INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Triangles to Capture Social Cohesion

Adrien Friggeri — Guillaume Chelius — Eric Fleury

N° 7686

8 July 2011

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--7686--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE – RHÔNE-ALPES

Triangles to Capture Social Cohesion

Adrien Friggeri , Guillaume Chelius , Eric Fleury

Thème : Réseaux et télécommunications
Équipe-Projet DNET

Rapport de recherche n° 7686 — 8 July 2011 — 16 pages

Abstract: 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 this article we take an orthogonal approach by introducing a novel point of view to the problem of overlapping communities. Instead of quantifying the quality of a set of communities, we choose to focus on the intrinsic community-ness of one given set of nodes. To do so, we propose a general metric on graphs, the cohesion, based on counting triangles and inspired by well established sociological considerations. The model has been validated through a large-scale online experiment called Fellows in which users were able to compute their social groups on Facebook and rate the quality of the obtained groups. By observing those ratings in relation to the cohesion we assess that the cohesion is a strong indicator of users subjective perception of the community-ness of a set of people.

Key-words: No keywords

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

Des Triangles pour Capturer la Cohésion Sociale

Résumé : Bien que la problématique de détection de communautés dans les réseaux sociaux ait attiré une attention grandissante à travers les sciences ces dernières années, aucun consensus formel n'a été atteint sur la nature de ce qui définit une communauté. Nous introduisons ici un point de vue novateur au problème de communautés recouvrantes. Au lieu de quantifier la qualité d'un ensemble de communautés, nous nous concentrons sur l'aspect intrinsèquement communautaire d'un ensemble donné de nuds. Pour ce faire, nous proposons une métrique générique sur les graphes, la cohésion, se fondant sur la notion de triangles et inspirée par des résultats établis en sociologie. Ce modèle a été validé à travers Fellows, une expérience à large échelle sur Facebook dans laquelle les utilisateurs avaient la possibilité de calculer de manière automatique leurs groupes d'amis puis de noter la qualité de ceux ci. En observant ces notes et la cohésion des groupes obtenus, nous concluons que la cohésion est une bonne évaluation de la perception subjective de l'aspect communautaire d'un ensemble de nœuds par un utilisateur.

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

Introduction

> The term community relates to a wide range of phenomena and has been used as an omnibus word loaded with diverse associations.
> *The Social Science Encyclopedia,* Adam Kuper

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 1955, George Hillery, Jr. analyzed 94 different sociological definitions of the term *community* [1] both from a quantitative and qualitative standpoint only to conclude that their only common defining feature was that they all dealt with people. Despite this fact, there were other traits on which the majority of definitions agreed, and he stated that *"of the 94 definitions, 69 are in accord that social interaction, area, and a common tie or ties are commonly found in community life"*. Fast-forward half a century, through the emergence of network science in the last two decades, the *communities community* has expanded to encompass scientists coming from backgrounds as diverse as, among others, computer science, theoretical physics or biology who brought along their own ideas and baggage on what should be called a community.

In this context, where social networks are modeled as graphs of individuals linked when they share a social connection in real life, all authors concur on the intuitive notion that a community is a relatively tightly interconnected 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 *tightly interconnected* and *less links*. The important aspect to consider, however, is that the defining concept of community in network science resides in topological features of the network. In real life, however, one rarely describes a group of people as *"this set of 10 people of density 0.8, featuring on average 2 outbound link per individual"*, understandably preferring clearer – and yet less formal – labels such as *'family'*, *'people at work'* or *'the poker group'*.

The whole idea behind community detection in social networks is due to the observation that there is a correlation between the topology of the network and some kind of labels which relate to social interactions[1], and that therefore it should be possible to infer the socio-semantic structure of the network by observing some of its topological traits.

Social networks are a peculiar beast in the sense that they only exist as descriptions of a fragment of what one would call *The Social Network*, an unmeasurable, exhaustive and dynamic multigraph of all social interactions at mankind scale. For example, Zachary's famous karate club dataset [2] is nothing more than Zachary's description of a subset of all social interactions, limited in terms of people (members of a karate club in a US university), nature (friendship) and time (at some point in time in the 1970s).

It is therefore important to keep in mind that any structural properties of communities are constrained by the nature of the network. The emergence of online social networks such as Facebook and Twitter in the last years and the

[1] For obvious reasons, this assertion does not hold for arbitrarily defined groups, consider for example the set of people of even height or any other group sharing a randomly distributed feature: chances that this group present a distinctive topological structure which separates it from the rest of the network are pretty slim.

availability of high computational power has led to a unique situation where there are not only rich datasets to study but also the ability to do so. But the richness of these networks lead them not only shine by their size but also their intricate complexity, as they encompass social links which may vary both in nature and in intensity. For example, people add close friends as well as professional acquaintances on Facebook, treating both categories as equals – in Facebook's terms, all social links are friendships – effectively flattening a complex multi-graph into a very slightly less complex graph.

In that case, what meaning should one give to *less links*? It is obvious, for example, that excluding an employee's boss from their "family" group should not be detrimental to the group's community-ness, whereas excluding their mother should. And yet in both cases the topological implications are the same: a edge in the network links someone inside the group to someone outside. Thus, given that all links are not equal in the network, the considered topological features should go beyond the simple notion of edges in order to discriminate those type of cases.

In this article, we introduce in Section 1 the *cohesion*, a new graph metric, inspired by well established sociological results, which rates the intrinsic community-ness of a set of nodes of a social network, independently from the existence of other communities. We then describe in Section 2 the experimental setup of Fellows, a large scale online experiment on Facebook which we launched to prove the validity of the cohesion. Finally in Section 3 we exhibit the high correlation between the cohesion of social groups and the subjective perception of those groups by users.

1 Cohesion

Before introducing the cohesion, let us reflect on the way community detection has blossomed in the past few years. In 2004, at the junction of graph partitioning in graph theory and hierarchical clustering in sociology, Newman and Girvan proposed an algorithm to partition a network into several communities. In order to assess the quality of the partitions which were produced by their algorithm, they introduced the modularity [3], a quantity which measures "*the fraction of the edges in the network that connect vertices of the same type (i.e., within-community edges) minus the expected value of the same quantity in a network with the same community divisions but random connections between the vertices.*"

In the following years, the modularity attracted attention, with several heuristics being proposed to attempt to find maximal partitions modularity-wise – see for example the Louvain method [4]. During the same time, other have exhibited several shortcomings of the modularity itself: that it has a resolution limit and therefore that modularity optimization techniques cannot detect small communities in large networks, that some random networks are modular.

Going further, when partitioning a network, each node is affected to a unique community, which has the rather unfortunate side effect of tearing families apart: an individual cannot be at the same time part of *their family* and *their company.*

In order to overcome these limitations, it is natural to shift to a context of overlapping communities, in which the one-node-to-one-community constraint disappears. This however has an incidence on modularity. "*If vertices may* *belong to more clusters,*" says Fortunato in his 2010 review, "*it is not obvious how to find a proper generalization of modularity. In fact, there is no unique recipe.*" Naturally, other techniques such as clique the percolation method [5] which do not rely on modularity were introduced – clique percolation goes even further as the method does not evaluate the quality of communities.

Behind the beautiful simplicity of the modularity actually lie two subtly different measures. First, the modularity encompasses the individual and intrinsic quality of each community's *content* by comparing them to a null model. Second, but no less important, it implicitly judges the quality of the *division* in communities. While this makes sense in the context of a partition because both those aspects are linked – one cannot change the content of a community without affecting other communities – there is no equivalent notion in an overlapping context.

1.1 A Word on Judging Divisions

Judging the quality of the division largely depends on the data one wishes to study. While it is obvious that two completely disjoint communities $S_1 \cap S_2 = \emptyset$ form a good division of the network $(S_1 \cup S_2, E)$ and that two completely overlapping communities $S_1 = S_2 = S$ form a really bad division of the network (S, E), the intermediate overlapping cases are less trivial.

On the one hand, in some occurrences, there is a case for allowing small *fuzzy* overlaps in order to model an vertex-based interface between groups instead of purely edges. On the other hand, there also are extreme cases where communities should be allowed to overlap at a great extent – consider for example college classes – or even be allowed to be fully embedded one in another (*e.g.* a computer science lab might be a small community inside a bigger university community).

For those reasons, we assess that their is no swiss army knife of division rating: the tools used to rate the division in communities itself should be carefully crafted to fit the data analysis.

1.2 Rating the Content of Communities

It is however 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.

As hinted earlier, the purely vertex/edge based approach to community rating has flaws. As an example, the toy network in Figure 1 consists of a group of dark nodes and a group of light nodes. Both groups contain the same number (4) of nodes and the same number (6) of internal edges (connecting two nodes in the same group). Moreover, both groups have the same number (4) of external edges (connecting one node inside the group to one node outside). That is, with a network vision restricted to nodes and edges, both groups are virtually indistinguishable, and yet one would say that the dark group is a "good" community, whereas the light group is a "bad" community. The asymmetry between both groups arises when observing triangles – sets of three pairwise connected nodes – in the network: there are 6 outbound triangles, that is having two vertices inside the dark group and one vertex in the light group.

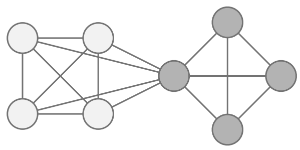

Figure 1: Two sets of nodes of identical size, featuring the same number of links both inside the set and towards the rest of the network. Despite those structural similarities, the darker set appears like a worse community than the lighter one.

The use of triangles does not only stem from the asymmetry they cause in the treatment of different group but is in line with the notions of *triadic closure* and *weak ties* introduced by Anatol Rapoport and Mark Granovetter [6,7]. Granovetter defines 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*".

From there triadic closure is the property on triplets u, v, w that if there exist a strong tie between u and v and between u and w then there is at least a weak tie between v and w. In the context of complex layered networks where ties can be of different nature – *blood-related*, *co-workers* – one can extend this notion by requiring the two strong ties to be of the same nature. In that case, when one observes a triangle in a network, there are chances than the three edges are of the same type, whereas edges which do not belong to triangles may be considered as weak ties, and as such serve as a bridge between communities and thus their exclusion from a community should not be detrimental to its quality. For the same reasons, only outbound triangles should negatively affect the quality of a group of nodes.

Building on this observation, we now formally define the cohesion. Let $G = (V, E)$ be a network and $S \in V$ a set of nodes. We define a triangle as being a triplet of nodes $(u, v, w) \in V^3$ which are pairwise connected, *ie.* such that $((u, v), (v, w), (u, w)) \in E^3$. In respect to S, $\triangle_i(S)$ denotes the number of triangles where all nodes belong to S and $\triangle_o(S)$ is the number of outbound triangles of S, that is having exactly two vertices in S.

$$\mathcal{C}(S) = \underbrace{\frac{\triangle_i(S)}{\binom{|S|}{3}}}_{\triangle \text{ "density"}} \times \underbrace{\frac{\triangle_i(S)}{\triangle_i(S) + \triangle_o(S)}}_{\text{isolation}} \tag{1}$$

From there, we define the cohesion of S in Equation 1 as a product of two factors. The first one is a triangular analog to the usual definition of density: it denotes the fraction of all possible triangles in a set of given size $|S|$ which are present in S. The second factor is an isolation factor where, intuitively, a penalty is awarded to the set when there exist outbound triangles, an example is given on Figure 2.

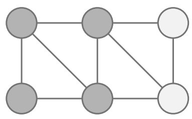

Figure 2: In this example, the set of dark nodes contains 4 nodes, features 2 inbound triangles and only 1 outbound triangles, leading to a cohesion $\mathcal{C} = \frac{1}{3}$.

The absence of impact of weak ties is naturally encompassed by the definition of the cohesion: given that the it only relies on counting triangles, deleting edges which do not belong to any triangles do not affect the number of triangles and therefore does not impact the value of the cohesion.

1.3 Evaluation on simple models

Random Networks In a random network $G(n, p)$, the expected number of triangles in a set S_k of size k is $\triangle_i = p^3\binom{k}{3}$ and the expected number of outbound triangle is given by $\triangle_o = p^3(n-k)\binom{k}{2}$. From there, the expected value (for large k and n) of the cohesion is given by $\mathcal{C}(S_k) \sim p^3\frac{k}{n}$. This exhibits the absence of expected community structure in random networks as the best possible community is the whole network.

Four groups The "four groups" test was introduced by Newman and Girvan to test the accuracy of a community detection algorithm. We here use the same framework to illustrate the pertinence of the cohesion. The setup is the following: consider a network of size $4n$ consisting of 4 groups of size n. Edges are placed independently between vertex pairs with probability p_{in} for an edge to fall inside a community and p_{out} for an edge to fall between communities. The cohesion of such a group is given, for large n, by $\mathcal{C} \sim \frac{p_{\text{in}}^5}{p_{\text{in}}^2+9p_{\text{out}}^2}$, which increases when p_{in} increases or p_{out} decreases as one would expect from a quality function.

2 Fellows

Defining a new metric of such a subjective notion as "how community-like is this set of nodes ?" raises the critical issue of its evaluation – or put another way, how does one defines the *quality of a quality function.* While in the previous section we exhibited that the cohesion makes sense on simple models, this is not enough to validate its use on real data. We now present Fellows [8], a large scale online experiment on Facebook which was conducted in order to provide an empirical evaluation of the cohesion. The gist of the idea behind Fellows is to quantify the accuracy of the cohesion by comparing it to subjective ratings given to communities by real persons.

2.1 The Experiment

'Fellows' is a single page web application which provides the user with a short description[2] of the experiment and its motivations. When a visitor wishes to take part in the experiment, they authorize the application to access their personal data on Facebook. From that point, the application connects to Facebook through the Facebook API [9] and downloads the list of their friends and interconnections between pairs of friends to reconstruct the social neighborhood of the user $\mathcal{N}(u)$. The application also publishes a message on the user's Facebook wall to invite their friends to participate. Using a simple greedy algorithm [10], similar in spirit rather than in metric to one previously introduced by Clauset [11], the application computes the user's groups of friends in their immediate social neighborhood by locally maximizing the groups' cohesion. It is important to note that all computation is done in JavaScript inside the user's browser and that no **identifiable** information is ever transmitted back to the application's server. Statistics on each of the groups are then sent to the server along with an anonymous unique user and session identifier (to be able to exclude users participating several times). The user's and their friends' birthdays and genders are also anonymously recorded.

Once those groups are computed, the application displays a list of names and pictures of friends which are present in the group featuring the highest cohesion (Fig. 3). The user is asked to give a numerical rating between 1 and 4 stars, answering the question "would you say that this list of friends forms a group for you?" They then have the opportunity to create a Friend List on Facebook, which is a feature which allows a better control on the diffusion of the information they publish on the social network. Once they submit the rating, it is uploaded to the application server where it is associated to the relevant group. In case the user has created a Friend List, the name they have given is also recorded. The user is then presented with another group and the process is repeated until either I) the user exits the application or II) all groups are rated and a message is displayed to thank the user for their involvement.

2.2 Progress

Fellows was launched on February 8$^{\text{th}}$, 2011. The authors published a link to the application on their Facebook walls and sent the URL to several active mailing lists. In less than a day, 500 users had taken part in the experiment and at the

[2] In English, French, Portuguese and Spanish.

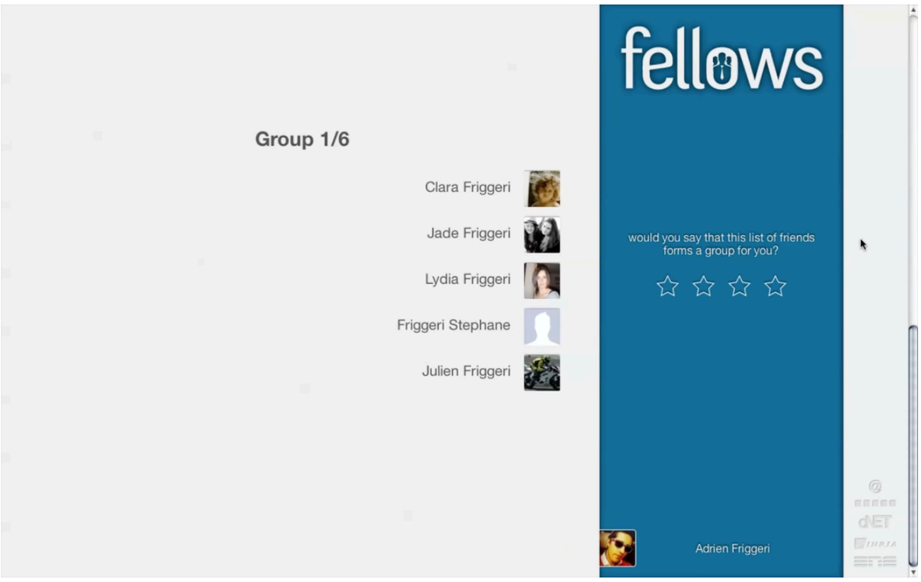

Figure 3: Screenshot of the application displaying a group.

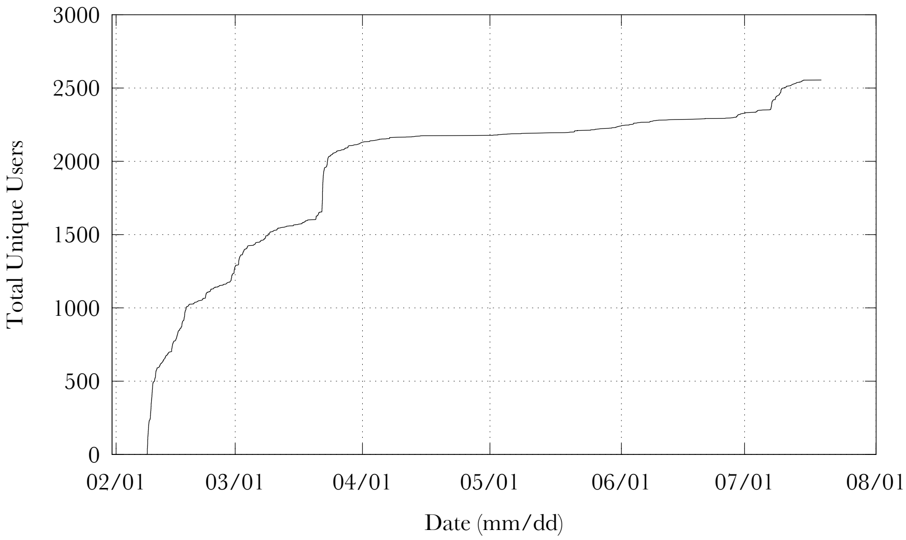

Figure 4: Evolution of total unique users through time.

time of writing, participations totaled 2635 persons (Fig. 4). Although unrelated to the evaluation of the cohesion, their are several facts which are interesting in the spread of the experiment. We observed a pattern of daily increase and nightly stagnation in the number of participants, corresponding to Western Europe timezone, which is coherent with data obtained from Google Analytics[3] indicating that the vast majority of Fellows' visitor came from France.

[3]A service from Google which provides detailed statistics of visitors access

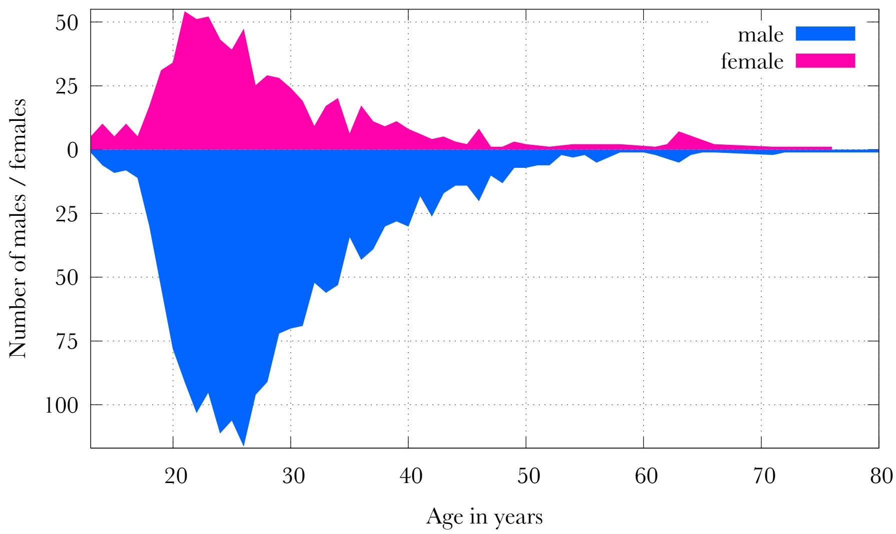

Figure 5: Densities of ages of male and female participants.

Moreover, the total number of unique users increases by bursts: observe how on March 23rd the number of users rises from ~ 1700 to ~ 2000 in a single day after having increased by 200 in two weeks. We have been able to trace back this sudden influx of participants to the publication of an article on a high traffic French blog on that date. Although this event was the most notable, we have been able to manually track down the origin of several different bursts – *e.g.* an email on a large mailing list on February 14th, a tweet by an *influent* twitterer on February 28th.

As stated above, when a user started the application for the first time, a message was automatically published on their Facebook wall to invite their friends to participate. Despite that fact, less than half the incoming traffic on the website came from Facebook. We conclude unfortunately that either the message was not appealing enough or that Fellows did not have the same viral potential as, for example, a double rainbow.

2.3 Population

In some cases, the participations were corrupted or incomplete – *e.g.* the user temporarily lost their internet connection. As a consequence, 78 participations had to be discarded, leaving 2557 valid contributions (1797 males, 698 females and 62 persons of unknown gender). The participants were on average 29.31 ± 8.99 years old – male subjects: 29.76 ± 8.80 yo, female subjects: 28.03 ± 9.24 (age distributions for male and female subjects are given in Figure 5).

On Facebook, the number of friends one might have cannot exceed 5000. The distribution of the number of friends is heterogeneous (Fig. 6), with 10% users having less than 74 friends and 90% users having less than 581, the median being at 237 friends.

3 Experimental Validation

In this section, we present the main contribution of this article, namely that the cohesion captures well the community-ness of a set of nodes. We first present statistics on the ratings which were obtained through the experiment and then exhibit how both cohesion and ratings are correlated.

3.1 Ratings overview

The 2157 valid subjects lead to the detection of 67750 groups. Given the fact that a user could stop the experiment at any time, 51161 groups received a rating – however, 78% of the subjects rated more than 90% of their groups. There are several explanations to those forfeitures, among others: I) that the user felt the groups they were presented with were of poor quality (the non-rated groups have on average a cohesion $\mathcal{C} = 0.108 \pm 0.107$) or II) that the user had too many groups to rate – although the number of groups is bounded, if a user has a lot of friends, that bound can be sufficiently high to discourage them.

Out of the 43589 rated groups, 25.1% received a rating of 1 star, 21.8% received 2 stars, 22.5% were rated 3 stars and 30.7% were awarded 4 stars. It is important to note here that the aim of the experiment was **not** to obtain the highest possible proportion of 4 stars ratings.

The first thing to notice is that the algorithm assigns all nodes of degree greater than 3 to at least one group. In practice, there is no reason that all nodes belong to at least one socially cohesive group: a social neighborhood might be constituted of an heterogeneous set of communities linked through weak ties and/or sparse meshes. Moreover, the social topology on Facebook and in the real world are not isomorphic, not only because people tend to add more distant acquaintances as Facebook friends, but also due to the presence of *non-human* profiles representing brands – incidentally, those would be better

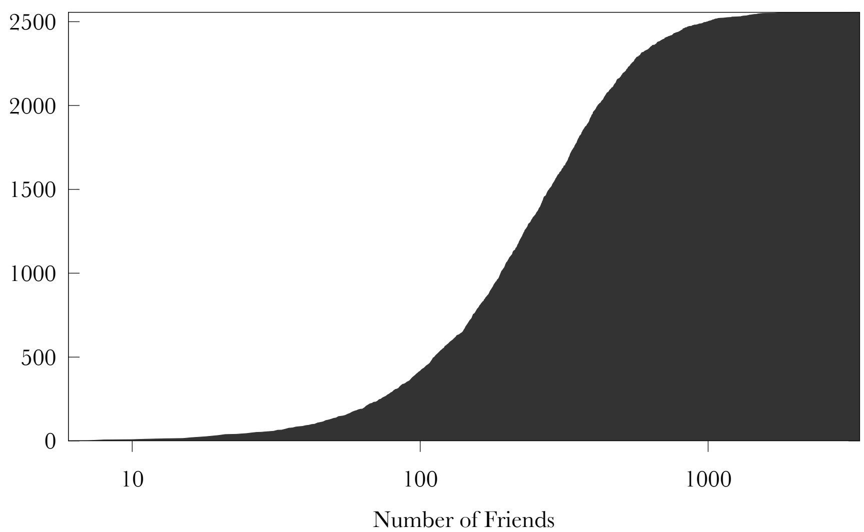

Figure 6: Distribution of users' number of friends.

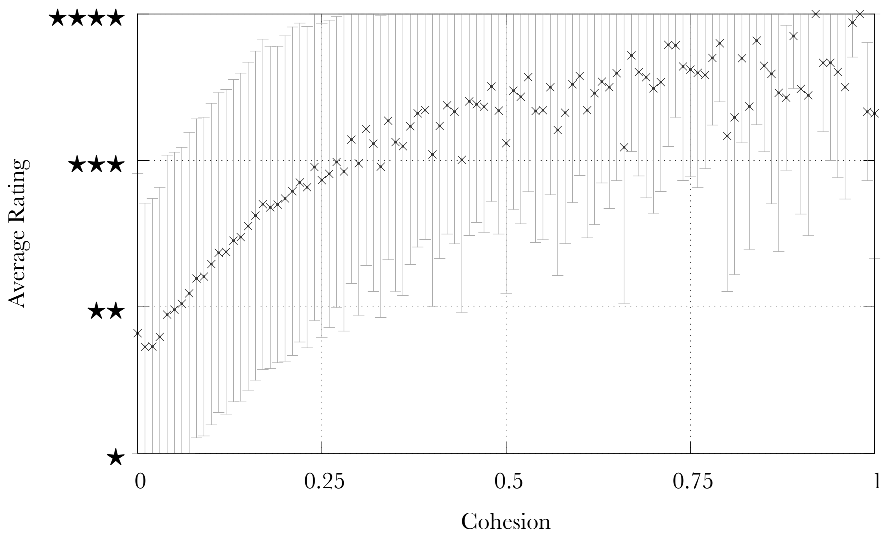

Figure 7: Average rating obtained by groups as a function of their cohesion.

represented as Facebook pages, but for some reasons some organizations prefer this structure.

Second, and perhaps more important, is that the aim of the experiment is not to evaluate the quality of the – rather simple – algorithm, but that of the underlying metric. In this context, obtaining low ratings is perfectly acceptable – and desirable – as long as they correlate to the cohesion.

3.2 Cohesion ∼ ★

We now exhibit the experimental links between a structural metric, the cohesion $\mathcal{C}$, and the subjective appreciation of a group's pertinence expressed as the average rating R given by users. On Figure 7, we discretize the cohesion of all groups in increments of 0.01 and we represent the average rating obtained by groups in the same increment. Both quantities are rank correlated (Spearman's correlation $\rho = 0.90$, p-value $= 9.1 \times 10^{-37}$). Thus, when the cohesion increases, so does the average rating, and conversely. Furthermore, $\ln \mathcal{C}$ and $\ln R$ are linearly correlated (Pearson's correlation $r = 0.97$, p-value $= 2 \times 10^{-61}$).

On Figure 8 we plot the distributions of cohesions of each of the four sets of groups of rating 1, 2, 3 and 4 stars. From this, we observe that the higher the rating, the higher the probability of obtaining high cohesions. Therefore, we conclude that the cohesion in a pertinent measure to evaluate the community-ness of a set of nodes, as it is highly correlated to its subjective evaluation.

Furthermore, it is interesting to look at the relation, if there is any, between the ratings and other graph metrics, such as the density of the considered set. On Figure 9 we plot the average rating obtained for groups of a given density. Groups having a density greater than $\frac{1}{3}$ tend to have the same average rating (between 2 and 3 stars). There seems however that for densities smaller than $\frac{1}{3}$ the rating increases with the density. To explain this fact, consider that

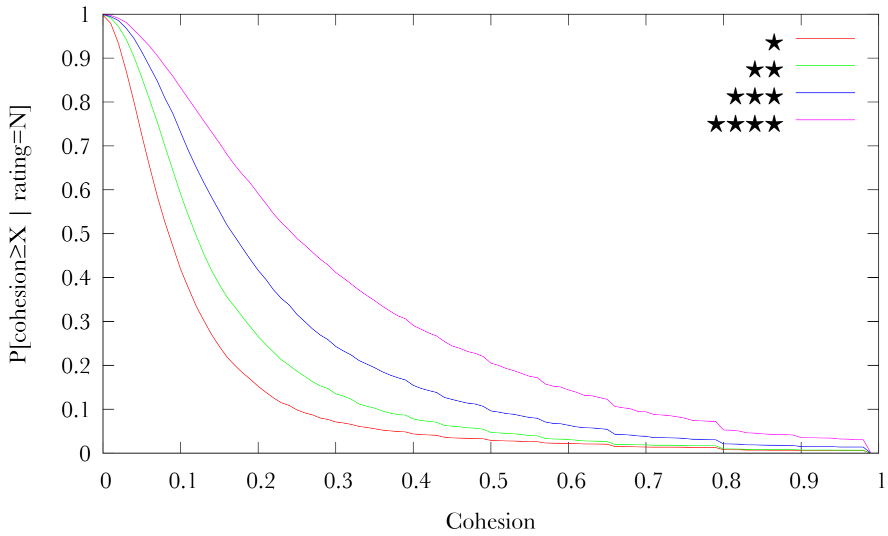

Figure 8: Normalized reversed cumulative distribution of cohesion for groups rated 1,2,3 or 4 stars ($\mathbb{P}[\text{cohesion} \geq X | \text{rating} = N]$).

$\mathcal{C}(S) < \frac{\triangle_i(S)}{\binom{|S|}{3}}$. Given that $\triangle_i(S) < m\sqrt{m}$ where m is the number of edges in S, there exist a bounding relation between density and cohesion as exhibited in Figure 10. Therefore, the lower ratings obtained by less dense groups can be explained by the fact that those have low cohesion, which itself is highly correlated to ratings.

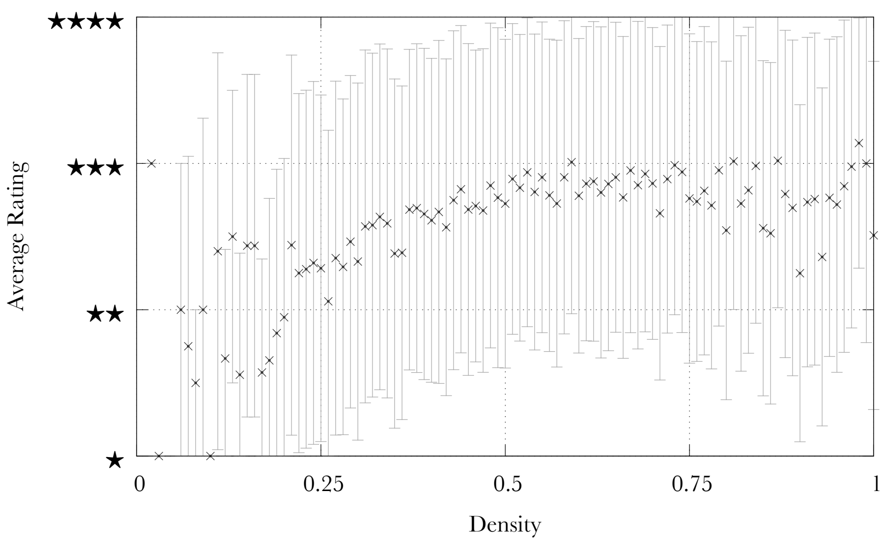

Figure 9: Average rating obtained by groups as a function of their density.

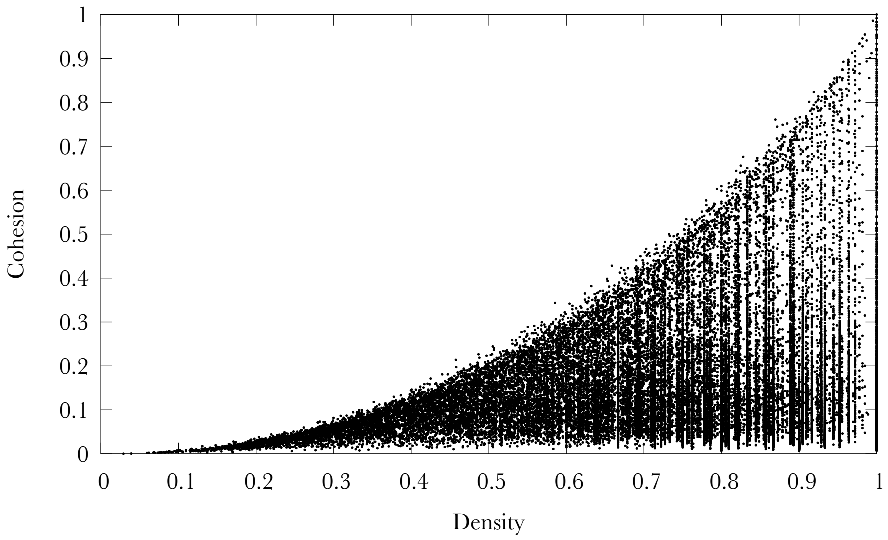

Figure 10: Density vs. cohesion.

For similar reasons, groups having a low clustering coefficient or low conductance display low ratings, because the clustering coefficient imposes a higher bound on the number of triangles in the set of nodes and the conductance imposes a higher bound on the number of outbound triangle. Yet again, high values of clustering or conductance do not yield high ratings, because the value of the cohesion can span a far greater range (*e.g.* a set with high clustering but a lot of outbound triangles might lead to a lower cohesion than that of a set with lower clustering but lower number of outbound triangles). As such, we assess that the cohesion leads to a more refined way of rating communities than by solely considering density, clustering or conductance.

4 Ongoing Work

4.1 Complexity

We conjecture that finding a subgraph of maximal cohesion in a given network is an NP-hard problem and are currently working on a proof. We define the problem Subgraph With Cohesion c as follows: Given a graph $G = (V, E)$ and a positive integer k, is there a subset S of V such that $\mathcal{C}(S) = c$. Counter-intuitively, the difficulty seems to arise from low rather than high cohesion values: here we show that Subgraph With Cohesion 0 is NP-complete but that Subgraph With Cohesion 1 can be solved in polynomial time. The problem for values of $c \in]0, 1[$ remains however open.

Subgraph With Cohesion 0: First note that $\mathcal{C}(S) = 0$ is equivalent to $\triangle_i(S) = 0$, thus the problem is equivalent to that of finding a triangle-free induced subgraph of G of size k. "Triangle-free" is a non-trivial and hereditary property and as such, per Lewis and Yannakakis [12], the problem is NP-complete.

Subgraph With Cohesion 1: In this case, a set S such that $\mathcal{C}(S) = 1$ is a clique which has 0 outbound triangles. We introduce the notion of triangle connectivity, an equivalence relation on edges of the network defined as such: two edges e and e' are said to be triangle connected if there exist a sequence of triangles $(t_i)_{0 \le i \le N}$ of G such that e is an edge of t_0, e' is an edge of t_N and $\forall i < N, t_i$ and t_{i+1} share a common edge. From there, if there exist a set S of cohesion 1, then all the edges of its induced subgraph must be in the same equivalence class and moreover the equivalence class cannot contain any other edges – if so, the associated subgraph would contain an outbound triangle for S. In conclusion, a set of size k with cohesion 1 exists if and only if there is an equivalence class containing $\binom{k}{2}$ edges. Given that it is possible to list all triangles in polynomial time and that by using a union-find algorithm one can compute all triangle connected equivalence classes in a time polynomial in the number of triangles, the problem of finding a set of nodes of size k having a cohesion 1 can be solved in polynomial time.

4.2 Extension to weighted networks

Besides complexity analysis, 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.

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

Conclusion

We have presented and justified the introduction of a novel measure, the cohesion, which quantifies the intrinsic community-ness of a set of nodes of a given network. We have then confronted the measure to real-world perception during a large-scale experiment on Facebook and found that the cohesion is highly correlated to the subjective appreciation of communities of Facebook users. Moreover, we have shown that there were no correlation between other metrics such as density and ratings. As such, we conclude that the use of the cohesion allows a good quantification of the community-ness of a set of nodes. Future works lie among others in the study of the cohesion from an algorithmic point of view and extensions to the metric to weighted networks.

References

[1] J. Hillery, George A., "Definitions of Community: Areas of Agreement," *Rural Sociology*, vol. 20, pp. 111 – 123, 1955.

[2] W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, pp. 452–473, 1977.

[3] M. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, 2004.

[4] 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.

[5] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[6] A. Rapoport, "Contributions to the Theory of Random and Biased Nets," *Bulletin of Mathematical Biophysics*, vol. 19, pp. 257–277, 1957.

[7] M. Granovetter, "The Strength of Weak Ties," *Amer. J. of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[8] A. Friggeri, G. Chelius, and E. Fleury, "Fellows, a social experiment," 2011. [Online]. Available: http://fellows-exp.com

[9] Facebook, "Graph API," 2011. [Online]. Available: http://developers.facebook.com/docs/api

[10] A. Friggeri, G. Chelius, and E. Fleury, "Egomunities, Exploring Socially Cohesive Person-based Communities," INRIA, Research Report RR-7535, 02 2011. [Online]. Available: http://hal.inria.fr/inria-00565336/en/

[11] A. Clauset, "Finding local community structure in networks," *Physical Review E*, vol. 72, no. 2, aug 2005.

[12] J. M. Lewis and M. Yannakakis, "The node-deletion problem for hereditary properties is np-complete," *Journal of Computer and System Sciences*, vol. 20, no. 2, pp. 219 – 230, 1980.

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