experiment prove that the cohesion accurately captures the quality of a community.

It is important to note that whereas the Q-modularity gives a score to a partition of a network, the *cohesion* serves as an intrinsic characterization of a subgraph embedded in a network, independently of any other subgraphs. As

such, we propose a definition of a community, namely a set of nodes with high cohesion. Moreover, even though the cohesion is a generic metric on subgraphs, it was primarily conceived to characterize social communities. Its inception relies on social considerations which formal extension to other types of network is beyond the scope of this report. Therefore, we make no claim towards or against its pertinence when used in the context of networks representing non social data.

This paper is organized as follows: in Section 2 we describe the construction of a new metric, the *cohesion*, to evaluate the *communityness* of a set of nodes. In Section 3 we present a user-centric way of thinking about communities: *ego*munities and introduce an algorithm, relying on *cohesion*, which computes those. In Section 4 we first present the egomunities of Facebook friends of a test subject and evoke preliminary results of an ongoing large scale experiment. Finally we highlight several applications and extensions.

2 Cohesion

Before delving into technicalities and formal definitions, we consider important to take a moment to reflect on the idea of community detection and highlight inherent differences between the problems of disjoint versus overlapping communities. We assess that the evaluation of the quality of a given set of communities in a network mainly boils down to the two following questions:

- **boundaries** : does the set of communities makes sense as a whole ?
- **inside** : is each community intrinsically sound ?

The main difference between disjoint and overlapping communities problems is that in the latter a node can belong to several communities. Although seemingly mundane, in the disjoint case this has for effect that "*belonging to the same community*" is an equivalence relation on nodes. As a consequence of this relation's transitivity, the two aforementioned questions are deeply linked when partitioning a network into communities.

This is actually the main idea behind Q-modularity, defined as follow: $Q = \operatorname{Tr}\mathbf{e} - ||\mathbf{e}^2||$ where $\operatorname{Tr}\mathbf{e}$ is the trace of the matrix $\mathbf{e}$, in which $\mathbf{e}_{i,j}$ represents the density of links going from community i to community j. Q increases when the communities are dense (*i.e.* are intrinsically sound) and decreases in presence of links between communities (*i.e.* when boundaries between communities are not well defined). In the case of disjoint communities, optimizing the Q-modularity leads to a balance between intrinsic and extrinsic qualities. Contrast this with the overlapping problem, where those two questions are decoupled as one can modify one community without affecting the others.

2.1 The volatility of boundaries

Of those two questions, within the scope of this paper, we evade the first one for the most part, as we believe that methods to quantify the quality of a set of communities should arise from choices adapted both to the analyzed data and to the type of results one wishes to manipulate.

Consider for example two overlapping cliques. It seems reasonable to consider two communities if the overlap is reduced to a single node, and one big community when the intersection contains all nodes but one in each clique. The intermediate case, however, is more of a gray area (Fig. 1). On the one hand,

Figure 1: Three couple of cliques. On the left, there are clearly two communities, and on the right only one. The middle case is more of a gray area.

it might be legitimate to consider only one community when two sets of nodes feature a high enough overlap. In the field of network visualization, for example, representing sets which intersect greatly each other could lead to visual clutter, rendering the visual output unreadable. On the other hand, there is a case for the opposite strategy, when the resulting communities should be fine grained.

As such, the rating awarded to a set of communities should be tailored on a case by case basis, in order to fit to the type of results which are sought.

2.2 Focus on the inside

It is possible to rate the quality of one given community embedded in a network, independently from the rest of the network. The idea is to give a score to a specific set of nodes describing wether the underlying topology is *community like*. In order to encompass the vastness of the definitions of what a community is, we propose to build such a function, called *cohesion*, upon the three following assumptions:

1. the quality of a given community does not depend on the collateral existence of other communities;
2. nor it is affected by remote nodes of the network;
3. a community is a "dense" set of nodes in which information flows more easily than towards the rest of the network.

The first point is a direct consequence of the previously exhibited dichotomy between content and boundaries. The second one encapsulates an important and often overlooked aspect of communities, namely their locality. A useful example is to consider an individual and his communities; if two people meet in a remote area of the network, this should not ripple up to him and affect his communities.

The last point is by far the most important in the construction of the cohesion. The fundamental principle is linked to the commonly accepted notion that a community is denser on the inside than towards the outside world, with a twist.

In [7], Granovetter defines the notion of *weak ties* as edges connecting acquaintances, and argues that "*[...] social systems lacking in weak ties will be fragmented and incoherent. New ideas will spread slowly, scientific endeavors will be handicapped, and subgroups separated by race, ethnicity, geography, or other characteristics will have difficulty reaching a modus vivendi.*". Furthermore, he states that a "*weak tie [...] becomes not merely a trivial acquaintance tie but rather a crucial bridge between the two densely knit clumps of close friends*". And finally, he assesses that *local bridges* – edges which do not belong to a triangle, that is a set of three pairwise connected nodes – are weak ties. For

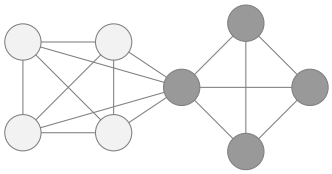

Figure 2: Two communities featuring the same number of links towards the outside world but clearly different from a communityness standpoint.

these reasons, we consider that the structural backbone of communities does not lie solely in the edges of the network, but rather in its triangles.

In Figure 2, two communities are represented in light and dark gray. Both contain the same number of nodes and edges towards the rest of the network. However, although it is sound to dismiss the lighter community as one of bad quality – as it is included in a larger clique – the darker one is what one would expect to be a community. Thus we are confronted with two sets of nodes, featuring the same sizes, inner and outer densities, and yet one is a good community and the other one is not. The difference between the two sets of nodes appears when looking at triangles: the light set features six *outbound* triangles – that is, triangles having an edge inside the community and a point outside – whereas the other set contains no such triangles.

Hence, we contend that the feature to consider when evaluating how well a community's border is defined is not merely the presence of outbound edges, but that of outbound triangles. Finally, we consider important to insist on the fact that this metric does not describe how good is a set of communities but merely the intrinsic quality of one community.

2.3 Definition

Given an undirected network $G = (V, E)$ and $S \in V$, we extend the notion of neighborhood $\mathcal{N}(G, u)$ of a node $u \in V$ to S, $\mathcal{N}(G, S) = \bigcup_{u \in S} \mathcal{N}(G, u) \setminus S$.

We first define two quantities, $\triangle_{\text{in}}(G, S)$ which is the number of triangles of G contained in S and $\triangle_{\text{out}}(G, S)$ the number of triangles "pointing outwards" — that is, triangles of G having two nodes in S and the third one in $\mathcal{N}(G, S)$. We then define the *cohesion* $\mathcal{C}$ of a subset of nodes S of a graph G:

$$\mathcal{C}(G, S) = \frac{\triangle_{\text{in}}(G, S)}{\binom{|S|}{3}} \frac{\triangle_{\text{in}}(G, S)}{\triangle_{\text{in}}(G, S) + \triangle_{\text{out}}(G, S)}$$

The first factor is the *triangular density* of the community, while the second one represents the proportion of triangles having a edge inside the community which are wholly contained by said community. Intuitively, a community has a high cohesion if it is dense in triangles *and* it cuts few outbound triangles. An example is given in Figure 3. If there is no ambiguity on the graph G, we will simplify the notation of $\mathcal{C}(G, S)$ and note it: $\mathcal{C}(S)$.

2.4 Properties

We assimilate the notions of *weak tie* and *local bridges*, and define a weak tie as an edge which does not belong to any triangle. Let $G_{\triangle} = (V, E_{\triangle})$ be the graph obtained by removing all weak ties from G.

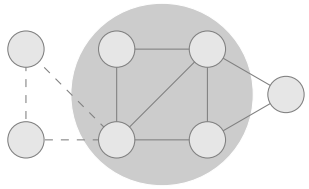

Figure 3: Cohesion of a set of nodes (circled) in a network. $\triangle_{\text{in}} = 2$, $\triangle_{\text{out}} = 1$ (the dashed triangle is not taken into account as it has no edge in the set), therefore $\mathcal{C} = \frac{1}{3}$.

Property 1. *For all $S \subseteq V$, $\mathcal{C}(G, S) = \mathcal{C}(G_{\triangle}, S)$.*

Proof. When removing weak ties, no triangles are added or removed, thus $\triangle_{\text{in}}(G, S) = \triangle_{\text{in}}(G_{\triangle}, S)$ and $\triangle_{\text{out}}(G, S) = \triangle_{\text{out}}(G_{\triangle}, S)$. Therefore, $\mathcal{C}(G, S) = \mathcal{C}(G_{\triangle}, S)$. □

This echoes the argument exposed in Section 2.2: as weak ties serve only as links between communities, removing them from the network does not affect communities quality.

Property 2. *Let $S \subseteq V$ and $S' \subseteq V$ be two disconnected sets of nodes ($\nexists e = (u, v) \in E$ s.t. $u \in S$ and $v \in S'$). If $\mathcal{C}(S) < \mathcal{C}(S \cup S')$ then $\mathcal{C}(S') \geq \mathcal{C}(S \cup S')$.*

Proof. Suppose $\mathcal{C}(S) < \mathcal{C}(S \cup S')$ and $\mathcal{C}(S') < \mathcal{C}(S \cup S')$. From there it comes:

$$\frac{\triangle_{\text{in}}(S)^2}{\binom{|S|}{3}} + \frac{\triangle_{\text{in}}(S')^2}{\binom{|S'|}{3}} < \frac{(\triangle_{\text{in}}(S) + \triangle_{\text{in}}(s'))^2}{\binom{|S|+|S'|}{3}}$$

Given that $\forall a, b > 1, \binom{a}{3} + \binom{b}{3} < \binom{a+b}{3}$,

$$\left(\binom{|S'|}{3}\triangle_{\text{in}}(S) - \binom{|S|}{3}\triangle_{\text{in}}(S')\right)^2 < 0$$

Hence the contradiction. □

As the cohesion of a group of nodes is a measure of its quality as a community, it is understandable that adjoining a really good community to a lower quality one might result in a group of nodes which is averagely good (consider for example a huge clique and a poor set of nodes, the union might be more *community-ish* than the latter alone). Property 2 can be understood the following way: if a community is disconnected, then one of its connected component has a better cohesion than all connected component taken altogether. As such, it makes sense to try to maximize the cohesion on connected subgraphs. From now on, unless otherwise specified, we consider all cohesions on a connected graph containing no weak ties.

We now present two analytical results. The first one is important as it exhibits that sets of nodes conforming to the common definition of communities – using edge densities – will obtain high cohesion. The second one shows that a large clique does not shadow a smaller one if the overlap between the two is smaller than a threshold depending on the size of the latter.

Compatibility. Let S be a random network with an edge probability p_{in} and suppose S is embedded in a network G, where an edge exist between each node

of S and each node of G with probability p_{out}. Then the cohesion of S in G is given by:

$$\mathcal{C}(S) = \frac{p_{in}^3}{1 + \frac{3p_{out}|G|}{p_{in}(|S|-2)}}$$

Figure 4 shows that when S has a higher inner (resp. outer) density, the cohesion increases (resp. decreases). This ensures that cohesion remains compatible with the classical view on communities: it gives a higher score to dense set of nodes featuring a low density to the outside world.

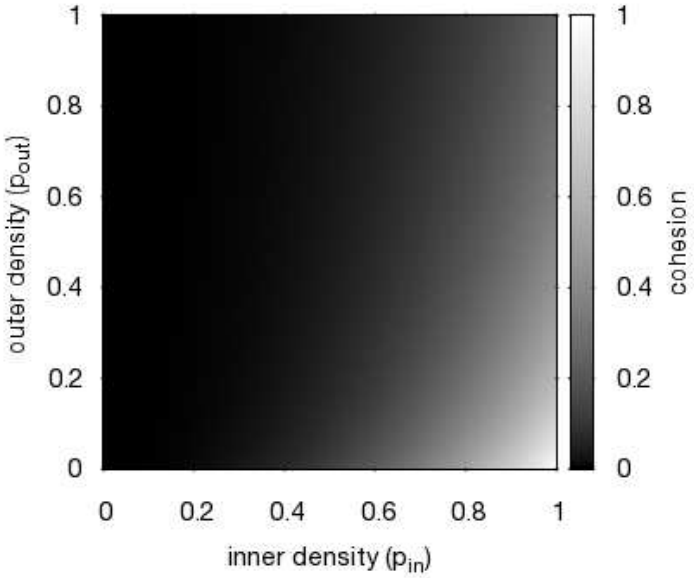

Figure 4: Cohesion of a set of 500 nodes connected to 500 external nodes as a function of inner and outer densities.

Non-shadowing. We now consider a network containing two cliques S_1 and S_2 of size $n_1 \geq n_2$, having p nodes in common. We have to the following cohesions :

$$\begin{aligned} \mathcal{C}(S_2) &= \frac{1}{1 + \frac{3(n_1-p)p(p-1)}{n_1(n_1-1)(n_1-2)}} \\ \mathcal{C}(S_1 \cup S_2) &= \frac{\binom{n_1}{3} + \binom{n_2}{3} - \binom{p}{3}}{\binom{n_1+n_2-p}{3}} \end{aligned}$$

In Figure 5, we represent in black the region where $\mathcal{C}(S_2) \geq \mathcal{C}(S_1 \cup S_2)$. What this figure shows is that, although S_2 might be much smaller than S_1, there is a threshold – greater than one common node – under which S_2 has better cohesion than the whole network, *i.e.* a large clique does not always absorb a smaller one. This ensures that cohesion does not suffer from resolution limit.

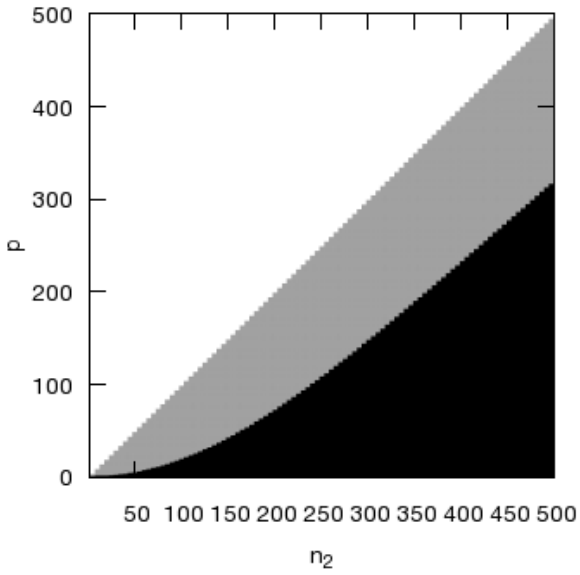

Figure 5: Regions where considering one community per clique (black) leads to a higher cohesion than considering only one big community (gray). $n_1 = 500$

.

3 Egomunities

3.1 Interlude

As most recent works have focused on *how* to detect communities, we deem necessary to bring back the *why* in the equation. It adds constraints to the structure and type of communities one wishes to obtain: community detection, in our opinion, has several purposes. First, as stated by Newman in his seminal paper [9], the "*ability to find and analyze such groups can provide invaluable help in understanding and visualizing the structure of networks*". Hence, paraphrasing, detecting community is a way to *simplify* a complex topological structure in order to facilitate its visualization and analysis.

If an algorithm produces an order of magnitude more than n communities in a network of size n – which incidentally cannot happen in the case of disjoint communities but might be the case when considering overlapping sets of nodes – the volume of data to deal with is not reduced but expanded and no simplification occurs. This is striking when trying to visualize a network: the aim of regrouping nodes into clusters is to reduce the clutter, not to pile up a great deal of communities one on top of the other. However, graph compression is not the only application of community detection.

Another possible use case lies in traits inference and social recommendation. The past few years have witnessed the emergence of so-called *online social networks*, such as Facebook, LinkedIn, Twitter, etc. which have proven invaluable as a source of data to study the structure of social interactions. The main benefit of using such social networks is that they not only reproduce the underlying social topology but add meta-information in the form of interests, events, etc.

They are however inherently limited by the fact that all information they contain are subject to what the user reveals about himself. Therefore, although the interpersonal links tend to be pretty exhaustive – in terms of *who knows who* – the information associated with each user is not. This can be easily explained: whereas adding a connection to another user is a matter of an instantaneous and simple click, entering one's centers of interest is time consuming and is often done in an incremental manner.

However, as it is common knowledge that *birds of a feather flock together*, it is possible to exploit the community structure of the network to infer what an individual might be interested in. Consider for example a person and all their acquaintances, if 1% of those notified they liked going to a specific restaurant, not much can be deduced. If however those 1% represent 90% of a tight and coherent social community, chances are that the considered individual has been to said restaurant. As such, community detection allows a refinement of the social neighborhood in order to infer more precisely what might be relevant to a given person, which has applications in terms of information discovery and advertising.

In this user centric context, the relevance of a community set is defined by the individual at the center in a subjective manner. In consequence looking for communities at a global level – the whole network – might not be the best approach. Consider for example a two spouses: both will have a *family* community, but might not include the same persons inside – both will include their children, their parents, maybe their in-laws, but when it comes to the other spouses cousins their perception of what their family is might differ.

3.2 Algorithm

For the aforementioned reasons, we introduce the concept of *egomunities*, namely person-based communities rooted in the subjective and local vision of the network by a given node. In a manner of speaking, we attempt to bring a possibly overlapping structure to the neighbors of the node. In this section we first present a greedy algorithm which, given a network and a node, uncovers all egomunities that this node belongs to. This is done by optimizing their cohesion (Algorithm 1). We then refine this algorithm by expanding into several optimizations.

Let G be a network and u a node of G, we focus on u's neighborhood $\mathcal{N}(G, u)$ and discard the rest of the network. The core idea is to group together neighbors in possibly overlapping egomunities, all containing u. To do so, we initialize an egomunity by selecting a node $v_0 \in N$ to serve as *seed* – thus the egomunity contains u and v_0. From that point we iterate and expand the egomunity by adding neighbors as long as it is possible to increase the cohesion. If there are several nodes which addition increases the cohesion, we choose to add the node v which addition maximizes the number of internal triangles $\triangle_{\text{in}}$ – and in the case more than one node satisfies this condition, we select the one which maximizes the number of outbound triangles $\triangle_{\text{out}}$. Once no more node can be added to the egomunity, we start over by selection the next seed from the sets of neighbors which haven't been assigned to an egomunity and repeat the process until all neighbors are in at least one egomunity.

The idea behind the algorithm is the following: each neighbor will be added, at some point in time, to an egomunity. As such, it is possible to use any

Algorithm 1 Greedy egomunities algorithm.

Require: G a graph, u a node
  $E \leftarrow \emptyset$
  $V \leftarrow \mathcal{N}(G, u)$
  while $V \neq \emptyset$ **do**
    $v \leftarrow$ node with highest degree in V
    initialize the egomunity $\epsilon \leftarrow \{u, v\}$
    $S \leftarrow \{v' \in \mathcal{N}(G, \epsilon) / \mathcal{C}(\epsilon \cup \{v'\}) > \mathcal{C}(\epsilon)\}$
    while $S \neq \emptyset$ **do**
      Add to ϵ the node $v \in S$ with the highest $\triangle_{\text{in}}(\epsilon \cup \{v\})$, in case of ties, chose the node with the highest $\triangle_{\text{out}}(\epsilon \cup \{v\})$
      $S \leftarrow \{v' \in \mathcal{N}(G, \epsilon) / \mathcal{C}(\epsilon \cup \{v'\}) > \mathcal{C}(\epsilon)\}$
    end while
    $V \leftarrow V \setminus e$
    add ϵ to the set of egomunities $E \leftarrow E \cup \{\epsilon\}$
  end while
  return the set of egomunities E

neighbor as a seed; however, by choosing a node with a high degree in the neighbors subgraph (that is, a node that forms a high number of triangles with the initial node) as a seed, we create a set of nodes with a low $\triangle_{\text{in}}$ and a high $\triangle_{\text{out}}$. The rationale behind the selection function in the greedy expansion phase is to maximize $\triangle_{\text{in}}$ as long as it results in a cohesion increase. We do not seek to directly maximize the cohesion as this could lead to cases where one node is selected because its addition decreases $\triangle_{\text{out}}$ too much, thus limiting the number of candidates at the next step. The exploratory phase can be seen as a growth of an egomunity first by selecting the inner nodes and then only the *corners*.

For obvious reasons, it is costly to compute the cohesion at each step – as it would require at least to enumerate all triangles in one egomunity ϵ, which might be as high as $\binom{|\epsilon|}{3}$. This gives a complexity of $\mathcal{O}(n^3)$ if $|\mathcal{N}(u)| = n$ just to compute the cohesion. However, it is possible to decrease the complexity by locally updating the cohesion when adding a new node v to ϵ :

$$\mathcal{C}(\epsilon \cup \{v\}) = \frac{(\triangle_{\text{in}}(\epsilon) + I_v)^2}{(\binom{|\epsilon|+1}{3})(\triangle_{\text{in}}(\epsilon) + \triangle_{\text{out}}(\epsilon) + O_v)}$$

Where $I_v = \triangle_{\text{in}}(\epsilon \cup \{v\}) - \triangle_{\text{in}}(\epsilon)$ and $O_v = \triangle_{\text{out}}(\epsilon \cup \{v\}) - \triangle_{\text{out}}(\epsilon))$ are the number of inbound and outbound triangles which would be added to ϵ when including v. We now describe algorithm to add a node to an egomunity, updating the cohesion and both quantities I_v and O_v for all impacted nodes. It is important to remember that here, all egomunities contain one node in common (the origin, u) and that because we restrict ourselves to the subgraph containing only u and its neighbors, $\mathcal{N}(\{u, v\}) = \mathcal{N}(v) \setminus \{u\}$.

We first initialize, for all nodes v, $I_v = 0$ (there would be no triangles in $e = \{u, v\}$) and $O_v = \deg(v) - 1$ (all triangles having an edge $\{u, v\}$ would be cut, which is exactly the number of common neighbors to u and v). Then, each time a node is added to the egomunity, only the values pertaining to its neighbors – not included in ϵ – need to be updated as described in Algorithm. 2 (Fig. 6).

Algorithm 2 Updating when adding v to an egomunity ϵ.

$\triangle_{\text{in}} \leftarrow \triangle_{\text{in}} + I_v$
$\triangle_{\text{out}} \leftarrow \triangle_{\text{out}} + O_v - I_v$
$\epsilon \leftarrow \epsilon \cup \{v\}$
for $v' \in \mathcal{N}(G, v) \setminus \epsilon$ **do**
  $n \leftarrow \mathcal{N}(G, v) \cap \mathcal{N}(v')$
  $I_{v'} \leftarrow I_{v'} + |n \cap \epsilon|$
  $O_{v'} \leftarrow O_{v'} + |n \setminus \epsilon| - |n \cap \epsilon|$
end for

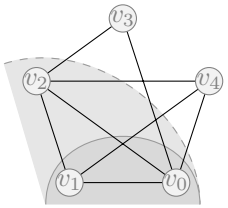

Figure 6: Updating the cohesion when adding v_2 to $\{v_0, v_1\}$, values for I and O are given inTable. 1.

node	(I_v, O_v) (before)	(I_v, O_v) (after)
v_2	(1, 4)	–
v_3	(0, 1)	(1, 0)
v_4	(1, 3)	(3, 0)

Table 1: Values for I and O before and after adding v_2 to $c = \{v_0, v_1\}$ as depicted on Fig. 6.

3.3 Two important heuristics

As said earlier, the cohesion is conceived to judge the quality of a given egomunity and not a set of egomunities, which is a totally different issue. The algorithm as defined above generates overlapping egomunities in an independent manner – in regard to previous output. We assess that in some cases, obtaining several groups of nodes which overlap too greatly might lead to irrelevant results and propose a simple yet effective way of merging egomunities.

We define the overlap $\text{overlap}(\epsilon_1, \epsilon_2) = \frac{\epsilon_1 \cap \epsilon_2}{\min(|\epsilon_1|, |\epsilon_2|)}$ and build an egomunity graph G_E which nodes are egomunities, and an edge (ϵ_1, ϵ_2) exists if $\text{overlap}(\epsilon_1, \epsilon_2)$ is greater than a threshold $o_{\{}\min\}$. Although several approaches might be thought of in order to carefully select which egomunities to merge (for example recursively computing egomunities on G_E), we have observed that a less cumbersome yet resilient method was to merge all egomunities pertaining to a same connected component in G_E.

This merging step raises another issue: given the fact that some egomunities might be merged, why bother and compute them separately in the first place? In the worst case, given a neighborhood of n nodes, the algorithm might output $\frac{n}{2}$ egomunities containing each $1 + \frac{n}{2}$ nodes. This is illustrated in Fig. 7, where up to 6 egomunities $\epsilon \cup \{v_i\}$ might be generated only to be merged after hand. Given that computing those distinct egomunities only is costly, we propose

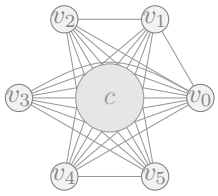

Figure 7: Network leading to overlapping egomunities.

another heuristic in order to reduce the useless calculations. After generating an egomunity, a last step is done in which all nodes v having a ratio $\frac{I_v}{O_v}$ greater than a given threshold are added to that egomunity.

We deem important to point out the fact that this algorithm cannot be trivially extended to compute overlapping communities on the whole network. An idea would obviously be to generate egomunities for all nodes and consider the resulting set of all egomunities as a set of communities for the network. This raises however an issue, which is that, in a network $G = (V, E)$ of size n, containing m edges, there is no *a priori* reason that the egomunities for a node u containing a node v would be the same as the egomunities generated for v and containing u. Therefore, it would require to compare two by two the different egomunities and decide whether they should be merged. As a node u of degree d_u can generate at most d_u communities, it would be necessary to compare $\sum_{u,v \in V} d_u d_v$ egomunities, which is $\mathcal{O}(nm)$.

4 Early results & future works

In the previous sections we have defined a metric, the *cohesion*, in order to quantify the *communityness* of a group of nodes and an algorithm which produces egomunities of high cohesion for a given node. In order to validate both the cohesion and the algorithm, we applied the latter to real world data. Through the Facebook Graph API [3], it is possible to extract the social neighborhood of a given individual. In this section we present preliminary results which we obtained by computing egomunities for specific Facebook users, first through a case study and then in the context of a large scale experiment. Finally, we describe some possible applications and extensions.

4.1 Case study

We used our algorithm to compute for a few users their egomunities of Facebook friends. We then interviewed those persons in order to determine if the egomunities we obtained had a subjective meaning for them. In this section, we present the results of one of those interviews.

Egomunities. The subject, a 32.5 year old male, had, at the time of the computation, 145 friends. Those friends were found to be distributed across 12 egomunities. 18 friends were not present in any egomunity (for example, friends having no friends in common with the subject), 94 were in only one egomunity, 26 in two egomunities, 3 in three and 4 in four different egomunities. Table 2 lists those egomunities along with there size and cohesion. A quick interview of the subject was conducted in order to figure if each group had a

social meaning to them and if so, how they would describe it. All but one group echoed a specific part of the subject's life. However, it is important to note that those egomunities only reflect the underlying Facebook network, which may be incomplete and differ from the real world social network.

description	size	cohesion
higher education	7	0.64
research (france)	5	0.61
elementary school	8	0.49
friends in Brazil	10	0.38
circle of friend	31	0.25
family	10	0.22
brazilian dancers/musicians 1	11	0.19
capoeira	13	0.17
dance	22	0.14
group of close friends	5	0.11
brazilian dancers/musicians 2	9	0.09
vague (mostly dance related)	52	0.07

Table 2: Egomunities ordered by cohesion. A short description of what people in the same group have in common is given.

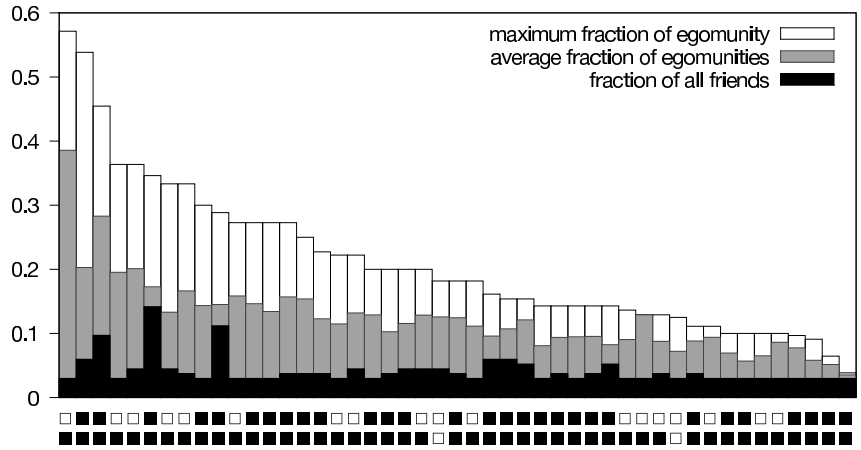

Figure 8: Proportions of all friends sharing a *like*, and average, maximum proportions on egomunities.

Traits inference. Ongoing work focuses on traits inference based on egomunity structure. For example, it is possible to access some information about a user on Facebook such as center of interest. We outline here an idea on how to exploit egomunities to mine more accurate information on a given individual. It is possible to gain some insight on one's centers of interest by observing their *likes*. Those are center of interests which were specified by the users and

might be shared between them. In Figure 8, we have extracted a subset of those interests satisfying the following criteria: *(i)* having been *liked* by at least 4 of the subject's friends and *(ii)* absent from the user's *likes*. Each column represents a particular interest and we plotted, for each of those: the proportion of all friends having this interest, the average proportion of friends sharing this interest in the subject's egomunities where the interest appears and the maximum proportion across egomunities. The abscissa also features two squares. On the top row, a full (resp. empty) square indicates that the subject was aware (resp. not aware) of the existence of the interest. The bottom row indicates whether it might be of interest to the subject. In this particular case, more than 95% of the *likes* were relevant to the subject (with no a priori knowledge on his centers of interests): 61.7% were already known but had not been specified on Facebook, and 34% were new *likes* which were of interest to him. Only two *likes* were of no interest to the subject, and it is notable that those do not feature a high maximum proportion across all egomunities. It is moreover interesting to observe that out of the 8 interests having the highest maximal proportion in an egomunity, the majority was unknown to the subject despite being of interest to him after hand.

4.2 The Fellows Experiment

Building on the Facebook, API we have launched Fellows [6], a large scale experiment on Facebook in which users are able to compute their egomunities and rate them. The data we are collecting from this ongoing experiment will allow us to statistically confront the cohesion model to individual perception of egomunities. In this section we present preliminary results obtained the data collected at the point of writing.

The participants were presented with an application which, once connected to Facebook, analyzes their social neighborhood and presents them with their egomunities computed by our algorithm. They are then asked to give a numerical rating between 1 and 4, answering the question "*would you say that this list of friends forms a group for you?*". We collect all egomunities with their cohesion and size. As the participant could stop rating at any time, some egomunities have no ratings.

The result we present here are extracted from data collected from 980 participants, which totaled 22,697 egomunities and of those, 14,634 were rated. On Figure 9 we group the rated communities by rating and represent the distributions of cohesion for each rating, this shows that on average egomunities with higher rating feature a higher cohesion. Conversely, in Figure 10, we plot the average rating obtained by egomunities grouped by cohesion slices of width $1/100^{\text{th}}$; in turn, this shows that on average, egomunities with higher cohesion tend to obtain higher ratings. Hence we conclude that *cohesion* provides an accurate quantification of an egomunity's quality, as perceived by its original node.

Considering the fact that gathering all *likes* for all users is intrusive and might put off some participants, we decided to focus on the inference of other traits, such as their age. Due to the presence of family, older co-workers, younger siblings, etc. the neighborhood of a given person features age disparities. However, all egomunities do not suffer from such heterogeneity: although someone may have some friends of disparate ages, it is likely that at least one of their

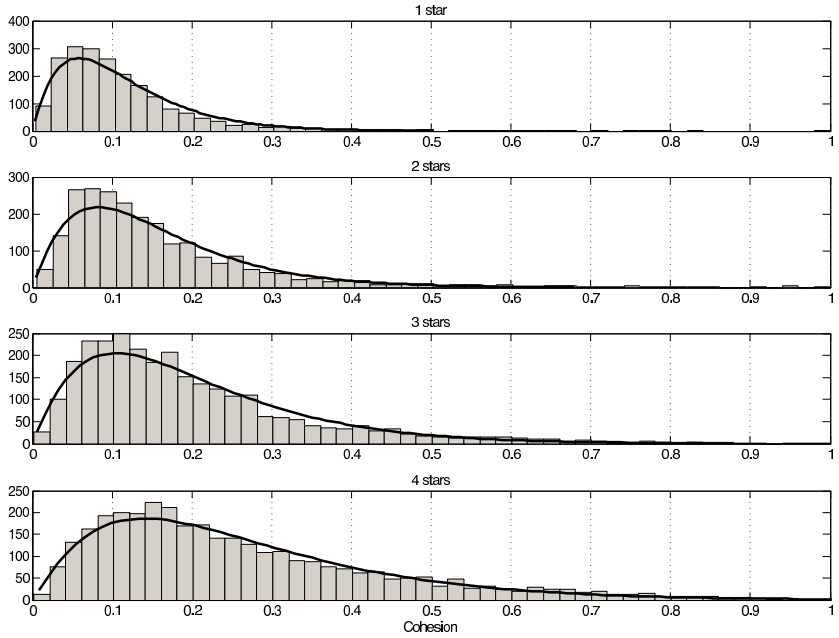

Figure 9: Density of cohesion for egomunities of rating 1, 2, 3 and 4.

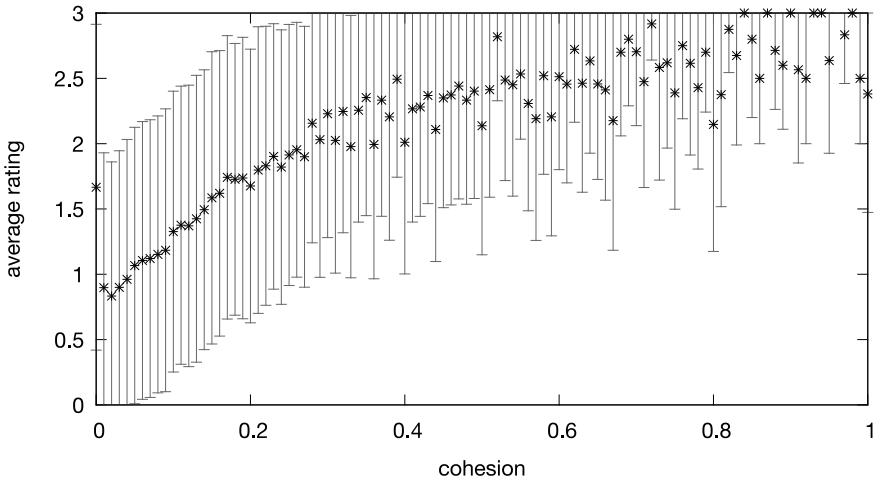

Figure 10: Average rating *vs.* cohesion.

egomunity features a low age variability (for example, an egomunity of classmates). Our idea is to exploit this fact to pinpoint more accurately the user's age.

From now on, we only take into account egomunities of size greater than 10 and of which the age standard deviation is less than 2.5 years (70.24% participants feature at least one such egomunity). Let a be the vector where a_i is the age of the i^{th} participant. We then define the *globally estimated age* g_i

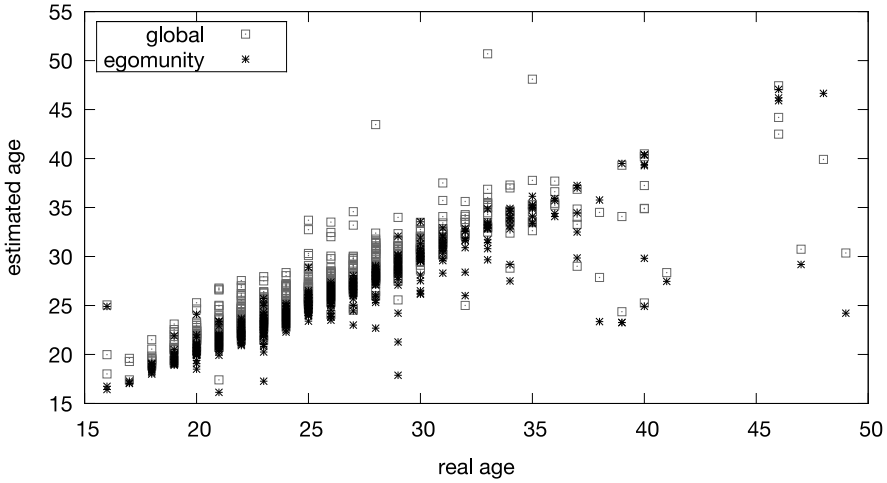

Figure 11: Subject age *vs.* estimated age on all friends and most homogeneous egomunity.

as the average of all the i^{th} participant's friends, the *egomunity based age* e_i as the average age of the members of the participant's egomunity featuring the lowest relative standard deviation. Figure 11 shows g and e in relation to a. Given that both quantities are correlated to a (Pearson correlation coefficients: $r_{a,g} = 0.859$ and $r_{a,e} = 0.894$), we assess that they can be used to infer real ages. However, the bias when considering all friends is of $+1.27$ years whereas it is only -0.296 years when using only the less variable egomunity. Both estimations feature similar variability ($\sigma_g = 2.9$ and $\sigma_e = 2.3$), but the average absolute error is of 1.96 years when using g whereas it is of 0.938 years in the other case. We conclude that the egomunity based age leads to a more accurate estimation of the participant's age.

We conclude that although it is possible to infer a person's age using all their friends, it is even more precise to do so using only a portion of their friends, pertaining to the egomunity featuring the lowest relative variability.

4.3 Extension to weighted networks

Besides traits inference, future works will also focus on the evaluation of weighted cohesion to quantify the quality of weighted social communities. In a simple unweighted model of social networks, when two people know each other, their is a link between them. In real life however, things are more subtle, as the relationships are not quite as binary: two close friends have a stronger bond than two acquaintances. In this case, weighted networks are a better model to describe social connections, this is why we deem necessary to introduce an extension of the cohesion to those networks.

The definition of the cohesion can, as a matter of fact, be extended to take the weights on edges into account. We make the assumption on the underlying network that all weights on edges are normalized between 0 and 1. A weight

$W(u,v) = 0$ meaning that there is no edge (or a null edge) between u and v, and a weight of 1 indicating a strong tie. We define the weight of a triplet of nodes as the product of its edges weights $W(u,v,w) = W(u,v)W(u,w)W(v,w)$. It then comes that a triplet has a strictly positive weight if and only if it is a triangle. We then define inbound and outbound weights of triangles and finally extend the cohesion.

$$\begin{aligned}
\triangle^w_{\text{in}}(S) &= \frac{1}{3} \sum_{(u,v,w)\in S^3} W(u,v,w) \\
\triangle^w_{\text{out}}(S) &= \frac{1}{2} \sum_{u\notin S,(v,w)\in S^2} W(u,v,w) \\
\mathcal{C}^w(S) &= \frac{\triangle^w_{\text{in}}(S)}{\binom{|S|}{3}} \frac{\triangle^w_{\text{in}}(S)}{\triangle^w_{\text{in}}(S) + \triangle^w_{\text{out}}(S)}
\end{aligned}$$

5 Conclusion

In this article we have presented a novel take to uncovering overlapping communities. Our approach lies in the use of egomunities: a person-based point of view of their neighbors' communities. To that effect, we define a metric, the *cohesion*, to quantify the intrinsic communityness of any subset of nodes of a network. We used the cohesion to design an algorithm which constructs egomunities. We applied this algorithm on data extracted from Facebook, both in the form of case studies and through a large scale ongoing experiment called Fellows, in which users are presented with their egomunities and are asked to rate them according to their own perception. The experiment provides us with data which already tend to validate the accuracy of cohesion as a community quality measure. Moreover, preliminary results are promising, as we were able to exhibit that the use of egomunities can lead to the construction of efficient estimators for several personal traits. Future work will rely on data collected during the Fellows[1] experiment to further our study on traits inference. Using the weighted cohesion, we will also investigate the influence of weights on egomunity detection.

[1] `http://fellows-exp.com/`

References

[1] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics-Theory and Experiment*, 2008.

[2] C. Castellano, F. Cecconi, V. Loreto, D. Parisi, and F. Radicchi. Self-contained algorithms to detect communities in networks. *EPJB*, 38(2):311–319, 2004.

[3] Facebook. Graph api, 2011. `http://developers.facebook.com/docs/api`.

[4] G. Flake, S. Lawrence, C. Giles, and F. Coetzee. Self-organization of the web and identification of communities. *Communities*, 35(3):66–71, 2002.

[5] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[6] A. Friggeri, G. Chelius, and E. Fleury. Fellows, a social experiment, 2011. `http://fellows-exp.com`.

[7] M. Granovetter. The strength of weak ties: a network theory revisited. *Amer. J. of Sociology*, page 46, Jan 1981.

[8] T. Nepusz, A. Petróczi, L. Négyessy, and F. Bazsó. Fuzzy communities and the concept of bridgeness in complex networks. *Phys. Rev. E*, 77(1):16107, 2008.

[9] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2), 2004.

[10] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[11] H. Shen, X. Cheng, and J. Guo. Quantifying and identifying the overlapping community structure in networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2009:P07042, 2009.

[12] M. Sozio and A. Gionis. 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*, pages 939–948, 2010.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399