Ego-munities, Exploring Socially Cohesive Person-based Communities

Adrien Friggeri, Guillaume Chelius, Eric Fleury

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Ego-*munities*
Exploring Socially Cohesive Person-based Communities

Adrien Friggeri — Guillaume Chelius — Eric Fleury

N° 7535

Février 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

Ego-munities
Exploring Socially Cohesive Person-based Communities

Adrien Friggeri , Guillaume Chelius , Eric Fleury

Thème : Réseaux et télécommunications
Équipe-Projet DNET

Rapport de recherche n° 7535 — Février 2011 — 18 pages

Abstract:

In the last few 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 compute such communities stem from the field of disjoint community detection by either extending the concept of modularity to an overlapping context or by attempting to decompose the whole set of nodes into several possibly overlapping subsets. In this report we take an orthogonal approach by introducing a metric, the cohesion, rooted in sociological considerations. The cohesion quantifies the community-ness of one given set of nodes, based on the notions of triangles - triplets of connected nodes - and weak ties, 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. As such, we introduce a numerical characterization of communities: sets of nodes featuring a high cohesion. We then present a new approach to the problem of overlapping communities by introducing the concept of ego-munities, which are subjective communities centered around a given node, specifically inside its neighborhood. We build upon the cohesion to construct a heuristic algorithm which outputs a node's ego-munities by attempting to maximize their cohesion. We illustrate the pertinence of our method with a detailed description of one person's ego-munities among Facebook friends. We finally conclude by describing promising applications of ego-munities 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

Ego-munautés
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 contribution of sociology, several definitions have been proposed by the physics and computer science communities [3][1].

Despite the lack of formal 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. Unfortunatly, 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" [8], which compares the density of links inside a given community to what would be expected if edges where distributed randomly across the network. This method has proven to give sensible results on several test networks and gained traction in the *communities* community. It it important to note that, given that maximizing the Q-modularity on general graphs is an established NP-hard problem, several heuristics have been proposed[4].

These 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 simultenaously belong to a family, a friends group and co-workers groups.

Due to the historical evolution of the field, to this day, most methods used to detect overlapping communities were inspired by, or adapted from, existing counterparts in disjoint community detection. If some of those methods take a literal approach to the issue and are built upon extensions to the modularity[7, 10], others have taken another path, such as clique percolation[9]. However, we assess that all those methods aim at finding all communities in a network.

In this report, we propose to take a step back and reflect on several questions, the first of which being our take on the longstanding problem of what defines a community as such. Drawing inspiration from well established sociological results, we introduce a metric, the *cohesion* based on the notions of triangles – triplets of connected nodes – and weak ties, instead of the classical view using edges to rate the *community-ness* of one given set of nodes. It is important to note that whereas the Q-modularity gives a score to a partition of a network, the *cohesion* serves as a formal, quantitive and intrinsic characterization of a subgraph embedded in a network, independantly any other subgraphs. As such, we propose a definition of a community, namely a set of nodes with high cohesion.

It is important to note that 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.

Furthermore, although the cohesion can be used to rate any set of nodes, in the latter half of this paper we introduce and focus on a specific type of user-centric communities, which we call *ego-munities*, and present a heuristic algorithm based on the cohesion to generate such ego-munities.

This paper is organized as follows: in Section 2 we introduce a new metric, the *cohesion*, to evaluate the *community-ness* of a set of nodes. In Section 3 we present a user-centric way of thinking about communities: *ego*-munities and introduce an algorithm which computes thos. Finally, in Section 4 we present the ego-munities of Facebook friends of a test subject and highlight several possible applications.

2 Cohesion

Before delving into technicalities and formal definitions, we consider important to take a moment to reflect on community detection and highlight our findings on the inherent differences between the problems of disjoint versus overlapping communities, as these have implications on the principles of related algorithms.

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 ?
- **content** : is each community intrinsically sound ?

The main difference between disjoint and overlapping communities problems is, as redundant as it seems, that in the latter a node can belong to several communities. Although seemingly mundane, in the case of disjoint communities 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 in communities.

This is actually the whole idea behind Q-modularity, defined as follow: $Q = \operatorname{Tr} \mathbf{e} - ||\mathbf{e}^2||$ where $\mathbf{e}$ is a matrix where $\mathbf{e}_{i,j}$ 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, 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

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.

choices tailored 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 one sole 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, 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. For example, the classical concept of *nuclear family* – a group consisting of a father and mother and their children – forms an obvious social community while at the same time being strictly included in a broader *extended family* containing other individuals.

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, independantly from the rest of the network. The idea is to give a score to a specific set of nodes describing 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.

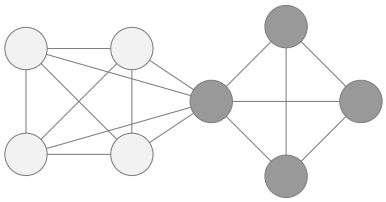

Figure 2: Two communities featuring the same number of links towards the outside world but clearly different from a community-ness standpoint

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. We digress slightly from the subject of community detection to that off information diffusion, only to explain the roots of our metric.

In [6], Granovetter defines the notion of *weak ties*, which are edges connecting acquaintances, and argues 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*". Furthermore, he states 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.*". And finally, he assesses that *local bridges* – edges which do not belong to a set of three connected nodes – are weak ties. For these reasons, we consider that the structural backbone of communities does not lie in the edges of the network, but rather in its triangles (three connected nodes).

In Figure 2, two communities are represented in light and dark gray. Both contain the same number of nodes and the same number of 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 four *outbound* triangles – that is, triangles having an edge inside the community and a point outside – whereas the other set contains no such triangles.

Therefore, 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, it is important to note that this metric does not describe how good is a set of communities but merely the intrinsic quality of each community.

2.3 Definition

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

We first define two quantities, $\triangle_{\text{in}}(G, S)$ which is the number of triangles of G contained by S and $\triangle_{\text{out}}(G, S)$ the number of triangles "pointing outwards"

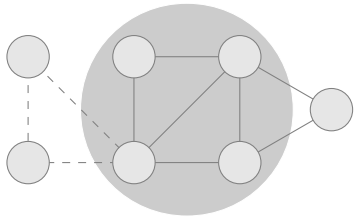

Figure 3: Cohesion of a set of nodes (circled) in a network. $\triangle_{\text{in}} = 2$, $\triangle_{\text{out}} = 1$ (the dashed triangle is not taken into account as it has only one node in the set), therefore $\mathcal{C} = \frac{1}{6}$

— that is, triangles of G having two nodes in S and the third one in $\mathcal{N}(G, S)$. We then define the *cohesion* $\mathcal{C}$:

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

The first factor is the *triangular density* of the community, while the second one represents the proportion of triangles having a base 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.

2.4 Properties

We call *weak tie* as an edge which does not belong to any triangle $G_{\triangle} = (V, E_{\triangle})$ be the graph obtained by removing all weak ties from G.

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

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, $\mathcal{C}(G, S) = \mathcal{C}(G_{\triangle}, S)$. □

Property 2. *Let S and S' be two disconnected sets of nodes. 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. □

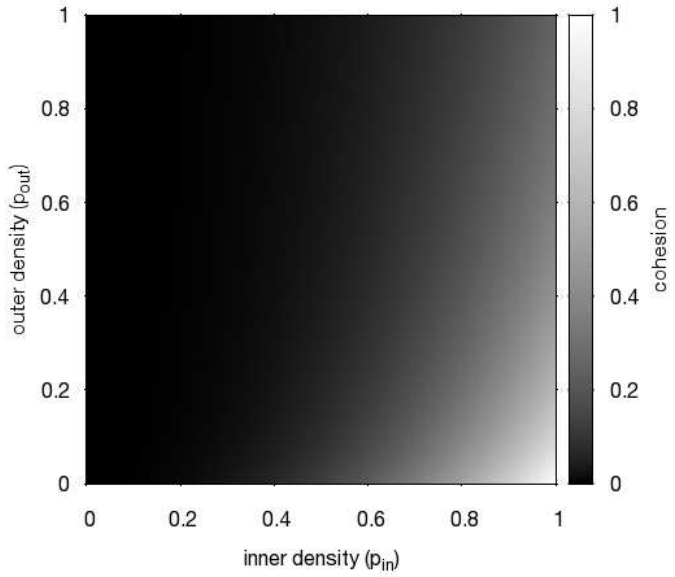

Figure 4: Value of the cohesion for a community of size 500 connected to a outside world of size 500 as a function of inner and outer densities

.

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 alltogether. 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.

2.5 Analytical results

In this section, we present two analytical results. The first one is important as it exhibits what one would expect from a community quality metric, namely that it gives a higher score to dense set of nodes featuring a low density to the outside world. The second one shows that a large clique does not shadow a smaller one if the overlap between the two in smaller than a treshlod depending on the size of the latter.

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:

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

Figure 4 shows that, all other things being equal, the denser S, the higher the cohesion. Conversely, when s has a higher outer density, the cohesion decreases, which is what one would expect from a community's quality. This is important as it ensures that sets of nodes conforming to the common definition of communities (using edge densities) will obtain high cohesion.

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 :

$$\mathcal{C}(S_2) = \frac{1}{1 + \frac{3(n_1-p)p(p-1)}{n_1(n_1-1)(n_1-2)}} \quad (1)$$

$$\mathcal{C}(S_1 \cup S_2) = \frac{\binom{n_1}{3} + \binom{n_2}{3} - \binom{p}{3}}{\binom{n_1+n_2-p}{3}} \quad (2)$$

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 under which S_2 is considered as better cohesion than the whole network, *i.e.* a large clique does not always absorb a smaller one.

3 *Ego-munities*

3.1 Interlude

As most recent work 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 [8], the "*ability to find and analyze such groups can provide invaluable help in understanding and visualizing the structure of networks*". Thus, paraphrasing, detecting community is a way to *simplify* a complex topological structure in order to facilitate its visualization and analysis.

Therefore, 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 recommandation. 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

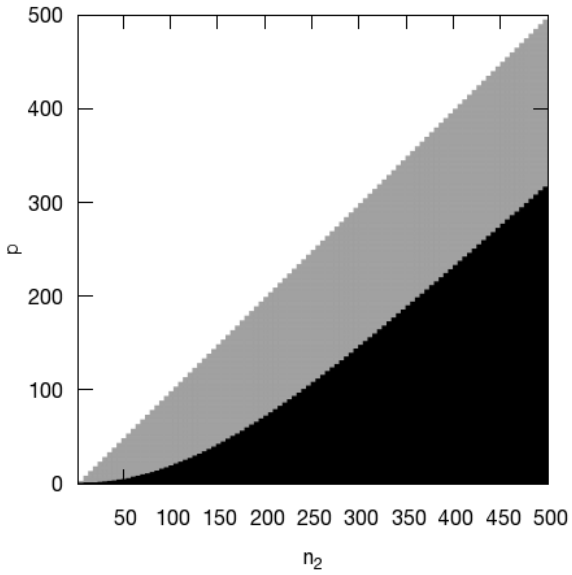

Figure 5: Region where considering one community per clique leads to a higher cohesion than considering only one big community (black). $n_1 = 500$

.

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 interpersonnal 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 instantenous 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 knownledge 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 *ego-munities*, namely person-based communities rooted in the subjective and local vision of the network by a given node, in our case his neighborhood. In this section we first present a greedy algorithm which, given a network and a node returns all ego-munities that a node belongs to from its point of view by attempting to heuristically optimize 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 $\mathcal{N}(u)$ and discard the rest of the network. The core idea is to group together neighbors in possibly overlapping ego-munities, all containing u. To do so, we initialize an ego-munity by selecting a node $v_0 \in N$ to serve as *seed* – thus the ego-munity contains u and v_0. From that point we iterate and expand the ego-munity 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 ego-munity, we start over by selection the next seed from the sets of neighbors which haven't been assigned to an ego-munity and repeat the process until all neighbors are in at least one ego-munity.

The idea behind the algorithm is the following: each neighbor will be added, at some point in time, to an ego-munity. As such, it is possible to use any 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 ego-munity first by selecting the inner nodes and the 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 ego-munity e, which might be as high as $\binom{|e|}{3}$. 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 e :

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

Where $I_v = \triangle_{\text{in}}(e \cup \{v\}) - \triangle_{\text{in}}(e)$ (resp. $O_v = \triangle_{\text{out}}(e \cup \{v\}) - \triangle_{\text{out}}(e)$) is the number of inbound (resp. outbound) triangles which would be added to e when

Algorithm 1 Greedy ego-munities algorithm.

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

Algorithm 2 Updating when adding v to an ego-munity c

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

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.

including v. We now propose an algorithm to add a node to an ego-munity, and update the cohesion and both quantities I_v and O_v for all impacted nodes.

It is important to remember that in our case, all ego-munities contain one node in common (the origin, u). Moreover, because we restrict ourselves to the subgraph containing only u and its neighbors, $\mathcal{N}(\{u, v\}) = \mathcal{N}(v) \cup \{u\}$.

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 ego-munity, only the values pertaining to its neighbors – not included in e – need to be updated as described in Algorithm. 2.

An example is given in Fig. 6, with values for I and O in Table. 1 before and after adding v_2 to $c = \{v_0, v_1\}$, v_0 being the origin node.

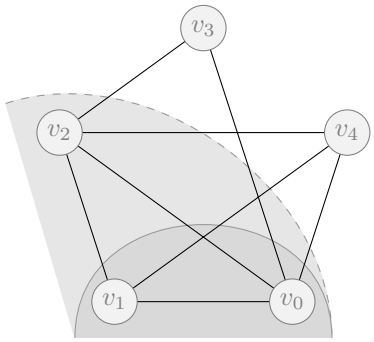

Figure 6: Updating the cohesion online.

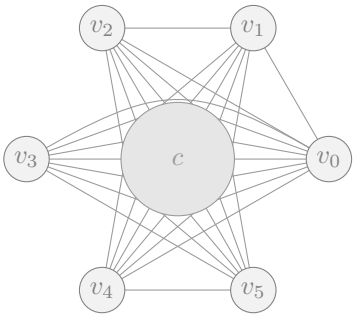

Figure 7: Hedgehog network leading to greatly overlapping ego-munities.

3.3 Two important heuristics

As said earlier, the cohesion is conceived to judge the quality of a given ego-munity and not a set of ego-munities, which is a totally different issue. The algorithm as defined above generates overlapping ego-munities in an independant 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 ego-munities.

We define the overlap $\text{overlap}(e_1, e_2) = \frac{e_1 \cap e_2}{\min(|e_1|,|e_2|)}$ and build an ego-munity graph G_E which nodes are ego-munities, and an edge (e_1, e_2) exists if $\text{overlap}(e_1, e_2)$ is greater than a threshold $o_{\{}\min\}$. Although several approaches might be thought of in order to carefully select which ego-munities to merge (for example recursively computing ego-munities on G_E), we have observed that a less cumbersome yet resilient method was to merge all ego-munities pertaining to a same connected component in G_E.

This merging step raises another issue: given the fact that some ego-munities 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 $\frac{n}{2}$ ego-munities containing each $1 + \frac{n}{2}$ nodes. This is illustrated in Fig. 7, where up to 6 ego-munities $e \cup \{v_i\}$ might be generated only to be merged afterhand.

As computing those distinct ego-munities is costly, we propose another heuristic in order to reduce the useless calculations. After generating an ego-munity,

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: Ego-munities ordered by cohesion. A short description of what people in the same groupe have in common is given.

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 ego-munity.

4 Preliminary results and future works

In the previous sections we have defined a metric, the *cohesion*, in order to quantify the *communityness* of a group of nodes. As a matter of fact, the gist of our proposal is a model of a social community: a set of nodes, in a network, featuring a high cohesion. Which, as such, there are no formal and direct ways of proving true: it can only be evaluated.

We started off by confronting the model to real world examples. Through the Facebook Graph API [2], it is possible to extract the social neighborhood of a given individual.

We used the access to this social data to launch Fellows [5], a large scale experiment on Facebook in which users are able to compute their ego-munities 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 ego-munities.

Moreover, we used the algorithm to compute for a few users which we interviewed in order to determine if the ego-munities we obtained had a subjective meaning for them. In the following sections, we present the results of one of those interviews and describe some possible applications and extensions.

4.1 Ego-munities

This being said, we show here an example of output for a given individual. The subject, a 32.5 years old male, had, at the time of the computation, 145 friends. Those friends were found to be distributed across 12 ego-munities. 18 friends were not present in any ego-munity (for example, friends having no friends in common with the subject), 94 were in only one ego-munity, 26 in two ego-munities, 3 in three and 4 in four different ego-munities.

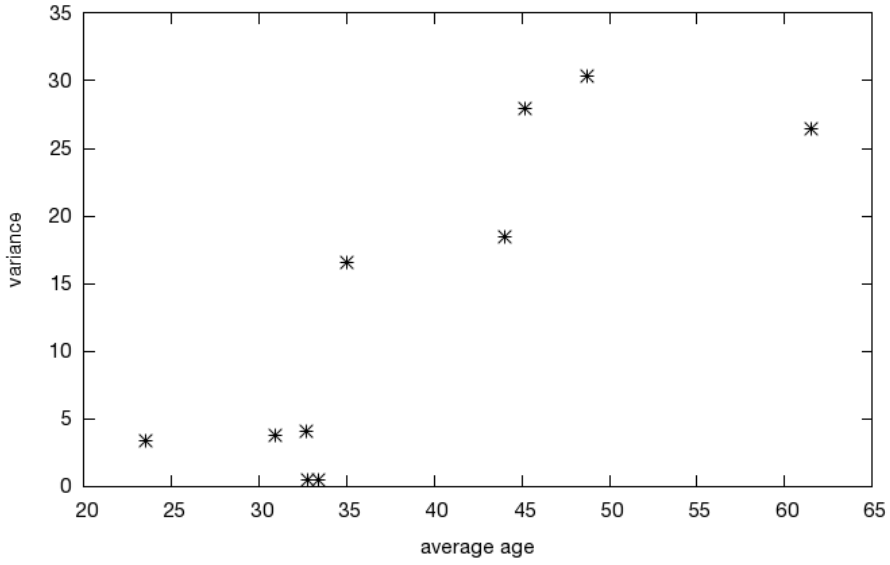

Figure 8: Average age vs. standard deviation in each ego-munity.

Table 2 gives a list of those ego-munities along with there size, 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. It is important to note that the ego-munities only reflect the underlying network, which may differ from the real world social network. Moreover, there is no apparent correlation between size, cohesion and density.

4.2 Traits inference

Ongoing work focuses on traits inference based on ego-munity structure. For example, it is possible to access some information about a user on Facebook such as their age, their gender or even center of interest. We outline several ideas on how to exploit ego-munities to mine more accurate information on a given subject.

Consider for example the age of an individual. Due to the presence of family, teachers, older co-workers, younger siblings, etc, the neighborhood of a given person features age disparities. In the case of the aforementioned subject, the average age of all his friends is 33.53 ($\pm$ 11.38) years old. However, by observing those same metrics group by group, it is possible to exclude heterogeneous groups such as family and focus on more homogeneous groups (typically, friends groups) and obtain a clearer idea of their age (Fig. 8). In this case, the group with the lesser variability has an average age of 32.75 ($\pm$0.43) years old.

Not only does Facebook gives access to users ages, it is also 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 9, we have extracted a subset of those interests which were *liked* by at least 4 of the subject's friends. Each column represent a particular interest

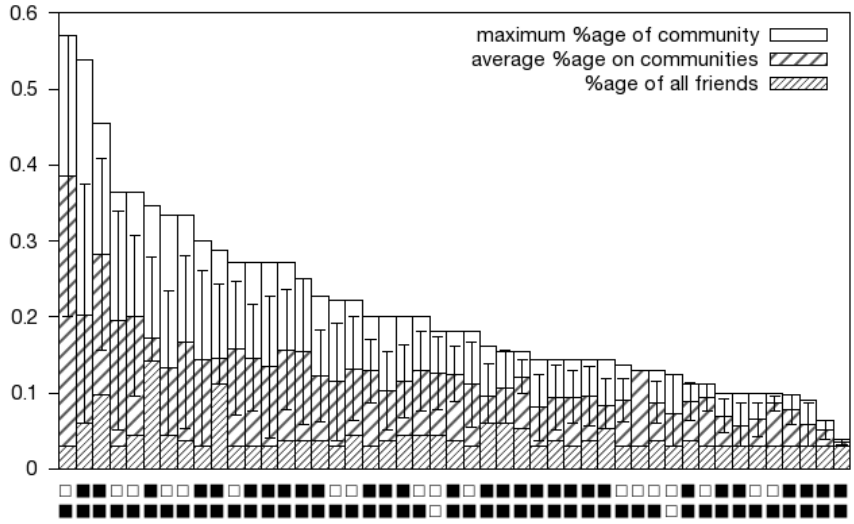

Figure 9: likes.

	known	unknown
interested	0.617	0.340
not interested	0	0.043

Table 3: Foo

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 ego-munities where the interest appears and the maximum proportion across ego-munities. The abscissa 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 wether it might be of interest to the subject. The repartition of those indicators in given in Table 3.

Only two *likes* were of no interest to the subject, and it is notable that those featured a low maximum proportion across all ego-munities. It is moreover interesting to observe that out of the 8 interests having the highest maximal proportion in an ego-munity, the majority was unknown to the subject despite being of interest to him afterhand.

4.3 Extension to weighted networks

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 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 extended to take the weights on edges into account. We make the following assumption on the underlying network, 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) betwenn 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 comes from this that a triplet has a strictly positive weight if and only if it is a triangle.

We can then define weighted equivalent to $\triangle_{\text{in}}$ and $\triangle_{\text{out}}$ and finally extend the cohesion.

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

Besides traits inference, future works will also focus on the evaluation of weighted cohesion to quantify the quality of weighted social communities.

5 Conclusion

In this article we presented a new metric, the *cohesion*, to quantify the communityness of a set of nodes. We used the cohesion to build an algorithm which constructs ego-munities, that is communities as seen by a node in its neighborhood. We applied this algorithm on a social network extracted from Facebook and shown that it lead not only to sound results but that those results could be used to infer information on the user or to provide him with interest recommandation. We have launched Fellows, a large scale experiment on Facebook, in which each user will be able to compute their ego-munities and rate them according to their perception. We believe such an experiment will validate our model and provide us with data which we will be able to use to develop traits inference algorithms.

References

[1] C Castellano, F Cecconi, V Loreto, D Parisi, and F Radicchi. Self-contained algorithms to detect communities in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):311–319, 2004.

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

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

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

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

[6] Mark Granovetter. The strength of weak ties: a network theory revisited. page 46, Jan 1981.

[7] 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.

[8] MEJ Newman and M Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):26113, 2004.

[9] 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.

[10] HW Shen, XQ Cheng, and JF Guo. Quantifying and identifying the overlapping community structure in networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2009:P07042, 2009.

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