A Quantitative Theory of Social Cohesion

Adrien Friggeri

THÈSE

en vue de l'obtention du grade de

DOCTEUR

DE L'ÉCOLE NORMALE SUPÉRIEURE DE LYON

UNIVERSITÉ DE LYON

Discipline: Informatique

Laboratoire de l'Informatique du Parallélisme

École Doctorale InfoMath

Présentée et soutenue publiquement le 28 août 2012

par M. Adrien FRIGGERI

A Quantitative Theory of SOCIAL COHESION

Directeur de thèse:

M. Eric FLEURY

Co-directeur de thèse:

M. Guillaume CHELIUS

Après l'avis de:

M. Vincent BLONDEL

M. Santo FORTUNATO

Devant la commission d'examen formée de:

M. Vincent BLONDEL (Rapporteur)

M. Guillaume CHELIUS (Directeur)

M. Eric FLEURY (Directeur)

M. Santo FORTUNATO (Rapporteur)

Mme. Anne-Marie KERMARREC (Présidente)

M. Pierre MERCKLÉ (Membre)

A Quantitative Theory of Social Cohesion

A Quantitative Theory of Social Cohesion

Adrien Friggeri

To a Dreamer,

Contents

As I was growing increasingly frustrated with the lack of framework to set apart good from bad communities, I was struck by a profound feeling that triads and not just edges were somehow intricately related to the answer to my problem.

Although intuition shaped this idea, it took me two years to connect the dots and appreciate its pertinence, beautiful nature and complex implications. This thesis is a didactic account of those findings.

I cannot thank Éric Fleury and Guillaume Chelius enough for their support and for granting me the freedom to pursue whatever crossed my mind.

Lyon, June 2012

Résumé

L'analyse des réseaux sociaux est un domaine marqué par une ambiguité constante entre les outils et l'objet d'étude. En tant que praticiens d'une science focalisée l'ensemble des interactions humaines, un volume difficile à appréhender, nous sommes voués à developer la théorie parallèlement à l'études des données. Pire encore, la demande pour une analyse quantitative augmentant, le domaine est confronté à la mesure de notions de plus en plus intangibles et subjectives.

Ceci est similaire à ce que traversèrent les sciences naturelles au cours du XVIIème siècle, à l'exception du fait qu'Isaac Newton fut capable de décrire la gravitation et les lois du mouvement sans avoir besoin de définir et quantifier au préalable la notion même de longueur. L'analyse des réseaux sociaux a certes un certain nombre de statistiques à sa disposition, statistiques qui peuvent servir à décrire tout ou partie du réseau. Exception faite de quelques unes de celles ci, telles le coefficient de clustering ou la densité, il n'y a en général aucune relation immédiate entre la quantité et un fait social observable. En conséquence de quoi, une large partie de notre travail consiste non seulement à trouver des manières de quantifier certains aspects des interactions sociales de manière à extraire de l'information du réseau, mais aussi à prouver que les choix effectués dans l'élaboration de la quantification sont raisonnables et nous permettent de mesurer précisément un phénomène.

D'un point de vue plus philosophique, il est intéressant de mentionner la question de la nature de ces mesures ou statistiques sociales. Sont-elles ce qu'Auguste Comte appelait des "lois invariables" de la nature, ou au contraire rien de plus que des gadgets mathématiques qui ont le bon goût de corréler fortement à un fait social ? Je n'essayerai pas de répondre à cette question, bien qu'il me semble nécessaire de noter qu'il existe à l'heure actuelle bien peu de lois acceptées en tant que tel, un fait déjà noté par Merton[1] et qui demeure vrai à cette date :

> Malgré les nombreux volumes traitant de l'histoire des théories sociologiques, et malgré la pléthore d'investigations empiriques, les sociologues […] sont amenés à discuter les critères logiques des lois sociologiques sans pour autant citer un seul example satisfaisant à ces critères.

Ceci donne un aperçu de la raison pour laquelle il existe si peu de ces lois sociologiques, étant donné la contradiction avec l'une des première propriétés des lois de la nature telles qu'exprimées par Feynman[2] : une telle loi doit rester vrai en toute instance observable, et en au moins une qui doit avoir été rapportée. J'ai attaché un soin tout particulier, quand c'était possible, d'apporter une preuve mathématique à mes assertions, ou à tout le moins les justifier en m'appuyant sur des données empiriques.

La détection de communautés est une branche de l'analyse des réseaux sociaux – et plus généralement de l'analyse des réseaux –

1 Robert K Merton. *On Theoretical Sociology*. Free Press. New York, 1967.
2 Richard Phillips Feynman. *The character of physical law*. Modern Library. 1967.

qui a attiré une attention considérable ces dernières années. Les communautés sont omniprésentes dans les réseaux sociaux, par exemple on pense intuivement aux familles, groupes d'amis ou collègues de travail en tant que communauté. Il y a par ailleurs une multitude d'application[1] à la détection de communautés. De manière évidente, l'application la plus visible est liée à l'analyse purement structurelle d'un réseau social de manière à déterminer de quelle façon il est organisé en communautés interconnectées et recouvrantes.

Prenous par exemple une société confronté à des problèmes d'ambiance ou de communication. En détectant les communautés dans le réseau social de communications entre les employés, et en comparant celui ci à l'organigramme officiel de la société, il devrait être possible d'identifier des équipes dysfonctionnels et en tirer les conséquences nécessaires – par exemple, séparer une équipe, ou embaucher une personne ayant certaines caractéristiques qui lui permettrait de s'intégrer à l'interface entre plusieurs communautés.

L'explosion des medias sociaux en ligne tels que Facebook ou Twitter s'est accompagnée d'inquiétudes concernant notamment la confidentialité des données et le respect de la vie privée. En utilisant des outils de détection de communauté pour déterminer algorithmiquement le *Gruppengeist*, il serait possible de mettre en correspondance une publication – statut, photos partagées, etc. –

1 Bien que la détection de communautés ait été utilisée dans l'analyse de réseaux plus généraux, je me concentrerai sur les réseaux *sociaux*. Il est utile de préciser que les travaux présentés dans cette thèse *pourraient* s'appliquer à l'étude de réseaux d'autres nature, mais les résultats présentés ici étant soutenus par des données sociales, je n'ai pas de preuve à apporter à une telle généralisation.

et le groupe d'amis auquel elle est destinée. De la même manière, si l'on considère un réseau social où deux personnes sont liées si elle partagent des centres d'intérêt similaires, la connaissance de leurs communautés peut être non seulvement utilisée de manière à recommander de nouveaux contenus, mais aussi à mettre en relation des individus partageant des passions communes.

Dans le Chapitre 2, je présente une particularité du domaine qui a trait à la diffulté à définir son objet d'étude, ceci malgré – ou à cause de – sa simplicité à illustrer par des exemples. En l'occurrence, un grand nombre d'algorithme a été développé de manière à trouver des communautés, sans pour autant définir formellement ce qu'était une communautés – un certain nombre peuvent prétendre introduire une définition quantitative, dans la limite où l'on considère que donner une formule sans la justifier empiriquement est une définition.

En juin 2010, j'ai naïvement commencé à réfléchir à quelques concepts qui, je pensais, permettraient de trouver des communautés locales et recouvrantes de manière plus précise. Sans rentrer dans des détails superflus, dont la plupart m'échappent aujourd'hui, disons simplement que j'accouchai d'une floppée d'algorithmes desquels je n'avais aucun moyen de déterminer lequel était le pire. Rétrospective, il me semble évident que mes chances de trouver un bon algorithme de détection de communautés était faible, ne sachant alors pas ce qu'était une bonne communauté.

Mes premiers instincts furent de me servir de la modularité, la manière standard de juger de la pertinence d'une partition en communautés. Mais non seulement cette notion n'avait

aucun sens en terme de communautés recouvrantes, celle ci ne s'appliquait tout simplement pas au cas d'une communauté en elle même: la modularité n'a que faire de savoir si une communauté est bonne ou non, elle ne s'intéresse uniquement à la qualité d'un bon ensemble de communautés. Pire encore, il n'y avait alors aucun moyen acceptable de juger de la qualité d'une communauté, je devais donc en trouver un par moi même.

La réponse m'arriva subitement, et l'utilisation de triades sociales en lieu et place d'arêtes pour définir les aspects structurels des communautés m'apparurent soudain comme une évidence. Dans le Chapitre 3, je décris la progression intellectuelle qui me mena à concrétiser cette intuition, des sources sociologiques aux contraintes qu'une telle métrique devait suivre et finalement à la définition de la mesure elle même – la cohésion, continuant ensuite à explorer des considérations et propriétés mathématiques et terminant sur son évaluation sur des réseaux artificiels.

Néanmoins, comme je le disais précédemment, il y a une longue tradition dans le domaine de proposer des métriques et statistiques de graphe, balayant leur justification d'un simple "on dirait que ça fonctionne". Dans le Chapitre 4 je décris pourquoi et comment j'ai choisi de suivre le chemin inverse et ai lancé une expérience de grande envergure de manière à valider l'utilisation de la cohésion. Bien qu'ayant mes intuitions, ce n'est qu'après avoir constaté une corrélation forte entre la cohesion et la perception subjectives des communautés que j'ai été certain que la cohésion était un indicateur solide du caractère communautaire d'un groupe de personnes.

Ayant à ma disposition une manière quantitive de définir ce

qu'était une communauté – à savoir un groupe cohésive de personnes – il était tout naturel de s'intéresser à son utilisation pour élaborer un algorithme de détection de communautés. Le problème est néanmoins complexe, et je montre dans le Chapitre 5 que le problème de trouver une communauté maximalement cohésive est $\mathcal{NP}$-dur. Je montre par ailleurs que le problème dual de trouver des communautés minimalement cohésive – à taille fixée – est $\mathcal{NP}$-dur lui aussi, ce qui peut avoir une incidence dans la détection de communautés molles de manière à identifier les faiblesses socio-structurels d'un réseau social.

Étant donné qu'il n'était donc a priori pas possible de trouver un algorithme polynomial pour optimiser la cohésion, j'ai développé C^3, une heuristique que je présente dans le Chapire 6. C^3 a plusieurs avantages comparé aux algorithmes existants: premièrement, il s'appuie sur la cohésion, et en tant que tel il chercher à maximiser une quantité que l'on sait à présent être corrélée à la qualité des communautés. Par ailleurs, il détecte des communautés recouvrantes, ce qui dans le contexte des réseaux sociaux a plus de sens que l'approche inverse – la partition en communautés – qui elle assigne un individu a une communauté, approche absurde s'il en est car une personne devrait avoir la possibilité d'appartenir à la fois à sa famille et à son groupe d'amis. Finalement, C^3 s'appuie uniquement sur des informations locales et donc n'introduit pas de fausse dépendance de la communauté sur l'ensemble du réseau.

Pour valider l'utilisation de C^3, je décris ensuite son application à des réseaux d'agrément de votes de sénateurs des États-Unis, pour chacune des sessions du Congrès. Dans le Chapitre 7, je montre que l'algorithme extrait des communautés

pertinentes, liées aux partis politiques. Il est intéressant de noter que l'on observe une continuité temporelle dans ces communautés, quand bien même celles ci sont calculées de manière indépendantes à chaque pas de temps. Par conséquent, je conclus que C^3 est une heuristique qui non permet de détecter des communautés pertinentes. Par la suite, je décris ces résultats d'un point de vue politique et historique.

Un des aspects douloureux de l'analyse des réseaux sociaux réside, à cause du volume de données à prendre en compte, est la complexité de la représentation visuelle, intelligible et informative des résultats. Dans le Chapitre 8, je présente une extension aux algorithmes de placements physiques qui est spécialement étudiée pour visualiser des communautés prédéfinies. Ces algorithmes classiques utilisant un modèle physique pour placer les sommets sur un plan en simulant des forces d'attraction et de répulsion entre les sommets, cette extension ajoute un nouveau type de force élastique de manière contraindre les sommets d'une même communauté au sein d'un cercle. Non seulement l'algorithme RubberBand est conçu pour être compatible avec tout type d'algorithmes de cette sorte – et peut donc être adapté dans le futur à de nouveaux algorithmes, mais il mène de manière consistante à des résultats plus lisibles concernant la visualisation de communautés, ce que nous verrons au travers d'une évaluation quantitative.

Le chapitre final de ma thèse porte sur un aspect souvent ignoré de l'analyse des réseaux sociaux. Contrairement à la sociologie traditionnelle et à la psychologie dont le point focal est l'individu, l'analyse de réseaux sociaux favorise l'étude des interactions entre acteurs au détriment de leurs traits personnels.

En utilisant C^3 et la visualisation décrite précédement, je m'attacherai dans le Chapitre 9 à étudier conjointement des données décrivant la personnalité d'utilisateurs de Facebook et leur réseau social de manière à mettre en lumière les liens entre psychologie et topologie du réseau. Nous verrons qu'il y a un impact fort de l'extraversion sur l'entourage social, et que cette notion est connectée au nombre d'amis, nombre de communautés, leur taille et leur recouvrement.

~

Les contributiosn de cette thèse sont multiples et ont trait à une variété de domaines. En son cœur, elle introduit la cohésion, une mesure quantitative définissant la notion sociologique de communauté. Cette notion est ensuite étudiée du point des vue des mathématiques et de la théorie des graphes. Algorithmiquement, la cohésion est une quantité qui est $\mathcal{NP}$-dure à la fois à maximiser et à minimiser. De manière à détecter des communautés cohésives, j'introduis C^3, une heuristique que je valide en notant sa pertinence dans l'étude des groupes d'agréments du Sénat états-uniens. Je propose ensuite RubberBand, une extension aux algorithms de placements physiques classiques utilisés en dessin de graphes, qui permet une visualisation efficace de la structure d'un réseau. Finalement, j'utilise ces contributions pour analyser et représenter l'impact de traits psychologiques sur la structure sociale. En un sens, cette thèse vit à l'intersection entre Informatique, Sociology, Visualisation de données, Politologie et Psychologie, bien que sa contribution majeure soit issue de la théorie des graphes sous forme d'une approche quantitative de la cohésion sociale.

Prologue

O chestnut-tree, great-rooted blossomer,
Are you the leaf, the blossom or the hole?
O body swayed to music, O brightening glance,
How can we know the dancer from the dance?
Among School Children
William Butler Yeats

Social network analysis is a field marked by a constant ambiguity between the tools and the object. As practitioners of a young science with a focus set on the difficult to grasp collection of human interactions, we are vowed to develop the theory and at the same time conducting our studies. Worse yet, as the demand for more quantitative analysis arises, the field has to devise ways of measuring more and more intangible and subjective notions.

This is somehow similar to what happened in the natural sciences in the XVIIth century, except that Isaac Newton was able to describe universal gravitation and the laws of motion without needing to go through the hassle of finding a way of defining and quantifying what a length is. Social network analysis does have statistics which can be used to describe all or part of the network, but except for few of them such as clustering or density, there are usually no immediate relationships between the quantity and an observable social fact. As such, a large part of our work consists

not only in finding ways to quantify aspects of social interactions to extract information from the network, but also to prove that our choices in building the quantification are sound and allows us to capture accurately a given phenomenon.

On a more philosophical note, the question of the nature of those social metrics and statistics is worth mentioning: are they what Auguste Comte used to refer to as "invariable laws", or nothing more than useful mathematical gizmos which just happen to correlate to a social fact ? I will not attempt to answer this question, although it must be stated that few such laws exist, a fact already noted by Merton[1] and remains true today:

> Despite the many volumes dealing with the history of sociological theory and despite the plethora of empirical investigations, sociologists [...] may discuss the logical criteria of sociological laws without citing a single instance which fully satisfies these criteria.

This provides insight into why there are so few examples of sociological laws, as it is in contradiction with one of the first properties of the laws of nature as identified by Feynman:[2] such a law should be true in all observable instances, let alone in at least one, which should be reported. I took particular care to provide, when possible, a mathematical proof for all statements or at the very least back those with empirical data.

Community detection is an area of social network analysis – and more generally of network analysis – which has attracted

1 Merton, op. cit.
2 Feynman, op. cit.

important attention this last years. Communities are ubiquitous in social networks, for example one intuitively thinks of families, groups of friends or co-workers as communities, and there is a multiplicity of applications[1] to community detection. For obvious reasons, the most visible application is related to the purely structural analysis of a social network to determine how it is architectured into interconnected and overlapping communities.

Consider for example a company which has been having some morale or communications issues, by detecting the communities in the social communication network between employees and comparing those with the official organization of the company, it ought to be possible to identify dysfunctional teams and take appropriate action – *e.g.* split up some teams or hire individuals having certain traits which would insert them at the interface between communities.

With the boom of online social networking websites such as Facebook and Twitter, there has been an growing concern about privacy-related issues. By coupling community detection to algorithmically be able to identify the *Gruppengeist*, one could match publications such as status updates or uploaded photos to be displayed only to the relevant community of friends. Continuing in this vein, if we consider an interest social network where two individuals are linked if they share similar interests, the knowledge

1 Although community detection has been used in the analysis of more general networks, I shall focus on *social* networks. It should be noted though that the work presented in this thesis *could* be used in the study of other complex networks, but since the validation was done using social data, I have no solid evidence to back such claims.

of its communities can be exploited to not only recommend new content to someone but also introduce like-minded people to one another.

On a more dramatic note, social network analysis has been used in law enforcement and counter-terrorism. For example, journalist Chris Wilson argues that social network analysis was instrumental in Saddam Hussein's apprehension and provides a lengthy, well-researched narrative to back it up.[1] Here is how Maddox, an U.S. interrogator, recounts how suspects unintentionally divulged useful information in their efforts to protect what they believed was more critical information:

> In piecing together a trail through his network, Maddox says detainees often simply told him what he wanted to know. "They're not going to tell me about the insurgency", he explained. "But they'll talk about who's drinking buddies with who". In thinking that they were deflecting the interrogators, lower-level operators were in fact leading Maddox closer to his target. These detainees, in a way, were making precisely the same mistake that the American military made at the start of the Iraq war. Institutional information about the insurgency wouldn't bring coalition troops closer to Saddam's hiding place. The social information that these lower-level Musslits provided was much more

1 Chris Wilson. *Searching for Saddam*. Slate Magazine. Feb. 2010. URL: http://www.slate.com/articles/news_and_politics/searching_for_saddam/2010/02/searching_for_saddam_5.single.html.

> valuable. Maddox wanted to know the names of Saddam's friends, not his former colleagues.

It can be envisioned that in the future, community detection techniques might be used in order to identify structural inconsistencies between intelligence data and what would be expected from typical communities. In turn, this could lead to the uncovering of yet undiscovered and unidentified criminals.

As we shall see in Chapter 2, one of the most striking peculiarity of the field of community detection surely is that its object of study has remained vague despite – or because – being so simple to illustrate through examples. Without spoiling the pleasure of the reader, suffice it to say that countless community detection algorithms have been developed to find something which was not formally defined – some actually may be considered to have introduced quantitative definitions, insofar as giving a formula without any experimental validation qualifies as such.

In June of 2010, with the naivety of the newcomer, I started working on some concepts which I thought would lead to an accurate detection of local overlapping communities. Without going into superfluous details – much of which I barely remember anyway – let us just say that I came up with a flurry of algorithms of which I had no way of telling the least bad apart. In retrospect it seems obvious that my chances of finding a good community detection algorithm were slim given that I did not know what a good community was.

My first instincts were to rely on the modularity, which was the standard way of rating a partition into communities, but not

only did it made no sense when used on overlapping communities, the notion did not apply to a community by itself: modularity does not care if a community is good or bad, only if a partition is a good enough set of communities. Worse even, there was no acceptable way of judging the quality of a community, hence I had to come up with my own.

At some point in time I had an epiphany, and the use of social triads rather than only links to define the structural aspects of communities suddenly entered my mind. In Chapter 3, I describe the intellectual progression which I followed to concretize this intuition, from the sociological underpinnings, to the constraints to which such a metric ought to obey, to the actual definition of the metric – the cohesion, to finally some mathematical considerations, properties and evaluation on artificial networks.

However, as stated earlier, there has been a long tradition in the field of proposing graph metrics and statistics with nothing more that "it seems to work" as a justification of their use. I recount in Chapter 4 why and how I chose to take the opposite course and launch a large-scale experiment relying on Facebook to validate the use of the cohesion. Although I had my intuitions, it is only after observing a high correlation between the cohesion and the subjective perception of communities that I could be certain that the cohesion is a strong indicator of the communitiness of a group of people.

Having a quantitative way of defining a community – that is, a strongly cohesive group of people – it was only natural to attempt to make use of the metric in a community detection algorithm. The problem is however a complex one, to the point that I show in Chapter 5 that the problem of finding such maximally

cohesive subgraphs is $\mathcal{NP}$-hard. I should also mention that the dual problem of finding communities of low cohesion is also $\mathcal{NP}$-hard, which might have an impact on some applications relying on the cohesion to identify the socio-structural weaknesses of a network.

Since finding a polynomial algorithm to optimize the cohesion was out of the picture, I developed C^3, a heuristic algorithm which I present in Chapter 6. C^3 has several advantages compared to existing community detection algorithms: first and foremost it relies on the cohesion and as such attempts to maximize a quantity which is known to be correlated to the quality of communities. It then detects overlapping communities; in the context of social networks this makes the most sense since the opposite approach, disjoint communities, only assigns an individual to one community, which is absurd as people should surely be allowed to belong to a group of friends as well as their families. Finally, only taking into account local information, C^3 does not introduce a false dependency of the community on the whole of the network.

To validate the use of C^3, I will then describe its application to voting agreement networks of United States Senators for each session of the U.S. Congress. In Chapter 7 we shall see how the algorithm extracts relevant communities, tightly related to political parties. More noteworthy is that, when detecting the communities independently on each agreement network, we shall observe that there is a continuity in the communities despite the absence of any temporal consideration in the algorithm. As such, we will conclude that although C^3 is a heuristic algorithm it does provide relevant and cohesive communities, and will relate those result to historical and political facts.

One of the growing pain points in social network analysis resides, due to the sheer volume of data involved, in the complexity of representing intelligible and informative results. In Chapter 8 I shall introduce an extension to classical force-directed layout algorithms which is tailored to visualize predefined communities. Since those classical layouts use a physical model to place nodes on a plane by simulating attraction and repulsion between pair of nodes, the extension adds a new type of force modelled upon rubber bands to constrain nodes of a same community inside enclosing circles. Not only was RubberBand designed to be compatible with any type of force-directed algorithm – and as such can be adapted to most of existing and future algorithms – but we will also see through quantitative benchmarks that it consistently leads to a better community visualization than if it was not used.

The final chapter of my thesis is devoted to an oft ignored aspect in social network analysis, whereas in traditional sociology and psychology the focus is primarily set on the individual, social network analysis tend to favour the study of relationships and interactions between actors without taking into account their personal traits. Having at hand both an overlapping community detection which was shown to provide meaningful communities and a way of visually illustrating the community structure of a network, in Chapter 9 I shall cross a dataset providing subjects' personality traits with social network information from Facebook to observe the way in which psychology shapes the topological structure of the network. We shall see that there is a deep impact of the trait of extraversion on the community structure of subjects' social entourage as it is connected to the number of friends, the number of communities and size and compartmentalization of

thereof. This influence is particularly visible when representing ego-networks using RubberBand.

The contributions of this thesis are multiple and relate to a broad variety of fields. At its core, it provides the cohesion, a quantitative way of defining the sociological notion of community; notion which is then analyzed from an mathematical and graph theory point of view. Algorithmically, the cohesion is a non-trivial graph metric which is $\mathcal{NP}$-hard both to maximize and minimize. In order to be able to detect cohesive communities, I propose C^3, a heuristic algorithm which I validate by noting the relevance of its output on dynamical data describing the agreement groups of the United States Senate. I then introduce RubberBand, an extension to classical force-directed layout algorithms, which allows for the efficient visual representation of a network's community structure. C^3 and RubberBand are then used to analyze and represent the impact of psychological traits on social structure. In a sense, this thesis lies at the intersection between Computer Science, Sociology, Data Visualization, Politology and Psychology, although its main contribution is a graph theoretic and quantitative approach to social cohesion.

The Sociological Construct

From Hillery to G=(V,E)

Φύσει μέν ἐστιν ἄνθρωπος ζώον
πολιτικόν
Politics
Aristotle

Social community is a subjective notion and although the term is used rather intuitively, it appears that no formal consensus has been reached on the nature of what qualifies a community as such. Despite this fact, the concept has been actively studied in sociology and has attracted a lot of attention in the last fifty years in the field of Social Network Analysis.

An Omnibus Word

In the last decades, the term community has expanded to encompass a wide array of meanings. From the media often referring to the "local community", to the emergence of so-called "online communities", the term has been alleviated its meaning to the more general term *group*, or dare I say, in our times and place, a community often seems to be no more than a bunch of people.

Although the focus of their study is on educational communities, Grossman et al. raise the issue that "this confusion [on the meaning of community] is most pronounced in the ubiquitous "virtual community", where, by paying a fee or typing a password, anyone who visits a web site automatically becomes a 'member' of a community".[1]

The term community is not, however, devoid of meaning, and its existence and use bare reason. Before expliciting the actual structure which is of interest, I deem important to observe what intention people, scholars and laymen alike, wish to convey when employing the word community.

In her thesis,[2] Stuckey faced the same semantic issue and, in a section appropriately titled *The difficulty of defining community*, introduces the term by referring to its etymology – from the Middle English "communite" meaning *citizenry*, from Latin "communitas", *fellowship* and "communis", *common* – and states that the use of the term has always been open to interpretation, although the notion of "something shared" is deeply rooted in the term.

In the *community* entry in the Social Science Encyclopedia,[3] Azarya introduces the following interrogation: does the notion of community refers to a group of people, or does it refer to the sense

1 Pamela Grossman, Sam Wineburg, and Stephen Woolworth. *What makes teacher community different from a gathering of teachers?* Tech. rep. Seattle, Dec. 2000.

2 Bronwyn E. Stuckey. "Growing online community: core conditions to support successful development of community in Internet-mediated communities of practice." PhD thesis. University of Wollongong, 2007, p. 49.

3 Adam Kuper and Jessica Kuper. *The Social Science Encyclopedia*. Ed. by Adam Kuper and Jessica Kuper. 2nd ed. Routledge world reference. London: Routledge, 2003, pp. 195–197.

of belonging to a group. Interestingly, the fact that the term confusion appears in both Azarya's and Grossman's citations should be taken as an indication of the difficulty of grasping what community means.

> A preliminary confusion arises between community as a type of *collectivity* or social unit, and community as a type of *social relationship* or sentiment. The root of the problem could be traced to Tönnies's *Gemeinschaft*, which uses the term to describe both a collectivity and a social relationship. Subsequently, most scholars have used community to connote a form of collectivity (with or without *Gemeinschaft* ties), but some, such as Nisbet, have kept the community-as-sentiment approach alive in their emphasis on the quest for community and their concern with the loss of community in modern life. These approaches are clearly mixed with some nostalgia for a glorious past in which people were thought to be more secure, less alienated and less atomized.[1]

Further, Azarya describes Schmalenbach's theory of the *bund* which provides an alternative to Tönnies's Community and Society (*Gemeinschaft und Gesellschaft*). According to Schmalenbach,[2] *Gemeinschaft* implies relationships which emerge naturally from day-to-day interactions, whereas emotional ties in community-as-sentiment are better describes by what he called communion

1 Ibid., p. 195.

2 Hermann Schmalenbach. "The sociological category of communion". In: *Theories of society*. Theories of society, 1961, pp. 331–347.

(*bund*) as they involve a sense of belonging, for example in political or religious groups, as opposed to *ad-hoc* groups that do not feature such an *esprit de corps*.

To the extent that Azarya then focuses on the collectivity aspect of community rather than the sentiment, and given that this thesis aims for a quantified approach of social cohesion in terms of network science, I shall restrict myself to the more observable community-as-social-unit – that is, I wont attempt to distinguish a prayer group from a poker group – since the scope of this thesis is set on structural aspect of the social network which in our definition does not encompasses the semantics of interactions between actors of the network.

It is however important to note that the sense of belonging is distinct from the recognition of a community as such from one of its members, in the sense that "I can recognize this is my poker group, because we play poker together, although I might not *feel* I belong to the Great Entity of the Poker Group".

The vagueness of the term is such that in the past 70 years a subgenre of sociological papers focused on reviews of the different definitions of the term have emerged. In 1955, George Hillery, Jr. analyzed[1] 94 different sociological definitions of the term community both from a quantitative and qualitative standpoint.

> All except three of the definitions clearly mention the presence of a group of people, *i.e.*, persons in social

1 George A. Hillery Jr. "Definitions of community: Areas of agreement". In: *Rural sociology* 20.2 (1955), pp. 111–123.

> interaction.
> […]
> The definitions which include social interaction fall into six categories: those which mention the presence of some *geographic area*, those which mention the *presence of some common characteristic other than area*, and four residual categories.
> […]
> As should be obvious from the classification scheme […], all of the definitions cannot be correct – *i.e.*, community cannot be all of the definitions in their entirety. For example, the logical principle of contradiction denies anything the quality of being an area and not being an area.[1]

The most striking aspect of his analysis is not that he was actually able to find 94 different definitions, which, in the authors own words, "are not all of the definitions of the community. However, it is believed that the picture given is a fairly representative one", but rather that their commonality is reduced to the most general idea of what a community might be: once again, a bunch of people.

> There is one element, however, which can be found in all of the concepts, and (if its mention seems obvious) it is specified merely to facilitate a positive delineation of the degree of heterogeneity: all of the definitions deal with people. Beyond this common basis, there is no agreement. To be sure, people

1 Ibid.

> who are grouped together will normally engage to some extent in social interaction, but definitions were found (in the class termed "Ecological Relationships") which deny social interaction as an essential feature of community life.
> […]
> Since these [ecological relationships] are the only definitions in this survey which exclude social interaction from the concept of community, and since social interaction is at least one major concern of all of the other definitions, these two must be considered the most radical deviants.[1]

The essence of Hillery's study is that the only area of agreement on the nature of community, at that time, was that communities dealt with *people*. He did however find that 69 definitions suggested that community is related to the notion of *area*, *common ties*, and *social interaction*.

Quarter of a century later, Poplin[2] carried a similar analysis on 125 sociological definitions only to find consistent evidence that Hillery's observation on the core definition of community stood the trial of time, although he noted that language, description and terminology had evolved.

Finally, Stuckey built on Hillery and Poplin's studies and examined 25 contemporary definitions of community. Most of her findings are in adequation with the previous works, although

1 Ibid.

2 Dennis E. Poplin. *Communities: A survey of theories and methods of research*. Macmillan. 1979.

she also noted that "definitions included conditions specific to the context in which the community may be developed"[1] such as teaching community or commercial orientation. Moreover, it is interesting to note that she had to tweak the notion of *area* to "go beyond the geographical, to be considered a temporal location and, to include area as a virtual place online" in order to fit into Hillery's categories of communities.

The geographical aspect of communities has also been noted by Azarya in the Social Science Encyclopedia,[2] where he introduces the territorial and non-territorial approaches, only to state that the territorial approach has dimmed with the emergence of newer communication technologies. In that context, it is legitimate to question the relevance of the concept of *area*, given that it might be considered as a subset of *common ties*, in the sense that the notion of area extended to include virtual "places" is nothing more than another trait shared by the members of the community.

> Community, in the sense of type of collectivity, usually refers to a group sharing a defined physical space or geographical area such as a neighborhood, city, village or hamlet; a community can also be a group sharing common traits, a sense of belonging and/or maintaining social ties and interactions which shape it into a distinctive social entity.
> […]
> The non-territorial approach has gained force as a

1 Stuckey, op. cit., pp. 51–54.
2 Kuper and Kuper, op. cit.

> result of modern advances in communication which have reduced the importance of territorial proximity as a bases for human association.[1]

From these reviews, spanning more than fifty years, two points are of importance. First, that although the study of community has evolved tremendously in the last seventy years, the meaning of the term has remained stable – relative to the stability of the sociological structural context. And second, that although the notion of community is difficult to agree upon, most scholars concur on the notions of *social interaction* and *common ties*.

Birth of Social Network Analysis

Although graph theory dates back to the XVIIth century, with Euler's *Seven Bridges of Königsberg*, the notion of understanding society through the use of networks did not appear before centuries, and the fact that the term *sociology* was not popularized until 1839 should provide an idea as to the why. Without going into specific details on the history of social network analysis – I refer the interested reader to Freeman's wonderfully exhaustive *Development of Social Network Analysis*[2] – I deem interesting to evoke how the understanding of community has developed.

In the XIXth century, Auguste Comte proposed one of the earliest statement of a structural vision of society. He argued that sociology was a field with two major aspects, *statics* and *dynamics*,

1 Ibid., p. 196.

2 Linton C Freeman. *The development of social network analysis*. a study in the sociology of science. Booksurge Llc, 2004.

where the former is focused on the study of the "laws of social interconnection".

Comte was instrumental in the advent of social science as a macroscopic understanding of social interactions as opposed to the observation of social behaviors such as role and position at the individual level – this has to be considered in the context of the development of natural sciences in the XIXth century. However, despite the existence of both a mathematical framework and the intuition of a macroscopic view of society, it was not until the XXth century that social network analysis emerged as a scientific field.

The original impulse which led to the birth of social network analysis was given in 1934 by Jacob Levy Moreno, a passionate although puzzling man whose strength was that he had a profound structural intuition of the role social structure had on individuals. In 1934 was published *Who shall survive?*,[1] the outcome of a sociometric study which involved systematic data collection and analysis, in which he first used the term "network" in the sense it is used today.

Moreno was a psychiatrist and a psycho-sociologist. At heart, his position was that spontaneity and creativity are the forces which drive human progress. In order to facilitate the treatment of his patients, he developed the "psychodrama" in which patients were pushed to gain insights into their lives through spontaneous dramatization and role playing.

1 Jacob Levy Moreno and Helen Hall Jennings. *Who shall survive?: A new approach to the problem of human interrelations.* Washington, DC, US: Nervous and Mental Disease Publishing Co., 1934.

Most important from our perspective, Moreno's sociometry emerged as an tool to understand individual feelings in respect to one another in the context of the "sociodrama" – a psychodrama where the subject is replaced by a group which expresses its issues as a whole. By observing the attractions and rejections between group members, the sociometric test aims to exhibit the underlying sociological structure of the group.

Although the notion of group was somehow already present in Moreno's view, he viewed the group as a collection of interacting patient rather than a unit which had its own existence in the network. As a consequence, the idea of algorithmically using the data to distinguish communities inside a network was not central to sociometry.

As an alternative to Moreno's sociogram – the representation of sociometric data by a graph representation – Leo Katz proposed a matrix based visualization,[1] arguing that it made possible "a more detailed analysis of group structure" (Fig. 2.1). His paper gave rise to considerable discussion, including a ferocious defence of the sociogram by Moreno, and lead to a follow up article[2] in which Katz explicit the advantages of the matrix based approach.

His argument was that the sociogram was an ambiguous representation, whereas a matricial approach was more objective. Moreover, he describes an algorithm where groups of people are

1 Elaine Forsyth and Leo Katz. "A matrix approach to the analysis of sociometric data: preliminary report". In: *Sociometry* 9.4 (1946), pp. 340–347.

2 Leo Katz. "On the matric analysis of sociometric data". In: *Sociometry* 10.3 (1947), pp. 233–241.

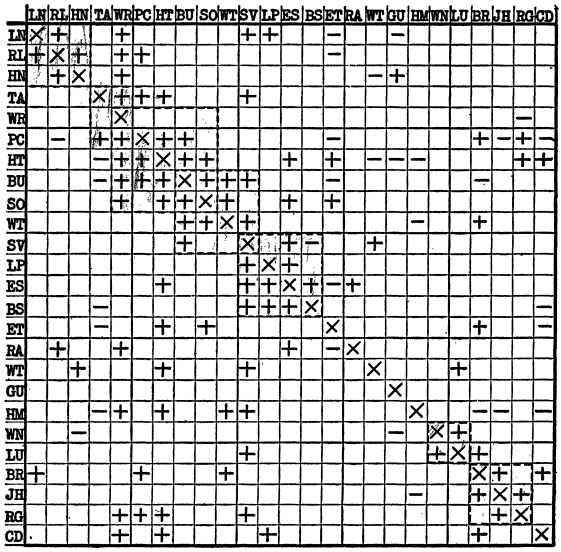


Figure 2.1 Katz's example of a matrix representation of a network, with clusters aligning on the diagonal.

constructed by selecting one actor, and then iteratively adding other actors as long as the newcomer is tied to at least half the members of the cluster.

Given that the ordering of rows and columns in a matrix is arbitrary, it is possible to rearrange the matrix by grouping together actors in same clusters. In that case, the resulting matrix is approximately in block diagonal form, which visually brings to light the subgroups of the network as illustrated in Figure 2.1. Finally, Katz asserts that the clusters he obtains are *cliques* in the sociological sense – or communities. Thus was born the concept of community detection.

> When minimization is complete, the resulting form has clusters of positive choices along the main diagonal and negative choices have gravitated to the extreme upper right and lower left hand corners. Each cluster of positive choices corresponds to a few individuals (their identifications appear in the main diagonal of the cluster) whose choices are concentrated within the cluster. Such individuals form a clique in the sociological sense. Thus cliques in the group correspond to the clusters obtained in the formal manipulation.[1]

The Advent of the Partition

The next step in the evolution of community came from Harvard in the 1970s, where White and his students participated in the renewed interest in social network analysis. In 1976, White – unsatisfied with both classical sociology and with Katz methodology – introduced the notion of *blockmodels*.[2]

> All sociologists' discourse rests on primitive terms – "status", "role", "group", "social control", "interaction" and "society" do not begin to exhaust the list – which *require* an aggregation principle in that their referents are aggregates of persons, collectivities, interrelated "positions", or

1 Ibid.

2 Harrison C. White, Scott A. Boorman, and Ronald L. Breiger. "Social structure from multiple networks. I. Blockmodels of roles and positions". In: *American journal of sociology* (1976), pp. 730–780.

> "generalized actors". However, sociologists have been largely content to aggregate in only two ways: either by positing categorical aggregates (*e.g.*, "functional subsystems", "classes") whose relation to concrete social structure has been tenuous; or by cross-tabulating individuals according to their attributes (*e.g.*, lower-middle-class white Protestants who live in inner city areas and vote Democrat). Both methods have "often led to the neglect of social structure and of the relations among individuals".[1]

Blockmodeling, as defined by White, is not an algorithm but rather an network analysis methodology built upon the notion of structural equivalence. Given a set S of actors and $\{R_i\}_{i=1}^{m}$ m binary relations on S, $u, v \in S$ are said to be *structurally equivalent* if and only if the following criterion is satisfied. For all $w \in S$ and any relation R_i:

$$\begin{cases} uR_iw \Leftrightarrow vR_iw \\ wR_iu \Leftrightarrow wR_iu \end{cases}$$

That is, u and v are structurally equivalent if they have identical in-neighbors and out-neighbors in all networks defined by R_i. However, real data does not exhibit large numbers of structurally equivalent pairs of actors, for a variety of reasons. To compensate this fact, this notion of equivalence had to be relaxed. Instead of requiring that all actors in a same class are structurally equivalent, in a blockmodel two actors in a same block only share the absence of ties towards some identical other actors. More formally,

1 Ibid.

a blockmodel is a *lean fit* to a given matrix M if and only if there exist a permutation of M, leading to a permuted matrix M^*, and a subdivision of M^* such that:

1. zeroblocks, which are in White's terminology sub-matrices containing only zeroes, in M^* correspond to 0 in the blockmodel;

2. blocks containing at least a 1 in M^* correspond to a 1 in the blockmodel.

Note that the relaxation comes from the *at least*, in the previous constraint. Structural equivalence is achieved when there if a 1 in the blockmodel if and only if there are only 1 in the corresponding block of M^*

Blockmodeling departs from classical sociometry in two ways. First, it eludes the need for an a priori aggregation of individuals – *i.e.* there is no need to project multiple networks onto a unique one – which frees the sociologist from having to combine potentially heterogeneous data into a single dimension of interaction. One might however argue that this projection is hidden inside the permutation algorithm and that it deprives the analyst of using other projections.

Second, given the fact that actors are grouped by the correlation of their ties, it can serve as a way of identifying other social constructs, such as hierarchical or cooperative structures, rather than only communities.

Among the case studies White reports, the analysis of *A Monastery in Crisis* gives a good illustration of how *pattern*

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

Figure 2.2 Image for Esteem network in White's blockmodeling of the monks interactions. Each row/column represents a block containing several monks, the value 1 or -1 are assigned depending on the values in the network submatrix containing only those monks.

detection differs from *community detection*. From data obtained in Sampson's account of social interactions in an isolated monastery,[1] eight different networks are constructed based different type of relationships between the 18 monks – *like*, *esteem*, *influence*, *praise*, *antagonism*, *disesteem*, *negative influence* and *blame*.

Where this approach is different from the sociometric clustering is in the use of configuration "templates" to interpret the resulting image. For example, considering the esteem network, they obtain three blocks which are ordered in a *linear hierarchy* (Fig. 2.2).

This hierarchy implies that those in the first block only have esteem for themselves, those second block have esteem for themselves and those in the first block, and finally, those in the last block have esteem for everybody, which is "certainly plausible in a monastery"[2] – in the sense that there is a clear hierarchy among individuals, with an inner circle, regular members of the cloister

1 Samuel F Sampson. "Crisis in a cloister". PhD thesis. 1969.

2 White, Boorman, and Breiger, op. cit.

and novices which for the outer shell of the group. Finally, White is able to recognize Sampson's observations.

> The concrete social structure suggested is […]: a top-esteemed block unambivalently positive toward itself, in conflict with but conceding influence to a second, more ambivalent, block, to which is attached a block of losers.
> […]
> Sampson's Loyal Opposition is wholly contained in the first block; the Young Turks are exactly the men in the second block; the Outcasts are wholly contained in the third block. Sampson's Waverers 8 and 10 are in the Loyal Opposition block, whereas Waverer 13 is in the Outcast block.[1]

The question of nature of those three groups remains open: should those be considered as *communities*, or are those groups of people something totally different which only exist under the prism of hierarchy? This legitimate question does not limit itself to this case study and can be replicated in any context.

Consider for example a large family, containing children and parents. Given the asymmetry between the two sets (parents & children), the children might exhibit similar ties which in turn are different than those of their parents – if not only because the former share a child-parent relationship towards the latter. Does this mean that *parents* and *children* are part of two different social communities, or does this blockmodeling merely distinguish *role* inside the *family community*?

1 Ibid.

It is therefore essential to understand that White's work does not limit itself to community detection, but to the identification of social sub-structures which might be caused both by interaction and role.

Although interesting, the process of generalizing social network analysis to the identification of those patterns – rather than that of communities – had an unfortunate side-effect which ended up setting the tone for the next thirty years of the study of communities:

> Persons not in cliques are usually disregarded [in sociometry] (*i.e.*, treated as outside the effective sociometric system). In contrast, blockmodeling requires searching for a complete partition, such that sets of persons can be structurally important regardless of whether the sets resemble cliques.[1]

Though White's approach of communities-as-partition is, as we will see further, questionable, it does introduce the important idea that there should be no *a priori* on the structure one is looking for in the network, nor in the nature of the interactions.

From Blockmodels to Modularity

In the context of social partitioning, concomitant to White's blockmodeling, Breiger introduced concor,[2] an algorithm

1 Ibid.

2 Ronald L. Breiger, Scott A. Boorman, and Phipps Arabie. "An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling". In: *Journal of Mathematical*

which recursively divides a network into sub-clusters.

Given an $n \times m$ matrix M_0 whose columns are noted v_j, $1 \leq j \leq m$, the algorithm first computes an $m \times m$ matrix M_1, where the $(i, j)^{\text{th}}$ value is the correlation coefficient between the two vectors v_i and v_j. This same algorithm is then iteratively applied to M_1 to obtain M_2, M_3, …matrices, all of size $m \times m$. Breiger observed, on empirical data, that this process always converged to a matrix $M_\infty = \lim_i M_i$ which can be permuted into a two blocked matrix that can be summarized in the following form:

$$\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

From this bipartite split arises a natural 2-partition of the actors in M_0 and the algorithm can be recursively applied on each of those blocks to obtain a finer division. Breiger however does not address the question of where to stop this recursive split and exhibit a partition, but rather presents complete hierarchical clustering trees.

Community detection – in the modern sense – was born at the intersection of sociology, physics and computer science. In 2004, Newman suggested[1] a novel way of partitionning a network: instead of separating nodes with low *similarity* – such as pairs of non-equivalent nodes – he proposed the idea of discriminating on edge *betweenness*, where "betweenness" is defined as a way to favor

Psychology 12.3 (1975), pp. 328–383.

1 Mark E.J. Newman and Michelle Girvan. "Finding and evaluating community structure in networks". In: *Physical Review E* 69.2 (2004), p. 26113.

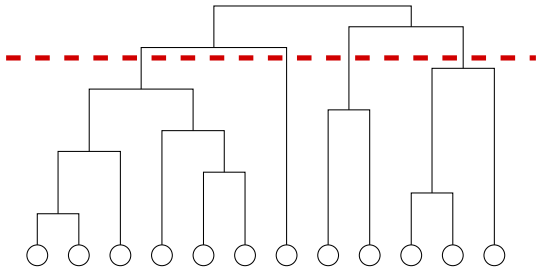


Figure 2.3 Newman's example of the output of his original algorithm. Each circle at the bottom represent a vertex, which are joined into larger communities as we move up the tree. The dashed line indicates the scale at which the partition is "frozen", *e.g.* where the modularity is optimal.

edges who structurally lie between communities such as, among others:

- *shortest-path betweenness*, defined as the number of shortest paths in the network running through a given edge;
- *random-walk betweenness*, the expected number of times that a random walk between a particular pair of vertices will pass down a given edge, summed over all pairs of vertices.

The algorithm is then straightforward: first compute the betweenness of each edges, identify the edge with the highest betweenness, remove it and reiterate the process – including the computation of edges' betweenness. This in turn gives a hierarchical tree of deleted edges and thus conversely a tree of clusters.

The most important contribution of the paper, however, is

not the use of the betweenness in the computation of the community structure but lies in the evaluation of the algorithm's output.

> In practical situations the algorithms will normally be used on networks for which the communities are not known ahead of time. This raises a new problem: how do we know when the communities found by the algorithm are good ones? Our algorithms always produce some division of the network into communities, even in completely random networks that have no meaningful community structure, so it would be useful to have some way of saying how good the structure found is. Furthermore, the algorithms' output is in the form of a dendrogram which represents an entire nested hierarchy of possible community divisions for the network. We would like to know which of these divisions are the best ones for a given network – where we should cut the dendrogram to get a sensible division of the network.[1]

To that effect, Newman introduces the modularity, a function which associates to a partition of a graph a score between -1 and 1. Consider a division of a network in k communities, and let $\mathbf{e}$ be the symmetric $k \times k$ matrix where $\mathbf{e}_{i,j}$ is the fraction of all edges of the network which lie between the communities i and j.

$$Q = \mathrm{Tr}\, \mathbf{e} - ||\mathbf{e}^2|| \tag{2.1}$$

1 Ibid., pp. 7-8.

The modularity as defined by Newman (Eq. 2.1) is the difference between the fraction of edges which lie between vertices of a same cluster and the expected fraction of such edges if they were randomly distributed across the network – but with the same community structure and same degree distribution. If the fraction of edges inside a community is not better than random, then $Q = 0$. However, values approaching $Q = 1$ indicate a strong community structure, as most edges lie inside the communities.

The strength of the metric is that, compared to Breiger's concor, it is possible to pinpoint inside the dendrogram the exact point where the partition should be in order to maximize its "communitiness".

The focus of this chapter is not to list exhaustively all algorithms which were built upon the modularity – the interested reader should turn to Fortunato's extensive review of the field,[1] which does not only cover the modularity but community detection as a whole – but just as White's partitionning had transformed the field into that of the partition, Newman's modularity so much refined what "community" had to be about that its consequences have to be evoked.

Rapidly, people realized that if modularity was used to rate the quality of a partition, then the best possible algorithms would surely be those attempting to optimize it. Although Brandes et al. showed in 2007 that finding a partition with maximal modularity was $\mathcal{NP}$-complete,[2] Blondel et al. came up the year later with the

1 Santo Fortunato. "Community detection in graphs". In: *Physics Reports* 486.3-5 (Jan. 2010), pp. 75–174.

2 Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer,

Louvain method, a fast and efficient heuristic,[1] which has since become a *de facto* standard in terms of graph partitionning.

All isn't well in the land of modularity, as the metric exhibits several flaws which have been largely documented. Of those, I shall mention three, the first two are tied to mathematical properties of the expression, and the third one is a more fundamental disagreement.

First was raised the question of the meaning of modularity in the context of random networks.[2] Understandably, partitions which are selected according to the actual presence of modules have a legitimate reason to have high modularity, but their is no apparent justification for the fact that there exist random networks which may be partitioned in order to achieve a high value of Q.

Next is the issue of resolution limit. Fortunato has shown[3] that it is possible to construct networks which present an unambiguous community structure which the modularity fails to uncover due to the fact that it cannot "identify modules smaller than a scale which depends on the total size of the network and on the degree of interconnectedness of the modules".

Zoran Nikoloski, and Dorothea Wagner. "On finding graph clusterings with maximum modularity". In: *Graph-Theoretic Concepts in Computer Science* (2007).

1 Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* (2008).

2 Jörg Reichardt and Stefan Bornholdt. "When are networks truly modular?" In: *Physica D: Nonlinear Phenomena* 224.1-2 (2006), pp. 20–26.

3 Santo Fortunato and Marc Barthélemy. "Resolution limit in community detection". In: *Proceedings of the National Academy of Sciences* 104.1 (2007), p. 36.

Finally, there remains the notion that communities in social networks are disjoint – which is inherent to the concept of partition. Although the idea of disjoint sub-structures might have made sense in the context of White's blockmodel which aim for pattern detection, it is absolutely unreasonable in terms of *social community* detection.

Ending the Social XOR

Despite the difficulties to define the term community, it is far easier to find examples of communities just by observing our social entourage. We have *families* and *co-workers*, *friends from high-school* and *friends from college* and more specialized communities such as the *poker group* or the *tennis club*, etc.

The mere possibility of citing all those possible communities without risking a social collapse points to the major flaw of *communities-as-partition*: its disjoint nature. Human have complex social interactions in different circles which cannot be translated in terms of partition, except by "forcing" each actor into one and only one social group. Note that the disjoint approach still has important uses, in particular in the terms of *pattern detection*, as it allows an efficient compression of the network which enhances the readability of interactions between patterns.

> The ability to find and analyze such groups can provide invaluable help in understanding and visualizing the structure of networks.[1]

1 Newman and Girvan, op. cit.

However the fact that one actor belongs to one and only one community is simple nonsense from the point of view of *social* network analysis, as would any method which would force one to "choose" between their family and friends.

In order to model more accurately the nature of social communities, the concept of *overlapping communities* – in which the one-node-to-one-community constraint disappears – arose quite naturally. Vastly different methods have been suggested to identify these enhanced type of communities.

In the continuity of Newman's work, several authors have exploited the modularity in some way to try and capture the overlap of communities. It should first be noted that as defined in Eq. 2.1, the modularity is of no use in the detection of overlapping communities, as it is clear that the way to achieve the highest possible value of Q it suffices to consider one community per edge. Nevertheless, there were attempts to tweak the definition or use of the modularity in order to counter that intrinsic effect of the metric – a task which is nor trivial nor intuitive, as Fortunato warns in his 2010 review.

> If vertices may belong to more clusters, it is not obvious how to find a proper generalization of modularity. In fact, there is no unique recipe.[1]

Among other modifications, it has been proposed that the modified modularity take into account the number of communities a node belongs to[2] or that the each node should

1 Fortunato, op. cit.

2 Hua-Wei Shen, Xue-Qi Cheng, and Jia-Feng Guo. "Quantifying and identifying the overlapping community structure in networks". In: *Journal of Statistical*

have a belonging coefficient towards its communities which should be included into the modularity.[1] Pushing further, some authors did not rely on the actual metric but on the instability of the partitions generated by the Louvain method to identify cores of communities.[2]

The fundamental flaw of these approach is – in my opinion – that community detection has recently evolved around a local maximum centered on the modularity. Although the metric has a legitimate use in order to partition a network while at the same time minimizing the densities of cross-cluster edges, their is no apparent reasons – other than historical – for its use in the detection of overlapping communities.

One of the most popular technique not inspired by the modularity is Palla's Clique Percolation Method[3] which builds upon the assumption that the inner density of communities leads to the formation of cliques – in the graph theory sense of a complete graph, as opposed to Moreno's sociological clique. Palla's basic intuition is that it is unlikely that vertices belonging to distinct communities are part of a large clique. Two k-cliques, defined as cliques of size k, are said to be adjacent if they share $k-1$ nodes and

Mechanics: Theory and Experiment 2009 (2009), P07042.

1 Vincenzo Nicosia, Giuseppe Mangioni, Vincenza Carchiolo, and Michele Malgeri. "Extending the definition of modularity to directed graphs with overlapping communities". In: *Journal of Statistical Mechanics: Theory and Experiment* 2009 (2009), P03024.

2 Qinna Wang and Eric Fleury. "Uncovering Overlapping Community Structure". In: *Complex Networks*. 2010.

3 Gergely Palla, Imre Derényi, Illés Farkas, and Tamas Vicsek. "Uncovering the overlapping community structure of complex networks in nature and society". In: *Nature* 435.7043 (2005), pp. 814–818.

the union of adjacent k-cliques is called a *k-clique chain*.

The notion of adjacency of k-cliques is an equivalence relation which extends the concept of connectivity – the case $k = 2$ is actually the exact definition of connectivity. This method has however one drawback, as it has a free parameter k which has to be set manually, and therefore communities of size smaller than k are invisible. On the other hand, when setting a small k, the opposite problem arises: suppose for example that there exist in a network two 100-cliques which share 10 nodes, if k is smaller than 10, those two cliques will appear as only one, if k is larger than 10, it means that no community with less than 10 members will be found.

Finally, another approach has revolved around the idea of finding a quantity descriptive of what a community is in order detect a community by locally optimizing that quantity. Baumes et al. proposed[1] several ways of weighing a cluster in order to rate its *communitiness*: the density of the set of nodes, the proportion of edges insides the cluster in respect to the number of edges having one extremity in the cluster and finally the ratio of density located inside the cluster in respect to that of its neighborhood. They then propose a heuristic which expands a seed – a set of nodes – of a graph in order to maximize the chosen weighing function by iteratively adding and removing nodes to the community as long as the metric increases.

1 Jeffrey Baumes, Mark Goldberg, Mukkai Krishnamoorthy, Malik Magdon-Ismail, and Nathan Preston. "Finding communities by clustering a graph into overlapping subgraphs". In: *International Conference on Applied Computing (IADIS 2005)* (2005), pp. 97–104.

Incidentally, Clauset independently proposed a similar greedy algorithm[1] to optimize a metric similar to Baumes second weight, with a twist. Clauset notes that the important vertices inside a community are those at its boundary – *i.e.* having neighbors both inside and outside the community. Therefore he specifically targets the fraction of links from the boundary to the community with respect to those from the boundary to the rest of the network. The algorithm he proposes to optimize this quantity is identical to Baumes' IS algorithm, although it is specialized to target his *local modularity* – it should be noted that he also introduces a parameter to limit the size of the communities.

Worth mentioning in this section is the work by Moody[2] in which he defines a quantity he names "structural cohesion" but which is nothing more than "node connectivity". His assertion is that most subgraph metrics fail to encompass the nature of what a community is, and that the fact that a community features a solid structure should be reflected into its mathematical model. However his proposal does not take into account the relative isolation of a community from the rest of the network.

As we have seen, the history of community *and its intersection with social network analysis has been bumpy. Strong disagreement have been noticed both in the meaning to give to the term community, as*

1 Aaron Clauset. "Finding local community structure in networks". In: *Physical Review E* 72.2 (Aug. 2005).

2 James Moody and Douglas White. "Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups". In: *American Sociological Review* 68 (2003), pp. 103–127.

well as in the ways of identifying them as such in a network. However, their seems to be a slow convergence towards the idea that, in a network-centric sense, a community is a set of densely packed actors isolated from the rest of the network, without constraints imposed on the relationship between communities – ergo, *communities may overlap. Of course, this assertion is dependant on the actual meaning given to both the terms* densely *and* isolated.

Social Cohesion

Quantifying Communities

> Ego will have a collection of close friends, most of whom are in touch with one another – a densely knit clump of social structure.
>
> *The Strength of Weak Ties*
> Mark Granovetter

Both this chapter and the following ought to be the single most important contribution of this thesis to the field of community detection and sociology as a whole. Here I shall describe the reasoning behind the construction of the *cohesion*, a weighing function or metric which rates the quality – community-wise – of a set of nodes in a network. I shall then proceed to describe several remarkable mathematical properties before ending on some thoughts about the fundamental difference between judging the quality of a community and the quality of a set of communities.

Derivation of the Metric

As I have described in the previous chapter, most works on communities have revolved around the idea of dividing a network into several communities, yet more recently some have taken the dual approach of attempting to define quantitatively the notion of community. This thesis is inscribed in the latter school of thought, for the simple reason that the raw social network data is usually complex enough to avoid adding a second layer of difficulty to the problem by trying to look for something one has not even formally defined.

Let us consider the requirements and constraints one should impose to such a quantitative definition. In order to start out with a simple case, let $G = (V, E)$ be an undirected unweighted network. Our aim is to construct a quality function

$$\mathcal{C}_G : V(G) \to [0, 1]$$

which we shall call the *cohesion*, such that $\mathcal{C}_G(S) = 0$ when S is not a community and $\mathcal{C}_G(S) = 1$ when S is a really good community.

The first constraint, which is implicitly given above, is that the cohesion of a set of community should not depend on the collateral existence of other communities in the network: if not, the cohesion would not represent an intrinsic quality of the community, but rather the way it behaves in respect to other communities.

As I have recalled in the first chapter, Hillery, Poplin or Stuckey, who have reviewed the notion of communities, agree

that the notion of community is based on shared traits and social interaction. In this context, the existence of other communities in interaction with the one of interest is never mentioned. I would go as far as to say that the idea that their exist a link between the quality of a community and the presence of a set of communities only stems from the tradition of partitionning a network into disjoint communities.

Next, we should impose on C that it should be *local*. Obviously, the people in a group as well as their relationships impact its cohesion. Almost as important, people who are not in the community but have a relationship with some members influence the overall cohesion. Thus we are considering, in the elaboration of the metric, the members of the group and its immediate neighborhood. The question now remains of the impact the rest of the network should have on the group's cohesion. I submit that their is no such impact, and this for two major reasons.

First and foremost, it is intuitive and legitimate to restrict ourselves to the social neighborhood of the set which is of interest. A useful example is to consider an individual and the communities he belongs to; if two people meet in a remote area of the network, this should not directly affect his communities. Another way to see it is that the notion of community is an information which exists emergently in the topology of the network and does not ripple to those who do not share a connection with the community.

On a practical note, not related to the actual definition of the community but to its use in social network data is an argument already raised by Clauset.[1] More and more networks tend to be

1 Clauset, op. cit.

large in terms of size which makes it impractical to load all the data in memory just to compute the cohesion of a small set of nodes. In this context where the network might be prohibitively expensive or outright impossible to observe, it makes sense to restrict the data used to assess the cohesion of a set of nodes to its neighborhood.

More broadly, one may consider that a *social network* is nothing more than a description of a fragment of what one would call *The Social Network*[1], 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) and therefore can be considered as incomplete and severely limited snapshot. As such, given that the full knowledge of The Social Network is unattainable, it makes no sense to add an arbitrary dependency of the notion of community on the whole of a dataset which limits are due to measurement constraints rather than actual social facts.

Those previous points can be formally summarized by stating that the cohesion of a set S of nodes of G only ought to depend on the subgraph restricted to S and its neighborhood, as expressed in Equation. 3.1.

$$C_{G[S \cup \mathcal{N}(S)]}(S) = C_G(S) \tag{3.1}$$

1 No connection to a recent movie whatsoever.

2 Wayne Zachary. "An information flow model for conflict and fission in small groups". In: *Journal of Anthropological Research* 33 (1977), pp. 452–473.

Finally, now that we have pinpointed the part of the network which should impact the cohesion, it is almost time to give its formal definition. What intuitively comes to mind when thinking about a cohesive set of nodes is twofold: first, it should be *dense* for some adequate definition of density. Second it should be isolated from the rest of the network, in the sense their should be a clear boundary between the content of the group and the exterior world. Historically, both those notions of density and isolation have been quantified in terms of edge densities.[1]

This approach has however some limits when considering a community which may overlap. As an example, the toy network in Figure 3.1 consists of a group of squares and a group of circles. 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 circle group is a "good" community, whereas the square group is a "bad" community – what would be considered a good community would be the square group to which the leftmost circle would be added.

In order to explain the differences between those two groups, let us go back in time to 1973 and recall Mark Granovetter's work. A few years before his mentor introduced the notion of blockmodel, Granovetter – who was one of White's student – suggested

1 Clauset, op. cit.; Newman and Girvan, op. cit.

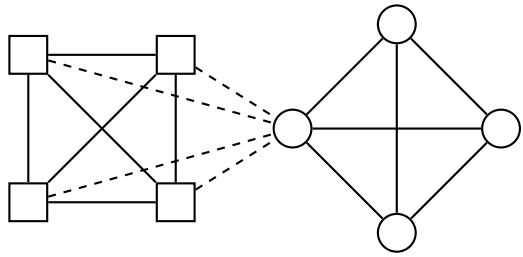


Figure 3.1 Toy Network featuring two sets of nodes of identical size, featuring the same number of links both inside the set and towards the rest of the network. Despite those structural similarities, the set of circles appears to be a better community than the set of squares.

a radical paradigm shift in the way social graphs were to be analyzed. By that time, of course, it was well understood that all relationships in a social network did not share nor the same nature – *e.g.* family ties, friendships – nor the same intensity.

Granovetter's contribution[1] did not lie in the fact that he distinguished strong ties – between close friends – and weak ties – between acquaintances, but rather that he was able to exhibit that the latter play an unequivocal role in the spread of information to the network.

> The macroscopic side of this communications argument is 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

1 Mark Granovetter. "The Strength of Weak Ties". In: *American journal of sociology* 78.6 (May 1973), pp. 1360–1380; Mark Granovetter. "The strength of weak ties: a network theory revisited". In: *Sociological Theory* 1 (1981), pp. 201–233.

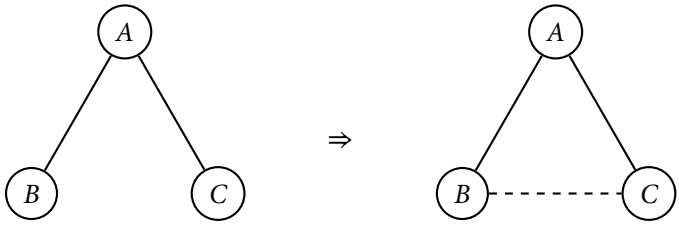


Figure 3.2 Description of the forbidden triad according to Rapoport: if *A*, *B* and *C* are such that *A* shares a strong tie with both *B* and *C*, then *B* and *C* must share a tie (be it a strong or weak tie).

> other characteristics will have difficulty reaching a modus vivendi.[1]

Furthermore, and equally important to our structural endeavour, he finally assesses that local bridges – edges which do not belong to a triangle, that is a set of three pairwise connected nodes – are weak ties. His reasoning is articulated around the notion of triadic closure and forbidden triads,[2] which can be summarized in the following way: if a node is strongly tied to two other nodes then those two nodes are tied (Fig. 3.2).

It is important to remember that in Granovetter's vision, people shared either strong ties – between close friends – or weak ties. Transposing this to the context of communities, we can extend this notion of triadic closure to take into account the type of ties which link two individuals: *e.g.* if *A* shares a strong **family** tie towards both *B* and *C*, then *B* and *C* share at least a weak tie (of any type).

1 Idem, "The strength of weak ties: a network theory revisited", p. 202.
2 Anatol Rapoport. "Contributions to the Theory of Random and Biased Nets". In: *Bulletin of Mathematical Biophysics* 19 (1957), pp. 257–277.

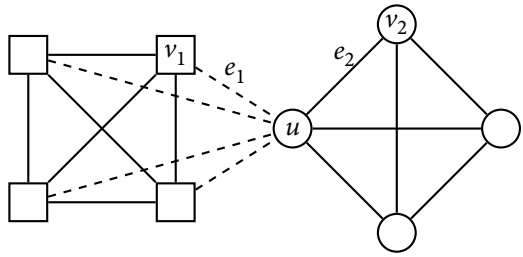


Figure 3.3 Detailed version of the toy network in Figure 3.1.

However, if A shares a strong *family* tie with B and a strong *co-worker* tie with C, this does not provide any information of the presence or absence of a tie between B and C.

Armed with this knowledge, let us go back to our previous toy network and understand how both groups differ. Consider the two edges e_1 and e_2 attached to u (Fig. 3.3), and suppose that both those edges are strong ties of the same type, then there should exist at least a weak tie between v_1 and v_2 but that is not the case. Therefore, we can assert that e_1 and e_2 are either of different types or that at least one of those edges is a weak tie.

As a consequence, when considering the group of circles, the fact that four dotted edges are crossed does not impact the cohesion of the set because all of those dotted edges are either weak ties (in respect to the circle community) or ties of different type. If those are weak ties, they are, in Granovetter's words, "a crucial bridge between the two densely knit clumps of close friend".[1]

1 Granovetter, loc. cit.

If those edges are of a different type, then they should not impact the cohesion for obvious reasons. On the other hand however, this reasoning does not hold when considering the square community, and this is where a fundamental structural difference between those two groups arises.

We have seen that the discriminating feature of those two groups lies in the presence of edges which nature can be characterized using our extended triadic closure property. As a consequence, the edges uv, where $u \in S$ and $v \in V(G) \setminus S$, such that for all $w \in S \setminus \{ u \}$ the edge uv is not present in the graph G, should not affect the cohesion. In that case, it is clear that the quantity which affects the cohesion of a group of nodes is not only linked to edges but to triads.

Given a graph G we first recall that a triangle is a set of three pairwise connected nodes. Let $S \subseteq V(G)$ be a set of nodes of G, we now introduce two quantities which capture the inner and outer triadic connections of the set. We call $\circledcirc_G(S)$ the number of *inbound* triangles of S, that is the number of triangles of the graph G totally contained in S. Similarly, we define the number of *outbound* triangles $\triangleq_G(S)$ as the number number of triangles in G which have exactly an edge in $G[S]$.

$$\begin{aligned} \circledcirc_G(S) &= \left|\{ u, v, w \in S \mid uv, vw, uw \in E(G) \}\right| \\ \triangleq_G(S) &= \left|\left\{ u, v \in S, w \in V(G) \setminus S \mid uv, vw, uw \in E(G) \right\}\right| \end{aligned}$$

Recall how earlier we remained vague while asserting that a community should feature a high *density* and be relatively isolated

from the rest of the network. We now have the tools to provide a formal quantitative definition for both of those terms.

As we have concluded that only taking edges into account is not sufficient to correctly capture the cohesion of a group, we shall use a *triangular density* – as a matter of fact, the transitivity of the subgraph – in order to describe the strength of the relationships inside the community.

To rate the isolation from the rest of the network at the boundary of the group, we shall introduce an *isolation coefficient* which is the ratio of inbound triangles to the total number of triangles having an edge inside the group – that is, we both count inbound and outbound triangles here. The idea behind this metric is that most triangles adjoining the group should be inside in order to have a clear cut boundary.

Given that a cohesive community should feature both high density/transitivity and isolation, we naturally derive from those two coefficients the expression of the cohesion $\mathcal{C}_G(S)$ of S in the graph G (Eq. 3.2).

$$\mathcal{C}_G(S) = \underbrace{\frac{\text{△}_G(S)}{\binom{|S|}{3}}}_{\text{transitivity}} \times \underbrace{\frac{\text{△}_G(S)}{\text{△}_G(S) + \text{◬}_G(S)}}_{\text{isolation}} \tag{3.2}$$

On Figure 3.4, for example, the set of circles contains four nodes which form two inbound triangles and only one outbound triangle to the square node on the lower right-hand corner – the square node on the upper right corner does not form any outbound triangle since it is only connected to one node of the set – which leads to a cohesion of $\mathcal{C} = \frac{2}{4} \times \frac{2}{1+2} = \frac{1}{3}$.

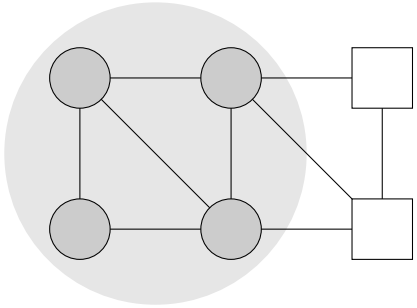


Figure 3.4 In this example, the set of circles has a cohesion $C = \frac{1}{3}$.

Going back to our constraints, we can fairly easily check that if S is a clique in G disconnected from the rest of the network, $C_G(S) = 1$, therefore the cohesion is 1 when the group is a really good community. Conversely, if S is a graph with no edges, which is the worst possible case for a community, then $C_G(S) = 0$. Moreover, the definition of the cohesion only takes into account S and its neighborhood, from which it comes the locality: $C_{G[S \cup \mathcal{N}(S)]}(S) = C_G(S)$.

Finally, the expression is based on the notion of triangles, and explicitly excludes edges which are not part of triangle and thus satisfies the last constraint. If we go back to the toy network in Figure 3.1, we now see that the set of circles is far more cohesive than the set of squares ($C_\square = \frac{2}{5} < 1 = C_\bigcirc$).

This provides a quantitative way of discriminating between groups of nodes which have seemingly identical characteristics when only taking into account the number of nodes and edges both inside and across groups.

Mathematical Properties

Now that we have established the expression of the cohesion and observed that it complies with the constraints and requirements we have detailed in the first section of this chapter, let us explore some interesting mathematical properties of the metric.[1]

First off, remember how Granovetter says that edge which do not belong to triangles – the so-called *local bridges* – are weak ties,[2] and recall how weak ties are edges which lie between communities. As such, the presence or absence of such edges should not impact the cohesion of a set of nodes, which gives us Theorem 1.

Theorem 1 Let G_Δ be the graph obtained by removing all local bridges from G. That is, the set of nodes of G_Δ is $V(G_\Delta) = V(G)$, and the set of edges $E(G_\Delta)$ is $E(G)$ restricted to edges which do belong to a triangle of G, $E(G_\Delta) = \{\, uv \in E(G) \mid \exists w \in V(G)\ s.t.\ uw \in E(G) \wedge vw \in E(G) \,\}$. Then, for all set of nodes $S \subseteq V(G)$:

$$\mathcal{C}_G(S) = \mathcal{C}_{G_\Delta}(S)$$

Proof When removing the local bridges which, by the definition given above, do not belong to any triangle, no triangles are added or removed. Therefore $\circledcirc_G(S) = \circledcirc_{G_\Delta}(S)$ and $\triangleq_G(S) = \triangleq_{G_\Delta}(S)$. Given that the size of the set remains constant, it comes that $\mathcal{C}_G(S) = \mathcal{C}_{G_\Delta}(S)$. □

1 The notations used extensively in this section as well as throughout this thesis are those of Diestel's textbook(Reinhard Diestel. *Graph theory*. Springer Verlag, Feb. 2006).

2 Granovetter, "The Strength of Weak Ties", pp. 1364-1365.

To follow in the same vein, let us consider an network having at least two disjoint connected components. Theorem 3 tells us that the most cohesive groups lie inside each of the components.

Lemma 2 Let $S_1, S_2 \subseteq V(G)$ be two disconnected sets of vertices $((S_1 \times S_2) \cap E(G) = \varnothing)$. If $\mathcal{C}(S_1) \leq \mathcal{C}(S_1 \cup S_2)$ then $\mathcal{C}(S_2) > \mathcal{C}(S_1 \cup S_2)$.

Proof Suppose that $\mathcal{C}(S_1) \leq \mathcal{C}(S_1 \cup S_2)$ and $\mathcal{C}(S_2) \leq \mathcal{C}(S_1 \cup S_2)$,

$$\frac{\text{△}(S_1)^2}{\binom{|S_1|}{3}} \leq \Big(\text{△}(S_1) + \text{⧋}(S_1)\Big)\, \mathcal{C}(S_1 \cup S_2)$$
$$\frac{\text{△}(S_2)^2}{\binom{|S_2|}{3}} \leq \Big(\text{△}(S_2) + \text{⧋}(S_2)\Big)\, \mathcal{C}(S_1 \cup S_2)$$

By summing the two lines it comes that:

$$\frac{\text{△}(S_1)^2}{\binom{|S_1|}{3}} + \frac{\text{△}(S_2)^2}{\binom{|S_2|}{3}} \leq \Big(\text{△}(S_1) + \text{⧋}(S_1) + \text{△}(S_2) + \text{⧋}(S_2)\Big)\, \mathcal{C}(S_1 \cup S_2)$$

Now, given that S_1 and S_2 are disconnected, we have:

$$\text{△}(S_1) + \text{△}(S_2) = \text{△}(S_1 \cup S_2)$$
$$\text{⧋}(S_1) + \text{⧋}(S_2) = \text{⧋}(S_1 \cup S_2)$$

Therefore,

$$\begin{aligned}\frac{\text{△}(S_1)^2}{\binom{|S_1|}{3}} + \frac{\text{△}(S_2)^2}{\binom{|S_2|}{3}} &\leq \Big(\text{△}(S_1 \cup S_2) + \text{⧋}(S_1 \cup S_2)\Big)\, \mathcal{C}(S_1 \cup S_2) \\ &\leq \frac{(\text{△}(S_1) + \text{△}(S_2))^2}{\binom{|S_1|+|S_2|}{3}}\end{aligned}$$

Furthermore, given that $|S_1|, |S_2| > 1$, the following holds:

$$\binom{|S_1|}{3} + \binom{|S_2|}{3} < \binom{|S_1| + |S_2|}{3}$$

From there it comes:

$$\frac{\text{△}(S_1)^2}{\binom{|S_1|}{3}} + \frac{\text{△}(S_2)^2}{\binom{|S_2|}{3}} < \frac{(\text{△}(S_1) + \text{△}(S_2))^2}{\binom{|S_1|}{3} + \binom{|S_2|}{3}}$$

Which simplifies to:

$$\left(\binom{|S_2|}{3} \text{△}(S_1) - \binom{|S_1|}{3} \text{△}(S_2)\right)^2 < 0$$

Hence the contradiction. Therefore, for all $S_1, S_2 \subseteq V(G)$, disconnected:

$$C(S_1) \leq C(S_1 \cup S_2) \Rightarrow C(S_2) > C(S_1 \cup S_2)$$

□

Theorem 3 Let S be a non-connected set of vertices of G. Then there exist a connected set $S' \subseteq S$ with higher cohesion $C(S') > C(S)$.

Proof Let $S_1, S_2 \subseteq V(G)$ such that $S_1 \cup S_2 = V(G)$ and S_1, S_2 disconnected, then at least one of S_1 or S_2 has a higher cohesion than S per Lemma 2.

If the set with higher cohesion is connected, the result is immediate. If not, the same reasoning applies to that set, which leads to the conclusion: there exist a connected subset with higher cohesion. □

A common model of networks which is often used is the Erdős–Rényi model of random graph. Although not being an accurate description of social networks, it is accepted as a null model to which compare actual social graphs. I deem it interesting to observe the behavior of the cohesion on such networks. Let us recall that in this model, the graph denoted by $G_{n,p}$ is obtained by randomly connecting n nodes, each edge having a uniform probability of appearing p independent from every other edge.

Theorem 4 Let $G_{n,p}$ be a random graph, and $S \subseteq V(G_{n,p})$ a set of nodes, then:

$$C(S) = p^3 \frac{|S| - 2}{3n - 2|S| - 2}$$

Proof In S, each triad has a probability p^3 of being a triangle, therefore the expected number of triangles in S and the expected number of triangles of S are given by

$$\text{Ⓐ}(S) = p^3 \binom{|S|}{3}$$
$$\text{⏢}(S) = p^3 (n - |S|) \binom{|S|}{2}$$

Finally, it comes that the cohesion of S is

$$\begin{aligned} C(S) &= p^3 \frac{\binom{|S|}{3}}{(n - |S|)\binom{|S|}{2} + \binom{|S|}{3}} \\ &= p^3 \frac{|S| - 2}{3n - 2|S| - 2} \end{aligned}$$

□

From Theorem 4, given that the function $k \to p^3 \frac{k-2}{3n-2k-2}$ is increasing, we observe that the most cohesive group in a random network is the whole actual network. Put the other way round, it means that the random graph $G_{n,p}$ does not contain, on average, any group which is more cohesive that the whole network. This is to be linked with the fact that in an Erdős–Rényi random graph the density is expected to be homogeneous across the network.

The "four groups" test was introduced[1] by Newman and Girvan to test the accuracy of their community detection algorithm. I shall here use the same test to exhibit the pertinence of the cohesion.

Consider a network G of size $4n$ consisting of 4 groups of size n. Edges are placed independently between pairs of nodes with a probability p_{in} for an edge to fall between two nodes of a same group and a probability p_{out} to fall between nodes of two different groups.

Theorem 5 The limit of the cohesion of a group S in the "four groups" test, when $|S| = n \to +\infty$ is given by

$$\lim_{n \to +\infty} C(S) = \frac{p_{\text{in}}^5}{p_{\text{in}}^2 + 9p_{\text{out}}^2}$$

Proof Using the same reasoning that in the proof of Theorem 4,

1 Newman and Girvan, op. cit.

we have

$$\text{△}(S) = p_{\text{in}}^3 \binom{n}{3}$$
$$\text{⛴}(S) = p_{\text{in}} p_{\text{out}}^2 3n \binom{n}{2}$$

From there it comes that

$$C(S) = p_{\text{in}}^3 \frac{p_{\text{in}}^3 \binom{n}{3}}{p_{\text{in}}^3 \binom{n}{3} + p_{\text{in}} p_{\text{out}}^2 3n \binom{n}{2}}$$

And therefore,

$$\lim_{n \to +\infty} C(S) = \frac{p_{\text{in}}^5}{p_{\text{in}}^2 + 9p_{\text{out}}^2}$$

□

What Theorem 5 shows is that, as one would expect, the cohesion of those groups increases when there density increases (higher p_{in}) and decreases when the inter-group density increases (higher p_{out}), as illustrated in Figure 3.5.

To sum up, this section has listed a few mathematical properties of the cohesion, which in a sense exhibit that the metric is compatible with what one would commonly refer to as a community, in a intuitive structural way. The two first theorems establish important facts regarding the links between the cohesion and the presence or absence of links in the topology. Theorem 1 does so by showing that the exclusion of edges which lie between communities are not detrimental to cohesion of a group and Theorem 3 states that the most cohesive groups are connected. In layman's term, if one

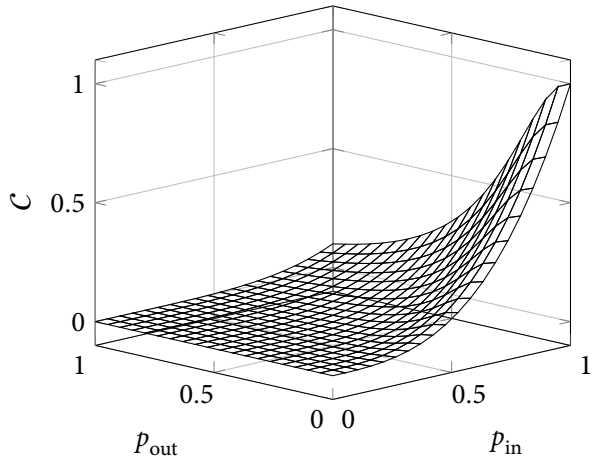


Figure 3.5 Cohesion of the groups in the Newman-Girvan four group test, as a function of the intra-group probability p_{in} and the inter-group probability p_{out}.

brings a stranger that no one knows in a group, the resulting group will have a lower cohesion.

Finally, I have presented two results which relate arbitrary structures which are often used to evaluate algorithms and metrics. Theorem 4 highlights how an Erdős–Rényi random graph is not expected to feature highly cohesive subgroups and Theorem 5 makes the link between the notion of cohesion and the popular testing framework suggested by Newman and Girvan.

Division and Content

The notion of overlapping community introduces a shift in the way one thinks about rating communities. For example, 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 in which edges are randomly rewired with the constraint that the degree distribution is preserved. Second, but no less important, it implicitly judges the quality of the *division* into communities.

This makes sense in the context of a partition because both those aspects are linked, when the content a of community is changed – *i.e.* a node is moved to an other community – the boundaries between communities are affected and therefore the whole division into communities is changed. There is however no equivalent notion in an overlapping content , since a node can be added to several communities.

In the overlapping context, judging the quality of the division largely depends on the data one wishes to study. On one side of the spectrum, the case of two completely disjoint communities $S_1 \cap S_2 = \emptyset$ with very high cohesion form an obvious good division of the network $(S_1 \cup S_2, E)$. On the other side of the spectrum, two totally overlapping communities $S_1 = S_2 = S$ form a really bad division of the network (S, E). However, the intermediate overlapping cases are far less trivial.

The fact is that, in some occurrences, there is a case for allowing small *fuzzy* overlaps in order to model an vertex-based interface

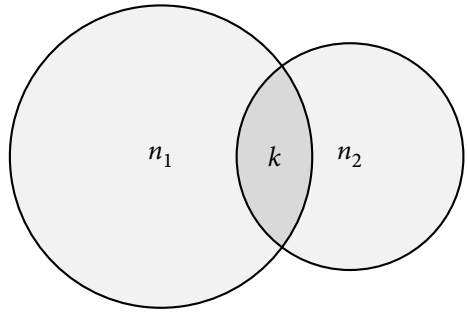


Figure 3.6 Two overlapping cliques C_1 and C_2 of size n_1 and n_2, where the overlap is $|C_1 \cap C_2| = k$.

between groups instead of purely edges. The most minimal example of this is that one individual can be part of two groups which share no other members. On the other hand, there also are cases where communities should be allowed to overlap at a great extent – consider for example college classes. Even more extreme, in some cases, it might be desirable that communities be allowed to be fully embedded one in another, for example a computer science lab is a smaller community inside a larger university community.

It is important to bare in mind that major part of the argument above was made with having in mind highly cohesive communities. Because the cohesion takes into account both the content and the boundary of a set of nodes, a sufficiently large overlap between two sets of nodes can lead to the fact that the union of both those sets are of higher cohesion.

Consider for example a network consisting of two overlapping cliques C_1 and C_2 of sizes $|C_1| = n_1$ and $|C_2| = n_2$, such that there

intersection is of size $|C_1 \cap C_2| = k$ (Fig. 3.6).

We therefore have the following cohesions:

$$\begin{aligned} \mathcal{C}(C_1) &= \frac{\binom{n_1}{3}}{\binom{n_1}{3} + (n_2 - k)\binom{k}{2}} \\ \mathcal{C}(C_2) &= \frac{\binom{n_2}{3}}{\binom{n_2}{3} + (n_1 - k)\binom{k}{2}} \\ \mathcal{C}(C_1 \cup C_2) &= \frac{\binom{n_1}{3} + \binom{n_2}{3} - \binom{k}{3}}{\binom{n_1+n_2-k}{3}} \end{aligned}$$

The *overlap* coefficient of two sets U, V is given by $O(U, V) = \frac{|U \cap V|}{\min|U|,|V|}$. On Figure 3.7 we have represented the three possible cases as a function of $\frac{n_1}{n_2}$ and $O(C_1, C_2)$. Four cases are possible, although here only three actually appear: in light gray (resp. dark gray) is represented the area where the larger (resp. smaller) clique is more cohesive that the union of the two cliques.

In the intermediate gray area, the overlap between the two cliques is sufficiently low in order for both cliques to have a higher cohesion than their union. In particular, note how in the case where both cliques are of the same size, they are still more cohesive that the union even if both cliques share more than half there nodes. This illustrate the fact that very highly cohesive groups are allowed to overlap at a large extent without one being fully absorbed by the other.

The light gray area is where the union of both cliques have a higher cohesion than that of the smaller one but not of that of the larger one. Let us admit that, for obvious reasons, there are no other subset of this network which have a higher cohesion. We

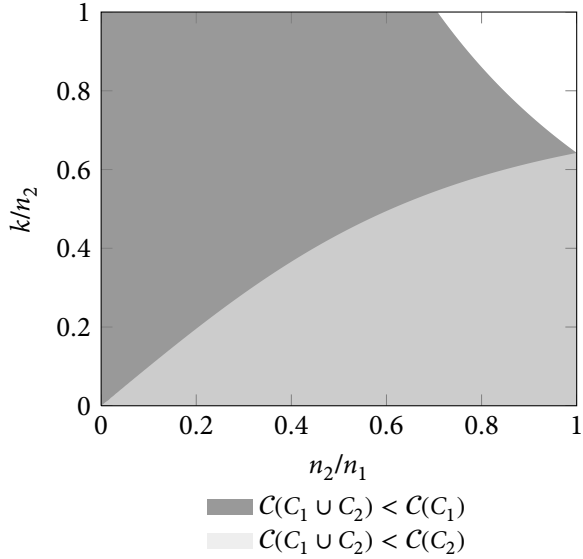


Figure 3.7 Overlapping cliques cohesion threshold. Given two cliques of size n_1 and n_2, having k common nodes, the cohesion of the union is higher to the cohesion of both cliques in the darkest area. The cohesion of the union in only higher than that of the smallest clique in the darker area and both cliques have a higher cohesion than the union in the lighter area. Here, $n_1 = 1000$.

then obtain that the two best communities of the network are the large clique and the union of the two cliques, which does lead to a very important overlap between the best communities but at the same time preserve the large clique as a community.

Finally, in the white area, both cliques are less cohesive than their union. The best way of covering the network with communities is then to choose the whole network as a community. Note

that this only happens if two conditions are met: both cliques are of relatively similar size and the almost entirely overlap.

It is however important to realize that this does not mean that the cohesion has a resolution limit, as opposed to the modularity. As we have seen earlier, communities which are not connected, or only connected by weak ties, are always more cohesive than their union.

Let us go further and now consider a case where we relax the conditions. The setup is similar to the previous case, except that instead of considering two overlapping cliques we now consider two overlapping Erdős–Rényi random graphs $S_1 = G_{n_1,p_1}$ and $S_2 = G_{n_2,p_2}$ sharing k nodes. Furthermore, we suppose that the probability that an edge appears between two nodes of $S_1 \cap S_2$ is $p = \min(p_1, p_2)$. We obtain the following cohesions:

$$\begin{aligned} C(S_1) &= \frac{p_1^5 \binom{n_1}{3}}{p_1^2 \binom{n_1}{3} + p^2(n_2 - k)\binom{k}{2}} \\ C(S_2) &= \frac{p_2^5 \binom{n_2}{3}}{p_2^2 \binom{n_2}{3} + p^2(n_1 - k)\binom{k}{2}} \\ C(S_1 \cup S_2) &= \frac{p_1^3 \binom{n_1}{3} + p_2^3 \binom{n_2}{3} - p^3 \binom{k}{3}}{\binom{n_1+n_2-k}{3}} \end{aligned}$$

The first thing to notice is that if $p_1 = p_2$, then the results stated previously about the clique still hold and that we obtain the same thresholds as in Figure 3.7. Actually, the only factor which impacts the variation of the thresholds is the ratio of p_1 to p_2, that is the relative density & transitivity of both groups.

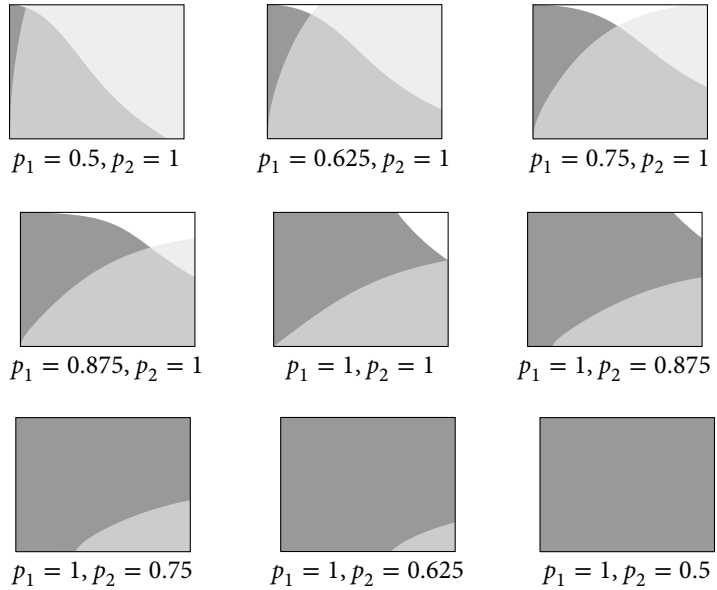


Figure 3.8 Overlapping Erdős–Rényi random graphs cohesion threshold

On Figure 3.8 are represented the different possible cases for different values of p_1 and p_2 – the ratio $\frac{p_2}{p_1}$ increases from top-left to bottom right. We have voluntarily limited ourselves to those values, as more extreme ratios $\frac{p_1}{p_2}$ lead to results where the cohesion of one group is larger than that of the union and the cohesion of the other group is smaller, independently of the values of n_2 and k. In other terms, in those omitted cases there clearly are two communities, one of which is the denser group, and the other one is the whole network. Back to our figure, we first observe that the sparser a group is relative to the other one, the smaller the area where that

group is more cohesive than the union. This makes sense, as the transitivity of the denser group will tend to compensate that of the sparser one. Note however that conversely, this area is larger for the less denser group, which means that it wont absorb the sparser group unless the overlap is large enough.

One final point to consider is the white area, which we recall is the area where the union of both groups has higher cohesion that each of them taken independently. What we can conclude from this analysis is that although the overlap between two sets of nodes can impact the cohesion of one of them, the cases where the union has a higher cohesion only happen for a restricted area of values of p_1, p_2, n_2 and k.

From a set of both intuitive and sociologically backed constraints we have constructed a local graph metric which we call the cohesion. *The metric aims to measure the extent to which a set of nodes is a community by measuring its density and isolation in terms of closed triads. Furthermore, we have exhibited notable properties of the cohesion, in particular that it attains its optimal value in connected components, that it ignores weak ties which are irrelevant community-wise and that it does not suffer from a resolution limit. Finally, we have established the closed-form expression of the cohesion in several cases, among others that of interacting cliques and Erdős-Rényi random graphs.*

Fellows, a Social Experiment

Real-World Validation

> The true method of knowledge is experiment.
>
> *All Religions are One*
> William Blake

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*. Given the subjective nature of the the notion of community, we have chosen to turn to an empirical and subjective source to confirm that the cohesion does actually capture to which extent a set of nodes is a community. In this chapter, we shall present Fellows[1], a large scale online experiment on Facebook which was conducted in order to provide an empirical evaluation of the cohesion. The main motivation behind Fellows was to quantify the accuracy of the cohesion by comparing it to subjective ratings given to communities by real persons, rather than only relying on arbitrary benchmarks.

1 http://fellows-exp.com

Experimental Setting

Fellows was a single page web application which provides the subject with a short description in several languages – English, French, Portuguese and Spanish – of the experiment and its motivations. When a visitor wished to take part in the experiment, he authorized the application to connect to his Facebook account and granted access to his personal data on Facebook. At that point, the application used the Facebook API[1] and downloaded the list of the subject's friends and the set of interconnections between the pairs of his friends – that is, for each of his different friends, the list of friends they have in common – in order to reconstruct his social neighborhood.

Using a simple greedy algorithm, similar in spirit rather than in metric to one previously introduced by Clauset,[2] the application computed the subject's communities of friends in their immediate social neighborhood by recursively adding new friends to a community as long as the community's cohesion increases.

It is important to note that all computation was implemented in JavaScript and ran inside the subject's browser. This had two consequences: first, should the experiment turn viral and attract a lot of users, it could scale using a bare bones HTTP server. Second, it also meant that no identifiable information was ever transmitted back to the application's server – the exact pieces of data we gathered were an anonymous unique identifier for each subject, along with their age and gender, if those were available.

1 Facebook. *Graph API*. Tech. rep. 2011.
2 Clauset, op. cit.

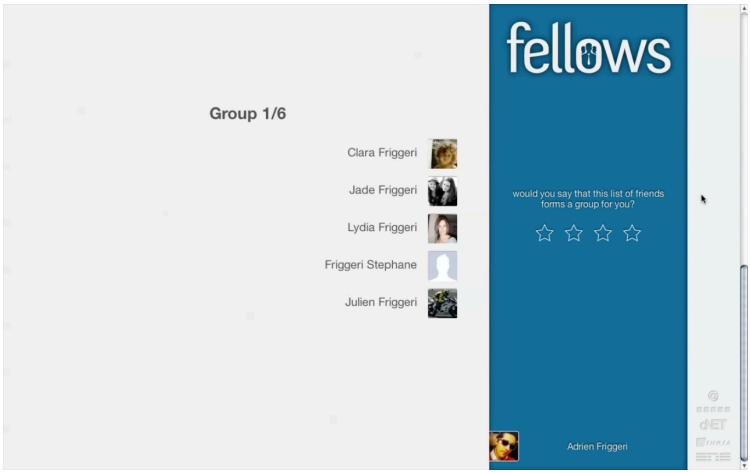


Figure 4.1 Screenshot of the Fellows application displaying a community and asking the subject to input a rating.

Statistics on each of the communities were then sent to the server along with an anonymous unique subject identifier and session identifier – to be able to exclude subjects participating several times. Birth date and gender were also anonymously recorded when available, both for the subject and his friends.

After those communities were computed, the application displayed a list of names and pictures of friends which were present in the community featuring the highest cohesion (Fig. 4.1). The subject was the asked to answer the question "would you say that this list of friends forms a group for you?" – or the relevant translation in one of the other languages – by giving a numerical rating between 1 and 4 stars, those options were labelled, respectively, "absolutely not", "a little", "yes" and "absolutely".

The subject then had 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: when the subject would then publish *e.g.* a status update or a picture, he would have the possibility to restrict those of his friends who would be able to see his posting by selecting a Friend List. We added this feature as an incentive for the subject to take part in the experiment.

Once the subject had submitted the rating, it was uploaded to the application server where it was associated to the relevant community. In case the user had created a Friend List, the name he had given was also recorded. The user was then presented with another community and the process was repeated until either i) the user exited the application or ii) all communities were rated, in which case a message was displayed to thank the user for their involvement.

The experiment was launched on February 8th, 2011. We 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 time of writing, participation totaled 3310 persons (Fig. 4.2). Although unrelated to the evaluation of the cohesion, there 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[1] indicating that of the 15,987 people who visited the website, 8,895 came from

1 A service from Google which provides detailed statistics of visitors access

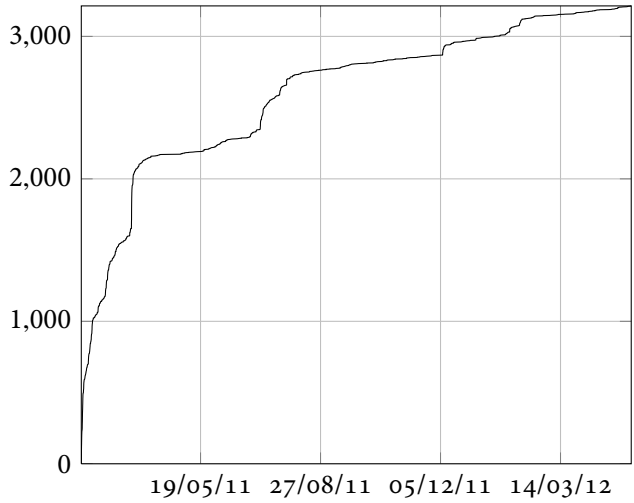


Figure 4.2 Evolution of the total number of participants in the Fellows Experiment through time.

France, 2,098 from the United States, 468 from the United Kingdom and no other country sent more than 350 visitors.

Moreover, the total number of unique users increases by bursts: observe how on March 23rd, 2011 the number of users rises from $\sim$ 1700 to $\sim$ 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 other bursts – for example an email relaying the experiment on a large mailing list on February 14th, 2011, a tweet by an *influent* twitterer on February 28th, 2011 or a comment posted on an high-traffic blog on July 4th.

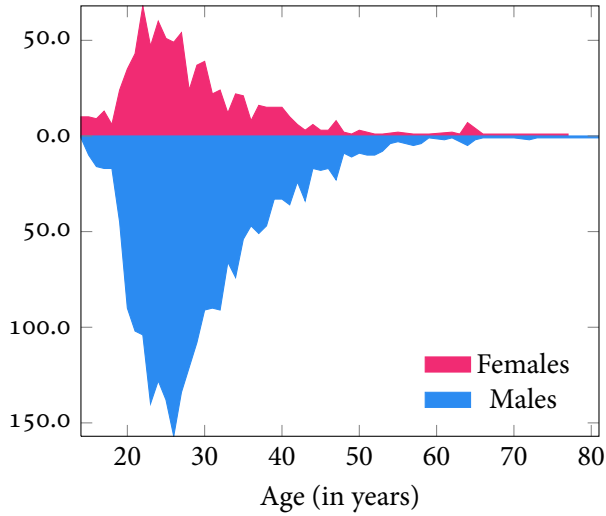


Figure 4.3 Densities of ages of male (blue) and female (magenta) subjects.

As stated above, during the first months of the experiment, when a subject 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, during that period, 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. This facility was however restricted by Facebook during the course of the experiment – it was considered as a form of spam – which further decreased the visit count.

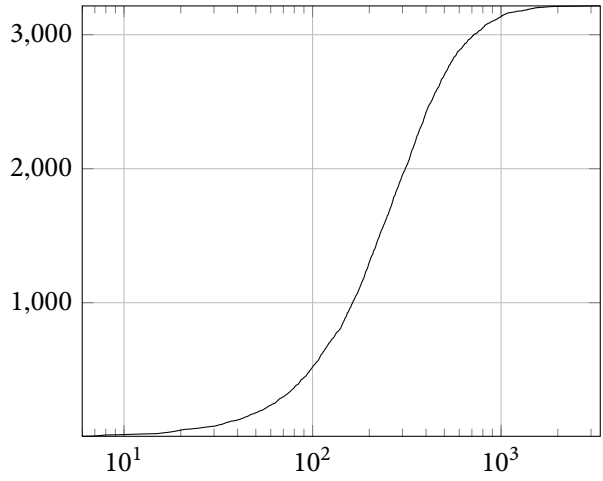


Figure 4.4 Distribution of the number of friends of the subjects.

In some cases, the contributions were corrupted or incomplete – *e.g.* the user temporarily lost their internet connection. Thus, 94 contributions had to be discarded, leaving 3216 valid contributions (2303 males, 830 females and 83 persons of undeclared gender). The participants were on average 29.99 ± 8.92 years old – male subjects: 30.31 ± 8.69 yo, female subjects: 28.92 ± 9.34 (the age distributions for male and female subjects are given in Figure 4.3).

On Facebook, the number of friends one might have cannot exceed 5000 – a limitation added so that users would only add "real" friends. The distribution of the number of friends is heterogeneous (Fig. 4.4), with 10% users having less than 74 friends and 90% users having less than 612, the median being at 244 friends

and the average number of friends is 310.

These numbers have to be contrasted with those obtained on the globality of the network, it was reported by Ugander *et al.* that the average number of friends of a Facebook user was of 190 and that the median of friends was of 99 as of May 2011.[1] The main reason why we observe those differences in metrics is that there is a unavoidable indirect bias toward higher degree in any such experiment. Given that the incentive of the study was to create Friend Lists on Facebook, those most likely to take part in Fellows were people who make an active use of Facebook and have a larger number of friends to categorize. On the other hand, we consider that this bias has a low impact on our conclusions for the simple reason that the people who sporadically use Facebook have a higher probability of having incomplete and therefore non representative networks.

Cohesion and Ratings

This section presents the most fundamental result around which my thesis is articulated, namely that the cohesion captures well the extent to which a set of nodes in a network is a community. Humbly, I consider it to be the climax of my work during this thesis, as everything up to this point builds up to it and everything after only holds because of it.

This being said, the 3216 valid subjects lead to the detection of 86,691 communities, computed as stated before by attempting

1 Johan Ugander, Brian Karrer, Lars Backstrom, and Camero Marlow. "The anatomy of the facebook social graph". In: *Arxiv preprint arXiv:1111.4503* (2011).

to maximize their cohesion. Because a subject could stop the experiment at any time, only 62,863 of those communities actually received a rating, yet it is notable that 76% of the subjects rated more than 90% of their communities. There are mainly two explanations to those forfeitures, first that the user felt the communities they were presented with were of poor quality – the non-rated communities have on average a cohesion $C = 0.105 \pm 0.103$ – and second, that the subject had too many communities to rate – although the number of communities was bounded, if a subject had a lot of friends, that bound could have been be sufficiently high to discourage him.

Out of the 62,863 rated communities, 25.0% received a rating of 1 star, 21.3% received 2 stars, 22.2% were rated 3 stars and 31.4% 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, as this would have just given a way to evaluate the quality of the rather simple greedy algorithm. Our interest lies in the evaluation of the cohesion, and, in that context, obtaining low ratings is perfectly acceptable – and desirable – as long as they correlate to the cohesion.

We shall now exhibit the experimental link between a purely structural metric, the cohesion C, and the subjective appreciation of a community's pertinence expressed as the average rating R given by users. On Figure 4.5, we discretize the cohesion of all communities in increments of 0.01 and we represent the average rating obtained by communities in the same increment. Both quantities are rank correlated (Spearman's correlation $\rho = 0.91$, p-value $= 4.3 \times 10^{-38}$). Thus, when the cohesion increases, so

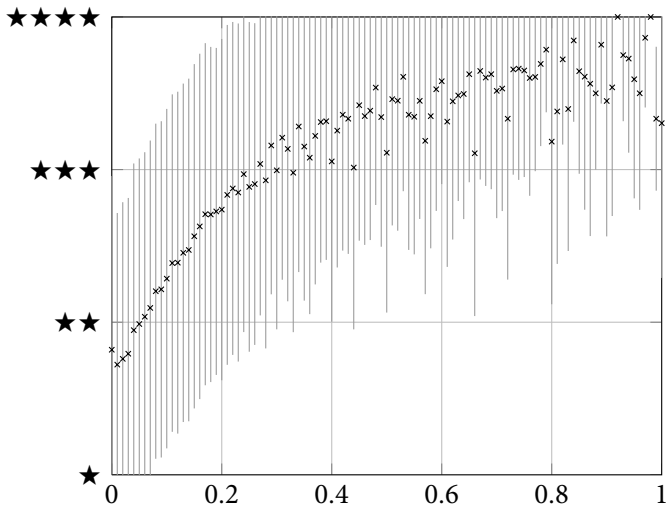


Figure 4.5 Average rating obtained by communities as a function of their cohesion.

does the average rating, and conversely. Furthermore, $\log C$ and $\log R$ are linearly correlated (Pearson's correlation $r = 0.97$, p-value $= 3.2 \times 10^{-63}$). A consequence of the strength of this correlation is that we can provide an equation linking the ratings to the cohesion in the form $R \propto C^{\frac{1}{3}}$.

On Figure 4.6 we plot the distributions of cohesions of each of the four sets of communities 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 communitiness of a set of nodes, as it is highly correlated to its subjective evaluation.

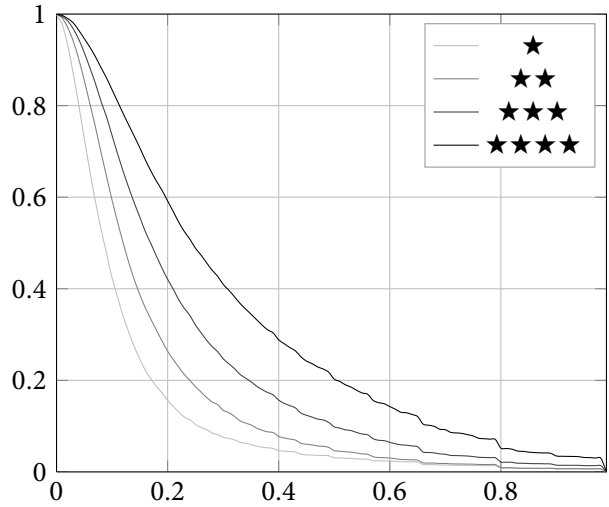


Figure 4.6 Normalized reversed cumulative distribution of cohesion for communities rated 1,2,3 or 4 stars ($\mathbb{P}[\mathcal{C} \geq X|\text{rating} = N]$).

Furthermore, it is interesting to look at the relation between the ratings and other graph metrics, such as the density of the considered set. On Figure 4.7 we plot the average rating obtained for communities of a given density. Groups having a density greater than 0.5 tend to have the same average rating (between 2 and 3 stars). There seems however that for densities smaller than 0.5 the rating increases with the density. To explain this fact, consider that the cohesion of a set S is bounded by its transitivity. Given that $\text{△}(S) < m\sqrt{m}$, where m is the number of edges in S, there exist a bounding relation between density and cohesion as exhibited in Figure 4.8. Thus, the lower ratings of the sparser communities can be explained by the fact that those have low cohesion, which itself

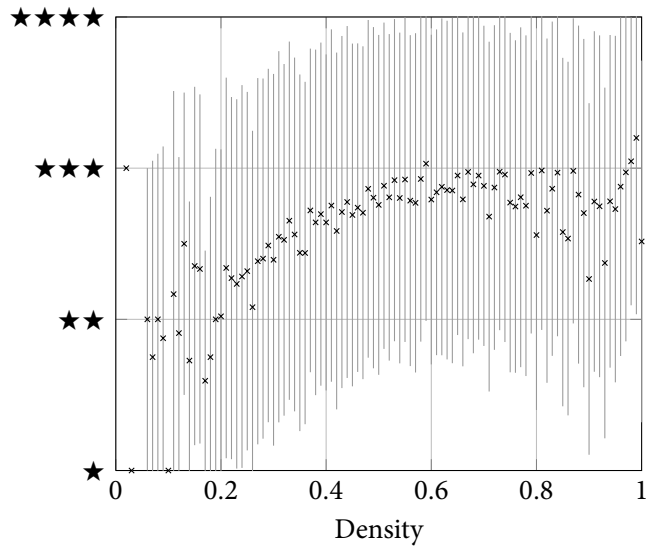


Figure 4.7 Average rating obtained by communities as a function of their density.

is highly correlated to ratings. Conversely there are communities with high density but low cohesion, which explains why high density does not imply high ratings.

For similar reasons, communities 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. For example, a community with high clustering but a lot of outbound triangles might

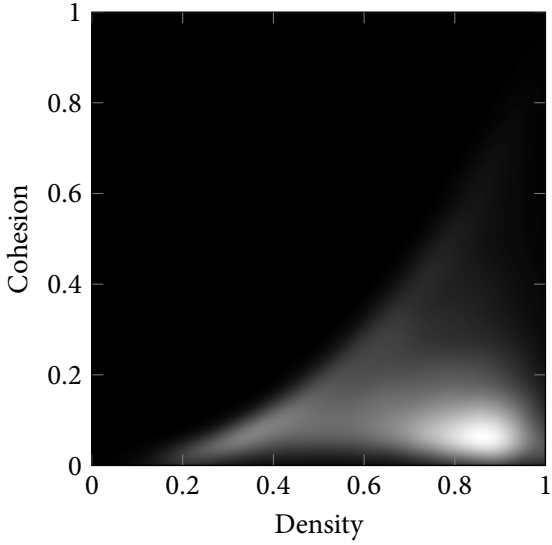
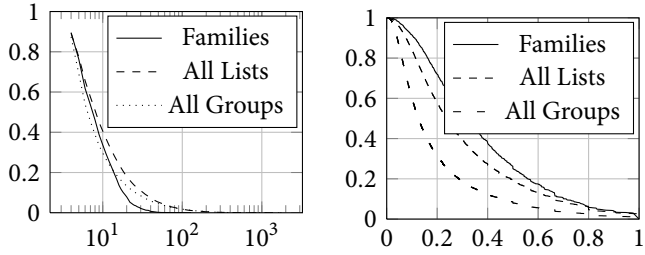


Figure 4.8 Kernel density estimation of cohesion as a function of density.

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.

Although we have said earlier that the aim of Fellows was no to evaluate the quality of the simple greedy algorithm, it is nevertheless interesting to try and understand why it did not lead to high ratings all the time. The answer to that is complex, but a direction is to notice is that the algorithm assigns all nodes of degree greater than 3 to at least one community. In practice, there is no reason that all nodes belong to at least one socially cohesive community:



(a) Size Reverse CDF (b) Cohesion Reverse CDF

Figure 4.9 Distribution of (left) the number and (right) cohesions of people in families, lists, and all data.

for example, 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.

Another interesting bit of information which was revealed in Fellows is related to the structures of families. Since subjects had the possibility to create Friend Lists on Facebook, and to do so had to input a descriptive label which was subsequently recorded on the server, we extracted the set of communities which had been given a label hinting that the people in the community formed a family and observed their sizes and cohesion. This gathering of those family communities was done by first selecting the subset of lists with name containing "fam" – as it matches "family", "famille", "familia", etc. – and then manually filtering this subset to remove any false positive.

On Figure 4.9(a), we have represented the reverse cumulative distribution of sizes of such Family communities, along with the same metric applied to all communities which led to the creation of a list, and the whole dataset. Notice how there are relatively fewer large families than large communities in the two other subsets. More interestingly though is the fact that there when considering all communities there are a higher proportion of smaller communities than for families. As a matter of fact, except for a couple of outliers, family sizes seem to be normally distributed around a mean size of 9.

Similarly, we have represented on Figure 4.9(b) the reverse cumulative distribution of cohesions for those communitie. We observe that, as we could have expected, families form more cohesive communities than those in all other cases. This not only means that family members tend to add each other on Facebook, but that there are few non family mutual friends between two family members. In other words, the data suggests that people do not mix friendship and family together much.

Through the use of a large-scale Facebook experiment that attracted 3310 to take part in the experiment, we have seen that the cohesion is highly correlated to the subjective perception of communities by users. As such, we conclude that the cohesion is a good measure of the extent to which a set of nodes in a network form a social community. Anecdotally, we have also observed that families, in a sense the epitome of communities, tend to be the most cohesive groups in a social neighborhood, which further comforts our use of the cohesion to rate communities.

A Glimpse of Complexity

Party Planners are Here to Stay

> It is a mistake to think you can solve
> any major problems just with potatoes.
>
> *Life, the Universe, and Everything*
> Douglas Adams

Earlier we have seen that the cohesion captures quantitatively the subjective perception of the quality as a community of a set of nodes in a graph. In this context, it is only natural to attempt to find the most cohesive subgraphs of a given network. We shall first prove that this problem is $\mathcal{NP}$-hard by reducing the clique problem. In a second time, we will establish that the dual problem of finding the less cohesive groups – with a given size – of a graph is $\mathcal{NP}$-hard.

Max-Cohesion is $\mathcal{NP}$-hard

In this section we examine the problem of finding a set of vertices $S \subseteq V(G)$ of maximum cohesion in a graph G, *i.e.* for all subset $S' \subseteq V$, $\mathcal{C}(S') \leq \mathcal{C}(S)$. We have shown in Theorem 3 that the set of vertices with maximum cohesion in a given network is connected, therefore our problem simplifies to finding a connected

set of vertices with maximum cohesion in G, which associated decision problem we will call Connected-Cohesive.

Connected-Cohesive

Input	A graph $G = (V, E)$, $\lambda \in \mathbb{Q}$, $\lambda \in [0, 1]$
Question	Is there a subset connected S of V such that $C(S) \geq \lambda$?

We shall now proceed to show that Connected-Cohesive is $\mathcal{NP}$-complete. First note that given a set S of vertices of G, it is possible to verify that S is a solution of Connected-Cohesive by computing its cohesion, its size, its connectivity and the minimum degree of its vertices, all in polynomial time. Therefore Connected-Cohesive is in $\mathcal{NP}$.

In order to show that Connected-Cohesive is $\mathcal{NP}$-complete, we will show that we can reduce Clique to it. We recall that Clique is the well known problem of finding a complete subgraph of size k in a given graph.

Clique

Input	A graph $G = (V, E)$, $k \in \mathbb{N}$, $k \leq \lvert V \rvert$
Question	Is there a subset S of V such that $\lvert S \rvert = k$ and the subgraph induced by S is a clique?

Before going into the details of the proof, let us provide a rough sketch. When looking for a clique of size k in a graph, we are looking for the existence of a structure in the graph which does not a contain a *non-edge* – that is, a pair of nodes u, v such that $uv \notin E$. Therefore, in order to penalize the choice of such

a *non-edge*, we will transform an instance of Clique into an instance of Connected-Cohesive by connecting all such u and v and adding a large number of triangles resting on uv. That way, all sets containing a *non-edge* will have an arbitrarily low cohesion, far lower than that of a clique from the original graph, and we will hence be able to find cliques of given size using a cohesion threshold.

More formally now, let $(G = (V, E), k \in \mathbb{N})$ be an instance of Clique[1]. We can assume that G is connected (if not, we use the following reasoning separately on each connected component of G).

In order to construct an instance $(G' = (V', E'), \lambda)$ of Connected-Cohesive, we take several steps. First, for all non connected nodes u and v in G – *i.e.* for all *non-edges* – we add $3n\binom{n}{3}$ nodes $(w_i^{uv})_{1 \leq i \leq 3n\binom{n}{3}}$ and then connect u and v to each of the w_i^{uv}, effectively adding $3n\binom{n}{3}$ triangles including the edge uv to the network. Finally, we construct a ring with the w_i^{uv}, by connecting together w_i^{uv} and $w^{uv}_{(i+1) \bmod 3n\binom{n}{3}}$ – in the following, we shall refer to such a structure stemming from uv as a *sprout* and write $Sp(uv) = (w_i^{uv})_{1 \leq i \leq 3n\binom{n}{3}}$ the sprout resting on uv. This process is described in Algorithm 5.1 and illustrated by Figure 5.1.

Lemma 6 There exists a clique of size k in G if and only if there exists a connected group of vertices of G' with cohesion $\lambda \geq \frac{k-2}{3n-2k-2}$.

1 We consider here that $|G| > 2$ and $k > 2$, although this is not exactly Clique, this problem is clearly $\mathcal{NP}$-complete, given that the complexity of Clique does not arise from those small values.

Input: $G = (V, E), k \in \mathbb{N}$
  $V' := V$
  $E' := E$
  for $uv \in V^2 \setminus E$ **do**
    $E' \leftarrow E' \cup \{uv\}$
    for $i = 1$ to $3n\binom{n}{3}$ **do**
      $V' \leftarrow V' \cup \{w_i^{uv}\}$
    for $i = 1$ to $3n\binom{n}{3}$ **do**
      $E' \leftarrow E' \cup \{uw_i^{uv}, vw_i^{uv}, w_i^{uv}w_{(i+1) \bmod 3n\binom{n}{3}}^{uv}\}$
  return $G' = (V \cup W, E'), \lambda = \frac{k-2}{3n-2k-2}$

Algorithm 5.1 Transforms an instance of Clique into an instance of Connected-Cohesive

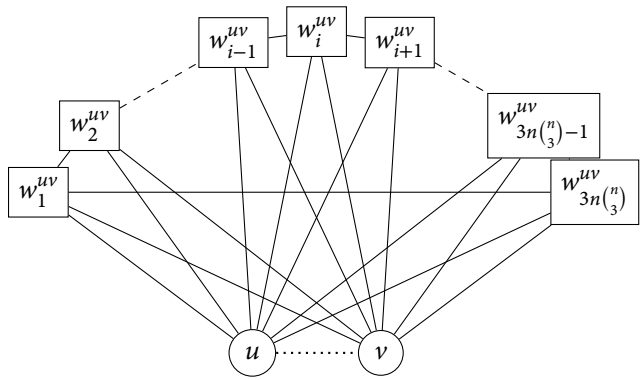


Figure 5.1 Illustration of adding a sprout to uv using Algorithm 5.1. At this step, we join u and v, add $3n\binom{n}{3}$ triangles resting on uv – *i.e.*, $3n\binom{n}{3}$ triangles of the form $(uvw_i^{uv})_{1 \le i \le 3n\binom{n}{3}}$. Finally, we join the extremities of those triangles in order to form a ring around uv.

Proof Let us first prove the implication. Let $K \subseteq V$, be a clique of size $|K| = k$ in G. Given that no node or edge are deleted when constructing G', G is a subgraph of G' and thus K is a clique in G' and $\triangle_{G'}(K) = \binom{k}{3}$.

Moreover, by construction, $G'[V]$ is a clique and for all u in K, the neighbors of u are also in V. Therefore, each edge in K forms one triangle with each vertex in $V \setminus K$, which leads to $\triangle\!\!\triangle_{G'}(K) = \binom{k}{2}(n-k)$. Finally, this gives a cohesion[1]:

$$\mathcal{C}_{G'}(K) = \frac{\binom{k}{3}}{\binom{k}{3} + \binom{k}{2}(n-k)} = \frac{k-2}{3n-2k-2}$$

Conversely, we shall write $\mathcal{C}_k = \frac{k-2}{3n-2k-2}$, we will now prove that if $S \subseteq V'$ is a connected set of vertices such that $\mathcal{C}(S) \geq \mathcal{C}_k$, then there is exist a subset $K \subseteq S$ which is a clique of size larger than k and $K \subseteq V$. In effect, $\mathcal{C}_k$ will serve as a threshold in cohesion above which we will be able to establish that there is a clique of size k in G.

We will now suppose that there is a connected set S such that $S \subseteq V'$ and $\mathcal{C}(S) \geq \mathcal{C}_k$. First note that $|S| \geq 3$, because by definition, if $|S| < 3$, $\mathcal{C}_{G'}(S) = 0$ which would lead to a contradiction.

We will split S into a union of two disjoint subsets $S = S_V \cup S_A$, $S_V \cap S_A = \emptyset$, where S_V contains the elements of S which are in V – in the original graph, and S_A the elements of S which are nodes w_i^{uv} which were added during the instance transformation – nodes

1 Note that since a n-clique is a $G_{n,1}$, this is the same equation, with $p = 1$, that we obtained in Chapter 3 for the cohesion of subset of an Erdős–Rényi random graph

which are in sprouts. Based on this decomposition, we will consider several cases and show that either they lead to the presence of a clique of size at least k in the graph, or to a contradiction:

$$\textsc{Case 1.} \begin{cases} |S_V| > 0 \\ |S_A| = 0 \\ \forall u, v \in S, uv \in E \end{cases}$$

In this case, S is a clique of G of size $|S|$. As we have seen previously, we have a cohesion:

$$C(S) = \frac{|S| - 2}{3n - 2|S| - 2}$$

Let f be the function defined as $f(x) = \frac{x-2}{3n-2x-2}$. From there it comes that:

$$\frac{\partial f}{\partial x} = \frac{3n - 6}{(3n - 2x - 2)^2}$$

Since $n = |G| > 2$, $\frac{\partial f}{\partial x} > 0$, and f if increasing. Therefore, since $C(S) \geq C_k$, then S is of size $|S| \geq k$, and thus contains a clique of size k.

$$\textsc{Case 2.} \begin{cases} |S_V| > 0 \\ |S_A| = 0 \\ \exists u, v \in S, uv \notin E \end{cases}$$

Here, S is included is G but is not a clique of G. It is however by design a clique in G'. Moreover, since we include at least a *non-edge* which belongs to $3n\binom{n}{3}$ triangles, the number of outbound triangles of S is at least equal to $\text{⧋} \geq \binom{|S|}{2}(n - |S|)3n\binom{n}{3}$. Therefore

the cohesion of S is bounded by:

$$\mathcal{C}(S) \leq \frac{\binom{|S|}{3}}{\binom{|S|}{3} + \binom{|S|}{2}(n - |S|) + 3n\binom{n}{3}}$$

For reasons similar to the previous case, this quantity increases as k increases, therefore the maximal value is obtained for $k = n$, it comes:

$$\begin{aligned}\mathcal{C}(S) &\leq \frac{\binom{n}{3}}{\binom{n}{3} + 3n\binom{n}{3}} \\ &\leq \frac{1}{1 + 3n} \\ &< \min_k \mathcal{C}_k\end{aligned}$$

As a consequence, $\mathcal{C}(S)$ is smaller than $\mathcal{C}_k$ for all values of k – in particular, it is smaller than $\min_k \mathcal{C}_k = C_3 = \frac{1}{3n-8}$ – hence the contradiction. Therefore, if S is a connected set such that $S \subseteq V'$ and $\mathcal{C}S \geq \mathcal{C}_k$, then S cannot be contained in V and include a *non-edge*.

Case 3. $\begin{cases} |S_V| = 0 \\ |S_A| > 0 \end{cases}$

Let us now consider the case where S is included in sprouts. Note that since S is connected, it is fully included in one sprout – if not, it would contain at least a node present in V, which would contradict $|S_V| = 0$. Therefore, there exist u, v such that $S \subseteq Sp(uv)$.

By construction, the sprout $Sp(uv)$ is a ring, and as such does not contain any triangles, therefore S does not contain any triangle and therefore $\mathcal{C}(S) = 0 < \min_k \mathcal{C}_k$, hence the contradiction.

Finally it comes that if S is a connected set such that $S \subseteq V'$ and $C(S) \geq C_k$, then S cannot be a subset of the nodes which were added during the transformation.

$$\textsc{Case 4.} \begin{cases} |S_V| > 0 \\ |S_A| > 0 \\ \forall u, v \in S_V, uv \in E \end{cases}$$

More generally, S consists in this case of a clique of G and a number of nodes which are contained in sprouts. We shall proceed in two steps here. First we shall suppose that all nodes in S_A are inside one sprout and then we will generalize this result by lifting this restriction.

Suppose that there exist an edge uv such that $S_A \subseteq (w_i^{uv})_{1 \leq i \leq 3n\binom{n}{3}}$. Since S is connected and that S does not contain any *non-edge*, it comes that exactly one of the two nodes u and v are contained in S. Without loss of generality, we shall suppose that this node is u.

From there we derive the number of triangles of S which are exactly those contained in S and those of the form $uw_i^{uv}w_{(i+1) \bmod 3n\binom{n}{3}}^{uv}$, where the two nodes w_i^{uv} and $w_{(i+1) \bmod 3n\binom{n}{3}}^{uv}$ are in S_A. The maximum value of the cohesion is attained when taking $|S_A|$ nodes in a sprout such that they form a chain $w_i^{uv} \dots w_{(i+|S_A|) \bmod 3n\binom{n}{3}}^{uv}$. In that case, the sum $\textcircled{\scriptsize\triangle}(S_A) + \textcircled{\scriptsize\triangledown}(S_A) = 3|S_A|$ – each of the triangles uvw_i^{uv} is outbound, and there are one outbound and one inbound triangle per edge in the chain, leading to $3|S_A|$. Therefore we can provide

the following bound on the cohesion of S:

$$\mathcal{C}(S = (S_V \cup S_A)) \leq \begin{cases} \frac{\binom{|S_V|}{3}+|S_A|-1}{\binom{|S_V|+|S_A|}{3}} \frac{\binom{|S_V|}{3}+|S_A|-1}{\binom{|S_V|}{3}+\binom{|S_V|}{2}(n-|S_V|)+3|S_A|} & \text{if } |S_A| > 1 \\ \frac{\binom{|S_V|}{3}}{\binom{|S_V|+1}{3}} \frac{\binom{|S_V|}{3}}{\binom{|S_V|}{3}+\binom{|S_V|}{2}(n-|S_V|)+3} & \text{if } |S_A| = 1 \end{cases}$$

The expression when $|S_A| > 1$ decreases when $|S_A|$ increases, therefore it is minimal when $|S_A| = 2$. Moreover, both the cohesions when $|S_A| = 1$ and $|S_A| = 2$ are strictly smaller than $\frac{\binom{|S_V|}{3}}{\binom{|S_V|}{3}+\binom{|S_V|}{2}(n-|S_V|)} = \mathcal{C}(S_V)$. The case when $|S_A| = 1$ is trivial since it has the same numerator and strictly higher denominator, whereas the case where $|S_A| = 2$ is more tedious but can it can be verified using a Computer Algebra System such as Mathematica that Eq. 5.1 holds for $|S_V| \geq 3$, $|V| \geq 3$.

$$\frac{\binom{|S_V|}{3}+1}{\binom{|S_V|+2}{3}} \frac{\binom{|S_V|}{3}+1}{\binom{|S_V|}{3}+\binom{|S_V|}{2}(n-|S_V|)+6} < \frac{\binom{|S_V|}{3}}{\binom{|S_V|}{3}+\binom{|S_V|}{2}(n-|S_V|)} \tag{5.1}$$

As a consequence, if S is a connected set which contains both nodes in the original graph and the sprouts, does not contain *non-edges* and is of cohesion $\mathcal{C}(S) \geq \mathcal{C}_k$, then $\mathcal{C}(S_V) \geq \mathcal{C}(S) \geq \mathcal{C}_k$. Therefore, the cohesion of the restriction of S to V also has a higher cohesion than the threshold $\mathcal{C}_k$, and from CASE 1 it comes that S_V (and as a consequence S) contains a clique in V of size larger than k.

It is easy to generalize to the presence of node in multiple sprouts by noticing that the quantities $\textcircled{\triangle}(S_A)$

and $\bigtriangledown(S_A)$ are linear in $|S_A|$. Therefore, if we consider that S_A is divided into $S_A = (sp_j)_{1 \leq j \leq s}$ disjoint subsets which are spread over s sprouts, then the previous formulæ still hold and $\bigcirc\!\!\!\!\bigtriangleup(S_A) = \sum_j \bigcirc\!\!\!\!\bigtriangleup(sp_j) \leq |S_A| - 1$ and $\bigcirc\!\!\!\!\bigtriangleup(S_A) + \bigtriangledown(S_A) = \sum_j (\bigcirc\!\!\!\!\bigtriangleup(sp_j) + \bigtriangledown(sp_j)) = 3|S_A|$. Using exactly the same reasoning as above, the cohesion of the resulting set increases when the size $|S_A|$ decreases, and as a consequence, $C(S_V) \geq C(S) \geq C_k$. Finally, we conclude that in this case too, S contains a clique in V of size larger than k.

$$\textsc{Case 5.} \begin{cases} |S_V| > 0 \\ |S_A| > 0 \\ \exists u, v \in S, s.t.\ uv \notin E \end{cases}$$

The last remaining case is when S contains both nodes of the original graph and the sprouts, and contains at least a *non-edge*. We build upon Case 4 in a way similar to how we have added *non-edges* to Case 1 in Case 2. From there, we obtain an upper bound of the cohesion in the form:

$$C(S) \leq \frac{\binom{|S_V|}{3} + 3|S_A| - 2}{\binom{|S_V|+|S_A|}{3}} \frac{\binom{|S_V|}{3} + 3|S_A| - 2}{\binom{|S_V|}{3} + \binom{|S_V|}{2}(n - |S_V|) + 3n\binom{n}{3}}$$

Given that there is at least one *non-edge*, this one belongs to at least $3n\binom{n}{3}$ triangles, and as such the sum of numbers of inbound and outbound triangles is at least $3n\binom{n}{3}$, hence the denominator. The $3|S_A| - 2$ at the numerator stems from the fact that each *non-edge* forms one inbound triangle with a node in S_A, and both the extremities of that *non-edge* forms a triangle with an edge between two connected nodes in S_A.

This quantity decreases like $\frac{1}{|S_A|}$, and its maximal value is attained for $|S_A| = 1$, that is:

$$\mathcal{C}(S) \leq \frac{\binom{|S_V|}{3} + 1}{\binom{|S_V|+1}{3}} \frac{\binom{|S_V|}{3} + 1}{\binom{|S_V|}{3} + \binom{|S_V|}{2}(n - |S_V|) + 3n\binom{n}{3}}$$

In turn, this quantity increases as $|S_V|$ increases, and therefore is maximal for $|S_V| = n$:

$$\begin{aligned}\mathcal{C}(S) &\leq \frac{\binom{n}{3} + 1}{\binom{n+1}{3}} \frac{\binom{n}{3} + 1}{\binom{n}{3} + 3n\binom{n}{3}} \\ &\leq \frac{1}{1 + 3n} \\ &< \min_k \mathcal{C}_k\end{aligned}$$

Hence the contradiction. Therefore, if S is a connected set such that $S \subseteq V'$ and $\mathcal{C}(S) \geq \mathcal{C}_k$, then S cannot contain both a *non-edge* and nodes in sprouts.

FINALLY, we have shown that all connected sets S such that $S \subseteq V'$ and $\mathcal{C}(S) \geq \mathcal{C}_k$ are either cliques of size k contained in V, or supersets of cliques of size k themselves contain in V. Therefore we conclude that if there is a connected set of cohesion $\mathcal{C}(S) = \mathcal{C}_k$ in G', then there is a clique of size k in G. □

THEOREM 7 CONNECTED-COHESIVE is $\mathcal{NP}$-complete.

Proof Per Lemma 6, there exists a clique of size k in G if and only if there exists a connected subset of vertices of G' of cohesion $\lambda \geq \frac{k-2}{3n-2k-2}$. Since a clique of size k has exactly cohesion $\frac{k-2}{3n-2k-2}$, if there is a set of cohesion larger than $\frac{k-2}{3n-2k-2}$ then there is a

clique of size k and thus a set of cohesion $\frac{k-2}{3n-2k-2}$. Conversely, if there is a set of cohesion $\lambda = \frac{k-2}{3n-2k-2}$, then this set is of cohesion $\lambda \geq \frac{k-2}{3n-2k-2}$ and thus there is a clique of size k in G. Therefore there is actually an equivalence between the existence of a clique of size k in G, and the presence of a set of cohesion $\frac{k-2}{3n-2k-2}$ in G'. Since the transformation from G, k to G', λ runs in polynomial time, Clique is reducible to Connected-Cohesive and therefore Connected-Cohesive is $\mathcal{NP}$-hard. Given that Connected-Cohesive is also in $\mathcal{NP}$, the problem is thus $\mathcal{NP}$-complete. □

The associated decision problem being $\mathcal{NP}$-complete, the problem of finding a set of vertices with maximum cohesion is $\mathcal{NP}$-hard. Note that the problem of finding a set of vertices of maximum cohesion containing a set of predefined vertices is also $\mathcal{NP}$-hard, by an immediate reduction.

k-Min-Cohesion is $\mathcal{NP}$-hard

Dual to the problem of finding a subset with maximum cohesion is that of minimizing the cohesion, which can be useful when trying to identify socially weak subgraphs in highly cohesive networks. We formulate the k-Min-Cohesion problem in the following way:

k-Min-Cohesion

Input	A graph $G = (V, E)$
Output	A subset $S \subseteq V$ such that $\lvert S \rvert = k$ and $\forall S' \subseteq V$, $\mathcal{C}(S) \leq \mathcal{C}(S')$

In order to prove that k-Min-Cohesion is $\mathcal{NP}$-hard, we will show that the problem of finding a set of nodes of size k with cohesion 0 is $\mathcal{NP}$-complete. First note that one can check in polynomial time that a set of nodes has cohesion 0, thus the problem is in $\mathcal{NP}$.

Now notice that if a set of nodes has cohesion 0, then in particular $\circledcirc(S) = 0$, which means that S does not contain any triangles. Conversely, if $\circledcirc(S) = 0$ then $C(S) = 0$, therefore finding a set S of size k such that $C(S) = 0$ is equivalent to the problem of finding a triangle free subgraph of size k. The property of being triangle free is hereditary: all subgraphs of a triangle-free subgraph is itself triangle free, and is non trivial: there are infinitely many triangle free subgraphs, therefore the problem of finding a triangle free induced subgraph is $\mathcal{NP}$-complete.[1]

Although this chapter might seem dry at first glance, it establishes two fundamental results in cohesion theory. Both the problems of, given a graph, finding the most cohesive sets and the less cohesive (with fixed size) sets are $\mathcal{NP}$-hard and as a consequence, there is no known fast algorithms to solve those problems. This has broad implications in social network analysis, as maximizing the cohesion is the way to find the best communities in the network. Similarly, the dual problem is interesting in itself, as it would allow to pinpoint the weak underbelly of a social network.

1 John M Lewis and Mihalis Yannakakis. "The node-deletion problem for hereditary properties is NP-complete". In: *Journal of Computer and System Sciences* 20.2 (1980), pp. 219–230.

C^3

Communitize, Cover, Combine

Pick me! Pick me! Me! Me!
Shrek
Donkey

Given that the problem of finding groups with maximal cohesion is $\mathcal{NP}$-hard, in this section we introduce a heuristic algorithm, Communitize which attempts to find an optimally cohesive group containing a given node. We then build on this algorithm to propose a second algorithm, Cover, which outputs a covering of a given network into communities. Then, to deal with the overlap between communities which might prove troublesome we introduce Combine to detect which communities should be merged into one community. Those three heuristics are the building blocks for C^3, our community detection algorithm.

Communitize around a node

As stated earlier, in Fellows we used a simple greedy algorithm in order to find a set of nodes with high cohesion containing one seed node. One of the main reasons why a greedy algorithm cannot lead

to optimally cohesive communities – other than the fact that the problem is $\mathcal{NP}$-hard – is due to the fact that if the original set of nodes resides inside a very cohesive community then chances are that at some point in the iteration of the algorithm there could be the need to lower the cohesion in order to go through a *barrier of transitivity*.

In order to overcome the limitation of this approach, we now present an enhancement of the greedy algorithm specially tailored for the problem of maximizing the cohesion. Suppose that we have a set of nodes S and we wish to add nodes to S in order to potentially find a new set $S' \supseteq S$ such that $C(S') \geq C(S)$. One obvious solution, and this is the greedy approach, is to start by adding to S a node which increases its cohesion.

When one thinks about the way the cohesion is defined, it is possible that no node in the neighborhood of S can increase its cohesion on its own, but that adding several nodes at the same time might. There are in fact two ways of increasing the cohesion, either increase transitivity or increase its isolation, or a combination of thereof.

One solution is then to explore the possibility that adding to S a node which increases its transitivity might lead to a higher cohesion in the end, if this turns out not to be as successful as hoped, then we can always revert back to S and try with another unvisited node. One way of understanding this heuristic is that the newly added node allows us to dig into the heart of the community, setting its cohesion aside temporarily.

Now let us consider the idea of adding a node to S in order to isolate S from the rest of the network. Unfortunately this is a rather

1: **function** isCandidate($S \subseteq V, u \in V$)
2: **return** false **if** $u \notin \mathcal{N}(S)$
3: **return** false **if** $u \in S$
4: **return** true **if** $\triangle(S) = 0$ and $\triangle(S) < \triangle(S \cup \{u\})$
5: **return** false **if** $\triangle(S) = \triangle(S \cup \{u\})$
6: **return** true **if** $C(S \cup \{u\}) \geq C(S)$
7: **return** true **if** $\triangle(S \cup \{u\}) \geq \frac{|S|+1}{|S|-2} \triangle(S)$
8: **return** false

Algorithm 6.1 Checks if a node u is a valid candidate to add to S

useless heuristic given that the main idea behind increasing the isolation is to construct the boundary of the community. In this context, there is no need to add a node if we have no intention of going further into that node's neighborhood.

We shall say that u is a *candidate* for S if adding it to S either increases the cohesion or the transitivity. Moreover, we add a few other constraints which are detailed in Algorithm 6.1. We first check that the node is a neighbor of our seed group because as we have seen before the community with maximal cohesion is connected and we then verify that the node has not already been added to the community.

After those two validations, we add a special case on Line 4 which acts when the group does not contain any triangles. In that case the node is a candidate only if we increase the number of outbound triangles by adding it, the rationale being that we are in presence of a sparse group which could benefit from new triangles. Therefore, adding a node which increases $\triangle$ allows us to have more triangles to choose from at a later step and go deeper into

what could be the heart of the community.

On Line 5, we forbid nodes who do not create any triangles inside the community to be added. The main reason the two previous constraint are added is to be able to deal with the cases where $|S| \leq 3$, in which case S does not contain any triangle and we need to bootstrap a beginning of community. Finally, we mark as candidates the nodes which increase the cohesion (Line 6) and the transitivity (Line 7).

At this point, we are now able to select candidates in the network to be added to a set of nodes. The task at hand is then to discriminate which of those nodes is the best possible candidate. To that effect, given a set of nodes S and its candidates C, we shall introduce an order $\preceq_S$ on the elements of C. We first define the importance $I(u)$ of a node u as a tuple consisting of the cohesion of the set if u was added, the transitivity of the set if u was added, the number of outbound triangles that u would add and finally the degree of u (Eq. 6.1).

$$I(u) = \left(C\left(S \cup \{u\}\right), \frac{\textcircled{\triangle}\left(S \cup \{u\}\right)}{|S| + 1}, \text{⛥}\left(S \cup \{u\}\right) - \text{⛥}\left(S\right), d(u) \right) \tag{6.1}$$

We then define the order $\preceq_S$ on C in terms of the lexicographical ordering of the values of I. Given two candidates $u, v \in C$, $u \preceq_S v$ if and only if $I(u) \preceq_{\text{lex}} I(v)$. This means that we first compare the cohesions after adding each node, if one node has a higher cohesion then it ranks higher than the other one. In case of equality, we move on to compare the transitivities, and so on.

The third and fourth component of I deserve an explanation. Suppose that u and v are such that adding both nodes to S lead to the same cohesion and transitivity, then two cases present themselves: first, the cohesion may be different from 0, in which case the number of new outbound triangles is the same, given that the transitivity is the same.

Second, and more interesting, is when the cohesion after adding u or v is 0, there $\triangle(S \cup \{u\})$ and $\triangle(S \cup \{v\})$ may be different, in which case we decide that the node which would create the larger number of outbound triangles is the best candidate.

Bare in mind that the underlying idea is to always increase the cohesion, in which case it can only be 0 in the first rounds of adding new nodes to the initial seed, thus adding nodes which create a lot of outbound triangles is a way of digging into the heart of the community.

Finally, the last component serve a similar purpose, in the event that the previous three metrics were equal, we choose to select as best candidate the node which addition will bring in the larger amount of potential candidates.

We now have the blocks to build the Communitize algorithm (Alg. 6.2) which we define recursively. Given a graph G and a set of nodes S we first establish the list C of nodes of G which are candidates with respect to S. Then, for each node u of C, chosen in decreasing order of $\leq_S$, if we have not already visited that node, we compute the best community containing $S \cup \{u\}$. Finally, we return the community with the highest cohesion in the set $\{S\} \cup \bigcup_{u \in C} \text{communitize}(S \cup \{u\})$.

```
1: function Communitize(S ⊆ V)
2:     B ← S
3:     C ← { u ∈ V | isCandidate(S, u) }
4:     for all u ∈ C in decreasing order of ≤_S do
5:         if u is not marked as visited then
6:             mark u as visited
7:             B′ ← Communitize(S ∪ {u})
8:             if C(B) ≤ C(B′) then
9:                 B ← B′
       return B
```

Algorithm 6.2 Communitize expands around a set of nodes to find the best enclosing community

There are several tweaks and enhancements which can be added to Communitize, which have been omitted for the sake of clarity. For example, instead of comparing only the cohesion at line 8, we can discriminate between communities of same cohesion by always choosing the larger one.

More importantly, computing the cohesion of a set of nodes has a non negligible cost, as it is done in $\mathcal{O}\left(\left|S \cup \mathcal{N}(S)\right|^3\right)$ – as it involves counting the number of triangles in a graph of size $\left|S \cup \mathcal{N}(S)\right|$. It is however not mandatory to recompute the cohesion at each step, as we can just track the variations induced by the addition or deletion of a node. To that effect, we need to keep count of the following quantities:

- the number of inbound triangles $\textcircled{\triangle}$;
- the number of outbound triangles $\textcircled{\triangle\!\!\!-}$;
- for each node u, the number of inbound triangles which would be added if u was added to S: $\textcircled{\triangle}_u = \textcircled{\triangle}(S \cup \{u\}) - \textcircled{\triangle}(S)$;
- and similarly, the number of outbound triangles which would

be added if u was added to S: $\triangle_u = \triangle(S \cup \{u\}) - \triangle(S)$. We can then write an updateSet function which adds or delete a node u to S and maintains the correct values of $\textcircled{\triangle}$ and $\triangle$. Let $\delta = 1$ when u is added and $\delta = -1$ when u is deleted. Note that if the list of neighbors $\mathcal{N}(u)$ are sorted (e.g. in degree order, or by identifiers), it is possible to iterate in linear time over the intersection of two neighborhoods. Overall, we can bring the complexity of the update down to $\mathcal{O}\left(\sum_{v \in \mathcal{N}(u)} d(u) + d(v)\right)$.

function updateSet($S \subseteq V, u \in V, \delta$)
  $\textcircled{\triangle} \leftarrow \textcircled{\triangle} + \delta \textcircled{\triangle}_u$
  $\triangle \leftarrow \triangle + \delta \left(\triangle_u - \textcircled{\triangle}_u\right)$
  for all $v \in \mathcal{N}(u)$ **do**
    for all $w \in \mathcal{N}(u) \cap \mathcal{N}(v)$ **do**
      if $v \in S$ **then**
        $\textcircled{\triangle}_w \leftarrow \textcircled{\triangle}_w + \delta$
        $\triangle_w \leftarrow \triangle_w - \delta$
      else
        $\triangle_w \leftarrow \triangle_w + \delta$

Algorithm 6.3 Updates the values of $\textcircled{\triangle}$, $\triangle$ of a set without counting all triangles

Finally, the list of candidates which may be added to the set can be updated in a similar fashion. When adding a node u to S, we update the candidate list and add new candidates which will all be neighbors of u that are not already present in the set of candidates. We compute the ordering of those new candidates in $\mathcal{O}(d(u) \log d(u))$ and merge them into the set of candidates in $\mathcal{O}(|C| + d(u))$. In practice, the cost of this operation is negligible compared to that of updating the cohesion.

By modifying the Communitize algorithm in order to create a community which is then modified in place using setUpdate, we obtain an algorithm which optimizes the cohesion by recursively adding nodes to an initial seed and returns the most cohesive group it encounters.

The complexity of adding or deleting a node stems from the update step which cost is $\mathcal{O}(\sum_{v \in \mathcal{N}(u)} d(u) + d(v))$. Each node is added and deleted from the community at most once, which leads to an overall worst case complexity complexity of $\mathcal{O}\left(\sum_{u \in V} \sum_{v \in \mathcal{N}(u)} d(u) + d(v)\right) = \mathcal{O}\left(\left|V\right| \left|E\right|\right)$. Note that if there is a bound on the degrees $d_{\max}$, then the complexity is reduced to $\mathcal{O}(|V| d_{\max}^2)$.

Cover a set of nodes

Although Communitize allows us to expand around a set of nodes in order to obtain a more highly cohesive set of nodes, it is not always desirable to obtain one and only one community. For example, let C_1, C_2 be two cliques with a sufficiently low overlap such that both are more cohesive than the union. Consider $u \in C_2 \setminus C_1$, if we simply compute the best community containing $C_1 \cup \{u\}$ we shall obtain a community S which is neither C_1 nor C_2, because S would contain $C_1 \cup \{u\}$, and thus S would be less cohesive than C_1 and C_2. More generally, we aim to obtain a *covering* of $C_1 \cup C_2$ in communities, *i.e.* a set (S_i) of communities such that $\bigcup_i S_i = C_1 \cup C_2$.

To that effect, we introduce a second algorithm (Alg. 6.4), which uses Communitize to expand around carefully selected

```
function Cover-Nodes(S ⊆ V)
    C ← ∅
    for all u ∈ S in increasing order of d(u) do
        if u is not marked as covered then
            c ← Communitize({u})
            for all v ∈ c do
                mark v as covered
            C ← C ∪ {c}
    return C
```

Algorithm 6.4 Cover-Nodes computes a set of cohesive communities such that each node in S is in a community.

nodes of the network. The idea is to choose one node u in S, find the best community containing u, mark all the nodes in that community as covered and repeat as long as there are non covered nodes. We choose to iterate over S by increasing degree as placing low degree nodes first into their communities allows us to more precisely capture the community structure around nodes with a higher degree – this choice was mainly done due to empirical observations that it lead to better communities.

Although the algorithm only computes one community for each node this does not lead to a partition because Communitize can add already covered nodes to a community. However, the algorithm as such misses some communities. For example, in the four triangles depicted in Figure 6.1, the algorithm will start by expanding around a node at the periphery, which will yield one of the outer cliques. It will then repeat the same thing twice and stop, resulting into three communities of size three, totally ignoring the presence of the middle clique which has already been covered.

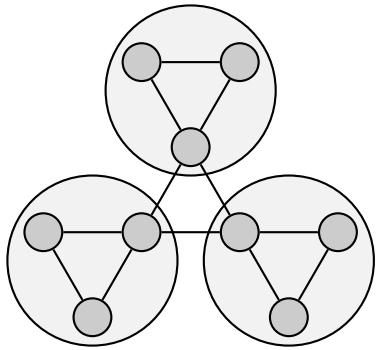


Figure 6.1 Those four cliques are incorrectly covered by Cover-Nodes which would only find the three encircled communities and ignore the clique at the center.

In order to compensate this last flaw, we propose a heuristic which deal with those kinds of effects. Instead of computing the best community containing a given node u we compute the best communities containing $\{u, v\}$, where v is a non covered neighbor of u. If all of the neighbors of u are covered, we then compute the best communities containing $\{u, v\}$ for all v neighbors of u. In both cases we obtain a set of community, of which we choose the one with the highest cohesion (Algorithm 6.5)

The final improvement we bring to Cover is to reduce the size of the considered network. As we have seen in the previous chapters, maximal cohesion is attained inside networks which are connected. In practical cases however – that is, in most real world use cases, except some rare pathological cases – this condition may be strengthened by considering what we shall call *triangle-connected* components.

```
function Cover(S ⊆ V)
    C ← ∅
    M ← ∅
    for all u ∈ S \ M in increasing order of d(u) do
        C_u ← ∅
        for all v ∈ 𝒩(u) \ M do
            c ← Communitize({u, v})
            C_u ← C_u ∪ {c}
            M ← M ∪ c
        if C_u = ∅ then
            for all v ∈ 𝒩(u) ∩ M do
                c ← Communitize({u, v})
                C_u ← C_u ∪ {c}
                M ← M ∪ c
        C ← C ∪ {arg max_{c∈C_u} C(c)}
    return C
```

Algorithm 6.5 Cover is an enhanced version of Cover-Nodes which bypasses the non-detection of some communities in the original algorithm

Let t_1, t_2 be two triangles of G, we will say that t_1 and t_2 are connected if both triangles share a common edge. Now, given a pair of edges $e_1, e_2 \in E$, we say that they are triangle-connected if there exist a sequence $(t_i)_{0 \leq i \leq k}$ of triangles such that e_1 is an edge of t_0, e_2 is an edge of t_k and for all $0 \leq i < k$ the triangles t_i and t_{i+1} are triangle-connected. One may notice that it is a reformulation of Palla's clique percolation applied with $k = 3$[1] and as such is a natural extension of connectivity taking into account triangles rather than edges.

1 Palla, Derényi, Farkas, and Vicsek, op. cit.

```
function Threeway Union(x, y, z)
    x_R ← Find(x)
    y_R ← Find(y)
    z_R ← Find(z)
    if x_R = y_R = z_R then return
    if x_R.rank ≥ y_R.rank and x_R.rank ≥ z_R.rank then
        y_R.parent ← x_R
        z_R.parent ← x_R
        if x_R.rank = y_R.rank or x_R.rank = z_R.rank then
            x_R.rank ← x_R.rank + 1
    else if z_R.rank ≥ z_R.rank then
        x_R.parent ← y_R
        z_R.parent ← y_R
        if y_R.rank = z_R.rank then
            y_R.rank ← y_R.rank + 1
    else
        x_R.parent ← z_R
        y_R.parent ← z_R
```

Algorithm 6.6 Threeway Union is a variation of Tarjan's original union which unifies three elements in one call.

To compute those triangle-connected components, we use a variation of Tarjan's Union-Find algorithm[1] which differs in the *union*: for each triangle (u, v, w), instead of unifying each pair of edges one at a time, we unify the three edges (uv, vw, uw) by doing a three way union in one sweep as described in Algorithm 6.6.

1 Robert Endre Tarjan and Cornell University. Department of Computer Science. *On the efficiency of a good but not linear set union algorithm*. Tech. rep. Nov. 1972.

We then use the Compact-Forward algorithm[1] to enumerate all triangles in a given graph and maintain a disjoint-set data structure using our Threeway-Union/Find algorithm. This gives us separate triangles connected components on which we apply Cover, merging the resulting sets of communities at the end.

Finally, we have presented in this section an algorithm which places each node of a given set in at least one cohesive community. Although the resulting algorithm has a worst case complexity bounded by $\mathcal{O}(|V||E|^2)$, this would be the case when each edge lead to a community containing all the graph without covering any other edge, which is impossible.

Combine similar groups

Recall that earlier we mentioned the issue of judging the quality of a covering versus that of a community, and that we concluded that there is no globally acceptable metric to that effect. We now present the last pillar of C³, which allows to parametrically control the amount of overlap which is authorized.

Let us consider a graph G. After running Cover we obtain a collection of communities $(S_i)_{0\leq i\leq k}$. Suppose that we are provided with a function *Ov* which rates at which extent two communities S_i and S_j overlap – examples of function commonly used to that effect are given in Table 6.1. Furthermore, suppose that we dispose of a maximum authorized overlap $o_{\max}$. We construct a weighted graph Γ where each node u_i corresponds to a community S_i, and

1 Matthieu Latapy. "Main-memory triangle computations for very large (sparse (power-law)) graphs". In: *Theoretical Computer Science* 407.1 (2008), pp. 458–473.

where there are edges between two nodes u_i and u_j if and only if $Ov(S_i, S_j) \geq o_{\max}$, the weight of the edge being $Ov(S_i, S_j)$.

Overlap	Contains	Jaccard	Dice	Cosine
$\frac{\lvert S_i \cap S_j \rvert}{\min\left(\lvert S_i \rvert, \lvert S_j \rvert\right)}$	$\frac{\lvert S_i \cap S_j \rvert}{\max\left(\lvert S_i \rvert, \lvert S_j \rvert\right)}$	$\frac{\lvert S_i \cap S_j \rvert}{\lvert S_i \cup S_j \rvert}$	$\frac{2\lvert S_i \cap S_j \rvert}{\lvert S_i \rvert + \lvert S_j \rvert}$	$\frac{\lvert S_i \cap S_j \rvert}{\sqrt{\lvert S_i \rvert \lvert S_j \rvert}}$

Table 6.1 Examples of overlap functions $Ov(S_i, S_j)$

Once the communities are laid out this way, the problem of finding sets of communities which overlap sufficiently reduces to a problem of "community" detection in the meta-graph. However, contrary to the graphs we have encountered until now, Γ is a weighted graph, therefore we have to adapt the definition of the cohesion in order to be able to recursively use C³ to find the meta-communities.

Fortunately enough, there are several ways the definition of the cohesion can be extended to take into account graphs where edges have weights in $[0, 1]$. Basically, it suffices to produce a function which allows to transfer the notion of weights from the edges uv, uw, vw to the triangle uvw. We suggest two such functions in Table 6.2.

	Product	2 out of 3
$W(uvw) =$	$W(uv)W(uw)W(vw)$	$\max \begin{cases} W(uv)W(uw) \\ W(uv)W(vw) \\ W(uw)W(vw) \end{cases}$

Table 6.2 Examples of weight functions on triangles

The first function naturally assigns the product of the weight of its edges to a triangle. However, when judging the overlap contained inside a triangle, it might be useful to use the second function which adds some transitivity to the overlap function. Using such weights of triangles, the definition of the weighted cohesion comes immediately, in the weighted version ⟁ is the sum of the weights of inbound triangles and ⟁ becomes the sum of the weights of outbound triangles.

We can then compute the communities of Γ by using Cover and recursively calling Combine on the result. Then, for each meta-community Σ, three cases are possible:

- $|\Sigma| = 1$: the community in Σ is not affected;
- $|\Sigma| = 2$: the communities S_1, S_2 in Σ are merged;
- $|\Sigma| \geq 3$: the communities in Σ are merged if $CC \geq C_{\text{merge}}$.

Given that for a graph of size n, Cover gives at most $n - 3$ communities, the size of the meta-graph is strictly smaller than that of the original graph, and as such Combine always finishes. In the end, we obtain cohesive communities while at the same time controlling the overlap between communities with two parameters $o_{\max}$ and C_{merge} – in practice, using a small value of C_{merge} is perfectly acceptable, we just need to avoid merging meta-communities of cohesion 0.

Continuing the algorithmic part of this thesis, since the maximization problem is $\mathcal{NP}$ complete, this chapter presents C³, a heuristic algorithm dedicated to finding cohesive communities in a network.

C^3

C^3 works in three steps: first it expands around a seed set in order to find a cohesive superset, then it repeats that procedure as long as there are uncommunitized nodes, and finally it merges carefully selected communities in order to constrain the maximal overlap between communities.

Dynamics of Communities

Evolution and Stability of the Agreement Groups in the United States Senate

> However [political parties] may now and then answer popular ends, they are likely in the course of time and things, to become potent engines, by which cunning, ambitious, and unprincipled men will be enabled to subvert the power of the people.
>
> *The Farewell Address*
> George Washington

Having introduced C^3 to compute cohesive communities, we will apply it in this chapter to real-world data, and this for two reasons. First our aim will be to validate that C^3 computes communities which make sense in terms of the semantic behind the data, and second we are also interested in the amount of non-trivial information C^3 is able to unearth from the data. Therefore, we present a case study of the evolution of the political groups in the United States Senate, as computed by C^3. Using publicly available data, we construct a graph of voting similarity between U.S. Senators for each of the 112 U.S. Congresses. We then apply

to each of those groups an extension of C^3 to weighted and signed graph in order to determine the communities of voters who are in agreement. Finally, building on the temporal aspect of communities which the algorithm does not take into account, along with a factual and historical analysis of the results, we will establish that C^3 is a valid method to compute overlapping communities.

The Dataset

The United States Senate is the upper house of the United States legislature. Contrary to the House of Representative which seats are up for election every two years, Senators serve terms of six years each. Those terms are however staggered so that approximately one-third of the Senate is renewed every two years. This period of two years is called a *United State Congress*.

Contrary to other countries, data concerning elected officials and the activity of the houses is openly available in the United States. We have used the GovTrack[1] website which provides both a list of all elected officials and votes both at the Senate and the Congress to construct graphs of agreement. As seen earlier, each Senator usually serves, except for unfortunate events, in at least three consecutive Congresses. Therefore we have decided to focus on the Senate rather than the House of Representatives as the former has the particularity of having a continuity of its members, which would allow to observe more precisely the evolution of political groups.

1 Civic Impulse, LLC. *GovTrack.us*. URL: http://www.govtrack.us/.

The data we have obtained consists, for each vote taking place at the United States Senate during a given Congress, of a list of those who have voted and the nature of their vote. For each of the 112 Congresses we shall construct an agreement graph $G_i = (V_i, E_i)$ where V_i is the set of the Senators active during the i^{th} Congress. Due to some ambiguity in the data, we could not restrict ourselves to the Senators and actually construct the graph of those who have casted at least one vote in Senate, this might also include other elected officials, such as the Vice-President, nevertheless, we shall qualify our actors of Senators, for clarity's sake.

All votes we have encountered were choices between two options which we shall arbitrarily denote A and B, therefore each Senator had either voted A, or B or did not vote. Usually those alternatives A and B were "Yea" and "Nay", indicating support or rejection of a proposal. For each Congress i we associate to the Senator s a vote vector $V^{i,s}$ of dimension the number of votes which have taken place during that Congress, such that:

$$V^{i,s}_k = \begin{cases} 1 & \text{if } s \text{ voted } A \text{ for the } k^{\text{th}} \text{ vote} \\ 0 & \text{if } s \text{ did not vote for the } k^{\text{th}} \text{ vote} \\ -1 & \text{if } s \text{ voted } B \text{ for the } k^{\text{th}} \text{ vote} \end{cases}$$

We can then compute the agreement (or weight) between two Senators during the i^{th} Congress as the cosine similarity between their votes:

$$W_i(s_1, s_2) = \frac{V^{i,s_1} \cdot V^{i,s_2}}{\| V^{i,s_1} \| \| V^{i,s_2} \|}$$

For example, if there were three votes during a session, and

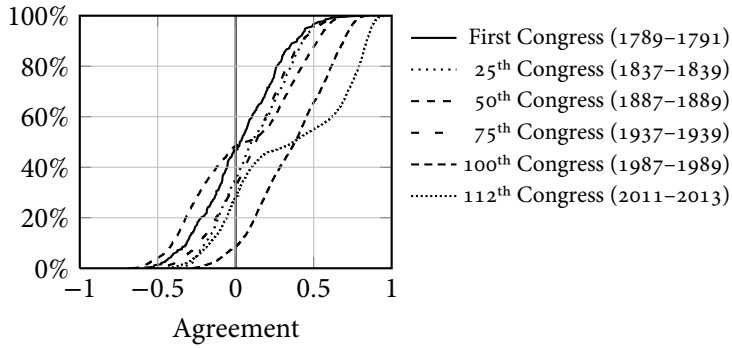


Figure 7.1 Cumulative distribution of the weights in Senate agreement graph for six different Congresses.

that two senators s_1 and s_2 voted[1] respectively (Nay, Yea, _, Nay) and (Yea, Yea, Nay, Yea), their vote vectors would be $V^{i,s_1} = (-1, 1, 0, -1)$ and $V^{i,s_1} = (1, 1, -1, -1)$. As a consequence, those two senators would have a similarity $W_i(s_1, s_2) = \frac{1}{2\sqrt{3}}$.

The highest agreement $W = 1$ is attained when s_1 and s_2 have identical vote vectors and the lowest agreement $W = -1$ is achieved when their vote vectors are opposed. The cosine similarity allows us to incorporate the absence of votes to the agreement metric and give them a smaller weight that between opposite weights. Given those agreement weights, we can now construct the edges of the agreement graph G_i, where we add an edge of weight $W_i(s_1, s_2)$ between s_1 and s_2 if $W_i(s_1, s_2) \neq 0$.

The cumulative distribution of those weights are given, as an example, on Figure 7.1. It is notable that in all cases more than

1 The underscore indicates here an absence of casted vote.

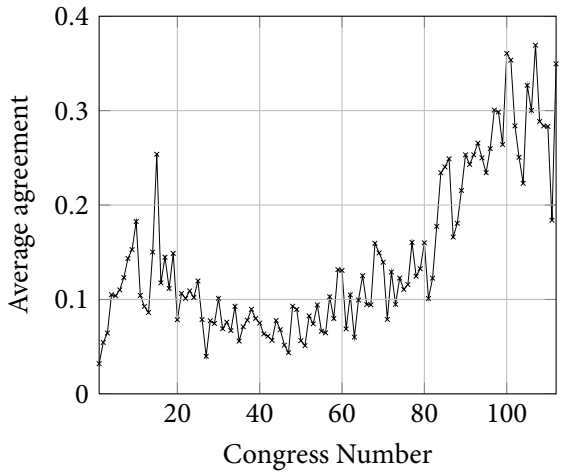


Figure 7.2 Evolution through time of the average agreement weight between Senators.

50% of the edges have a positive value. We can observe that the earlier Senates presented a certain balance in the distribution of the edges: there was a similar number of edges of positive and negative weights. More recent Senates have a bias towards agreement, as exemplified by the latest Senate (112^{th} Congress (2011–2013)) where 75% of edges have a positive weight and 45% of edges have a value greater than 0.5.

This trend is explicit when we observe how the average value of agreement (Fig. 7.2) evolves. The first thing to notice is that the average value is always greater than 0, which indicates that despite being from different political horizons Senators tend to agree more with each other than to disagree.

There are however variations in the evolution of the average agreement. We shall describe a few data points and provide some historical context which might shed some light on those variations although for obvious reasons we do not imply causation. During the first few Congresses, the average agreement increases in a context where the United States are a young nation. There is then a sudden drop during the Eleventh and Twelfth Congresses, which took place just before and during the war of 1812, the first major conflict between the United States and the British Empire since the end of the American Revolutionary War in 1783. During the Fifteenth Congress (1817–1819), the average agreement rises to more than 0.2. Coincidentally, at that time, the United States faced the "Panic of 1819", its first major financial crisis.

The agreement then steadily decreases, attaining its minimum during the 27th Congress (1841–1843), in the years of instability leading of the Civil War and the decreasing in average through the Reconstruction era. It is only with what Mark Twain dubbed the "Gilded Age" that the average agreement increases again, around the time of the 51st Congress (1889–1891). The next major increase occurs around the 83rd Congress (1953–1955). In 1953, major political changes occur both in the United States and the USSR, January is marked by the election of Dwight D. Eisenhower and after Joseph Stalin's death in March, Nikita Khrushchev became the Soviet leader. Those changes shifted the dynamic of the cold war which had been until then contained, in particular through the Korean war. At the same time, Joseph McCarthy started its communist witch hunt while heading the Senate Permanent Subcommittee on Investigations.

The Cuban Missile Crisis occurred during the 87th Congress (1961–1963) which marked a temporary decrease in agreement, although the increase would then continue until the 102nd Congress (1991–1993). In the past two decades, the agreement has swinged up and down, staying on average higher than 0.18. Notice how it has attained it has peaked at its maximum during the 107th Congress (2001–2003), which was marked by the 9/11 attacks.

Signed & Weighted Cohesion

The graphs G_i that we have obtained in the previous section have weights which vary between -1 and 1, therefore we need to extend the definition of the cohesion in order to take those negative edges into account. If an edge uv has a negative weight, it means that u and v are in disagreement and should not be added to a same community.

In terms of triangle, the consequence is that if a triangle contains at least a negative edge, then it should contribute negatively to the cohesion. We therefore introduce a the $\text{sgn}(uvw)$ function which gives us the sign of the contribution of a triangle:

$$\text{sgn}(uvw) = \begin{cases} -1 & \text{if } W(uv) < 0 \text{ or } W(vw) < 0 \text{ or } W(vw) < 0 \\ 1 & \text{in all other cases} \end{cases}$$

From there we can define the signed weight of a triangle $W_{sgn}(uvw) = \text{sgn}(uvw)W(uvw)$, where $W(uvw)$ can be any unsigned triangle weighing function, for example one given in Table 6.2. Here we shall choose to use the product function,

defined as $W(uvw) = W(uv)W(uw)W(vw)$.

Let us now extend the cohesion in order to take into account the signed weights of triangles and at the same time remain compatible with its unsigned version. If a group has negative △ and positive ⧋, it means that there is more disagreement inside the group than towards the rest of the network, and therefore the cohesion should be low. For similar reasons, if △ if positive and ⧋ is negative, the group has a high agreement with itself and is opposed to the rest of the network, which should result into a high cohesion. Intuitively, if △ and ⧋ are of opposite signs, the group is isolated from the rest of the network and the cohesion is reduced to the transitivity.

	$\text{△} < 0$	$\text{△} \geq 0$
$\text{⧋} < 0$	$\frac{\text{△}}{\binom{n}{3}} \frac{\text{△}}{\text{△}+\text{⧋}} \leq 0$	$\frac{\text{△}}{\binom{n}{3}} \geq 0$
$\text{⧋} \geq 0$	$\frac{\text{△}}{\binom{n}{3}} \leq 0$	$\frac{\text{△}}{\binom{n}{3}} \frac{\text{△}}{\text{△}+\text{⧋}} \geq 0$

Table 7.1 Impact of signed triangles weights on the Cohesion.

Finally, there is the case when both △ and ⧋ are negative. In that case the expression of the isolation factor and thus that of the cohesion remain the same. The formulas for the signed and weighted cohesion are given in Table 7.1. This new definition of the cohesion can be used directly in C^3 without adapting the algorithm. We shall however filter the resulting list of communities to remove those which have a negative cohesion, as those would not

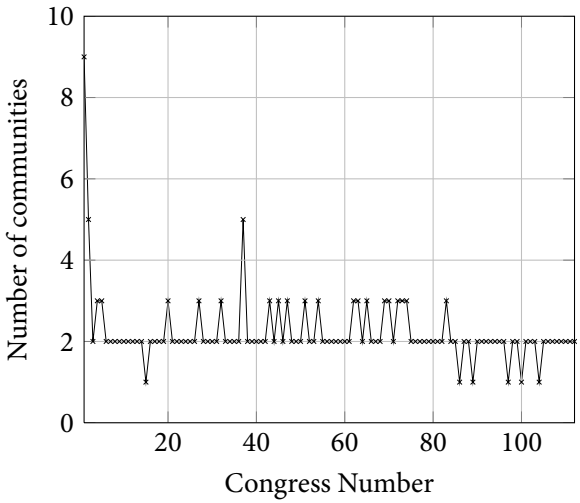


Figure 7.3 Evolution through time of the number of communities in the Senate agreement graphs.

qualify as "good" communities.

Of History, Dynamics and Stability

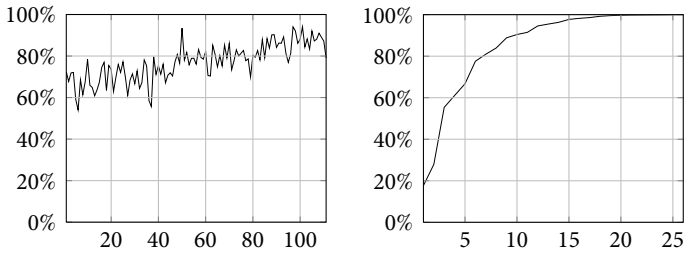
For each Congress we have computed using C^3 the communities of its agreement graph, using the extended cohesion presented in the previous section. On Figure 7.3 we have displayed the evolution of the number of communities of agreement through time. In all but three cases, there are between one and three communities. The First Congress (1789–1791) has 10 different communities and both the Second Congress (1791–1793) and the 37^{th} Congress (1861–1863) have 5 communities.

Concerning the two first Congresses, one has to bear in

mind that there were no national political parties prior to the Presidential Election of 1796. The United States were a young nation and did not have a two-party system, it basically had George Washington, a president without a party. At that time, those who served in Congress are best described as either Administration (*i.e.* those associated with Vice President John Adams) or Opposition (*i.e.* those who surrounded Secretary of State Thomas Jefferson) but it is important to note that that era is more an era of faction rather than parties, and thus alliances would shift at a fast pace in this early era of U.S. political history, which is visible when looking at the number of communities during the four first Congresses.

By the start of the Fifth Congress (1797–1799), two national political parties had emerged from the two aforementioned factions. Those who had be supporter of the Washington Administration became known as the Federalists because they were partisans of a strong Federal Government which could oppose the importance of individual states. Those who had been in the Opposition became know as the Republicans because they insisted on defending the sovereignty of the States against the Federal Government.

Notably, the third extremal point (5 communities for the 37th Congress (1861–1863)) coincides with the beginning of the American Civil war in 1861 and the larger number of communities reflects the political turmoil at the time. From there one, the United States have had a two-party system, which is visible in the number of communities in the agreement graph which except the three previously mentioned exceptions vary between 2 and 3. Furthermore, since the 84th Congress



(a) Senate Continuity (b) Senators numbers of Congresses

Figure 7.4 Left: Evolution through time of the proportion of Senators remaining in office between two Congresses as a function of the Congress number. Right: Cumulative distribution of the number of terms per Senator.

(1955–1957), there was no more than two communities – and there were five occurrences where there was only one community. This diminution in the number of communities is a direct consequence of the previously mentioned increase in agreement among Senators.

Let us now study the dynamics of those graph. The first thing to notice is that, as we have said earlier, at each Congress, only a third of the seats are up for election, which means that there should be a certain continuity in the members of the Senate. On Figure 7.4(a), we have plot the proportion of the members of the Senate during a given Congress which remained in position during the following Congress. This continuity C_i is expressed as:

$$C_i = \frac{|V_i \cap V_{i+1}|}{|V_i|}$$

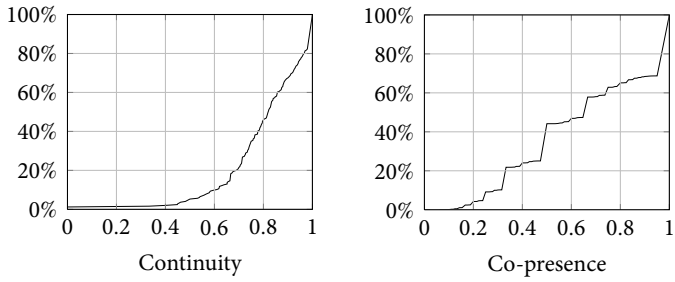
As we expected there is a high continuity between Senate sessions, in most cases larger than 2/3 – the cases where it is lower can be explained by the passing of some Senators or other unfortunate events. It is also interesting to notice that this score tends to increase as time passes, which is a consequence of the fact that, more and more, Senators keep there seats for more than one term.

Figure 7.4(b) represents the cumulative distribution of the number of terms served by each Senator. More than a third of all Senators have served more than five terms and almost half of the Senators have served at least three terms. Contrast this with the fact that a U.S. President cannot be elected for more than two terms since the passage of the Twenty-second amendment. The notion that there should be a term limit in Congress was brought forth by the Republican Party in the 1990s but the proposal fell through in the House.

We shall now quantify the evolution of the communities in two different ways. First, similar to the way we have defined continuity for the whole Senates, we shall define a metric of continuity between to communities of two consecutive Congresses. Let us consider the communities $(S_{i,j})$ and $(S_{i+1,j})$ of the i^{th} and $i+1^{\text{th}}$ Congresses. We shall define then continuity between two communities $S_{i,j}$ and $S_{i+1,k}$ as:

$$c(S_{i,j}, S_{i+1,k}) = \frac{\left|S_{i,j} \cap S_{i+1,k}\right|}{\left|S_{i,j} \cap \bigcup_l S_{i+1,l}\right|}$$

That is, the ratio of Senators present in both groups compared to those present in the oldest one and which are also active in



(a) Group Continuity (b) Co-presence in same groups

Figure 7.5 Left: Cumulative distribution of the values for community continuity between a community and its successor. Right: Cumulative distribution of the co-presence ratio for pairs of Senators present in at least one community together.

the second Congress. The idea is to compare apple-to-apple the dispersion on the Senators present in a given group, this is why we restrict ourselves to those who are present in both Congresses i and $i+1$. For each community S_1, we shall say that is *successor* is the community S_2 for which $c(S_1, S_2)$ is maximal.

Figure 7.5(a) displays the cumulative distribution of community continuities between each group and its successor. It is particularly notable that in more than 90% of cases, half of the members present in a community are also present in its successor – once again, only counting those present in both sessions.

Another way of looking at this question is at the level of the Senators themselves. We shall say that two Senators are *co-present* to a

certain degree if they belong to at least one community same community. Let u and v be two senators, we define their *co-presence* as the number of times they appear in the same community divided by the number of times they are active in a same Congress. Given that a Senator might be in several different communities, we shall count one presence for each community, and thus a Senator can be virtually present more than once during one given Congress. Let $(S_{i,j})$ be the set of communities for the graph G_i, we write $\mathcal{T}_{u,v}$ the number of times u and v appear in a same community:

$$\mathcal{T}_{u,v} = \sum_i \sum_j \mathbb{1}_{S_{i,j}}(u)\, \mathbb{1}_{S_{i,j}}(v)$$

Where $\mathbb{1}$ is the indicator function. We similarly define $\mathcal{S}_{u,v}$, the number of times u and v appear in the same session. We can then write the co-presence of u and v as:

$$P(u, v) = \frac{\mathcal{T}_{u,v}}{\mathcal{S}_{u,v}}$$

Figure 7.5(b) represents the cumulative distribution of the value of the co-presence for a selected subset of pairs of Senators. We have voluntarily excluded the pairs who never appear in the same Congress, that is $\mathcal{S}_{u,v} = 0$ as it would make no sense to compare their communities. Next, we have also removed the pairs which are never in the same community ($\mathcal{T}_{u,v} = 0$), as their inclusion bring no information on the stability of the communities.

Finally we have chosen to exclude the pairs who only appear in one Congress and belong to the same community ($\mathcal{T}_{u,v} = \mathcal{S}_{u,v} = 1$), although their inclusion would artificially increase the cumulative distribution we believe it would provide no insight on the

actual dynamic aspects of communities given that the information is only extracted from one same time slice. Their remains the pairs of Senators who appear in at least two Congresses and who are at least once in the same community. More than 30% of pairs of Senators are stable through time in respect to their communities, *i.e.* they have a co-presence of 1 and therefore always appear together in the same community. Moreover, 75% of the pairs have a co-presence greater than 0.5, meaning they are in the same community in the majority of the Congresses they are active in.

We have described the evolution of the number of communities of the agreement graphs through time, which we explained by referring to the history of the United States Political system and we have exhibited the continuity in Senate membership between Congresses. We have then shown that the communities which were found using C^3 present a certain stability, as they present a high continuity and that Senators appearing together in one community tend to be in the same communities during other Congresses. It is most notable to observe this kind of stability given that the data analysis done on each graph was made independently from the other graphs, which leads us to validate the use of C^3 to compute the communities of agreement.

The Blurry Line Between Parties

Until now, we have justified the number of communities by referring to political parties. Fortunately, we have access to the political affiliation of each Senator in our graph which means we can validate that intuition. We shall say that a party is *dominant* in a

community if it has the largest representation, and we shall call the *domination ratio* of a community the quotient of the number of members of the dominant party in the community divided by the size of the community.

On Figure 7.7 we have represented, for each session, the average of the domination ratio over all communities. Most communities have an average domination ratio of 70%, *i.e.* in most cases one can identify the community to the political party.

In particular, let us look more precisely at a subset of the data, ranging from the 105th Congress (1997–1999) to the 112th Congress (2011–2013). The sizes of the communities as well as the number members of each party represented in the community are given in Table. 7.2 and a visual representation of those communities are given in Figure 7.6. First notice that although the communities are allowed to overlap, there are only five cases where we witness an overlap, three of which being because one individual is part of the two communities, one because seven are shared between the two groups of the 111th Congress (2009–2011) and the largest overlap is attained during the 108th Congress (2003–2005) where 10 Senators are part of both communities.

As stated before, each of those communities has a clearly dominant party, be it the Democrat Party or the Republican Party. It is however interesting to observe three things. First, even though the communities have a large majority belonging to a same party, there are members of the other party which are more in agreement with there opponents than their political family, and this even in cases where there is no overlap between communities.

105th Congress (1997–1999)	100 members			
	45	Democrat	55	Republican
106th Congress (1999–2001)	102 members			
	45	Democrat	56	Republican
			1	Democrat
107th Congress (2001–2003)	101 members			
	49	Democrat	49	Republican
	2	Independent	1	Democrat
	1	Republican		
108th Congress (2003–2005)	100 members			
	40	Democrat	51	Republican
	1	Independent	18	Democrat
109th Congress (2005–2007)	101 members			
	44	Democrat	55	Republican
	1	Independent	1	Democrat
110th Congress (2007–2009)	102 members			
	50	Democrat	41	Republican
	2	Independent		
	11	Republican		
111th Congress (2009–2011)	110 members			
	63	Democrat	35	Republican
	2	Independent	6	Democrat
	11	Republican		
112th Congress (2011–2013)	101 members			
	51	Democrat	43	Republican
	2	Independent		
	6	Republican		

Table 7.2 Political party breakdown of the communities of the Senate agreement graph for the 8 last Congresses.

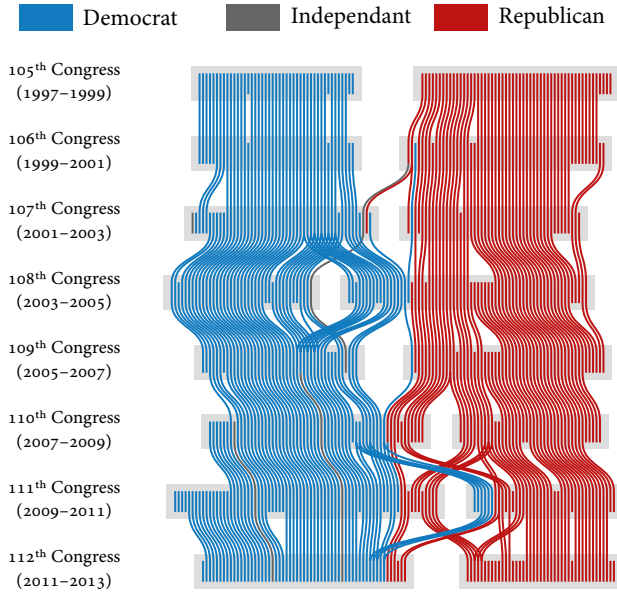


Figure 7.6 Groupflow diagram of the communities in the Senate agreement graph during the last 8 sessions, with the political party breakdown. Each path represents a Senator and its color indicates the Political party it belongs to. Each box is a community and the rows of boxes represent the different Congress sessions. Notice that the overlaps between communities are visible here (*e.g.* where a Senator path forks into two different communities).

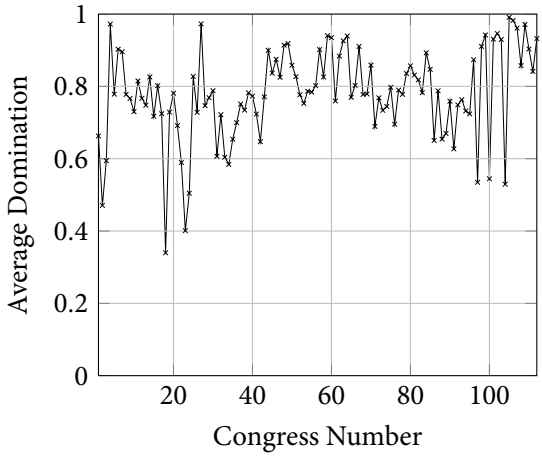


Figure 7.7 Evolution through time of the average domination of the communities by one political party.

Consider for example Zell Miller, although a member of the Democratic Party, he backed the Republic President George W. Bush over the Democratic nominee in the 2004 presidential election. Moreover he has been a frequent critic of the Democratic Party since 2003 and has supported Republican Candidates several times. Moreover, Miller also served as the national co-chair to the campaign of Republican presidential candidate Newt Gingrich in 2012. It is then only natural for him to be grouped with Republicans rather than Democrats in the 106th, 107th and 108th Congresses, which would not be the case had we only taken into account his political party.

During the 108th Congress (2003–2005) there is a large overlap between the two communities. We have found that most of the 18 Democrat Senators in the Republican community were members

of the Democratic Leadership Council, a non-profit corporation which aim was to steer away the Democratic Party from the leftward turn it had taken. This has to be replaced into the context at the time, which was marked by the beginning of the war in Iraq.

Finally, there are some Congresses where there is a large number of Republicans in the otherwise Democrat dominated community. For example, in the 110th Congress (2007–2009) there are 8 Republicans which are more in the Democrat community of agreement. Most of whom are either moderates or closer to the Democrats than to their own party.

Charles Hagel was critic of the Bush Administration which he described as "the lowest in capacity, in capability, in policy, in consensus—almost every area" of any presidency in the last forty years. *George Voinovich* has been known to oppose lowering taxes and frequently joined the Democrats on tax issues. *John Warner* is a moderate Republican and has centrist stances on many issues, to the point that he once faced opposition of other members of his own party when he decided to run for re-election.

Olympia Snowe and *Susan Collins* are the two Republican Senators from Maine who are both regarded to be leading moderates within their party. *John McCain* used to be described as a moderate although he began adopting more orthodox conservative views since his loss in the 2008 Presidential Election. *Gordon Smith* was placed in the exact ideological center of the Senate by a National Journal congressional rating.

Norm Coleman had been a Democrat until he switched parties in December 1996, although still being considered one of the most

liberal Republican in the Senate. Finally, *Arlen Specter* is a Democrat but was a Republican until 2009. Although he switched to the Democratic Party during the 111th Congress, the data shows us that his votes had switched long before that.

By computing the groups of agreement of Senators throughout the history of the United States we have observed that those groups are relatively stable through time. Furthermore we have seen that, although most communities are largely dominated by a party, there are some cases where Senators from other parties are present. By looking into the political profile of those seemingly displaced individuals, we have found that that in we could explain the community by a misalignment between the Senators and their official affiliation. This adds to the validation of C^3 to compute communities of agreement, given that no party information nor temporal data was used to calculate the communities.

Visualization

Springs and Rubber bands

> Graphics must not quote data out of context.
>
> *The Visual Display of Quantitative Information*
> Edward Tufte

Network analysis is a field which usually deals with large amounts of data. Whereas have previously introduced a method to compute overlapping communities in such large networks, there is an essential need to distillate the resulting analysis in a form which is easily apprehensible. This is of the utmost importance when studying overlapping communities which might be present in large quantities and of varying sizes and overlap. According to the saying, a picture is worth a thousand words, consequently it is natural to attempt to visualize the topology of a network as architectured by social communities as it might provide a better understanding of the social structure.

Use the Force, Luke

This past decades, several generic graph drawing algorithms have been proposed, among which the force-directed layout described

by Fruchterman and Reingold[1] is one of the most well known. More generally, force-directed and energy-based algorithms have gained in popularity, among other reasons, due to their simplicity, both conceptual and in terms of implementation, the fact that they usually lead to visually pleasing results, and that there are known optimizations to their efficiency.[2]

In data analysis, graph drawing generally serves one of two purposes related to the intended use of the visualization. *A priori*, it allows to make use of visual intuition in order to provide a basic understanding of the structure of a network. *A posteriori*, once the analysis has been conducted and results have been established, graph drawing is useful to provide an concise representation of what has been uncovered in the data. Those two purposes require different strategies in terms of layout, as the use of an *exploratory* visualization might pollute the information which should be conveyed in a *illustratory* visualization. For example, it has been shown[3] that the perception of node centrality and other social network features are deeply affected by the graph layout.

Given that our aim to obtain a visualization of social communities, its purpose falls in the latter illustratory category. Keeping this objective in mind, one of the less desirable features of

1 Thomas M. J. Fruchterman and Edward M. Reingold. "Graph Drawing by Force-directed Placement". In: *Software: Practice and Experience* 21.11 (Jan. 1997), pp. 1129–1164.

2 Pawel Gajer, Michael Goodrich, and Stephen Kobourov. "A Multi-dimensional Approach to Force-Directed Layouts of Large Graphs". In: *Graph Drawing*. Ed. by Joe Marks. Springer Berlin / Heidelberg, 2001, pp. 211–221.

3 Jim Blythe, Cathleen McGrath, and David Krackhardt. "The effect of graph layout on inference from social network data". In: *Graph Drawing*. Ed. by Franz Brandenburg. Springer Berlin / Heidelberg, 1996, pp. 40–51.

force-directed methods is that visual clusters tend to emerge and at the same time the actual communities – computed using another method – are exploded during the positioning of the nodes.

In terms of clustering, research has mainly focused on the aspect orthogonal to ours. Since force-directed layout algorithms have the tendency to create visual clusters, there has been an interest in methods which enhance that effect, going as far as actually computing clusters based on the visualization.[1]

Others have proposed algorithms which take into account a community structure while drawing the graph. Those methods however enforce constraints which are unreasonable in the visualization of social communities. For example, it has been proposed[2] that communities be computed by maximizing the modularity,[3] the nodes then being placed with a force-directed algorithm which assigns a weaker spring force to edges between communities, and the communities represented by drawing the convex hull containing the nodes of each communities.

It has also been suggested[4] that communities obtained by modularity optimization could be visualized through the use of a multi-hierarchical force-directed layout. The meta-graph of

1 Andreas Noack. "An Energy Model for Visual Graph Clustering". In: *Graph Drawing*. Ed. by Giuseppe Liotta. Springer Berlin / Heidelberg, 2004, pp. 425–436.

2 danah boyd and Jeffrey Heer. "Vizster: Visualizing Online Social Networks". In: *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on* (Sept. 2005), pp. 32–39.

3 Newman and Girvan, op. cit.

4 Amanda L Traud, Christina Frost, Peter J Mucha, and Mason A Porter. "Visualization of communities in networks". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 19.4 (2009), p. 041104.

the communities – where each node represents a community and those are linked if there is an edge from a node in one community to a node in the other one – is drawn using the classical force-directed layout. Only then are the nodes drawn *inside* the community using several heuristics concerning their actual position. In both the previous cases, the use of the modularity leads to a partition of a network into communities, and unfortunately none of those can be extended to visualize overlapping communities.

Another approach has been to create a dummy node for each cluster in the network,[1] add an attractive force between the dummy node and members of the cluster and a high repulsive force between dummy nodes. Although this method yields good results on partitions as it concentrates clusters around a specific location, it has the tendency to stretch and rip apart overlapping communities due to the high repulsion between the dummy attractors.

Recently, there have been proposals to render overlapping communities using force-directed layouts. Although impressive, the method suggested by Simonetto *et al.*[2] is impractical at a large scale. Given that the first pass of the rendering consist in a planarization of the intersection graph, the different communities affect the layout globally. As a consequence, in

1 Ralf Brockenauer and Sabine Cornelsen. "Drawing clusters and hierarchies". In: *Drawing graphs* (2001), pp. 193–227; Yaniv Frishman and Ayellet Tal. "Dynamic drawing of clustered graphs". In: *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on* (2004), pp. 191–198.

2 Paolo Simonetto, David Auber, and Daniel Archambault. "Fully Automatic Visualisation of Overlapping Sets". In: *Computer Graphics Forum* 28.3 (2009), pp. 967–974.

networks containing a large number of overlapping communities the rendering becomes time consuming.

Laying Out

Although unfit to our current purpose, the works described previously however indicate a global direction in the continuity of which we place our contribution. This proposal builds upon the knowledge of communities in order to affect the placement of the nodes and the strength of the force between nodes to add constraints specific to the visualization of communities. First, we introduce a new attractive force which constrains communities in a circle of a given radius and second we present a slight modification to the force of attraction between pairs of connected nodes.

At its most basic level, a force-directed algorithm[1] is a graph drawing algorithm which uses a specialized physics based model to compute the position at which the nodes should be placed. For clarity's sake, from now on we will use the term *node* and write u when referring to the actual node $u \in V$ of a graph $G = (V, E)$, and *particle* u° when referring to the object representing this node in the physical model. To simplify the notations, we will identify the particles to their position vector, allowing us to write $\|u^\circ - v^\circ\|$ to denote the distance between particles u° and v° in the drawing.

In the models we choose to extend, particles evolve in a 2d space and can attract or repel one another – traditionally, there

1 Fruchterman and Reingold, op. cit.; Jacomy Mathieu, Heymann Sebastien, Venturini Tommaso, and Bastian Mathieu. *ForceAtlas2, A Graph Layout Algorithm for Handy Network Visualization*. Tech. rep. Aug. 2011.

is a force of repulsion between all pairs of particles and an attraction between pairs $u^\circ v^\circ$ such that $uv \in E$. In this section, without loss of generality, we will extend any type of force-directed layout which applies an attractive force $\mathcal{F}$ between two particles u° and v° such that $uv \in E$ – *e.g.*, in the classical Fruchterman-Reingold algorithm, the attraction force is defined in analogy to Hooke's law of elasticity: $\mathcal{F}(u^\circ v^\circ) = k\|u^\circ - v^\circ\|$.

Baring in mind that our aim is to represent communities, we complete classical force-directed layouts to take predefined clusters into account. Rather than laying out a graph $G = (V, E)$, we shall render a graph covered in communities $(G = (V, E), C)$ where $C = \{c_i | c_i \subseteq V\}$ is a set of communities.

To do so, we draw our inspiration several earlier works[1] and for each community $c \in C$, we introduce a dummy and virtual particle c° in the visualization – dummy because c° is not associated to any node belonging to the graph, and intangible because we won't actually draw that node, only its position is of interest to us.

We then add another attractive force which only applies to pairs of particles $u^\circ c^\circ$ such that the node u belongs to the community c. Furthering the physical analogy, whereas the force between pairs of particles associated to nodes is modeled after springs, we model the attraction between a node particle and a community particle as a rubber band. If a node $u \in c$, then the force of attraction between u° and c° is null if both particles are close enough, and becomes that of a spring (Fig. 8.1 when the

1 Brockenauer and Cornelsen, op. cit.; Frishman and Tal, op. cit.

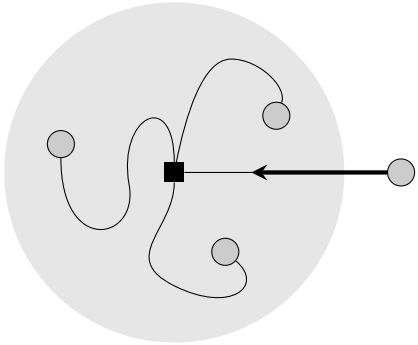


Figure 8.1 Rubber band force: the dummy community particle is represented as a square and each node in the community is represented by a circle. An attractive force is exerted only on nodes outside the constraining radius, here represented by a large circle, whereas all other nodes are connected to the community node by "loose rubber bands".

distance between particles crosses a certain threshold r:

$$\mathcal{R}(u^\circ c^\circ) = \max(0, \|u^\circ - c^\circ\| - r(c))\mathcal{F}(u^\circ c^\circ)$$

In effect, this allows us, when the algorithm converges, to constrain all nodes of a community c in a circle of radius $r(c)$ around the dummy community particle c°. In the following, we set this maximum rubber length to a value which takes into account both the size of the community and its cohesion: $r(c) = \lambda(1 - \log \mathcal{C}(c))\sqrt{|c|}$. The first factor, λ adjusts the scale of the drawing and the second factor, $(1 - \log \mathcal{C}(c))$ implies that highly cohesive communities should be constrained in a tighter circle whereas non cohesive communities can spread much more.

For example, when $C(c) = 1$, which is the case when c is an isolated clique in the network, the radius is equal to $r = \lambda\sqrt{|c|}$. To the contrary, if we assume that c has very low cohesion, then $r \to +\infty$ and the community "dissolves" in the drawing of the network. Finally, we also take into account the size of the community, for the simple reason that for all $u \in c$ there need to have enough place to position the particle u° inside a circle of radius r, therefore we make sure that the area of the bounding circle grows with $|c|$.

Note that what has been described until now does not affect the actual layout algorithm in terms of spatial complexity, given that we only add dummy particles to the rendering and adjust the actual values of the forces exerted on each particle. As such, we can use existing layout algorithms – *eg.* Fruchterman-Reingold or LinLog – to compute the position of our particles without incurring a large cost in terms of space. We have added a dummy node for each community, but in practice the number of communities is bounded by n, and we add $\sum_i |c_i| = \mathcal{O}(n)$ edges between those dummy nodes and the members of their communities, therefore, $n_{\text{rubber}} = \mathcal{O}(n)$ and $m_{\text{rubber}} = \mathcal{O}(m + n)$.

Since the whole idea behind the cohesion is that in graphs akin to social networks, the fundamental structure is shaped by triangles rather than solely by edges, we have chosen to modify the attraction force in order to contract triangles in the drawing[1]. We add a dependency of the attractive force on $\triangle(uv)$, the number of triangles an edge uv belongs to and we shall use

1 Note that this enhancement is generic and can be used even when representing communities which are not computed by maximizing the cohesion

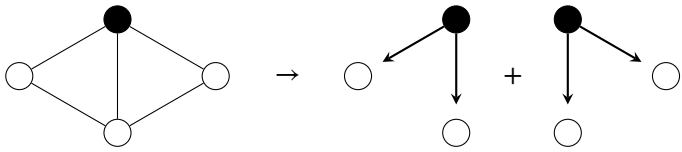


Figure 8.2 Triangle-weighted edges: given the network on the left, the forces exerted on the black node are displayed on the right.

$\triangle_{\max} = \max_{uv \in E} \triangle(uv)$ as a normalizing factor. We then define the triangle weighted attraction force as:

$$\mathcal{T}(u^{\circ}v^{\circ}) = \left((1-\alpha) + \alpha \frac{\triangle(uv)}{\triangle_{\max}} \right) \mathcal{F}(u^{\circ}v^{\circ})$$

where α is a parameter which allows to adjust the relative importance of edges and triangles. The reason why we choose to focus on the number of triangles is that in the most extreme case, when $\alpha = 1$, the force which applies to each node can be decomposed into a sum of forces exerted by triangles as exemplified in Figure 8.2. In turn, this leads to a contraction of all triangles towards their centroid. We have observed that a value of $\alpha = \frac{7}{8}$ gives a strong priority to the contraction of triangles while allowing edges which do not belong to triangles to exert an attraction.

Rendering Communities

Once the position of the nodes have been computed with a force-directed layout algorithm, we visually render the graph. Contrary

to what has been proposed by Simonetto *et al.*,[1] in order to enhance the representation of communities we have chosen to assign one color to each community. Since a node can belong to several communities, we divide each node into colored sectors, one per community it belongs to, which we orient towards the appropriate community. Our rationale is that this approach is more resilient to overlap, because in the case of convex hulls, the visual complexity arises when a large number of communities overlap whereas in our approach the rendering becomes hard to read only when both the number of overlapping communities and of communities each node belong to is large – that is, when the number of sectors is large.

To visually differentiate communities, we assign colors in a manner such that overlapping communities are of sufficiently different colors to be visually distinguished. Using the HSB color model, we generate k colors, where k is the number of communities – at this point no color is actually assigned to communities, but are just generated to be assigned at a further step. To do so, we uniformly choose $\frac{k}{2}$ hues and for each hue we create two colors, one vivid and bright, the other one less saturated and darker. The coordinates of the i^{th} color is given by:

$$col_i = (\frac{\lfloor i/2 \rfloor}{\lfloor k/2 \rfloor} 360^\circ, \frac{3 + (i \bmod 2)}{4}, \frac{3 + (i \bmod 2)}{4})$$

Once the colors are generated, they must be assigned to the communities. We use the perceptual color difference ΔE^*_{94}[2] to

1 Simonetto, Auber, and Archambault, op. cit.

2 M Melgosa, J J Quesada, and E Hita. "Uniformity of some recent color metrics

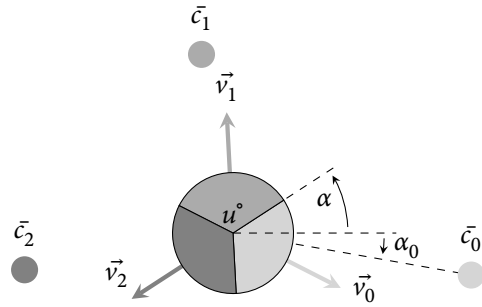
quantitatively capture the visual difference between colors. As such, the problem of assigning the color to communities is a problem of maximizing the minimum color difference between two overlapping communities, which is a generalization of the antibandwith problem.[1] To do so, we initialize all communities as colorless and use a breadth first search heuristic[2] to walk the graph of communities where each community is a node and two communities are linked if they overlap. We assign, to each community we visit, the color which maximizes the minimum distance to its neighbors' colors – formally, $\text{color}(c) = i$ such that $\max_{c' \in \mathcal{N}(c)} \Delta E^*_{94}(col_i, \text{color}(c'))$ is minimal.

Independently from the actual choice of colors, the other aspect of the rendering to take into account is the placement of the different sectors representing communities inside a node. Let u be a node which belongs to k communities, we divide the disc representing the particle u° in k equal sectors. This means there are k regions with a $\frac{2\pi}{k}$ angle and that the i^{th} sector runs from $\alpha + \frac{2i\pi}{k}$ to $\alpha + \frac{2(i+1)\pi}{k}$, where α is a rotation parameter with respect to the coordinate system. In order to maintain a visual coherence, we rotate the node in order to orient the sectors towards the center of mass of each community c_i.

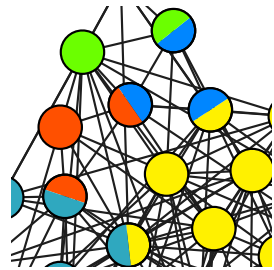
tested with an accurate color-difference tolerance dataset". In: *Appl. Opt.* 33.34 (Dec. 1994), pp. 8069–8077.

1 Yifan Hu, Stephen Kobourov, and Sankar Veeramoni. "On maximum differential graph coloring". In: *Graph Drawing* (2011), pp. 274–286; Manuel Lozano, Abraham Duarte, Francisco Gortázar, and Rafael Martí. "Variable Neighborhood Search with Ejection Chains for the Antibandwidth Problem". In: ().

2 Richa Bansal and Kamal Srivastava. "Memetic algorithm for the antibandwidth



(a) Node orientation



(b) Zoom on a rendering

Figure 8.3 Left: The node u belongs to three communities, and is divided in three equal sectors which we rotate to orient each sector towards the center of mass of the related community. Right: portion of a rendering which displays the correct orientation of nodes.

Each sector i has a bisector $\vec{v}_i$ (Fig. 8) which forms an angle $\alpha + \frac{(2i+1)\pi}{k}$ with the horizontal. Let us write α_i the angle formed by the horizontal and the segment joining the center of the disc u° and the center of mass $\bar{c}_i$ of c. To orient the sectors towards the

maximization problem". In: *Journal of Heuristics* 17.1 (2011), pp. 39–60.

appropriate center of masses, we choose the value of $\alpha = \frac{1}{k}\sum_i \alpha_i - \frac{2i\pi}{k}$ that minimizes the sum of squares of angular distances – that is such that $\sum_i(\alpha + \frac{(2i+1)\pi}{k} - \alpha_i)^2$ is minimal (an example of the result is given in Figure 8.3(b)).

Visualization and Benchmarks

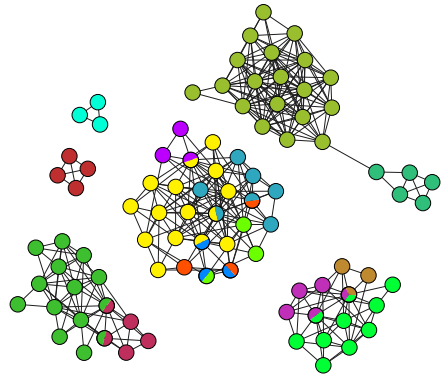
To illustrate and validate our approach, we have tested it using a set of ego-networks collected from Facebook – we call ego-network a subgraph restricted to the friends of a given individual, Ego, and where edges are present between two of Ego's friends if those two are friends together. Note that Ego himself is not present in his ego-network since it would bring no information as he would be connected to everyone.

Our enhancements, triangle weighted edges and rubber band force – which we will collectively refer to as "RubberBand" or "RB" – were implemented in Gephi[1], a network visualization software. Specifically, we have extended two force-directed algorithms: ForceAtlas2[2] which is an implementation of an optimized variant of Fruchterman-Reingold, and LinLog.[3] Our test case consists of 40 ego-networks of various sizes, ranging from 20 to 1066 nodes. We have applied the C^3 community detection algorithm to each of those ego-networks and obtained 2675 communities – between 4 and 215 per graph – containing from 3 to 127 nodes.

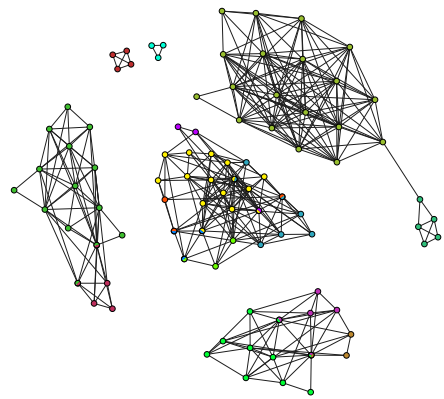
1 http://gephi.org
2 Mathieu, Sebastien, Tommaso, and Mathieu, op. cit.
3 Noack, op. cit.

Figure 8.4 Visual Comparison of Layouts Rendering.

Classical Layouts.



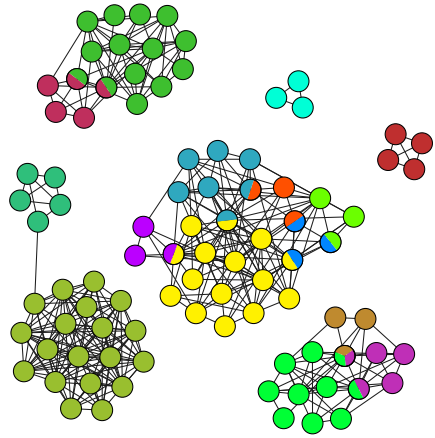
(a) Classic ForceAtlas2



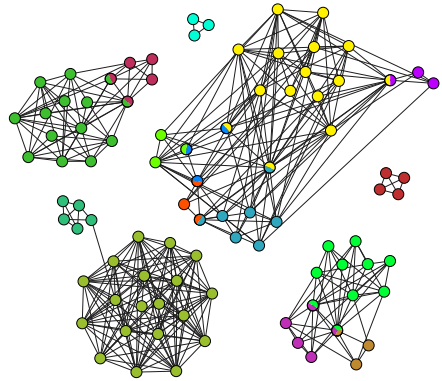
(b) Classic LinLog

Figure 8.4 Visual Comparison of Layouts Rendering.

RubberBand Layouts.



(c) Rubber ForceAtlas2



(d) Rubber LinLog

We have represented on Figure 8.4, using the sector coloring described in the previous section, the output of ForceAtlas2 (Fig. 8.4(a)), RB-ForceAtlas2 (Fig. 8.4(c)), LinLog (Fig. 8.4(b)) and RB-LinLog (Fig. 8.4(d)) for a graph of size 91 containing 15 communities. All four rendering have been scaled to fit into a square frame of same size, and thus, given that they are more compact, the RB-layouts are more readable than their classic counterparts.

Moreover, when looking specifically at the color of the nodes, one notices that whereas in the case of RubberBand all nodes in a community are close to one another, the communities at the center in ForceAtlas2, for example, are mixed up.

In order to assess more formally the contribution of the rubber band force to the visual rendering of communities, we conducted two quantitative benchmarks. Before going into the details, let us first recall that the goal of the enhancement is to allow a better visual depiction of pre-existing communities.

In other words, what we aim to observe are compact regions containing mainly nodes from one community only. We introduce two metrics on communities to capture the differences between using a classic force-directed layout and a force-directed layout enhanced with rubber band forces and triangle weighted edges.

Let c be a community, we will call $H(c)$ the convex hull, in the computed layout, of all particles u° associated with a node of c. To capture the extent to which a community is rendered compactly, we introduce the notion of spatial density of community, which

we define, depending on the layout $\mathcal{L}$, as:

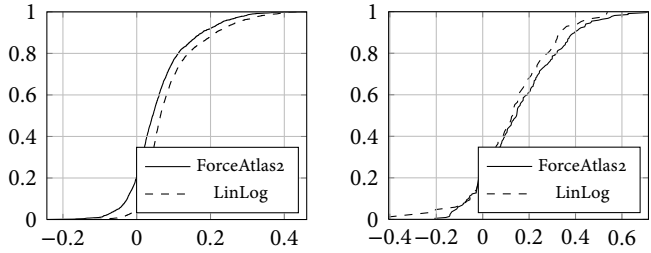
$$D_{\mathcal{L}}(c) = \frac{\text{number of nodes in } c}{\text{area}(H(c))/\text{area}_{\text{node}}}$$

Where $\text{area}_{\text{node}}$ is the area occupied by one particle in the drawing – this allows us to reason with a unitless and scale independent density. Note that this metric is related the *density metric* introduced by Frishman *et al.*[1] but more precise given that we take into account the area of the convex hull rather than that of the bounding box.

We introduce $\delta_{D_{\mathcal{L}}} = D_{\text{RB-}\mathcal{L}}(c) - D_{\mathcal{L}}(c)$, the difference of densities obtained by each communities in both a layout $\mathcal{L}$ and the same layout using RubberBand. We can see on Figure 8.5(a) that in the case of ForceAtlas2, 80% of communities are denser when rendered with RubberBand than without. This is even more pronounced when looking at the results obtained by LinLog, which has a natural tendency to spread nodes apart, as 95% of communities are denser when using the RubberBand than when using the classic LinLog. Which leads us to conclude that communities are more compact when using RubberBand than without, yet it is not sufficient to assess that we have attained a good layout for the communities.

If we were to use the density as sole indicator of the efficiency of a layout algorithm in representing communities, the solution yielding the smallest would be to pack all nodes in the same spot, which is all but what we wish to observe. Therefore, we introduce

1 Frishman and Tal, op. cit.



(a) CDF of δ_D (b) CDF of δ_K

Figure 8.5 Cumulative distribution of the difference (left) in densities between communities using layouts with and with RubberBand and (right) in consistencies between non optimally consistent communities using layouts with and with RubberBand.

a second metric which rates the visual consistency of each community, the idea being that there should be a penalty given to a rendering which mixes up communities. Formally, we shall define the visual consistency as:

$$K_{\mathcal{L}}(c) = \frac{\text{number of nodes in } c}{\text{number of nodes rendered in } H(c)}$$

The rationale is that in the case of of consistent communities, there should be few or no intruders in the convex hull – *i.e.* $K = 1$ – and the more nodes which do not belong to c are placed in $H(c)$, the less consistent the community will be.

First off, it is important to note that using ForceAtlas2, there were 66.1% of all communities which had a consistency $K = 1$ both with and without using RubberBand, and similarly, in the case of the LinLog layout, there were 79.8% such communities.

Given that in this case there is no room for optimization we conclude that in the majority of cases the addition of the rubber force and triangle weighted edges is at least as good as using one of the classical layout algorithms. For the sake of exhaustivity, there are a negligible 2.3% of communities using ForceAtlas2 and 1.5% using LinLog which have a consistency $K = 1$ without RubberBand and a consistency $K < 1$ with RubberBand, but in those cases we consistently observe that this value of K is greater than 0.75.

Where it becomes interesting though is in the cases where either consistency is not null. On Figure 8.5(b) we have represented the cumulative distribution of the differences $\delta_{K_{\mathcal{L}}}(c) = K_{\text{RB-}\mathcal{L}}(c) - K_{\mathcal{L}}(c)$ of consistencies of communities between the layouts with and without RubberBand. Observe how in 75% of cases where the communities are more consistent – that is, $\delta_{K_{\mathcal{L}}}(c) > 0$ – in RB-ForceAtlas2 than in ForceAtlas2, and in 70% of cases RB-LinLog leads to more consistent communities than when using only LinLog.

In conclusion, most of the times a RubberBand enhanced layout algorithm is as good as when not using RubberBand – that is, in those cases it is actually not possible to obtain a better consistency. And in cases where both values are not optimal, then adding RubberBand to the layout increases the consistency of communities in a large majority of cases.

We have built an enhancement to classical force-directed layout algorithms, called RubberBand, which mainly consists of two additions to those classical algorithms. To each community we add a dummy

attractor node which exerts a rubber band like force to the members of the community. The use of a rubber band rather than a spring force constrains the community into a circle of parametrized radius centered around the dummy attractor. And second, we take into account the number of triangles an edge belongs to in order to adjust the attraction to contract triangles. Using benchmarks on a set of real data, we observed that RubberBand yields drawings where communities are represented more compactly and more consistently than when using the same layout algorithms without the RubberBand enhancement. Therefore we conclude that RubberBand achieves the desired goal of representing pre-existing communities.

Personality and Structure

Psychological Aspects of Social Communities

Through others, we become ourselves.
The Genesis of Higher Mental Functions
Lev Vygotsky

Until now, we have mostly focused on the structure of the network without taking into account the characteristics of the individual involved – a usual approximation in social network analysis. In this chapter, we shall attempt to take those personal characteristics into account and identify how individual differences in psychological traits affect the community structure of social networks. Rather than neglecting either structural or psychological properties of an individual, we seek to understand how social network topology is shaped by the psychological attributes of interacting individuals. By doing so, we take an individualized approach to the study of social networks and view the actor as an individual who actively transforms the structure of his or her social network differently depending on his own specific properties.

When Sociology meets Psychology

In the analysis of social networks, particular attention has been dedicated to the structural properties of the direct neighborhood of an individual. This ego-centered approach has been used broadly in psychology and sociology to help at better understanding the relationship between an individual and its proximate social circle, and how individuals are integrated in social life.[1] The position of a person in a network and, complementary, the shape of its ego-network is the source of its social capital. By definition, social capital is considered to be "the sum of the resources, actual or virtual, that accrue to an individual or group by virtue of possessing a durable network of more or less institutionalized relationships of mutual acquaintances and recognition".[2]

A dense, interconnected network of often strong ties is associated to the notion of bonding social, enabling the flow of information within the network and containing an element of trust.[3] In comparison, open networks, containing many intransitive triads, are an indicator of bridging social capital, as an individual bridges structural holes between disconnected others, thereby facilitating knowledge sharing across the system.[4]

An important aspect missing in the latter structural studies is a

1 Stanley Wasserman and Katherine Faust. *Social Network Analysis*. Cambridge University Press. Nov. 1994.

2 Pierre Bourdieu and Loíc J D Wacquant. *An invitation to reflexive sociology*. University Of Chicago Press, July 1992.

3 Ronald S Burt. *Brokerage and Closure*. An Introduction to Social Capital. Oxford University Press, Aug. 2005.

4 Ibid.

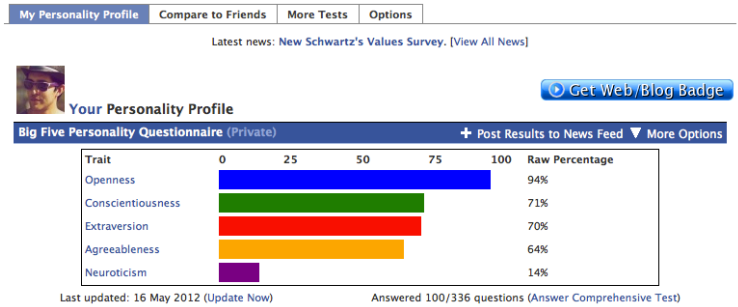


Figure 9.1 Screenshot of part of the myPersonality application.

characterization of the variability and differences among individuals, and the effect of non-structural attributes on link formation. A well-known example is homophily[1] stating that similarity, e.g. in terms of status or interests, fosters connection, as similar people tend to select each other, communicate more frequently and develop stronger social interactions.

Since our interest lie in the study of the way psychological traits are linked to topological features of the social network, we shall analyze the results of an online psychological test called myPersonality, in relation to the social entourage of the test subjects on Facebook. myPersonality is a Facebook application which allows users to take a variety of personality and ability tests (Fig. 9.1). Users also have the possibility to opt in and give their consent

1 Miller McPherson and Lynn Smith-Lovin. "Birds of a feather: Homophily in social networks". In: *Annual review of sociology* 27 (2001), pp. 415–444.

to share their personality scores and their Facebook profile information for scientific purposes. It should be noted that the psychological and structural data come from two different sources – respectively myPersonality and Facebook – and as a consequence are *a priori* independent one from the other.

The application undertakes a series of measures to avoid that users respond in a careless or mischievous way,[1] and thus to assure the highest quality of its database, *e.g.* by removing unreliable results through numerous validity tests. It has been shown that the quality of the responses is at least as high as in traditional pen-and-paper studies, with the significant advantage of reaching a much broader and less biased audience. The benefit in using the myPersonality data is that is allows to tap in a unique source which contains both psychological traits of the subjects and link this psychological profile to social information – list of friends and friendships between friends – extracted from their social graph on Facebook.

Personality is measured by the so-called five-factor model of personality,[2] which associates to each individual five scores corresponding to five main personality dimensions. Each dimension, labeled as OCEAN, can be summarized as follows:

– Openness, for spontaneity and adventurousness, denotes an appreciation for emotion, a sensitivity to beauty and intellectual curiosity;

1 Tom Buchanan and John L Smith. "Using the Internet for psychological research: Personality testing on the World Wide Web". In: *British Journal of Psychology* 90.1 (1999), pp. 125–144.

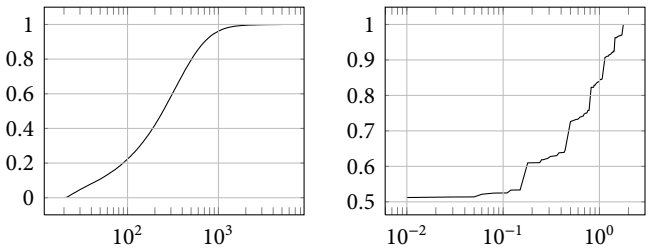
2 Paul T Costa and Robert R McCrae. *NEO Personality Inventory Revised NEO-PI-R Test Manual*. SAGE Publications. May 2005.

- Conscientiousness, for ambition and persistence, denotes a tendency to act dutifully and a planned rather than spontaneous behavior;
- Extraversion, for sociability and excitement seeking, denotes an energetic and spontaneous personality and the tendency to seek stimulation in company of others;
- Agreeableness, for trustingness and altruism, denotes a tendency to be compassionate an cooperative towards others;
- Neuroticism, for emotional liability and impulsiveness, denotes a personality prone to experiencing negative emotions easily, *e.g.* anger, anxiety and depression.

The *social* data – in terms of network and structure – comes from Facebook, and as such we have, for each subject, both the list of their friends and the information about pairs of those who are also friends on Facebook. We adopt an egocentered approach and, rather than considering the impact of all individuals on the whole Facebook social network, we will focus on the way the individual shape the social structure of their ego-network.[1]

The ego-network approach has a remarkable property, as there is a direct correspondence between classical network metrics on the original network and the ego-network. Obviously, the degree d – or number of friends – of a subject in the original network is equal to the size of the ego-network. More interestingly, the clustering coefficient of the subject in the original network is equal to the density of his ego-network. We will equivalently mention

1 We recall that we defined, in a previous chapter, that the ego-network of a subject is the subgraph containing only Ego's friends, excluding Ego.



(a) Degrees CDF (b) Extraversions CDF

Figure 9.2 Cumulative distributions of subjects' (*left*) degrees and (*right*) extraversions.

the *degree of a subject* and the *size of the ego-network*, and similarly the *clustering coefficient of a subject* and the *density of the ego-network*.

Our final dataset consists of a sample of 44,096 Facebook users who have taken the big five personality test. For each of these subjects, myPersonality provides their five personality traits, and thanks to Facebook we have their age, gender and ego-network. We will focus on the users whose number of friends on Facebook is comprised between 50 and 2000 (excluding 35 users with degree greater than 2000 and 3259 users with degree smaller than 50). The final sample contains 49,623 users, whose cumulative degree distribution is represented in Figure 9.2(a). Given that in the following we shall mainly focus on the trait of extraversion, we have represented on Figure 9.2(b) the cumulative distribution of our sample subjects' extraversion.

Having this data at hand, we are able to study the relationship between personality and the structural properties of

ego-networks. In that respect, previous research has consistently shown that extraversion is associated with the size of the ego-network and with greater social status.[1] Other dimensions have also been argued to have an effect on social network topology,[2] but findings tend to be inconsistent in the literature, and no significant correlation has been found in a recent study on a large sample of users of myPersonality.[3]

In studies interested in other aspects of the ego-network, it has been observed that extroverts are not emotionally closer to individuals in their network,[4] despite an increased size. It has also been shown that self-monitors, the chameleons of the social, are more likely to have a high centrality.[5] Brokerage also appears to be related to personality. People whose networks bridge structural

1 Cameron Anderson, Oliver P John, Dacher Keltner, and Ann M Kring. "Who attains social status? Effects of personality and physical attractiveness in social groups." In: *Journal of Personality and Social Psychology* 81.1 (2001), p. 116; Diane S. Berry, Julie K. Willingham, and Christine A. Thayer. "Affect and personality as predictors of conflict and closeness in young adults' friendships". In: *Journal of Research in Personality* 34.1 (2000), pp. 84–107; Rhonda Swickert, Christina J Rosentreter, James B Hittner, and Jane E Mushrush. "Extraversion, social support processes, and stress". In: *Personality and Individual Differences* 32.5 (2002), pp. 877–891; Daniele Quercia, Renaud Lambiotte, Michal Kosinski, David Stillwell, and Jon Crowcroft. "The personality of popular Facebook users". In: *Proc. ACM Conf. Comput. Support. Cooperat. Work* (2012).

2 Lauri A Jensen Campbell and William G Graziano. "Agreeableness as a moderator of interpersonal conflict". In: *Journal of personality* 69.2 (2001), pp. 323–362.

3 Quercia, Lambiotte, Kosinski, Stillwell, and Crowcroft, op. cit.

4 Thomas V. Pollet, Sam G. B. Roberts, and Robin I. M. Dunbar. "Extraverts Have Larger Social Network Layers". In: *Journal of Individual Differences* (2011).

5 Martin Kilduff and David Krackhardt. "Bringing the individual back in: A structural analysis of the internal market for reputation in organizations". In: *Academy of Management Journal* (1994), pp. 87–108.

holes are more likely to have an entrepreneurial personality.[1] In another study, it was shown that extraversion is positively associated with closed triads of strong ties and neuroticism with closed triads of weak ties.[2]

More generally, the five-factor model has been shown to predict a broad range of real-world behavior,[3] for instance how marriages turn out and people's taste in movies.[4] The five personality factors also relate to people's behavior in a broad variety of social contexts. It is likely that they predispose people's propensity to form more or fewer social ties, and may be related to the extent to which others form relationships with the focal actor. For instance, extroverts are expected to approach others more easily and engage in more social interaction. Moreover, the existence of different types of structural configurations has also been proposed, each associated to the social and psychological characteristics of an individual: people embedded in dense networks, people having several subsets of alters, etc..[5]

Let us now apply C^3 to the ego-networks we have previously described in order to obtain a covering of each ego-network with

1 Ronald S Burt, Joseph E Jannotta, and James T Mahoney. "Personality correlates of structural holes". In: *Social Networks* 20 (1998), pp. 63–87.

2 Yuval Kalish and Garry Robins. "Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure". In: *Social Networks* (2006).

3 Daniel Nettle. *Personality: what makes you the way you are*. Oxford University Press. 2007.

4 Olivia Chausson. "Assessing the impact of gender and personality on film preferences". In: *Cambridge University* ().

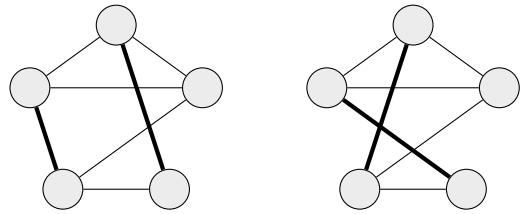
5 Barry Wellman and University of Toronto. Centre for Urban and Community Studies. *Studying personal communities in East York*. Sage. Beverly Hills, 1982.

overlapping communities for each Ego. Note that, given the definition of the cohesion, isolated nodes and more generally nodes which do not belong to a triangle are not considered to be part of any communities As a consequence, contrary to other community detection algorithms C^3 does not force people into communities and therefore the covering might not be complete.

As we wish to study the links between Ego's psycho-social characteristics and the community structure of his ego-network, we will restrict the set of users to those whose surrounding topology consist at least of 50% of friends present in one or more community (24,285 subjects). By using C^3 on our final data set, we have obtained 974,677 communities.

In order to compare our findings to a baseline, we introduce, for each subject, a random null model ego-network. Since we wish to study the impact of the psychological traits on the community structure of the ego-network, we need to control for other topological factors such as size and density of the ego-network. In order to do so, we will construct the null model ego-network G_R by randomly rewiring edges of the original ego-network G in the following manner: choose two distinct edges randomly, such that the ends of those two edges are four distinct nodes, and swap the ends of those two edges, as illustrated on Figure 9.3. We obtain the null model ego-network G_R after repeating this procedure m times, where m is the number of edges in G. During this random rewiring, each edge has been rewired on average twice, guaranteeing a the randomness of the graph G_R.

Note that at each step of the rewiring, no node or edge are added nor deleted, which guarantees that the null model ego-network has same size and density that the original



(a) before rewiring (b) after rewiring

Figure 9.3 Illustration of the rewiring process: first two edges are chosen and marked with thick lines and we then swap the ends of the two chosen edges.

ego-network. Furthermore, because at each step we swap the ends of edges, the degree of each node remains constant, and therefore the distribution of degrees of both G and G_R are identical. This means that the null model would be the ego-network of an individual which would have same degree and clustering coefficient that the original subject.

We have applied C^3 to each null model ego-network and obtained 1,709,883 communities. In the following section, where relevant, we will apply the same computations both to the original ego-network and the null model in order to highlight the effects due to the deeper community structure.

Psychology of Structural Features

As stated previously, it has been observed repeatedly that there is a linear Pearson correlation between the number of social connexions maintained by an individual and his extraversion. Note that we will consider the logarithm $\log(d)$ of the number of contacts (or

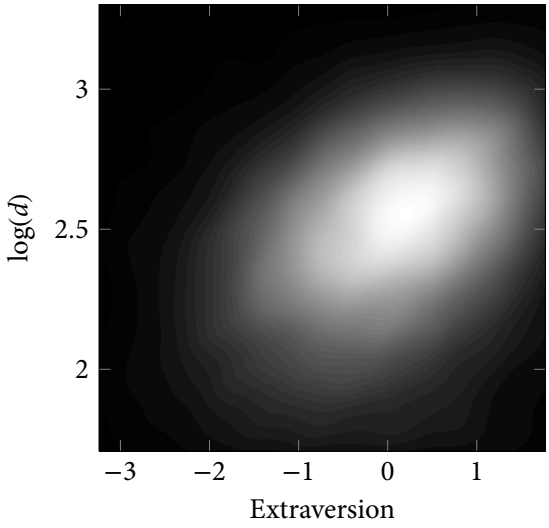
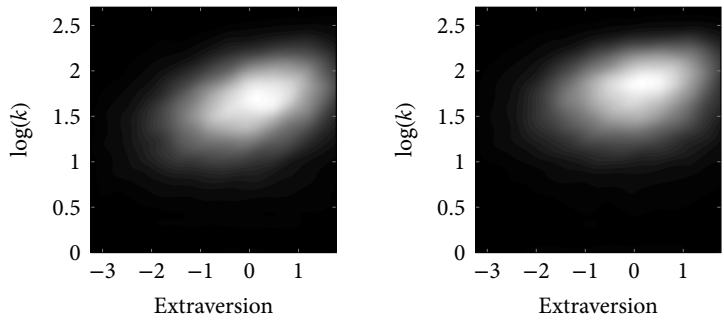


Figure 9.4 Kernel density estimation of degree – number of friends – as a function of extraversion.

degree) rather than the actual number of friends due to the high variability of its value. On Figure 9.4, we have represented a kernel density estimation of $\log(d)$ as a function of the extraversion. As expected, we observe a moderate correlation $r = 0.301$ (p-value $< 10^{-100}$) which indicates that the more extroverted users tend to add more friends on Facebook.

Given this correlation between number of friends and extraversion, it is only natural to look at the link between the number of communities in a subject's ego-network and extraversion. We also observe a moderate correlation $r = 0.293$ (p-value $< 10^{-100}$) between extraversion and number of communities (Fig. 9.5(a)). This result holds in part on the



(a) Original Data (b) Null Model

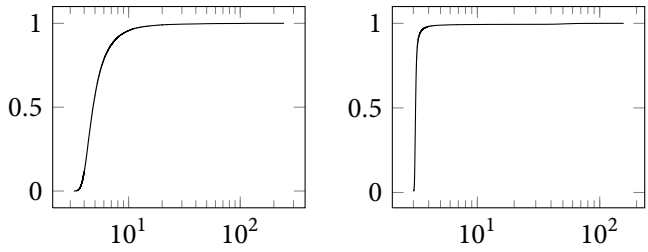
Figure 9.5 Kernel density estimations of the number of communities in the ego-network as a function of extraversion in the (*left*) original data and (*right*) null model.

null model where we observe a somehow smaller correlation $r = 0.196$ (p-value 1.2×10^{-207}) between extraversion and number of communities (Fig. 9.5(b)).

A possible explanation to this observation is that there is a strong correlation $r = 0.894$ (p-value $< 10^{-100}$) between the number of communities $\log(k)$ and his number of friends $\log(d)$. Interestingly, on the null model, we observe a smaller correlation $r = 0.788$ (p-value $< 10^{-100}$) between these two quantities.

It is important to point out that all the correlations are less important in the null model than in the original data. This leads us to conclude that there is a direct contribution of the actual community structure in the original ego-networks to correlations. That is, part of the correlation between extraversion and communities which cannot only be explained by the degree alone.

As expected, we have observed that more extroverted subjects



(a) Original Data (b) Null Model

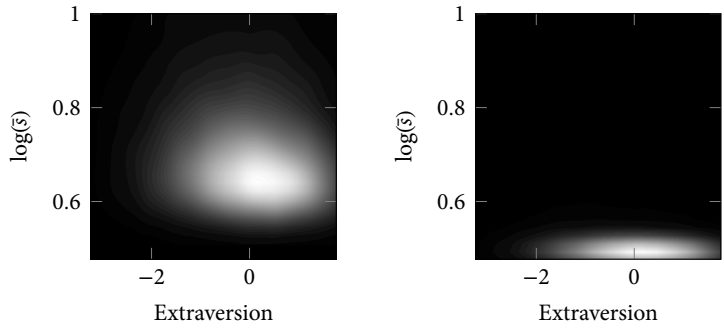
Figure 9.6 Cumulative distributions of average community size in (*left*) original data and (*right*) null model.

tend to have more friends on Facebook. More interestingly, we have also observed that the friends of more extroverted subjects are split across more communities. Our data suggest data suggest that extroverted people not only maintain more social relationships but also interact with a larger number of social groups.

Since there is a correlation between number of communities and extraversion, it is legitimate to look at the relationship between the size of those communities and extraversion. For each user, we will define the average size $\bar{s}$ of his communities (c) as:

$$\bar{s} = \frac{\sum_c C(c)|c|}{\sum_c C(c)}$$

We weight the sizes by cohesion in order to give more importance to good communities. The cumulative distribution of average sizes represented on Figure 9.6(a) shows that most communities have a rather small size (50% have $\bar{s} \leq 4$ and 95% have $\bar{s} \leq 10$).



(a) Original Data (b) Null Model

Figure 9.7 Kernel density estimations of average size weighted by cohesion as a function of extraversion in the (*left*) original data and (*right*) null model.

In the case of the null model, the average size of communities is evan smaller, which contrast with the original data, as 95% of the null model ego-networks have an average community size $\simeq 3$ (Fig. 9.6(b)).

There is a small negative correlation $r = -0.11$ (p-value 1.57×10^{-66}) between $\log(\bar{s})$ and extraversion, which shows that more extroverted subjects have smaller communities, and conversely (Fig. 9.7(a)). Strikingly, this has to be contrasted with the absence of correlation $r = -0.0574$ (p-value 4.16×10^{-19}) between average size and extraversion in the case of the null model (Fig. 9.7(b)). From this we conclude that correlations observed in the original data are mainly due to the community structure.

We have observed that introverted subjects tend to be in larger

groups and that extroverts tend to be in smaller groups. This observation suggests that introverts are more naturally inclined to blend into larger groups, rather smaller groups, to avoid being the center of attention. On the contrary, extroverts might prefer being part of a large number of smaller groups in order to have more chance of attracting attention.

Until now, we have analyzed the community structure in terms of number of communities of the subjects and in terms of number of people in those communities, that is the average size of those communities. We now focus on the the average cohesion – weighted by size – of a user's communities, which is defined as:

$$\bar{C} = \frac{\sum_c C(c)|c|}{\sum_c |c|}$$

The average cohesion $\bar{C}$ captures, for each user, to which extent he is part of socially cohesive environments. Interestingly, we have found that there is no significant correlation between this quantity and any of the big five personality traits. However, we have observed a small correlation ($r = 0.123$, p-value 1.96×10^{-82}) between the standard deviation of the cohesion σ_C and extraversion.

$$\sigma_C = \sqrt{\frac{\sum_c |c|(C(c) - \bar{C})^2}{\sum_c |c|}}$$

In the null model, this correlation is not present ($r = 0.0676$ p-value 7.05×10^{-26}) which suggests a relation between extraversion and the heterogeneity of social communities.

This correlation between cohesion variability and extraversion is related to the fact that that subjects who belong on average to

larger communities tend to evolve in social groups of similar cohesion – there is a strong negative correlation $r = -0.806$ (p-value $< 10^{-100}$) between σ_C and $\log(\bar{C})$. One mechanism may explain such a correlation: the number of large communities is much lower, and those are typically of low cohesion, given that a number of inbound triangles proportional to the cube of the size of the community would be needed in order to maintain a high cohesion.

We understand the positive correlation between the cohesion's standard deviation and extraversion in terms of social adaptability, which is key in the definition of extraversion: extroverts are members of different communities with highly varying cohesion, and are thus members of social communities which can be tight groups of close friends as well as more sparse communities of more distant acquaintances. On the other hand, as we have seen before, introverts tend to hide in larger groups and those groups tend to have an average cohesion, i.e. introverts tend to lack highly cohesive communities which would increase the variability of cohesion.

Another aspect of communities is the amount of overlap. As described earlier, one of the strengths of C^3 is that it computes communities without imposing the constraint that a subject belongs to one and only one community. As a matter of fact, C^3 does not impose the constraint that a subject belongs to at least one community either: a node might be present in 0, 1 or more communities. For the sake of clarity, nodes of an ego-network which are in at least one community will be referred to as *covered*.

A simple way to capture the notion that some of the covered

nodes are in more than one community is to look at the partition ratio $p = \frac{|\bigcup c|}{\sum |c|}$.

This partition ratio quantifies the extent to which the communities of an ego-network are disjoint from one another. It is equal to 1 when all covered nodes are exactly in one community, that is when covered nodes are partitioned into communities, and it decreases as more nodes are present in several communities, *i.e.* when the overlap increases.

We observe a small negative correlation $r = -0.132$ (p-value 1.56×10^{-95}) between δ and extraversion in the original data whereas the correlation $r = -0.0857$ (p-value 1.03×10^{-40}) is negligible on the null model.

We observe a negative correlation between extraversion and the partition ratio, which implies that there is a link between the compartmentalization of the ego-network and the subject's extraversion. More extroverted subjects tend to be in groups which are intricately linked to each other whereas less extroverted subjects tend to be in more distinct and separate social groups. This observation is compatible with the hypothesis that extroverts act as bridges between communities and introduce individuals form one community to those in another one.

Finally we explore the effect of age on network topology. Although not a psychological trait, age is part of the identity of the subjects and as such has an impact on the structure of their ego-network. We observe, as it was the case in,[1] that there is a negative correlation $r = -0.194$ (p-value 8.51×10^{-193}) between age and

1 Quercia, Lambiotte, Kosinski, Stillwell, and Crowcroft, op. cit.

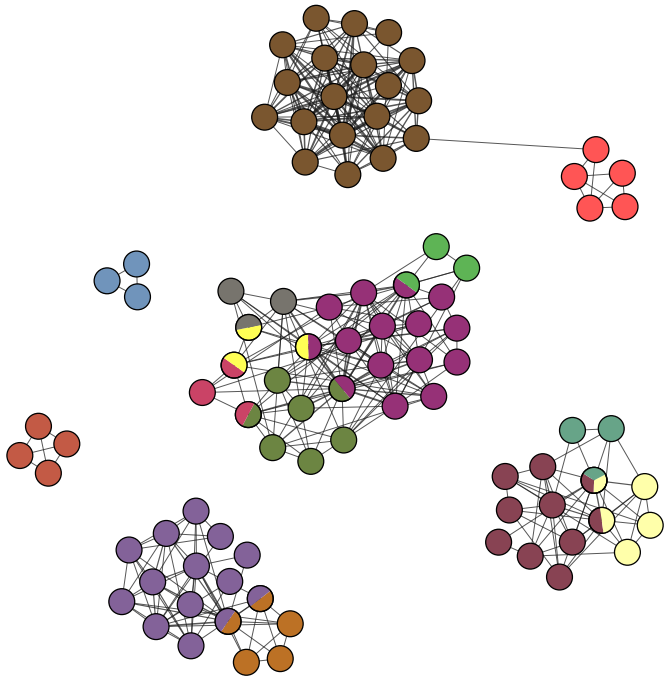
degree, which tends to indicate that the elder have a smaller social neighborhood on Facebook whereas the younger have more friends.

Our community based approach reveals a moderate correlation $r = 0.271$ (p-value $< 10^{-100}$) between age and the average cohesion $\bar{C}$ which is not present in the null model, where the correlation is $r = 0.0843$ (p-value 2.37×10^{-37}). This observation suggests that older individuals belong to denser communities on Facebook, whereas the younger are part of sparser ones.

Finally, let us mention a small correlation $r = 0.171$ (p-value 4×10^{-150}) between age and the Conscientiousness factor, which indicates that the older subjects exhibit less spontaneous behavior than younger ones. Intriguingly, though, we do not observe any correlation between degree and conscientiousness, nor between average cohesion and conscientiousness.

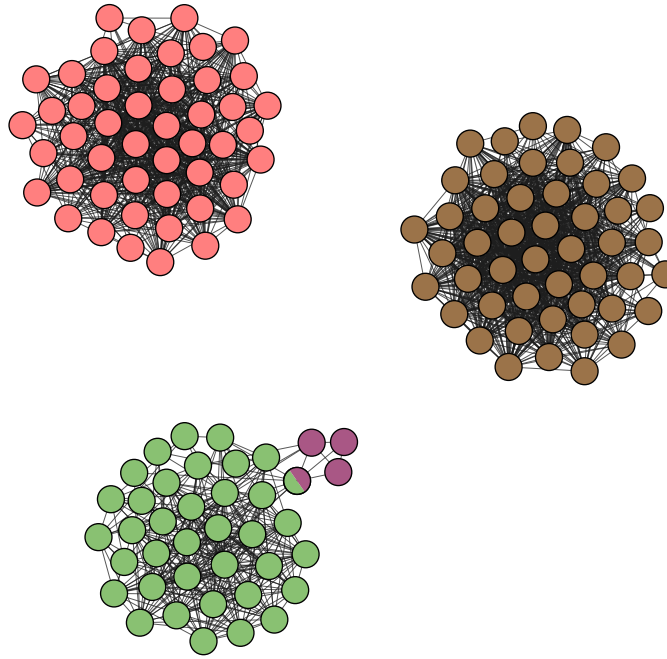
We have observed an impact of age on the structural properties of ego-networks, as the older a subject gets the less friends he has on Facebook. This observation can be explained by the fact that younger subjects are more active on Facebook and tend to add more friends than older ones. Our analysis also reveals that older subjects tend to be part of more cohesive groups than the younger ones, which might come from the fact that younger subjects are less careful before adding someone as a friend on Facebook.

We now focus on visualizations of ego-networks in order to illustrate the quantitative findings of the previous sections. Figure 9.8(a) shows a RubberBand drawing of the ego-network of a user, which we will call A, with high extraversion (ext = 1.33). Figure 9.8(b) shows the ego-network of user B who is more



(a) User A, 26 years old: high extraversion (ext = 1.33), 101 friends of which 91 are split across 15 communities of size varying between 3 and 19, and average cohesion $\bar{C} = 0.46$.

Figure 9.8 Examples of two ego-networks of subjects with different psychological traits and structural features.



(b) User B, 19 years old: low extraversion (ext $= -1.21$), 145 friends of which 136 are split across 4 communities of size 4, 37, 48 and 48, and average cohesion $\bar{C} = 0.31$.

Figure 9.8 Examples of two ego-networks of subjects with different psychological traits and structural features (cont.).

introvert (ext $= -1.21$). Let us recall that an ego-network is the subgraph containing the neighborhood of Ego, and that Ego and its links are not represented. Each circle in the visualization represents a friend of the subject and their community is represented by their color. If a friend belongs to several communities, the circle is divided into equal regions and each slice is colored with the color of a community. This procedure makes the overlapping community structure of the ego-network immediately visible.

Notice the differences between both networks. First, user A has a degree ($d_A = 101$) which is slightly less than B ($d_B = 145$) even if both numbers have the same order of magnitude. More interestingly, the organization of those friends into communities is strikingly different. B has four communities, three of which contain more than 35 friends. Moreover, two of those groups are totally isolated from the others. On the other hand, A exhibits 15 communities which are much more interconnected and of smaller size. This figure illustrates the different ways in which personality impacts the structure of the ego-network. More specifically, it shows that a user with low extraversion, such as B, is part of a few large cohesive groups which are compartementalized, whereas a user with high extraversion, such as A, evolves in more different social groups, of smaller sizes, varying cohesion and overlapping one another.

Crossing two datasets, one containing psychological traits of subjects and the other consisting of social ego-networks extracted from Facebook, we have shown several ways in which the personality of users

may affect there social entourage. After verifying that extroverts tend to have more friends than introverts, we have exhibited that extroverts' networks contain more communities than those of introverts. Moreover, we have observed that introverts tend to hide into larger communities, which we hypothesize is to avoid being the center of attention. We have also noticed that extroverts belong to communities of varying cohesion, linked to their greater social adaptability, and that those are less compartmentalized, a sign they might act as bridges between social groups, introducing unrelated friends to one another.

Epilogue

I was born not knowing and have had
only a little time to change that here
and there.
Letter to Armando Garcia J
Richard Feynman

The notion of community, although intuitive, has evaded a clear and quantitative definition for decades. This thesis introduces the cohesion, a network statistic built upon widely accepted sociological notions and aimed to capture the extent to which a set of nodes is a community. Following a large-scale experiment, the use of the cohesion was validated when it was observed that it correlated strongly with the subjective perception of communities. This work contributes at long last the aforementioned elusive definition: *a community is a strongly cohesive group of people*, with respect to the underlying social network.

Roughly the first half of this work lies at the intersection of Computer Science and Sociology and is devoted to theoretical aspects of the cohesion. From the distillation of its heritage, to its construction and finally to its mathematical and algorithmic analysis which lead to the proof that finding maximally cohesive groups is an $\mathcal{NP}$-hard problem.

On the other hand, the second half is is dedicated to concrete applications of the cohesion, be it in the form of the

C^3 community detection algorithm, the application of the algorithm, or its use to understand the links between network and personality. Although those applications were envisioned out of a desire to justify the use of the cohesion, they grew into solid results which can stand by themselves. The C^3 algorithm has proven repeatedly that it could extract meaningful communities from social networks and the RubberBand visualization algorithm was helpful in the graphical interpretation of those communities.

Branching out from Computer Science and Sociology, the use of C^3 on the United States Senate agreement groups has not only been useful in the justification of the algorithm but has helped understand the continuity and stability of political parties. Extending in another direction, by exhibiting previously unquantified structural differences between extroverts and introverts – such as the topological proof that introverts tend to find refuge in larger groups, or that extroverts act like bridges between communities – it has brought valuable information on the intricacies between psychology and sociology.

More than just a definition of community, the cohesion is a new framework which paves the way to a better understanding and analysis of social networks structure. Although this thesis provides the foundations of this framework, I believe only the surface has been scratched and so much remains to be uncovered.

From a theoretical standpoints, I have believe strongly that not only the problem of finding maximally cohesive groups is $\mathcal{NP}$-hard, but that moreover it is hard to approximate. However, this does not mean that C^3 is the most efficient algorithm to optimize

the cohesion and a future line of work will surely be to try and refine the algorithm.

On a more practical – or applied – note, there are several other directions which I believe are worth investigating. One of my first contacts with communities dates back to 2008, when I was working on information diffusion. At that time, I had launched an online experiment aimed at measuring the spread of a resource across the blogosphere[1] and of interest here is an observation I had during the progress of the experiment. As I was looking at the blogs which had been "infected" by the resource, I noticed that it actually seemed to be hopping from one semantic community to another – for example, it stagnated for a few days in a marketing community before suddenly spreading like wildfire in a job searching community, before slowing again and hopping to yet another community.

The fact that there are links between communities and diffusion seems obvious in retrospect, since cohesive communities are by definition ideally separated by weak ties which are, let us recall, the ties without which "new ideas [would] spread slowly[2]". There has classically been an assumption that there is somehow a relation between number of friends and influence, an idea which seems simplistic upon closer inspection. I do admit that someone having more friends might transmit an opinion to more people in his immediate neighborhood – due to the simple fact that there are more people to transmit the opinion to – but consider

1 Adrien Friggeri, Jean-Philippe Cointet, and Matthieu Matthieu Latapy. "A Real-World Spreading Experiment in the Blogosphere". In: *Complex Systems* 19.3 (2011).

2 Granovetter, "The strength of weak ties: a network theory revisited", p. 202.

two individuals, one having a large number of friends all part of the same community, the second one having far fewer friends but in separate communities. In a sense, the first can be considered from a topological standpoint as interchangeable with others in the community, whereas the second one has a more unique profile in that he acts as a structural hole between disparate communities. As such, it should be worth investigating the existence of links between the cohesion and influence.

Another field in which the use of communities might prove insightful is in terms of traits inference, which can be useful in a context of social recommendation. For example, on online social networks, some users add personnal information such as favorite brands, music or movies in addition to the actual social ties they maintain, which information is then used to suggests new centers of interest or target advertising. There are however some user who do not wish or know how to add such information to their profiles. As such, the knowledge of communities can be leveraged to infer this meta-data from their friends.

For instance, consider an individual having 100 friends, and suppose that 10 of his friends have stated they liked a specific restaurant. In the general case, this means that 10% of his friends share a common trait but does not bring a great deal of information. However, if those friends form a cohesive communities, then it is not only 10% of all friends, but 100% of a community, and as such the probability that the user would be interested – or already knows – this restaurant is far higher. Basically, this notion of inference relies on the analysis of the distribution of a trait among the social neighborhood of the user. In the cases where the trait is homogeneously spread among his

friends, not much information is gained, however if it is mostly located inside a community then the probability the topological structure of the network can enhance the deduction.

More broadly, this kind of reasoning could be used to identify the nature of the communities. As such, for the time being, once communities are computed using C^3, we know that the members of the community have something in common, but we do not know what. During the time where Fellows was in progress, a large number of users asked me how the application knew who were their family, their chess club or whatever community it had found. Invariably I answered that the application did not know that, since it only relied on the social ties of the network. However I believe that community detection to generate automatic lists of friends – such as mailing lists in email clients, or friends lists on Facebook – could benefit from an automatic labeling. For instance one could identify groups which are families by comparing surnames and checking for a high age variability among the group as families span several generation, similarly. More generally, a possible extension of this work would be in the identification of a minimal set of meta-data which use could lead to a characterization of a social community.

In conclusion, this thesis answers a long-standing open question of quantitative sociology by providing a justified, solid and experimentally validated mathematical definition of community. Although important theoritical results as well as applications to diverse fields have already been established in this thesis, it paves the way to a complementary approach to social network analysis which drills into the network halfway between the individual

and the global without suffering from the drawbacks of previous contributions since the cohesion has been shown to be a good indicator of communitiness, is based on solid sociological groups, is defined as a local metric and does not impose restrictions on the size or overlap between communities.

Publications

Although this thesis is focused on my work on social cohesion, I spent my first year in the DNET team working on the TubExpo project, a large-scale deployment of a wireless sensor network in two hospitals to monitor the interactions between Health Care Workers and patients. Due to physical constraints, the measure suffered a large amount of data loss – up to 80% of packets in some cases. I contributed to that project a signal processing treatment to reconstruct the collected data, which explains the presence of publications not directly related to communities and the cohesion.

Articles in Peer-Reviewed Journals

Lucet, Jean-Christophe, Cédric Laouenan, Guillaume Chelius, Nicolas Veziris, Didier Lepelletier, Adrien Friggeri, Dominique Abiteboul, Elisabeth Bouvet, France Mentré, and Eric Fleury. "Electronic Sensors for Assessing Interactions between Healthcare Workers and Patients under Airborne Precautions". In: *PLoS ONE* 7.5 (May 2012), e37893.

Friggeri, Adrien, Guillaume Chelius, Eric Fleury, Antoine Fraboulet, France Mentré, and Jean-Christophe Lucet. "Reconstructing Social Interactions Using an unreliable Wireless Sensor Network". In: *Computer Communications* 34.5 (2011), pp. 609–618.

Friggeri, Adrien, Jean-Philippe Cointet, and Matthieu Matthieu Latapy. "A Real-World Spreading Experiment in the Blogosphere". In: *Complex Systems* 19.3 (2011).

International Peer-reviewed Conferences/Proceedings

Friggeri, Adrien, Renaud Lambiotte, Michal Kosinski, and Eric Fleury. "Psychological Aspects of Social Communities". In: *2012 ASE/IEEE International Conference on Social Computing (SocialCom 2012)*. Amsterdam, Netherlands, Sept. 2012.

Friggeri, Adrien. "Agreement Groups in the United States Senate". In: *International Open Government Data Conference*. Washington DC, United States, July 2012.

Friggeri, Adrien, Guillaume Chelius, and Eric Fleury. "Fellows: Crowd-sourcing the evaluation of an overlapping community model based on the cohesion measure". In: *Interdisciplinary Workshop on Information and Decision in Social Networks*. Cambridge, United States, 2011.

— "Fellows: Crowd-sourcing the evaluation of an overlapping community model based on the cohesion measure". In: *Complex Dynamics of Human Interactions*. Vienna, Austria, 2011.

— "Triangles to Capture Social Cohesion". In: *2011 IEEE Third International Conference on Social Computing (SocialCom 2011)*. Cambridge, United States, 2011, pp. 258–265.

— "Egomunities, Exploring Socially Cohesive Person-based Communities". In: *NetSci 2011 The International School and Conference on Network Science*. Budapest, Hungary, 2011.

Lucet, Jean-Christophe, Guillaume Chelius, Cédric Laouenan, Adrien Friggeri, N. Veziris, D. Lepelletier, D. Abiteboul, Elisabeth Bouvet, Eric Fleury, and France Mentré. "Electronic Sensors for Measuring Interactions between Healthcare Workers (HCWs) and Patients (Pts): the Case of Tuberculosis (TB)". In: *2010 Interscience Conference on Antimicrobial Agents and Chemotherapy*. American Society for Microbiology. Boston, United States, 2010.

Local Peer-reviewed Conferences/Proceedings

Friggeri, Adrien, Guillaume Chelius, and Eric Fleury. "Communautés : Arrêtons de ne compter que les arêtes". In: *13es Rencontres Francophones sur les Aspects Algorithmiques de Télécommunications (AlgoTel)*. Ed. by Bertrand Ducourthial and Pascal Felber. Cap Estérel, France, 2011.

— "Trouver des communautés socialement cohésives est NP-dur". In: *13emes journées Graphes et Algorithmes*. Lyon, France, 2011.

Friggeri, Adrien and Guillaume Chelius. "Biais dans les mesures obtenues par un réseau de capteurs sans fil". In: *12èmes Rencontres Francophones sur les Aspects Algorithmiques de Télécommunications (AlgoTel)*. Ed. by Maria Gradinariu Potop-Butucaru and Hervé Rivano. Belle Dune, France, 2010.

Research Reports

Friggeri, Adrien and Eric Fleury. *Finding cohesive communities with C3*. Tech. rep. RR-7947. Apr. 2012.

Friggeri, Adrien, Guillaume Chelius, and Eric Fleury. *Egomunities, Exploring Socially Cohesive Person-based Communities*. Tech. rep. RR-7535. Budapest, Hongrie: ENS / LIP Laboratoire de l'Informatique du Parallélisme / INRIA Grenoble Rhône-Alpes, Feb. 2011.

Friggeri, Adrien and Eric Fleury. *Maximizing the Cohesion is NP-hard*. Tech. rep. RR-7734. 2011.

Friggeri, Adrien, Guillaume Chelius, and Eric Fleury. *Triangles to Capture Social Cohesion*. Tech. rep. RR-7686. 2011.

Bibliography

Anderson, Cameron, Oliver P John, Dacher Keltner, and Ann M Kring. "Who attains social status? Effects of personality and physical attractiveness in social groups." In: *Journal of Personality and Social Psychology* 81.1 (2001), p. 116.

Bansal, Richa and Kamal Srivastava. "Memetic algorithm for the antibandwidth maximization problem". In: *Journal of Heuristics* 17.1 (2011), pp. 39–60.

Baumes, Jeffrey, Mark Goldberg, Mukkai Krishnamoorthy, Malik Magdon-Ismail, and Nathan Preston. "Finding communities by clustering a graph into overlapping subgraphs". In: *International Conference on Applied Computing (IADIS 2005)* (2005), pp. 97–104.

Berry, Diane S., Julie K. Willingham, and Christine A. Thayer. "Affect and personality as predictors of conflict and closeness in young adults' friendships". In: *Journal of Research in Personality* 34.1 (2000), pp. 84–107.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* (2008).

Blythe, Jim, Cathleen McGrath, and David Krackhardt. "The effect of graph layout on inference from social network data". In: *Graph Drawing*. Ed. by Franz Brandenburg. Springer Berlin / Heidelberg, 1996, pp. 40–51.

Bourdieu, Pierre and Loíc J D Wacquant. *An invitation to reflexive sociology*. University Of Chicago Press, July 1992.

boyd, danah and Jeffrey Heer. "Vizster: Visualizing Online Social Networks". In: *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on* (Sept. 2005), pp. 32–39.

Brandes, Ulrik, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. "On finding graph clusterings with maximum modularity". In: *Graph-Theoretic Concepts in Computer Science* (2007).

Breiger, Ronald L., Scott A. Boorman, and Phipps Arabie. "An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling". In: *Journal of Mathematical Psychology* 12.3 (1975), pp. 328–383.

Brockenauer, Ralf and Sabine Cornelsen. "Drawing clusters and hierarchies". In: *Drawing graphs* (2001), pp. 193–227.

Buchanan, Tom and John L Smith. "Using the Internet for psychological research: Personality testing on the World Wide Web". In: *British Journal of Psychology* 90.1 (1999), pp. 125–144.

Burt, Ronald S. *Brokerage and Closure*. An Introduction to Social Capital. Oxford University Press, Aug. 2005.

Burt, Ronald S, Joseph E Jannotta, and James T Mahoney. "Personality correlates of structural holes". In: *Social Networks* 20 (1998), pp. 63–87.

Chausson, Olivia. "Assessing the impact of gender and personality on film preferences". In: *Cambridge University* ().

Civic Impulse, LLC. *GovTrack.us*. URL: http://www.govtrack.us/.

Clauset, Aaron. "Finding local community structure in networks". In: *Physical Review E* 72.2 (Aug. 2005).

Costa, Paul T and Robert R McCrae. *NEO Personality Inventory Revised NEO-PI-R Test Manual*. SAGE Publications. May 2005.

Diestel, Reinhard. *Graph theory*. Springer Verlag, Feb. 2006.

Facebook. *Graph API*. Tech. rep. 2011.

Feynman, Richard Phillips. *The character of physical law*. Modern Library. 1967.

Forsyth, Elaine and Leo Katz. "A matrix approach to the analysis of sociometric data: preliminary report". In: *Sociometry* 9.4 (1946), pp. 340–347.

Fortunato, Santo. "Community detection in graphs". In: *Physics Reports* 486.3-5 (Jan. 2010), pp. 75–174.

Fortunato, Santo and Marc Barthélemy. "Resolution limit in community detection". In: *Proceedings of the National Academy of Sciences* 104.1 (2007), p. 36.

Freeman, Linton C. *The development of social network analysis*. a study in the sociology of science. Booksurge Llc, 2004.

Friggeri, Adrien, Jean-Philippe Cointet, and Matthieu Matthieu Latapy. "A Real-World Spreading Experiment in the Blogosphere". In: *Complex Systems* 19.3 (2011).

Frishman, Yaniv and Ayellet Tal. "Dynamic drawing of clustered graphs". In: *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on* (2004), pp. 191–198.

Fruchterman, Thomas M. J. and Edward M. Reingold. "Graph Drawing by Force-directed Placement". In: *Software: Practice and Experience* 21.11 (Jan. 1997), pp. 1129–1164.

Gajer, Pawel, Michael Goodrich, and Stephen Kobourov. "A Multi-dimensional Approach to Force-Directed Layouts of Large Graphs". In: *Graph Drawing*. Ed. by Joe Marks. Springer Berlin / Heidelberg, 2001, pp. 211–221.

Granovetter, Mark. "The Strength of Weak Ties". In: *American journal of sociology* 78.6 (May 1973), pp. 1360–1380.

— "The strength of weak ties: a network theory revisited". In: *Sociological Theory* 1 (1981), pp. 201–233.

Grossman, Pamela, Sam Wineburg, and Stephen Woolworth. *What makes teacher community different from a gathering of teachers?* Tech. rep. Seattle, Dec. 2000.

Hillery Jr, George A. "Definitions of community: Areas of agreement". In: *Rural sociology* 20.2 (1955), pp. 111–123.

Hu, Yifan, Stephen Kobourov, and Sankar Veeramoni. "On maximum differential graph coloring". In: *Graph Drawing* (2011), pp. 274–286.

Jensen Campbell, Lauri A and William G Graziano. "Agreeableness as a moderator of interpersonal conflict". In: *Journal of personality* 69.2 (2001), pp. 323–362.

Kalish, Yuval and Garry Robins. "Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure". In: *Social Networks* (2006).

Katz, Leo. "On the matric analysis of sociometric data". In: *Sociometry* 10.3 (1947), pp. 233–241.

Kilduff, Martin and David Krackhardt. "Bringing the individual back in: A structural analysis of the internal market for reputation in organizations". In: *Academy of Management Journal* (1994), pp. 87–108.

Kuper, Adam and Jessica Kuper. *The Social Science Encyclopedia*. Ed. by Adam Kuper and Jessica Kuper. 2nd ed. Routledge world reference. London: Routledge, 2003.

Latapy, Matthieu. "Main-memory triangle computations for very large (sparse (power-law)) graphs". In: *Theoretical Computer Science* 407.1 (2008), pp. 458–473.

Lewis, John M and Mihalis Yannakakis. "The node-deletion problem for hereditary properties is NP-complete". In: *Journal of Computer and System Sciences* 20.2 (1980), pp. 219–230.

Lozano, Manuel, Abraham Duarte, Francisco Gortázar, and Rafael Martí. "Variable Neighborhood Search with Ejection Chains for the Antibandwidth Problem". In: ().

Mathieu, Jacomy, Heymann Sebastien, Venturini Tommaso, and Bastian Mathieu. *ForceAtlas2, A Graph Layout Algorithm for Handy Network Visualization*. Tech. rep. Aug. 2011.

McPherson, Miller and Lynn Smith-Lovin. "Birds of a feather: Homophily in social networks". In: *Annual review of sociology* 27 (2001), pp. 415–444.

Melgosa, M, J J Quesada, and E Hita. "Uniformity of some recent color metrics tested with an accurate color-difference tolerance dataset". In: *Appl. Opt.* 33.34 (Dec. 1994), pp. 8069–8077.

Merton, Robert K. *On Theoretical Sociology*. Free Press. New York, 1967.

Moody, James and Douglas White. "Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups". In: *American Sociological Review* 68 (2003), pp. 103–127.

Moreno, Jacob Levy and Helen Hall Jennings. *Who shall survive?: A new approach to the problem of human interrelations.* Washington, DC, US: Nervous and Mental Disease Publishing Co., 1934.

Nettle, Daniel. *Personality: what makes you the way you are*. Oxford University Press. 2007.

Newman, Mark E.J. and Michelle Girvan. "Finding and evaluating community structure in networks". In: *Physical Review E* 69.2 (2004), p. 26113.

Nicosia, Vincenzo, Giuseppe Mangioni, Vincenza Carchiolo, and Michele Malgeri. "Extending the definition of modularity to directed graphs with overlapping communities". In: *Journal of Statistical Mechanics: Theory and Experiment* 2009 (2009), P03024.

Noack, Andreas. "An Energy Model for Visual Graph Clustering". In: *Graph Drawing*. Ed. by Giuseppe Liotta. Springer Berlin / Heidelberg, 2004, pp. 425–436.

Palla, Gergely, Imre Derényi, Illés Farkas, and Tamas Vicsek. "Uncovering the overlapping community structure of complex networks in nature and society". In: *Nature* 435.7043 (2005), pp. 814–818.

Pollet, Thomas V., Sam G. B. Roberts, and Robin I. M. Dunbar. "Extraverts Have Larger Social Network Layers". In: *Journal of Individual Differences* (2011).

Poplin, Dennis E. *Communities: A survey of theories and methods of research*. Macmillan. 1979.

Quercia, Daniele, Renaud Lambiotte, Michal Kosinski, David Stillwell, and Jon Crowcroft. "The personality of popular Facebook users". In: *Proc. ACM Conf. Comput. Support. Cooperat. Work* (2012).

Rapoport, Anatol. "Contributions to the Theory of Random and Biased Nets". In: *Bulletin of Mathematical Biophysics* 19 (1957), pp. 257–277.

Reichardt, Jörg and Stefan Bornholdt. "When are networks truly modular?" In: *Physica D: Nonlinear Phenomena* 224.1-2 (2006), pp. 20–26.

Sampson, Samuel F. "Crisis in a cloister". PhD thesis. 1969.

Schmalenbach, Hermann. "The sociological category of communion". In: *Theories of society*. Theories of society, 1961, pp. 331–347.

Shen, Hua-Wei, Xue-Qi Cheng, and Jia-Feng Guo. "Quantifying and identifying the overlapping community structure in networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2009 (2009), P07042.

Simonetto, Paolo, David Auber, and Daniel Archambault. "Fully Automatic Visualisation of Overlapping Sets". In: *Computer Graphics Forum* 28.3 (2009), pp. 967–974.

Stuckey, Bronwyn E. "Growing online community: core conditions to support successful development of community in Internet-mediated communities of practice." PhD thesis. University of Wollongong, 2007.

Swickert, Rhonda, Christina J Rosentreter, James B Hittner, and Jane E Mushrush. "Extraversion, social support processes, and stress". In: *Personality and Individual Differences* 32.5 (2002), pp. 877–891.

Tarjan, Robert Endre and Cornell University. Department of Computer Science. *On the efficiency of a good but not linear set union algorithm*. Tech. rep. Nov. 1972.

Traud, Amanda L, Christina Frost, Peter J Mucha, and Mason A Porter. "Visualization of communities in networks". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 19.4 (2009), p. 041104.

Ugander, Johan, Brian Karrer, Lars Backstrom, and Camero Marlow. "The anatomy of the facebook social graph". In: *Arxiv preprint arXiv:1111.4503* (2011).

Wang, Qinna and Eric Fleury. "Uncovering Overlapping Community Structure". In: *Complex Networks*. 2010.

Wasserman, Stanley and Katherine Faust. *Social Network Analysis*. Cambridge University Press. Nov. 1994.

Wellman, Barry and University of Toronto. Centre for Urban and Community Studies. *Studying personal communities in East York*. Sage. Beverly Hills, 1982.

White, Harrison C., Scott A. Boorman, and Ronald L. Breiger. "Social structure from multiple networks. I. Blockmodels of roles and positions". In: *American journal of sociology* (1976), pp. 730–780.

Wilson, Chris. *Searching for Saddam*. Slate Magazine. Feb. 2010. URL: http://www.slate.com/articles/news_and_politics/searching_for_saddam/2010/02/searching_for_saddam_5.single.html.

Zachary, Wayne. "An information flow model for conflict and fission in small groups". In: *Journal of Anthropological Research* 33 (1977), pp. 452–473.

A Quantitative Theory of Social Cohesion

Adrien Friggeri

Community, a notion transversal to all areas of Social Network Analysis, has drawn tremendous amount of attention across the sciences in the past decades. Numerous attempts to characterize both the sociological embodiment of the concept as well as its observable structural manifestation in the social network have to this date only converged in spirit. No formal consensus has been reached on the quantifiable aspects of community, despite it being deeply linked to topological and dynamic aspects of the underlying social network.

Presenting a fresh approach to the evaluation of communities, this thesis introduces and builds upon the cohesion, a novel metric which captures the intrinsic quality, as a community, of a set of nodes in a network. The cohesion, defined in terms of social triads, was found to be highly correlated to the subjective perception of communitiness through the use of a large-scale online experiment in which users were able to compute and rate the quality of their social groups on Facebook.

Adequately reflecting the complexity of social interactions, the problem of finding a maximally cohesive group inside a given social network is shown to be $\mathcal{NP}$-hard. Using a heuristic approximation algorithm, applications of the cohesion to broadly different use cases are highlighted, ranging from its application to network visualization, to the study of the evolution of agreement groups in the United States Senate, to the understanding of the intertwinement between subjects' psychological traits and the cohesive structures in their social neighborhood.

The use of the cohesion proves invaluable in that it offers non-trivial insights on the network structure and its relation to the associated semantic.