Year : 2014

Character Networks and Centrality

Yannick Rochat

UNIL | Université de Lausanne

Faculté des sciences sociales et politiques
Institut de mathématiques appliquées

Character Networks and Centrality

Thèse de doctorat

présentée à la
Faculté de sciences sociales et politiques
de l'Université de Lausanne
pour l'obtention du grade de

Docteur en mathématique appliquée aux sciences humaines et sociales

par

Yannick Rochat

Directeur de thèse
Prof. hon. Henri Volken, Université de Lausanne

Co-directeur de thèse
Prof. Frédéric Kaplan, École polytechnique fédérale de Lausanne

Jury

Prof. Dominique Joye, Université de Lausanne
Prof. Jean-Gabriel Ganascia, Université Pierre et Marie Curie, Paris
Prof. Matthew Jockers, University of Nebraska, Lincoln

Lausanne
2014

UNIL | Université de Lausanne
Faculté des sciences
sociales et politiques

IMPRIMATUR

Le Conseil de la Faculté des sciences sociales et politiques de l'Université de Lausanne, sur proposition d'un jury formé des professeurs

- Henri VOLKEN, Co-directeur de thèse, Professeur honoraire de l'Université de Lausanne
- Frédéric KAPLAN, Co-directeur de thèse, Professeur à l'EPFL
- Dominique JOYE, Professeur à l'Université de Lausanne
- Jean-Gabriel GANASCIA, Professeur à l'Université Pierre et Marie Curie à Paris, France
- Matthew L. JOCKERS, Professeur à la University of Nebraska à Lincoln, USA

autorise, sans se prononcer sur les opinions du candidat, l'impression de la thèse de Monsieur Yannick ROCHAT, intitulée :

« Character Networks and Centrality. »

Lausanne, le 21 octobre 2014

Le Doyen de la Faculté

Professeur
Fabien Ohl

Abstract

A character network represents relations between characters from a text; the relations are based on text proximity, shared scenes/events, quoted speech, etc. Our project sketches a theoretical framework for character network analysis, bringing together narratology, both close and distant reading approaches, and social network analysis. It is in line with recent attempts to automatise the extraction of literary social networks (Elson, 2012; Sack, 2013) and other studies stressing the importance of character-systems (Woloch, 2003; Moretti, 2011).

The method we use to build the network is direct and simple. First, we extract co-occurrences from a book index, without the need for text analysis. We then describe the narrative roles of the characters, which we deduce from their respective positions in the network, i.e. the discourse.

As a case study, we use the autobiographical novel *Les Confessions* by Jean-Jacques Rousseau. We start by identifying co-occurrences of characters in the book index of our edition (Slatkine, 2012). Subsequently, we compute four types of centrality: degree, closeness, betweenness, eigenvector. We then use these measures to propose a typology of narrative roles for the characters.

We show that the two parts of *Les Confessions*, written years apart, are structured around mirroring central figures that bear similar centrality scores. The first part revolves around the mentor of Rousseau; a figure of openness. The second part centres on a group of schemers, depicting a period of deep paranoia. We also highlight characters with intermediary roles: they provide narrative links between the societies in the life of the author.

The method we detail in this complete case study of character network analysis can be applied to any work documented by an index.

Résumé

Un réseau de personnages modélise les relations entre les personnages d'un récit : les relations sont basées sur une forme de proximité dans le texte, l'apparition commune dans des événements, des citations dans des dialogues, etc. Notre travail propose un cadre théorique pour l'analyse des réseaux de personnages, rassemblant narratologie, close et distant reading, et analyse des réseaux sociaux. Ce travail prolonge les tentatives récentes d'automatisation de l'extraction de réseaux sociaux tirés de la littérature (Elson, 2012; Sack, 2013), ainsi que les études portant sur l'importance des systèmes de personnages (Woloch, 2003; Moretti, 2011).

La méthode que nous utilisons pour construire le réseau est directe et simple. Nous extrayons les co-occurrences d'un index sans avoir recours à l'analyse textuelle. Nous décrivons les rôles narratifs des personnages en les déduisant de leurs positions relatives dans le réseau, donc du discours.

Comme étude de cas, nous avons choisi le roman autobiographique *Les Confessions*, de Jean-Jacques Rousseau. Nous déduisons les co-occurrences entre personnages de l'index présent dans l'édition Slatkine (Rousseau et al., 2012). Sur le réseau obtenu, nous calculons quatre types de centralité : le degré, la proximité, l'intermédiarité et la centralité par vecteur propre. Nous utilisons ces mesures pour proposer une typologie des rôles narratifs des personnages.

Nous montrons que les deux parties des *Confessions*, écrites à deux époques différentes, sont structurées autour de deux figures centrales, qui obtiennent des mesures de centralité similaires. La première partie est construite autour du mentor de Rousseau, qui a symbolisé une grande ouverture. La seconde partie se focalise sur un groupe de comploteurs, et retrace une période marquée par la paranoïa chez l'auteur. Nous mettons également en évidence des personnages jouant des rôles intermédiaires, et de fait procurant un lien narratif entre les différentes sociétés couvrant la vie de l'auteur.

La méthode d'analyse des réseaux de personnages que nous décrivons peut être appliquée à tout texte de fiction comportant un index.

"My Princess," he said tenderly, "two great powers are on our side: the power of Love and the power of Arithmetic. Those two are stronger than anything else in the world."

The Island of the Nine Whirlpools
EDITH NESBIT

À Jacques et Véronique, mes parents…

Remerciements (Acknowledgements)

> "A recent article stated that Sir Robin Day, on seeing Ludovic Kennedy's autobiography *On my way to the club* (Collins, 1989) did what one might have expected: 'Naturally, the first thing he did was to look himself up in the index.'"
>
> *Indexes to works of fiction*
> PHILIP BRADLEY

Je vais tenter de remercier toutes celles et tous ceux qui ont rendu cette thèse possible par leur aide, par leur écoute ou par leurs encouragements. Ce travail s'est déroulé sur une période de temps prolongée, ces *Remerciements* le seront aussi afin d'inclure, idéalement, toutes ces personnes.

Tout d'abord, j'aimerais remercier les professeurs Henri Volken et Frédéric Kaplan, mes co-directeurs de thèse. J'étais alors mathématicien fraîchement diplômé quand Henri Volken m'a fait confiance en m'engageant à l'*Institut de Mathématiques Appliquées* (IMA) de l'UNIL pour l'assister ainsi que pour préparer une thèse au sein d'une faculté de sciences sociales et politiques. Il a cru en moi, m'a laissé une grande liberté, m'a poussé à aller présenter mes travaux dans diverses conférences, et est resté disponible et à l'écoute pendant tout ce temps. Puis, en 2012, j'ai rejoint le tout nouveau *Laboratoire en Humanités Digitales* (DHLAB) de l'EPFL dirigé par Frédéric Kaplan où j'ai eu la chance inespérée de pouvoir la terminer. Pendant ces deux années, Frédéric a été disponible à tout instant et j'ai pu bénéficier à son contact de conseils inestimables sur maints aspects de la recherche scientifique, et notamment sur la communication de celle-ci. Frédéric, j'ai pour toi une reconnaissance infinie. Merci à tous les deux pour votre confiance.

Mon comité de thèse était également composé des professeurs Jean-Gabriel Ganascia, Matthew Jockers et Dominique Joye qui m'ont fait l'immense honneur d'accepter de jouer ce rôle central dans mon doctorat, ont gracieusement mis à disposition un peu de leur précieux temps et m'ont fourni des retours constructifs qui ont permis d'affiner et de clarifier le texte ici présent. Merci beaucoup pour votre aide et votre intérêt.

De l'époque des débuts, j'aimerais remercier tous mes collègues à l'IMA pour leur sympathie et pour les innombrables discussions portant sur des sujets toujours variés. Et remercier en particulier Arnaud Thévenet et Jorge Peña avec qui j'ai tour à tour partagé le bureau 4126 et qui sont devenus des amis proches. Arnaud, ça va pas être possible de tout dire et décrire ici : merci pour ton amitié, ta folie, les incroyables moments passés ensemble et ceux à venir.

Chaque fois que j'écoute Cut Copy j'ai une pensée pour toi. Jorge, tu m'as convaincu que je devais croire en moi, tu as ouvert la voie et tu es encore une source d'inspiration. L'existence de cette thèse te doit beaucoup. Il y a aussi Jean-Pierre Müller, qui m'a sauvé un grand nombre de fois jusqu'au jour même du colloque de thèse, et Jean-Philippe Antonietti, qui m'a initié aux statistiques et a été un excellent adversaire au badminton. Et le plus important en ce qui les concerne, bien sûr, c'est de m'avoir fait découvrir le monde merveilleux de *R*. Merci à André Berchtold pour de précieux conseils dans une période de doute. On ne saura oublier les excellent-e-s membres des *soupers des assistant-e-s de l'IMA* (16 éditions) : Carolina Carvalho, Mélanie Glasson-Cicognani, Thomas Müller, et en particulier Julie Falcon. Merci Julie pour ton soutien dans la dernière ligne droite, et pour m'avoir montré que quatre pages de remerciements ça pouvait se faire.

Hors de l'IMA mais toujours à l'Université de Lausanne, je souhaite remercier les professeurs François Bavaud et Marco Tomassini pour leur ouverture, leurs encouragements et leurs excellents conseils, ainsi que François Allisson, Isis Giraldo, et Frédéric Schütz pour de nombreuses discussions, et pour la veille académique et intellectuelle. Merci à Martin Grandjean pour des discussions tout aussi intéressantes et pour les collaborations, notamment sur *@PegasusData*. Merci également aux membres du séminaire en humanités digitales UNIL–EPFL pour les échanges, et à Claire Clivaz pour son rôle crucial dans cette communauté.

À un moment où je semblais bon pour quitter le monde académique, le DHLAB m'a recueilli. En deux ans, dans un laboratoire passé de trois à quinze personnes, j'y ai vécu assez pour révolutionner mon univers jusqu'à la fin de ma carrière. Et j'ai pu y trouver les conditions parfaites pour terminer ma thèse. Tout d'abord, je tiens à remercier Alicia Foucart, soucieuse du détail et attentive à toujours maintenir la meilleure des ambiances, une personne d'une immense générosité, élément central et indispensable de notre laboratoire. Puis, par où commencer pour remercier Quentin Bonnard et Mélanie Fournier ? Quentin, merci pour tes expressions régulières, pour la *technique pomodoro*, pour les conseils à propos de la thèse, pour les discussions interminables sur tous les sujets imaginables. Merci Mélanie pour ta rigueur, pour les belles collaborations transdisciplinaires, pour ton amitié malgré la distance, pour l'excellente cohabitation quand le laboratoire n'était quasiment composé que de nous deux. Ensuite, merci à Cyril Bornet pour son écoute, pour son aide cruciale à plusieurs reprises, pour l'entente parfaite et pour ses mèmes de qualité supérieure. Merci à l'ensemble de mes excellent-e-s collègues qui mériteraient chacun et chacune au moins un paragraphe : Irene Bianchi, Vincent Buntinx, Giovanni Colavizza, Olivier Dalang, Dana Kianfar, Katerina Kunz, Isabella di Lenardo, Andrea Mazzei, Michael Mitchell, Federica Pardini, Catharina Pützschler, Dario Rodighiero, Fouad Slimane, Orlin Topalov. Je suis un peu court avec vous ici, mais pour la plupart nous allons continuer à travailler ensemble et j'aurai encore des occasions de vous rappeler ce que je vous dois.

Au Collège des Humanités de l'EPFL, j'aimerais remercier le professeur Thomas David pour ses bons conseils et sa disponibilité, Isabelle Vonèche pour sa bonne humeur et les services rendus, Claude Zwicky entre autres pour son aide lors des tracas administratifs. Merci à Claude Lutzelschwab pour des discussions très intéressantes à plusieurs moments importants. Un immense merci à Kevin "Genie Army" Baumer qui a lu intégralement cette

thèse et proposé d'y apporter plusieurs milliers (dizaines de milliers ?) de corrections. Kevin, je garde précieusement ton exemplaire corrigé en espérant ne plus jamais faire ces erreurs. J'en profite pour ajouter que de nombreux passages ont été revus et/ou augmentés après la lecture de Kevin : je suis donc le seul responsable des probables fautes résidant dans cette version finale.

J'ai eu l'occasion de faire de belles rencontres dans des conférences, séminaires et *summer schools* qui auront joué un rôle important par la suite. Pour des discussions, des conseils ou pour leur amitié, j'aimerais remercier Christian Hirschi, Karin Ingold, Uwe Serdült et Alexander Widmer d'*ASNA* à Zürich, Martin Everett de l'Université de Manchester, Claire Bidart et Tom Brughmans rencontrés lors de la summer school "*Étudier les réseaux sociaux*" à Porquerolles, les participantes et participants des deux *Digital humanities Venice fall school*, et finalement Cameron Bruckner, Ayse Gursoy, Geoffrey Rockwell, Maximilian Schich, Brandon Tinklenberg et Jeffrey R. Villines dont j'ai fait la connaissance à *TXDHC*, Houston.

J'ai commencé à utiliser *R* avec le package *igraph*, qui est encore au centre de mes travaux. Ce package impeccablement maintenu est l'oeuvre de Gábor Csárdi et de Tamás Nepusz. Leur disponibilité, leur capacité d'écoute, leur dévouement et leur modestie sont exemplaires. Participer à cette communauté a été un plaisir. En particulier, pouvoir rencontrer Gábor à Lausanne fut un des grands moments de ma thèse, tout comme l'organisation conjointe d'un workshop sur *igraph*. Merci Gábor pour cette générosité, cette humilité et cette exemplarité. Une autre grande source d'inspiration pendant ces années fut le laboratoire d'Ulrik Brandes à Konstanz. Pendant longtemps, ce fut le seul laboratoire dont les axes de recherche recoupaient les miens et j'ai eu la chance de rencontrer deux de ses membres, Bobo Nick et Uwe Nagel, que j'aimerais remercier ici pour des échanges qui m'auront inspiré ensuite. Je devais partir à Konstanz, cela ne s'est malheureusement pas fait : merci à Ulrik Brandes pour sa confiance à cette occasion et pour ses encouragements.

Parmi mes amies, amis et connaissances, j'ai pu compter sur d'innombrables soutiens que je n'ai pas encore cités. Certains étaient déjà passés par là et m'ont laissé bénéficier de leur expérience, pendant, après ou longtemps après leur doctorat : merci à Camilo Argibay, Jean-Marie Bolay, Frédéric Bongard, Karine Crousaz, Fabio Daolio, Anouk Flamant, Sharifa Himidan, Julian Kellerhals, Jérémie Knüsel, Nelly Niwa, Gregory Putzel, Gilles Steiner et Laura Vinckenbosch. Merci également à Ho Ngoc Huy pour les heures de défoulement à jouer au badminton. Merci à Benoît Perrier pour une correspondance stimulante via des emails et DMs en cascades. Merci à David Christoffel pour son intérêt dans ma recherche. Merci à Magali Philip d'avoir aidé à diffuser les travaux de *@PegasusData*. Merci à Florian Poupelin d'être, pour certains centres d'intérêt, le seul de mes amis à les partager. Merci à Joana Borrego, Xavier Company et Alberto Mocchi pour leur amitié allant bien au-delà de ce qui nous réunissait à l'origine. Merci à Vanessa Monney et Joël Terraz d'avoir été parmi mes tous premiers amis à l'UNIL. Merci à Francesco Gramiccia pour les pauses ludiques. Merci à Fabrice Lambelet pour les pauses café à l'UNIL et pour cette amitié qui dure depuis toujours. Merci à Adrien Schnarrenberger pour ses encouragements par l'entremise de pages du teletext, un soutien inestimable. Et finalement merci à Catherine Wheel, Do Make Say Think, Lali Puna et aux

Super Furry Animals pour leurs univers musicaux dans lesquels j'ai le plus souvent évolué pendant l'écriture.

La recherche aujourd'hui ne se fait plus seul, sans ouverture immédiate vers l'extérieur. Au contraire, on s'expose, on partage instantanément des résultats ou des impressions, et les retours sont rapides. J'ai eu la chance de recevoir à diverses occasions des feedbacks et encouragements de connaissances via le réseau social en ligne Twitter, le plus souvent elles-mêmes rencontrées par ce biais. Je vais remercier ici les personnes que j'ai encore en tête et prie celles omises de m'excuser. Ces encouragements, même très courts, m'ont en tout temps énormément apporté. Par nom ou par pseudo, un grand merci à : @antoinehawk, Eric Butticaz, Guillaume Cabanac, @Calamity_vt, Alex Dépraz, Chrystel Domenjoz, Marine Gout, Laurence Jobin, Fanny Lalleman, David Lemos, Maxime Mellina, @MlleFunambuline, Marc Münster, Nicolas Nova, Ludovic Péron et @shalf.

Pour terminer, j'aimerais remercier l'ensemble de ma famille pour son soutien inconditionnel, malgré les épreuves qui ont rendu ces dernières années difficiles. J'ai eu la chance de grandir entouré de tous mes Grand-Parents et de pouvoir bénéficier de leurs personnalités et de leurs connaissances pour construire la personne que je suis aujourd'hui. Je pense avec tristesse et gratitude à ma Grand-Mère Anne Rochat qui nous a quittés il y a deux ans, ainsi qu'à mon Grand-Père Jean Savoy qui nous a quittés l'année suivante. On ne peut imaginer combien je serai heureux d'accueillir ma Grand-Mère Émilie Savoy et mon Grand-Père Robert Rochat lors de la défense de ce texte en public dans quelques jours. Une pensée pour mon petit frère, Mikaël, et notre relation d'amour-haine qui durera toujours je l'espère.

Sarah, tu m'as montré qu'une thèse ça n'était pas stressant du tout en organisant ta défense privée huit jours avant notre mariage. Maintenant que j'y repense, c'était courageux mais surtout osé de ta part. Je t'aime pour les déclinaisons infinies de cette part de folie en toi. Je te remercie pour ta patience, ton écoute, tes conseils, tes encouragements, et pour ce quotidien que tu rends magique.

Et mes parents, à qui cette thèse est dédiée. Je ne peux terminer sans évoquer le décès de mon Père qui a eu lieu alors que je commençais à écrire. Mes parents m'ont toujours soutenu. Pendant mon enfance, ils ont répondu à toutes mes questions sans se lasser, ne restreignant par là jamais ma curiosité. Le présent document et le titre qui l'accompagne sont le fruit de leur engagement et de leurs enseignements. Chère Maman, je te trouve très courageuse. J'espère que cette bonne nouvelle qui m'arrive soulagera même un tout petit peu ta douleur. Tu es incroyable, et je ne saurai jamais te remercier pour ce que tu m'as offert. Cher Papa… tu me manques tellement. J'aurais voulu que tu voies ça.

Yverdon-les-Bains, 25 novembre 2014 Y. R.

Contents

Preface

This work gathers its methods, sources and intuitions from two broad academic disciplines which are commonly considered as independent from one another: mathematics and literary analysis. This is typical of the transverse discipline called *digital humanities*, within which this thesis falls. In this preface, we attempt to outline the goals of this work and the means used to reach them.

In brief, our goal is to better understand how the characters of a fiction novel are arranged within a literary text, and how one's position can interfere with the others' and vice versa. Once we have a text and have identified the characters in the text as the objects to study, then how do we extract the *structure* they form: more globally, the *character-system*? Here is our approach.

We begin with a text, more precisely a fiction novel, characters and all. The study of the character in a fiction novel has long been considered a subject of secondary importance in literary analysis. However, in 2003 Alex Woloch proposed a framework to study them, defining the *character-space* as the narrative environment of a character and the *character-system* as the union of all these spaces. Together, they form a representation, a subset of the novel, of its structure. The mathematical way to illustrate a structure—in our case a relational one since we are showing when the characters interact, share a scene or are simply mentioned close together—is the graph. It is composed of two things: nodes and edges. The nodes are unique elements, having no position and no required features. We then pair some of these nodes with edges. The mathematical object created by these nodes and edges is coined a *graph*.

Our approach is outlined in a mathematical fashion in figure 1.

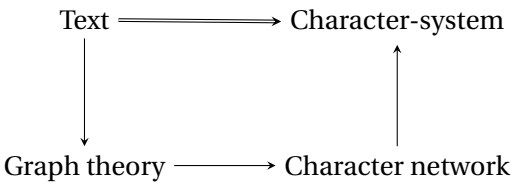


Figure 1: Overview of this work.

Born 278 years ago, the science of graphs—*graph theory*—has been extensively studied, and we can make the most of this. Thus, we model the character-system with a graph whose nodes are the characters and the *relations* between them are represented by the edges. We call this interpreted graph a *character network*, and we can apply to it methods from graph theory and its derived disciplines such as network science, social network analysis, etc.

So far, we have superficially covered the whole diagram in figure 1. In summary, to understand the character-system of a text, we represent it with a character network, borrowing techniques from graph theory, then analysing that object to draw conclusions on the character-system. We can now go into the details of the diagram, which will guide us through the obstacles we had to overcome and the different stages reached during our work.

In this thesis, we use a single text: *Les Confessions* from Jean-Jacques Rousseau. This is sufficient to cover our goals, which are to develop methods for the creation of character networks, to understand some of the pre-existing methods from other disciplines relevant to this context, to study them, to apply them, and to draw conclusions on the character-system based on the character network. We move into a recently developed field called *character network analysis*, and while we borrow notions and questions from literary analysis, we also require notions and methods from graph theory. The result of this encounter is an ensemble of concepts and tools that constitute a basis for the description of character network analysis. In the following, we explain the process of *a character network analysis* step-by-step. The numbers, whether in the text or in the diagrams, refer to the related chapters and sections.

In the first step, a novel is chosen (chapter 4), and our first task is to transform it into a network (section 2.1). To facilitate this, we use an index of characters to build a table of occurrences (chapter 3). An index of characters, in its simplest form (the characters and the pages on which they appear, with no subheadings), can itself be seen as a representation of the character-system. It is raw data, but it contains two necessary dimensions: one made of the characters, the other one based on the textual environment. Graph theory allows us to transform this object into a bipartite network—a network with two disjoint sets of nodes—and then to project it on one of the two sets, thus highlighting the structure of nearby apparitions in the textual environment—the cooccurrences. To do this, we developed two methods: firstly a *naive* one (section 6.1), which then lead to the development of a more sophisticated method that overcomes the flaws of an approach focused on the page by considering coocurrences on overlapping couples of pages (section 6.2). This transition from the text to the mathematical object is highlighted in the diagram shown in figure 2.

Now that a graph has been created, in the second step we learn to understand it (figure 3). For this purpose, we dig deeper into the meaning of the edges which symbolise the relations between characters (sections 2.2 and 7.1). Once interpreted, the graph is seen as a *network*; more precisely a *character network* (sections 5.3 and 5.4). In parallel, we explore the methods relevant to our study, in order to apply them properly. In particular, we discuss the concept of

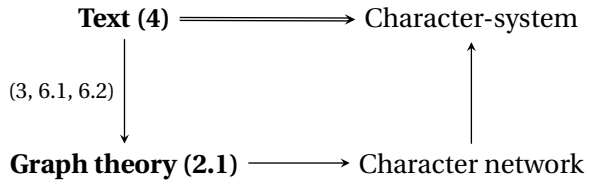


Figure 2: Step 1.

centrality (sections 2.3, 7.2 and 7.3) which has a wide range of forms and meanings. Centrality is at the core of our work, as it is a way to measure the *importance* of a node in the network. Importance being, for example, an expression for *popularity* (the number of relations) or a role played as an *intermediary* to enable the transition from one context or society to another in the discourse. We use centrality with characters just as social network analysis scientists use them with individuals. It is important to keep in mind that centrality has been used and discussed for years (the number of different definitions is never ending) but there is no unified theory encompressing all centrality indices. In this work, we remain cautious when using and interpreting them.

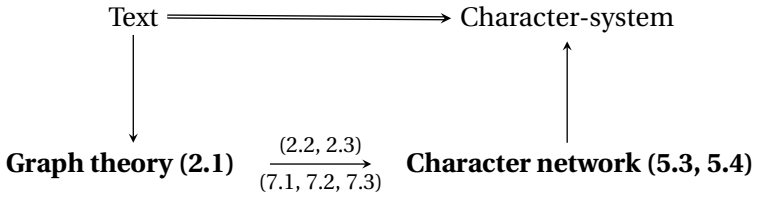


Figure 3: Step 2.

The next step—the third and last one (see figure 4)—eventually shows us a model of the character-system (sections 5.1 and 5.2). The character network is a representation of the plot via its characters. Using how the characters are positioned relative to one another—e.g. isolated, in the periphery or in the center—and how their occurrences are distributed—e.g. all in one chapter or spread over many chapters—we gain knowledge about their roles, and thus about the way the plot is structured. The author arranges the characters in a meaningful way, consciously or not, having them interact *orally* or *physically*, co-appearing in similar scenes, or, on the contrary, never in proximity. We hypothesise that, based on a character-system, a simple character network models the system's characters and their interactions. However, this does not cover all aspects of the character-system[1]. Depending on the source, or on the construction method, the character network can be augmented to include supplementary

[1] Prof. Matthew Jockers brought to our attention that a novel composed of only character would hijack the limits of the modelling. It is clear in such a case that character network analysis is not a relevant tool for modelling such

information about the character-system. For example, a decomposition of the network into subnetworks covering chapters (appendix B) or overlapping parts of the text (subsection 8.2.1), a topic modelling of the text surrounding characters' occurrences, or a sentiment analysis of the textual contexts in order to characterise the relations between characters. In the future, further work in these non-exhaustive directions will allow us to better assess how well the model matches Alex Woloch's ideas.

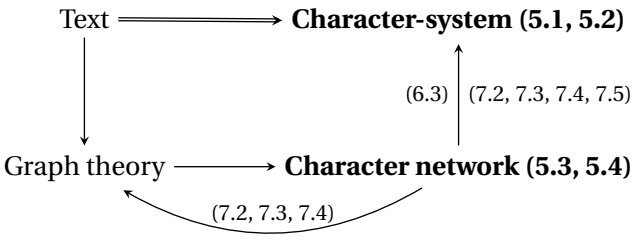


Figure 4: Step 3.

We use four classic centrality indices to differentiate the roles in the character network: degree, betweenness, harmonic closeness and eigenvector centrality. We show how these indices behave in the network and how the results correlate to the discourse as well as the story world. In the sections dealing with the computation and analysis of centrality measures (sections 7.2, 7.3, 7.4 and 7.5), we apply them to the network, then compare the results on characters according to one or more centrality measures. We also study concepts related to centrality, such as community detection and vitality. In this way, we describe the roles of the characters using a point of view based on their cooccurrences in the text, a form of narrative proximity.

The study of how centrality indices behave in a character network lets us improve, if even by a small margin, our knowledge of the network. This is symbolised in the diagram (see figure 4) by an arrow pointing backwards. Centrality measures, which all measure importance or, intuitively, the fact of being located in the center of the object modelled by the network, behave differently in this literary context, allowing us to characterise the novel's characters, and thus form a typology of roles.

This thesis offers an essential contribution to digital humanities: the definition of a methodology to conduct character network analysis of fiction novels. Moreover, it suggests a discovery: centrality measures behave differently depending on the types of characters in a character network. This last observation introduces some of the next steps. One will be to show the results using the method on other novels, while another will be to study character

a character-system, with the exception of showing that character relations are not a notable aspect of this novel's character-system.

network analysis systematically and on a much larger corpus. For the latter, we may need to define a larger, universal framework.

1 Introduction

> M. Slope, however, on his first introduction must not be brought before the public at the tail of a chapter.
>
> *Barchester Towers*
> ANTHONY TROLLOPE

In 1871, French author Émile Zola began a series of novels, *Les Rougon-Macquart*, that would end in 1893 with a total of twenty units. He set himself the task of describing an epoch, the *Second French Empire* (1852-1870), using the points of view of different characters in nearly each novel. The contexts and characters would span the aspects of the life at the time: political, cultural, social, urban and rural, etc. Despite not having all of the story lines ready from the start, not even the overall length of the series, Zola organised his project around a fictional family, for whom he had sketched a genealogical tree: all individuals had unique characteristics (some were hereditary traits) and were attributed a wide range of roles in society. The planned *character-system*—the set of all characters along with their relations—was the fundamental building block from which he would construct the rest of his extended narrative. In the preface of the first novel, *La fortune des Rougon*, he outlined the importance of the relations between the characters of his then developing work:

> Je veux expliquer comment une famille, un petit groupe d'êtres, se comporte dans une société, en s'épanouissant pour donner naissance à dix, vingt individus qui paraissent, au premier coup d'oeil, profondément dissemblables, mais que l'analyse montre intimement liés les uns aux autres. L'hérédité a ses lois, comme la pesanteur.
>
> Je tâcherai de trouver et de suivre, en résolvant la double question des tempéraments et des milieux, le fil qui conduit mathématiquement d'un homme à un autre homme. Et quand je tiendrai tous les fils, quand j'aurai entre les mains tout un groupe social, je ferai voir ce groupe à l'oeuvre comme acteur d'une époque historique, je le créerai agissant dans la complexité de ses efforts, j'analyserai à

> la fois la somme de volonté de chacun de ses membres et la poussée générale de l'ensemble.[1] (Zola, 1985, préface de *La Fortune des Rougon*, p.15)

Zola consciously created the fictive social network of family ties that would underlie his work to come. Since his goal was to give a representation of the society from that time, we may even consider this character network as an historical and social model of the given era. Zola also announced that he would build each narrative while considering the others waiting to be written or published, implying a narrative coherence among the twenty novels to come. Thus, with *Les Rougon-Macquart*, he built and used a multi-level relational structure between characters. The character-system is the representation of a society as well as the backbone of the narration. His approach was both literary and scientific.

A similar attention to structure was used by Gustave Flaubert in *L'éducation sentimentale* in 1869. In this novel, Flaubert used technical and metaphorical vocabulary to describe the role of the character Dussardier:

> Ils étaient, cependant, aussi liés qu'autrefois, et même ils avaient tant de plaisir à se trouver ensemble, que la présence de Dussardier les gênait. Sous prétexte de rendez-vous, ils arrivèrent à s'en débarrasser peu à peu. Il y a des hommes n'ayant pour mission parmi les autres que de servir d'intermédiaires; on les franchit comme des ponts, et l'on va plus loin.[2] [Gustave Flaubert, *L'éducation sentimentale*, chap. IV]

This is singular in that we cannot discern if this is the narrator's or the characters' point of view. However, we do know that the point of view is distant. In this case, the role of the cited character helps the narration move from one scene to another, binding a group of characters with another one. Equally, he helps some other characters to meet each other, before finally disappearing. This precise figure—or narrative role—happens inside and outside the story world. In social sciences, for instance, the group of characters would be modelled by social

[1] "I wish to explain how a family, a small group of human beings, conducts itself in a given social system after blossoming forth and giving birth to ten or twenty members, who, though they may appear, at the first glance, profoundly dissimilar one from the other, are, as analysis demonstrates, most closely linked together from the point of view of affinity. Heredity, like gravity, has its laws.

By resolving the duplex question of temperament and environment, I shall endeavour to discover and follow the thread of connection which leads mathematically from one man to another. And when I have possession of every thread, and hold a complete social group in my hands, I shall show this group at work, participating in an historical period; I shall depict it in action, with all its varied energies, and I shall analyse both the will power of each member, and the general tendency of the whole."

Translation by E. A. V. Merton, edited by Ernest Alfred Vizetelly, (publisher unknown), 1898, see http://www.gutenberg.org/ebooks/5135, accessed on 28/05/2014.

[2] "They were, however, on the same intimate terms as before, and they found so much pleasure in each other's society that Dussardier's presence was an obstacle to their free intercourse. Under the pretence that they had appointments, they gradually got rid of him. There are some men whose only mission amongst their fellow-men is to serve as go-betweens; people use them as if they were bridges, by stepping over them and going on farther."

Unknown translator, edited by Dora Knowles Ranous, New York: Brentano's, 1922, see https://archive.org/details/sentimentaleduca00flauiala, accessed on 28/05/2014.

networks, and the narrative role of Dussardier measured by the betweenness index, which quantifies how much an individual in the network helps the other individuals interact one with another. Flaubert reveals that this intermediary role may exist simultaneously, both in the story and in the discourse.

In these two examples, classic authors broach how they meticulously crafted the narrative environment of their novels. Specifically, they place characters at the centre of their narrations. They build relations between characters, and in doing so they imply the existence of the character networks behind their respective plots. The purpose of this doctoral work is to go the opposite direction: to extract the networks of relations between characters, and to study them. We deconstruct them in order to show the intended or unintended underlying structures that they form in narratives.

1.1 Literary studies and mathematics

Exchanges between the literary and mathematical worlds have existed for a long time. For example, at the end of nineteenth century, Lewis Carroll, author of *Alice's Adventures in Wonderland* and a mathematician, used to borrow ideas from mathematics for his writings (Devlin, 2010). In addition, years later in France, a group of writers and mathematicians would form the OuLiPo[3]: a movement interested in experimenting with combinatorics and word play in literary works. The possible arrangements of words in narratives are infinite, and the idea of creating or discovering rules for the writing process is a complex task—both creatively and scientifically. Therefore, we ask: Is it possible to show the underlying construction of the narrative process? Which methods must we use? Can we obtain visualisations of certain aspects of the narrative? What kind of analyses can we conduct on the resulting representations? Is a quantified approach relevant? How do we interpret the results?

Alex Woloch created a literary theory of the character focusing on the usual asymmetric organisation of the discourse, which opposes the respective importance of protagonists and minor characters (Woloch, 2003). From the introduction of his study, he defines characters' role and importance as dependent on their relations with the textual surroundings:

> Lykaon, Thersites, Achilles, Diomedes, Odysseus, the anonymous men fleetingly registered in the Catalogue of Ships: over and over again we are presented with different inflections of individuals into the total work of art. The rich diversity of these characters—the multitudinous ways in which the *Iliad* comprehends the human—depends on each character's structured position within the literary totality, or the narrative space that he occupies. In each instance, the character's referential personality—the unique sense and abiding impression that the character leaves us with—emerges in-and-through, not despite, his textual position and

[3] *Ouvroir de littérature potentielle* (Oulipo, 1988).

> the descriptive configuration that flows out from this position. (Woloch, 2003, p. 12)

The perception of a character in the reader's mind depends on the narrative entities close by (e.g. the other characters, the places, the time of the action.) An abstract view of the character as immersed in the narrative is provided by the distribution of its occurrences in the text—the occupied narrative space—and those of the characters nearby. The same is valid for all the characters: each one induces what Woloch coins a *character-space*, i.e.

> [...] that particular and charged encounter between an individual human personality and a determined space and position within the narrative as a whole [...] (Woloch, 2003, p. 14)

Taken altogether, they constitute a society of characters in the story world as well as in the discourse. In the second case, where the focus is on the resulting text, this is coined the *character-system*, i.e.

> [...] the arrangement of multiple and differentiated character-spaces—differentiated configuration and manipulations of the human figure—into a unified narrative structure [...] (Woloch, 2003, p. 14)

As the main objective of our work, this is what we circumscribe, extract, build, analyse and interpret. Our intention is to show the underlying relational dimension binding the characters in a narrative.

Researchers have shown an interest in quantified approaches to literary studies for a long time. This began with the studies of Roberto Busa, using computer science methods on the works of Thomas d'Aquin[4] from the mid-twentieth century. Since then, we have seen an increase in the use of quantified approaches on different scales, and in proportion to the rise of computational power (Moretti, 2000, 2005; Michel et al., 2011; Jockers, 2013).

The unabashed integration of digital methods in literary studies eventually allowed us to consider the use of relational methods from social science on literary works. Recently, Franco Moretti published an exploratory study attempting to *reverse engineer* the authors' narrative intentions—in his case *Hamlet*. His study, presented as a pamphlet to emphasise the exploratory and sometimes intuitive approach[5], focused on the characters, and led him to ask:

> [...] if you work on novels or plays, style is only part of the picture. What about plot – how can that be quantified? (Moretti, 2011)

[4] Incidentally, this is considered as the first work in the field of *digital humanities*—or at least *humanities computing*.

[5] "The theory proper requires a level of mathematical intelligence which I unfortunately lack; and it typically uses vast quantities of data which will also be missing from my paper. But this is only the first in a series of studies we're doing at the Stanford Literary Lab; and then, even at this early stage, a few things emerge." (Moretti, 2011, p. 2).

In order to answer this question, he used networks to model the interactions between the characters, then used methods from social network analysis (average path length, centrality, clustering) to interpret the resulting model. Globally, he questioned the use and reach of network analysis for literary studies. Resulting from this preliminary study, he emphasised the abstracted power of the model and summarised this original use of networks:

> [...] I am discussing Hamlet, and saying nothing about Shakespeare's words – but also, in another sense, much more than it, because a model allows you to see the underlying structures of a complex object. (Moretti, 2011, p. 4)

We adopt the same approach: we take a view on the novel and its plot[6] which does not require the text itself for the description nor the analysis. However, to some extent, we refer to the story and to the discourse for analysis and interpretation of the character-system.

1.2 Goals and methods

Our objective is to reveal the relational structure linking the characters in a discourse, using the text obtained from the narration of a story. We call the result a *character network*: the literary equivalent to the *social network* in social science. The character network is a model of the character-system defined by Woloch. It reveals how the narration puts forward some relations and minimises or even hides others. During the process of building the character network, we attach to each edge a measure of the intensity of each relation[7]. Our framework allows us to trace the state of a narrative relation through time. The importance of the characters is deduced from the structural knowledge gathered through the existing and missing relations. The framework defined to study character networks—we it call *character network analysis*—is the following (see section 5.4 for development):

1. Determination of a corpus.
2. Identification of the nodes.
3. Identification of the relations.
4. Extraction.
5. Analysis and interpretation:
   - 5.1 At the character-level.
   - 5.2 At the system-level.
   - 5.3 Of the network structure.

[6] In this work, the *plot* is the set of events and actions that take place along the narrative, being interlocked or not. The connection between the plot and the characters will not be discussed in detail: in the end, we shape a plot with a character-system, itself shaped by a character network. This approach has a possible drawback, missing here but that need to be addressed in a following study: modelling the plot of a novel with an uncrowded character-system will probably result in a degenerate object. A definition of the *plot* could be done in accordance with that case.

[7] In graph theory—the mathematical field behind—this is a weight attribute.

5.4 Of its temporality[8].

Firstly, the choice of a corpus is a step that must not be ignored. In most cases, we have to keep in mind that there may be omissions of narrative links at a higher level, for example when the analysis is restricted to a scene from a play, to a chapter of a novel, or to a novel belonging to a larger series of novels. However, this situation is admissible if the study is defined and designed to focus on the organisation of a given part, and thus does not intend to include bypassing relations.

The identification and extraction of the elements constituting the character network—the nodes—then ensue from the corpus. The extraction of the relations is a key step and many choices are possible. For example, we could link them when, from reading the book, we feel that they physically meet in the story world, or else we could automatically link two characters if they appear on the same page. The research question must designate the methods to use, e.g. automatic or manual extraction? with or without semantics? In some cases, the research may focus only on the method—e.g. automatic extraction—and later deal with the implication and interpretation of the links, or sometimes even ignore it. We discuss the matter of determining edges in chapters 5 and 6.

In this work, we build a character network from co-occurrences of characters as they appear in book indexes. This original approach bears its own design questions and constraints, which we handle in sections 6.1 and 6.2. The relations are based on text distance, with the page as unit. Thus, they depict a form of narrative proximity. The identification of the characters in the text, as well as the mining of their positions, is delegated to the preliminary work of editorial scholars and their setting up of the book index.

At this stage, the network exists: it has nodes and edges and is ready for the analysis. We adapt concepts and methods from social network analysis to the context of character networks: they are close objects, since in both cases nodes represent human figures and edges relations between them.

The choice of networks to model character-systems ensues from the ubiquitous presence of networks in today's research:

> To name but a few examples, 'network analysis' is carried out in areas such as project planning, complex systems, electrical circuits, social networks, transportation systems, communication networks, epidemiology, bioinformatics, hypertext systems, text analysis, bibliometrics, organization theory, genealogical research and event analysis. (Brandes and Erlebach, 2005, p. 1)

The importance and implication of network analysis in most fields of research, at least those with relational aspects in their data, show the pertinence of this long ignored discipline in

[8] In this work, we only sketch that point (see the conclusion, section 8.2).

applied research. This recently led to the creation of a journal called *Network Science*, in which the editors claim:

> We believe there is a distinctive science of networks that crosses traditional disciplinary boundaries. It is ready to be brought together in a coherent form that transcends disciplinary silos. (Brandes et al., 2013, p. 13)

This is the vast mathematical context within which character network analysis falls.

Node centrality[9] is one of the most important concepts we import from social network analysis into this study. It is expressed in graph theory language by families of measures computed from the network structure. In social sciences, centrality enables us to quantify the importance of an actor and the dimensions of the role it plays in the network. It can take into consideration attributes on relations, such as measures of intensity, to develop more fine-grained definitions. The property shared by most centrality measures is that they recognise, for a given set of nodes, the centre of the star-shaped graph as owning the position that maximises the centrality among all possible networks of that order[10] (see figure 1.1).

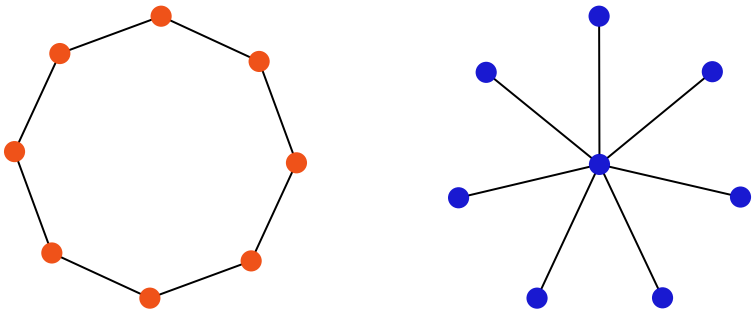


Figure 1.1: (Left.) In this ring-shaped graph, all the positions are structurally equivalent. (Right.) While in this star-shaped graph, a single node connects all the others. It is central.

Each centrality measure focuses on a dimension of the structural importance of the nodes, with input ranging from the immediate neighbourhood to the whole network. The concept of *centralisation* sums up these to the network-level, allowing us to evaluate to which extent a network possesses a centralised structure. Practically, centralisation is a function of a centrality measure: it provides as many angles on a network as there are measures of centrality.

Centrality is a widespread but catch-all concept. For more than fifty years, it has been a label of patchwork studies, not a mathematically consistent theory. In this context, we must

[9] Centrality can also be computed on edges, arcs, or groups of nodes.
[10] The order is the number of nodes.

choose complementary measures: we use centrality and the associated concepts—indexes and centralisation—to measure narrative aspects of the role of a character in the discourse. Chapter 7 is devoted to study how these measures behave and how they are to be interpreted. This step borrows concepts from narratology—the study of the transformation of a *story* (what is in the writer's mind) into a *discourse* (the textual form). We call this transformation process the *narration*[11]. With an appropriate data collection approach, character network analysis can answer questions about how the narration turns the "social network" of the story world into a character network in the discourse. Are they equal? Do they have more or fewer relations than their neighbours? Which individuals have been turned into influential or passive characters? What relations are rendered as prominent by the narration?

1.3 Reach

Our corpus for the exploration of character network analysis is the autobiographical novel *Les Confessions* by Jean-Jacques Rousseau, subject of chapter 4. It contains many relevant properties for the scope of this study:

- It is composed of a large set of characters, corresponding to a wide range of roles in society as well as in Rousseau's life.
- We possess external informations about the story[12] of Rousseau's life (Rousseau et al., 2011, 2012), which offer a second look on the results, after interpreting the character network analysis.
- The edition we use (Rousseau et al., 2012) comes with an index of names.
- The autobiography tells the story chronologically, thus allowing the use of the pages as a monotonous temporal scale.

We assert that character network analysis is applicable to every narrative conceived with characters as a constitutive element of the discourse. The theory we develop in chapter 5 and the analysis we conduct on *Les Confessions* in chapters 6 and 7 define character network analysis. We construct a bridge from the structuralist study of characters to a network analysis approach, illustrating it with an example that raises and also answers analytical problems and questions. In a discipline—narratology—that considers literature from a structural point of view, we still observe a gap with its approach to character studies:

> Whereas the study of the story's events and the links among them has been developed considerably in contemporary poetics, that of character has not. (Rimmon-Kenan, 2002, p.29)

The character-space and the character-system of Alex Woloch are parts of a methodology that allows the narratological study of the characters. His approach, despite using mathematical

[11] We base our understanding of narratology on (Bal, 1984; Rimmon-Kenan, 2002; Genette, 2007).
[12] And history (Genette, 2007).

language, is a work of literary theory. In our work, we deal with the problematic of the character from an opposite point of view: we adapt the system of Alex Woloch and translate it with networks and measures; we use them to dissect the network representation of the characters in the discourse and to some extent of the plot.

In a sense, we already know everything about Shakespeare's or Rousseau's works. They are among the most studied authors, along with a few hundred others to whom many scholars have devoted their careers, having adopted close-reading approaches. Thus, we may not want to use a network to model the facts that we already know from historical sources and literary studies, rather build it uniquely on the text, without these influences, and use it as the explanatory variable of an exploratory relational research. Franco Moretti's ambitions are going in this direction when he states about his character network analysis of *Hamlet* that:

> The idea behind this study [...] was, very simply, that network theory could offer a way to quantify plot, thus providing an essential piece that was still missing from computational analyses of literature.
>
> [...] What I took from network theory were less concepts than visualisation. [...] Basically, I used (or mis-used) the theory [...] as a way of arranging literary data that presupposed a principle of order - but not a full conceptual architecture. (Moretti, 2011, p.11)

He implies that researchers can have more ambition with network analysis than what he presents[13]. In our work, the use of concepts and descriptive statistics is expressed, for example, by the computation and analysis of centrality measures on the network extracted from *Les Confessions*. These measures highlight parts of the network, therefore parts of the narrative, where particular patterns can be seen, thus also showing interesting narrative combinations. Still, like many studies, it shows a particular focus on visual representation: it is composed of twelve pages of text, along with thirty pages of figures, all showing network visualisations.

Another use of network analysis in the context of literary studies is shown by the exposure of figures illustrating the underlying network of narrative relations, as well as the use of network data for an exploratory purpose. The positioning of nodes and relations with an appropriate layout algorithm offers a preliminary reading of the plot. This is an aspect of character network analysis we discuss with visuals highlighting characters with specific narrative roles or visuals of the plot relational backbone. The interpretation of a network visual is less rigorous than a statistical analysis, thus inferring results from them is an hazardous task which the researcher must be aware of.

For instance, the website *Moviegalaxies*[14] (see figure 1.2) offers a glimpse into the educational stakes of providing a network to "summarise" a plot—an admitted goal of its authors[15].

[13] The text is labelled as "pamphlet", thus allowing a slightly less rigorous tone.

[14] moviegalaxies.com

[15] moviegalaxies.com/research, accessed on 28/03/2014. See also (Harrington et al., 2013).

It consists of a library containing the character networks of 775 movies[16]. For each, the corresponding network shows the structure of the interactions between the characters, as well as categoric and numeric indicators, e.g. community detection, density, diameter, clustering and average path length. The objective is also to reveal the complexity of the scriptwriters' works. Finally, it gives the potential spectator a taste of what to expect from the movie before watching it. This library of movie plots is a tool we would love to see used for novel plots—and the present work is a step in that direction. For instance, their database is an invitation to build a typology of plot structures, and see how they are distributed through a classification by genres. We think that character network analysis will allow us to propose a typology of novels based on their character networks. In the future, it will help constitute an atlas of narratives and compare plot structures.

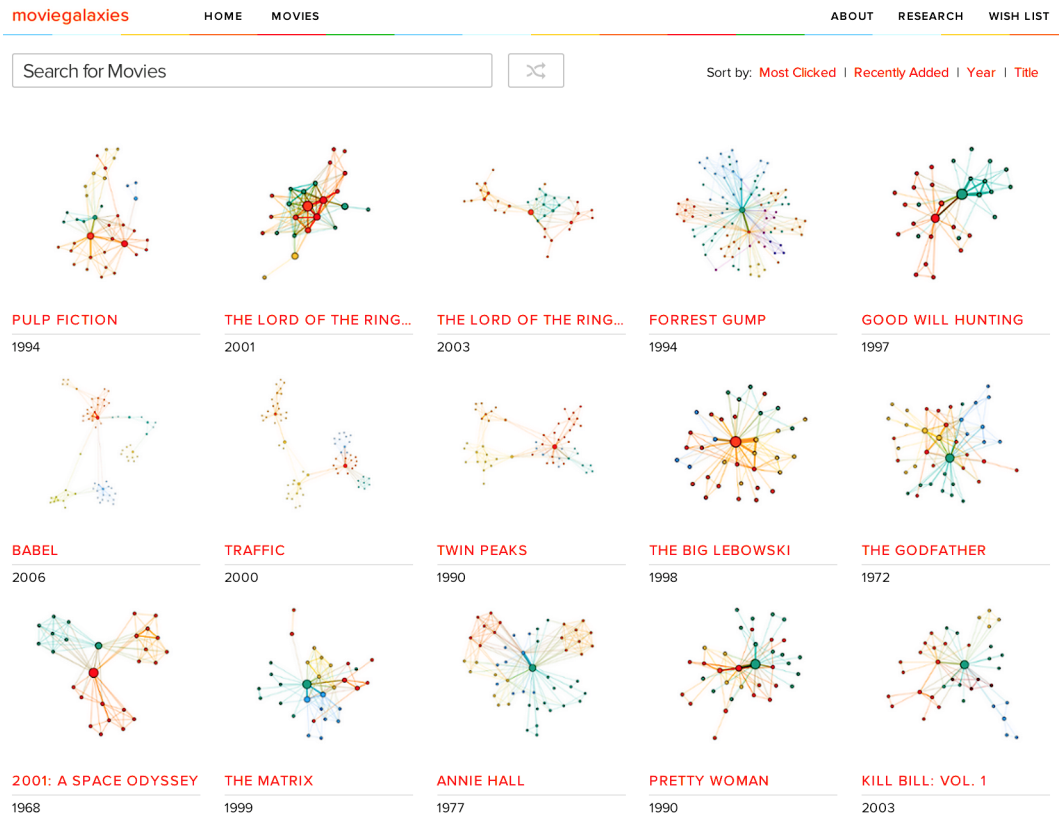


Figure 1.2: Screenshot of website *Moviegalaxies* interface[17].

We use an index to extract from a text the list and positions of characters, as well as the underlying relations between them. The indexation is usually processed by professional in-

[16] Two thirds of them date from the last twenty years, as of 28/03/2014. The character networks appear to have been automatically extracted from the *Internet movie script database* (imsdb.com).

[17] Accessed on 21/05/2014.

dexers. To some extent, for example in a play where all the names systematically appear at the beginning, the indexation process can be automatised. Then, relations are extracted from co-appearances. The index extraction is a fast method when compared to the manuel extraction of relations, which demands well described methods that are consistently applied on the text. The more complex the method, the higher the risk of inserting biases in the data. Nonetheless, for the purpose of exhaustivity, we review the methods of algorithmic extraction of the relations. They start with the recognition of the character, then with the potential relations. They have the advantage of extracting the semantic context: for example, it is possible to attach to the relations an attribute quantifying sentiment. However, they also produce mistakes, and may be limited in some cases, for example in *Les Confessions* when, in an intimate context with her, Rousseau calls his mentor Françoise-Louise de Warens "Maman[18]", a common word difficult to detect automatically and to attribute to the corresponding character. A network extraction process based on an index allows a character network analysis study like ours on every text owning one, which is for instance the case of the majority of biographies or literary works edited by the collection *Bibliothèque de La Pléiade*[19].

Character network analysis incorporates any valid method chosen for the extraction. It necessitates a literary work and a network representation to work on. Still, the understanding and description of the network extraction method is required for the results to be relevant. These stakes are important in relation to character studies because they offer—like the movie narratives mentioned earlier—indicators, or summaries, that imply a high range of applications in literary studies.

Applications of character network analysis are diverse, as is the case for social network analysis, where a research project studies a single static society, models the evolution of a society, or compares parallel societies. In the present work, we study a single text, but it is possible to compare literary works (Mac Carron and Kenna, 2012, 2013), or study the evolution of a plot via the evolution of the relations between the characters (Moretti, 2011).

The first case is close to a classic literary study of the text. We describe and analyse a novel with a new analytical tool; however, in the end it is a commentary on the discourse and the narration as it shows the big picture as well as highlights motifs in the narrative structure (centrality, community detection, principal component analysis). In addition to the networks, distributions of occurrences (see figures 4.4 and appendix C) or numerical indicators (see tables 4.1 and 4.2) also provide information used to interpret the character network. This new reading of a literary classic proposes scientific novelty in some of its results (roles of characters, properties of the narration), but also in its approach–quantified and relational. The range of literary genres is wide, e.g. plays (Moretti, 2011), novels (Agarwal et al., 2012), a set of fictional texts like the complete works or an author or texts belonging to an ensemble

[18] "Mum".
[19] Published by *Éditions Gallimard*.

(Mac Carron and Kenna, 2012, 2013), historical reports and articles from newspapers (Kaplan and Stylianou, 2013). In most cases, all these forms of writing have the following in common: narratives integrating characters, persons, protagonists, people dialoguing or just appearing side-by-side.

The second case, the comparison or classification of literary works, is already a frequent task in qualitative literary studies:

> Cependant on s'aperçoit rapidement que tout effort de classement, toute tentative typologique un peu élaborée aboutit presque toujours, surtout chez les théoriciens modernes, et surtout dans les études portant sur le roman, à classer les oeuvres narratives d'après leurs personnages[20] [...] (Hamon, 1998, p. 9)

A focus on how central the positions of the characters in narratives are allows us to reduce the question of establishing a typology of narrative works to a study of characters, following the observations of Hamon. The character-system, thus the novel's character network, is a dual representation of the narration—we use it to classify literary works. A first and immediate step to take after this work will be the gathering and treatment of all existing indexes of fiction novels. This does not even require opening the books[21], bringing us close to what Moretti coins *distant reading* (Moretti, 2013a). A systematic study on a large-scale of all written fiction novels, for example, would offer an overview of some of the existing literature and allow a classification of these works based on character networks.

In the third case, we divide the narration into parts and analyse how it evolves, how the author positions his characters, modifies the relations, and manipulates the perception the reader has of the story. We discuss how to apply this approach to *Les Confessions* in the *future works* section of our conclusion (see section 8.2). We believe that the character network analysis approach is relevant in literary studies, and highlights an unnoticed—or at least unmeasured—dimension of narratives.

1.4 Overview of contributions

The original contributions of our work are:

- The definition of character network analysis, along with a discussion on the use of quantitative methods in literary studies.
- The description of the extraction of character networks with an index.
- The analysis of the roles of characters inside the whole narrative, or in parts of it (e.g. chapters).

[20] "However, we quickly realise that every effort to rank almost every elaborated typological attempt nearly always ends, especially for modern theoreticians, and particularly in studies about the novel, in classifying narrative works with respect to their characters." [Own translation.]

[21] Except for the back of the book pages in order to scan them.

- A better understanding of some applied mathematics methods, especially centrality (graph theory on literary theory).

1.5 Outline

> "Et ensuite on se trouvera un… chouette hôtel chic. Et on régalera."
> "[En se tournant vers la caméra] Vous voyez ? Elle pense qu'à rigoler."
> "À qui tu parles ?"
> "Aux spectateurs."
> "Ha."
>
> *Pierrot le Fou*
> JEAN-LUC GODARD

Figure 1.3 shows the organisation and dependencies of the chapters.

Chapter 2 presents graph theory and social network analysis.

Chapter 3 is a short contextualisation of the work of indexing, with dedicated sections on indexing biography and autobiography.

Chapter 4 presents the life of Rousseau, the novel *Les Confessions*, and its role and importance in his life. It also describes the index (Rousseau et al., 2012).

Chapter 5 reunites the theoretical aspects of this work, and links literary and character studies with network analysis.

Chapter 6 describes how to extract characters and their narrative relations from an index, build the network model, and start an overall analysis. It provides a case study in the comparison of character networks extracted from two given chapters.

Chapter 7 makes use of centrality indices to depict and explain narrative roles of characters or groups of characters, based on the network model.

Chapter 8 presents our conclusions as well as ideas for future developments, such as using temporality in character networks to reconstruct the narration.

All citations in French or German were translated in English, with the exception of epigraphs.

Scripts and data used in this work are available on github:

https://github.com/yrochat/phd_thesis

Scripts are written in *R* (R Core Team, 2013) and rely heavily on the *igraph* package (Csárdi and Nepusz, 2006; Kolaczyk and Csárdi, 2014).

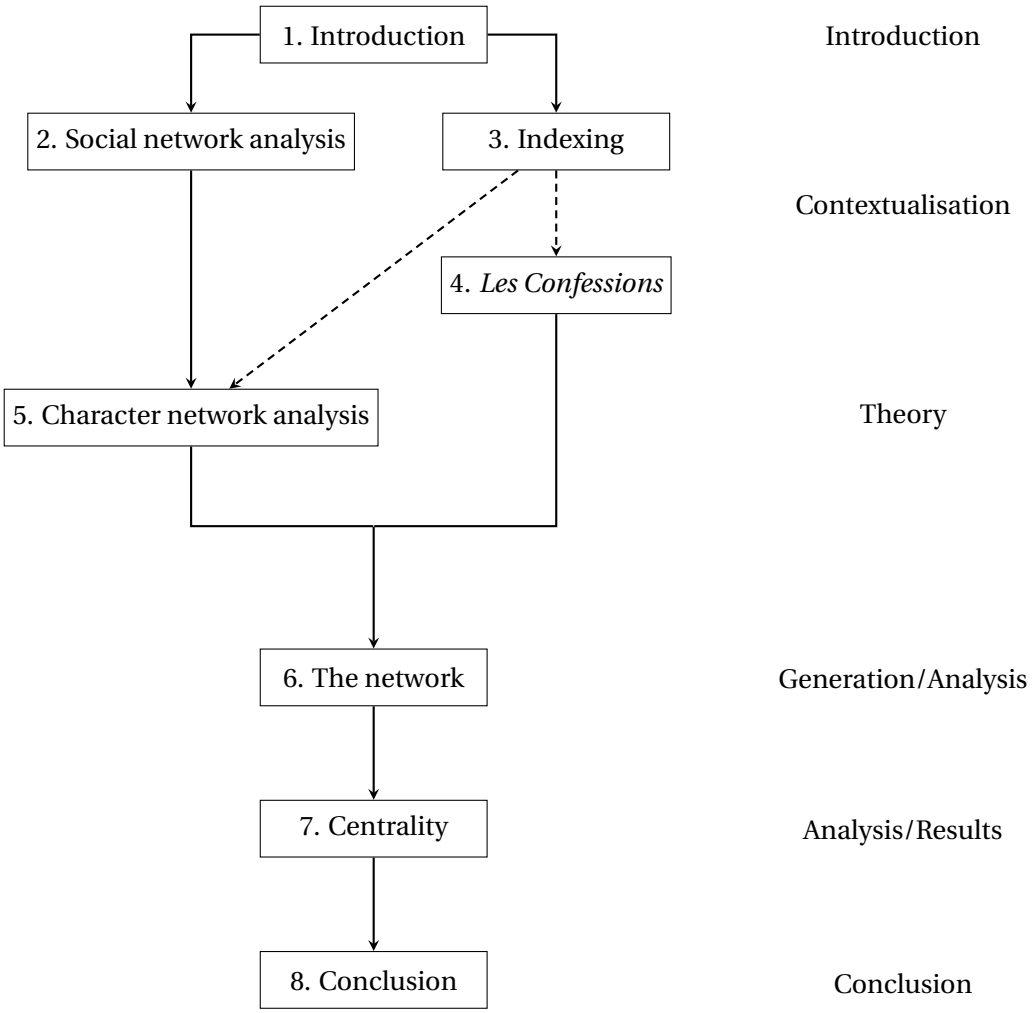


Figure 1.3: The organisation of the thesis.

2 Social network analysis

> "Je vous ai beaucoup aimé, mais je vous ai peu donné, mon pauvre ami."
> "Pardonnez-moi Françoise si au mépris des règles de ce genre littéraire, j'interromps une confession que j'aurais dû écouter en silence - m'écriai-je en essayant de plaisanter pour la calmer, mais en réalité mortellement triste."
>
> *Avant la nuit*
> MARCEL PROUST

Literary network analysis is the meeting of social network analysis and literary studies. Social network analysis is defined by (Hennig et al., 2012) as "[the study of] social phenomena by means of network representations." Networks went from entertaining combinatorial puzzles a few centuries ago to eventually becoming an independent and mature mathematical field called graph theory. In the mid-20th century, graph theory started being adopted by researchers in psychology and in social sciences; their experimentations gave birth to social network analysis, which continued to spread through humanities until the end of the century. Then, thanks to particularly wide use in statistical physics and life sciences, combined with the interest of influential scientific journals (Watts and Strogatz, 1998; Barabási and Albert, 1999; Oliveira and Barabási, 2005; Borgatti et al., 2009; Lazer et al., 2009), network models influenced new actors—and still are today. Recently, an attempt to recognise the interdisciplinary range of networks has been launched. The resulting field was coined *network science*[1]: it is now making its way into other disciplines that rely on relational data.

In this chapter, we discuss the graph theory (2.1) methods necessary for our network approach. We bring to light some famous results (small-world experiment, strength of weak ties, structural holes) (2.2), and pay particular attention to the concepts of centrality measures (2.3). The theory behind literary network analysis—and character networks in particular—will be explored separately in chapter 5.

[1] A journal with that name was launched in April 2013 (Brandes et al., 2013). The editorial team brings together influential scholars from very diverse fields: http://journals.cambridge.org/action/displayJournal?jid=NWS (Accessed on 06/03/2014.)

2.1 Graph theory

Social network analysis relies on graph theory to model and to explain relations in a social context. Graph theory is a mathematical discipline which originates from the "Seven Bridges of Königsberg[2]" problem, which Leonard Euler solved in the 18th century. The famous mathematician provided the solution to a puzzle that Königsberg's inhabitants were struggling with: the city was split by a river, and seven bridges connected the two banks and two islands in the city centre[3] (see figure 2.1, left). The inhabitants were curious if a path existed that would take them through the city, using each bridge only once, starting and ending at the same point. Leonard Euler proved that by representing the banks and bridges as vertices and edges—i.e. the "points" and "segments" of figure 2.1 (right)—there is no solution to this problem (Euler, 1736).

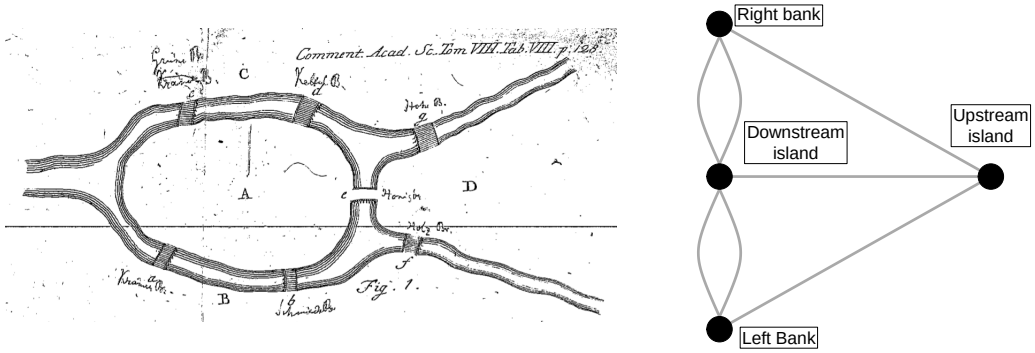


Figure 2.1: The seven bridges of Königsberg problem. (Left[4].) Geographical and (Right.) graph-theoretical representations of the two banks, two islands and seven bridges mentioned in the problem.

In the rest of this section, we introduce some required elementary concepts of graph theory. They are the low-level mathematical language behind network analysis, and thus are necessary for the constructing, computing and interpreting parts of chapters 5 to 7.

Definition 1 (Graph)**.** *A* graph *is a couple $G = (V, E)$ with V a set of vertices—equivalently called nodes—and E a set of edges. E is determined by an application*

$$e : V \times V \rightarrow E$$

defining each edge by binding together two elements of V. These two nodes are called the extremities *of the edge.*

[2] Today Kaliningrad.
[3] Some bridges have since disappeared.
[4] Left figure is taken from (Euler, 1736).

We see a *network* as an interpreted graph; it carries interpretations for both sets of vertices and edges on the contrary to a graph.

Definition 2 (Order, size)**.** *The number of nodes in a graph is the* order *of the graph. The number of edges is the* size *of the graph.*

Definition 3 (Density)**.** *The value* $\frac{2|E|}{|V|(|V|)-1}$ *is the* density *of the graph. This is the division of the number of edges in the graph to the theoretical maximal number of edges.*

Definition 4 (Adjacent, incident)**.** *Two vertices are* adjacent *if an edge exists between them. A vertex positioned at an edge's extremity and this edge are* incident.

Definition 5 (Loop)**.** *A* loop *is an edge with confounded extremities.*

Definition 6 (Multiple edge)**.** Multiple edges *are edges sharing the two same extremities.*

Definition 7 (Simple graph)**.** *A graph is* simple *if it owns neither loops nor multiple edges.*

Definition 8 (Directed graph, arc, undirected graph)**.** *A graph is* directed *if a direction is associated to the edges. In this case, they are called* arcs*. If no direction is defined, then the graph is* undirected.

All previous definitions are valid for directed graphs. Since the network we build in this thesis is undirected (see chapter 6), we do not consider directed cases for the rest of this chapter.

Definition 9 (Adjacency matrix)**.** *Let G be a simple graph of order n. The matrix* $\{a_{ij}\}_{1\leq i,j\leq n}$ *such that*

$$a_{ij} = \begin{cases} 1 & \text{if } (x_i, x_j) \in E, \\ 0 & \text{otherwise} \end{cases}$$

is called the adjacency matrix *of graph G.*

Definition 10 (Subgraph)**.** *A* subgraph $G' = (V', E')$ *of G is a graph defined on subsets of vertices and edges of* $G : V' \subset V$, $E' \subset E$*. Every edge of* E' *must have its extremities in* V'*.*

Definition 11 (Walk)**.** *A* walk *is a sequence of vertices and edges in alternation, starting and finishing with vertices.*

Definition 12 (Path)**.** *A* path *is a walk such that no vertex is crossed more than once.*

Definition 13 (Cycle)**.** *A* cycle *is a path that starts and finishes at the same vertex*[5]*.*

Definition 14 (Length)**.** *The* length *of a walk is the number of edges it contains.*

[5] More precisely, this is the definition of a *simple cycle.*

Definition 15 (Shortest path)**.** *A* shortest path *between two nodes is a path with minimal length between the nodes. Between two nodes, a given shortest path may not be the only such path.*

Definition 16 (Distance)**.** *The* distance *between two nodes is the length of one of their shortest paths.*

Definition 17 (Neighbour, open/closed neighbourhood)**.** *The adjacent nodes of a node are its* neighbours*. The subgraph induced by these nodes is the* neighbourhood *of the given node. It is* open *if the node is not included, otherwise it is* closed*.*

Definition 18 (Average path length)**.** *The* average path length *of a graph is the mean of the distances between all couples of nodes.*

Definition 19 (Connected graph)**.** *A graph is* connected *if a path exists between any two vertices.*

Definition 20 (Connected component)**.** *A maximal connected subgraph is called a* connected component*.*

Therefore, a graph with only one connected component is a connected graph.

Definition 21 (Bridge)**.** *A* bridge *is an edge whose deletion disconnects the connected component it belongs to.*

Definition 22 (Cut vertex)**.** *A* cut vertex *is a vertex whose deletion disconnects the connected component it belongs to.*

Definition 23 (Leaf vertex)**.** *A* leaf vertex *is a vertex incident to only one edge. That edge is thus a bridge.*

Definition 24 (k-vertex-connected)**.** *A graph is* k-vertex-connected *if it contains more than* k *vertices and no subset of* k *vertices exists such that deleting these* k *vertices makes the graph disconnected.*

Definition 25 (Star)**.** *A* star*, or* star graph*, is a graph such that all vertices except one are leaf vertices.*

Definition 26 (Isolate)**.** *In a simple graph, an* isolate *is a vertex incident to no edges. It forms a connected component composed only of itself.*

Definition 27 (Weighted graph)**.** *Let* G *a graph of size* m *and*

$$w : E \to \mathbb{R}^m$$

a function associating a real number to each edge of graph G*. Then* (G, w) *is called a* weighted graph*.*

Definition 28 (Weighted adjacency matrix)**.** *Let* (G, w) *be a simple and weighted graph of order* n*. The matrix* $\{a^w_{ij}\}_{1\leq i,j\leq n}$ *such that*

$$a^w_{ij} = \begin{cases} w_{ij} & \text{if } (x_i, x_j) \in E, \\ 0 & \text{otherwise} \end{cases}$$

is called the weighted adjacency matrix of graph G*.*

2.2 Social sciences

In 1932, psychiatrist Jacob Moreno showed that runaways occurring at a girls school were caused by social contacts more than by the girls' personalities (Freeman, 2004). The study showed that connections between these actors were more important than any personal attributes (e.g. age). Moreno used *sociomatrices* and *sociograms*—adjacency matrices, and graphs—to illustrate and prove that idea. This use of a network of social contacts is usually considered as the birth of social network analysis, though some people argue that it was in fact the earlier but similar and non-experimented ideas of Georg Simmel, or later contributions such as the works of Harrison K. White (Freeman, 2004; Borgatti et al., 2009; Mercklé, 2011). On the etymologic side, the first appearance of the term "social network" is in (Barnes, 1954).

Networks are used to explain social phenomena in a population or to highlight particular features from a social structure (Hennig et al., 2012). The network models a population—using vertices for the actors and edges for the relations between them. The relations can be of diverse natures: friendship, professional, intimate, etc. To model a society, social network analysis was not immediately accepted by the social sciences community. The use of both a new mathematical model and approach took time to gain acceptance:

> Few network analysts nowadays would query Seidman and Foster's (1981) comment that "it is useful to regard a social network as a graph, thus gaining access to the precise terminology of graph theory". Regrettably, for many analysts, though not for these two authors, this is apparently all that graph theory has to offer. (Barnes and Harary, 1983)

Using graph theory in accordance with social interpretation, social network analysis proposed studies and achieved results in what became a rich and long-lived interdisciplinary field. It adjusted elements from graph theory:

> A key task of social network analysis has been to invent graph-theoretic properties that characterise structures, positions, and dyadic properties (such as the cohesion or connectedness of the structure) and the overall "shape" (i.e. distribution) of ties. (Borgatti et al., 2009, p. 894)

and developed its own methods, measures, indices and properties:

> Perhaps the most fundamental axiom in social network research is that a node's position in a network determines in part the opportunities and constraints that it encounters, and in this way plays an important role in a node's outcomes. This is the network thinking behind the popular concept of social capital, which in one formulation posits that the rate of return on an actor's investment in their human capital (i.e., their knowledge, skills, and abilities) is determined by their social capital (i.e., their network location). (Borgatti et al., 2009, p. 894)

In this work, we are applying the same approach for literary works.

Today, after having exported many concepts into other fields that required the use of networks, social network analysis is being considered for a bigger project consisting of improving communication and network usage between scientific fields, thus letting ideas spread more effectively (Brandes et al., 2013).

The following are some of the most important results in social network analysis, exposed here to display examples of methods and associated results of this academic discipline. We will then focus on centrality in networks (section 2.3).

Small-world property Stanley Milgram was a psychologist who conducted two famous experiments in the 1960s. One of them consisted of asking random persons in Nebraska and Boston to each send an envelope to a given person in Massachussetts, with the stipulation that the recipient be a personal contact, who would then pursue the experiment with the same constraint. In the end, a small number of envelopes reached the final address, but the average number of steps needed from the start to the ending point was around six: they were showing that the underlying network of social contacts implies the existence of short[6] paths between individuals (Milgram, 1967; Travers and Milgram, 1969).

Strength of weak links In 1973, Mark Granovetter used an experiment to show that job seekers could more efficiently find new jobs through distant contacts rather than close ones (Granovetter, 1973). This result relies on the fact that close neighbours—those with whom one shares many links—bear many structural properties with the node under study. By revealing this contrasting effect, Granovetter also showed how similar qualitative attributes of actors are when their structural profiles coincide.

Structural holes Similar to Granovetter obtaining interesting results via the observation of a seemingly marginal network property, Ronald S. Burt formulated the concept of structural holes which highlight the importance borne by the absence of a relation (Burt, 1992). In the case of two distant clusters of nodes having common interests, links connecting those

[6] In comparison to the size of the network.

communities are of key importance. Structural holes are such empty dyads; creating a link—maybe a bridge—would bring power and control to its extremities.

2.3 Centrality

This idea is one of the main concepts in social network analysis, but it remains rather vaguely defined. We can see it as a general intuition:

> Centrality quantifies how important vertices (or edges) are in a networked system, [...][7]. There are a wide variety of mathematical measures of vertex centrality that focus on different concepts and definition of what it means to be central in a network. (Newman, 2010)

The use and understanding of centrality are somewhat immediate and intuitive, even if it relies on abstracted mathematical expressions, and on properties of the nodes and of the whole network.

> At the node level of analysis, the most widely studied concept is centrality—a family of node-level properties relating to the structural importance or prominence of a node in the network. (Borgatti et al., 2009, p. 894)

Since the origins of centrality at the end of the 1940s, there has been an important amount of research on that subject. In research, new variations of centrality measures are continuously defined. In fact, the concept itself relies on an intuitive definition, translated into mathematical expressions, but without a unique mathematical definition. The few attempts to build a mathematical centrality theory have failed, and the situation seems inextricable: the attempts at creating consistent axiomatic systems do not recognise most of the centrality measures (Sabidussi, 1966; Boldi and Vigna, 2013).

Some of the centrality indices find their origins in mathematics, like degree and eccentricity (Berge, 1958). In 1869, the mathematician Camille Jordan defines the *center* of a graph as the set of vertices minimizing the *eccentricity* value[8], *eccentricity* being the maximal distance from a given node to any other node in the network.

In 1950, Alex Bavelas, psychologist at the Massachusetts Institute of Technology, was studying diffusion in small groups of persons. In particular, he showed interest in the structural properties of the groups that were helping to spread information and how they were interfering with the quality of the transmission. For this purpose, he modelled the groups with networks. In a seminal article, Bavelas wrote:

> Do some patterns have structural properties that limit group performance? It may be that among several communication patterns, all logically adequate for

[7] "... and social network analysts in particular have expended considerable effort studying it."
[8] In the original text, Camille Jordan does not explicitly define *eccentricity*.

> the successful completion of a specified task, one gives significantly better performance than another. What effects can pattern, as such, have upon the emergence of leadership, the development of organization, the degree of resistance to group disruption, the ability to adapt successfully to sudden changes in the working environment? (Bavelas, 1950)

This is the origin of the question of centrality that consists of explaining some dimensions of the role of an actor on the basis of their position in the network and of the network's structure. In this article, Bavelas calls the proposed measure of uncovering these questions "relative centrality". For a given vertex, he computes it by summing the distances between all the vertices and dividing this value by the sum of the distances from the vertex to all others in the network. Mathematically, if d is the distance function, G the network and V the set of its vertices, the relative centrality c_R of node x is[9]:

$$c_R(x) = \frac{\sum_{x_i,x_j \in V} d(x_i, x_j)}{\sum_{x_k \in V} d(x, x_k)}. \tag{2.1}$$

The experiments had the participants distributed in predefined configurations, as in figure 2.2[10] (Bavelas, 1948, 1950; Leavitt, 1951). They had to share a piece of information among

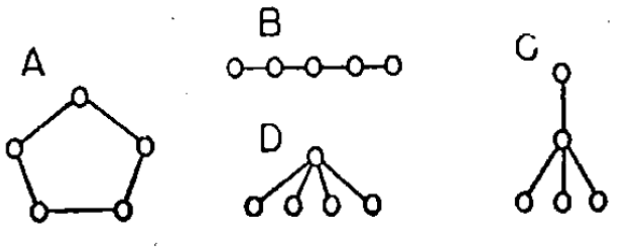


Figure 2.2: Figures from (Bavelas, 1950, p. 726). The author chose to study (A) a circle graph, (B) a line graph, (C) a mixed-case graph, and (D) a star graph.

them. In the end, they would indicate who had been the group leader according to their perception. The results obtained by the index of relative centrality regarding which were the best configurations and positions (see figure 2.3, left) were correlated with the observation (see figure 2.3, right).

According to Bavelas:

[9] Despite being right in expressing Bavelas calculation, this mathematical formula is anachronistic.

[10] There is a mistake with figure (C). It shows the same network as (D); this is not consistent with figure 2.3, which shows a network with one less leaf.

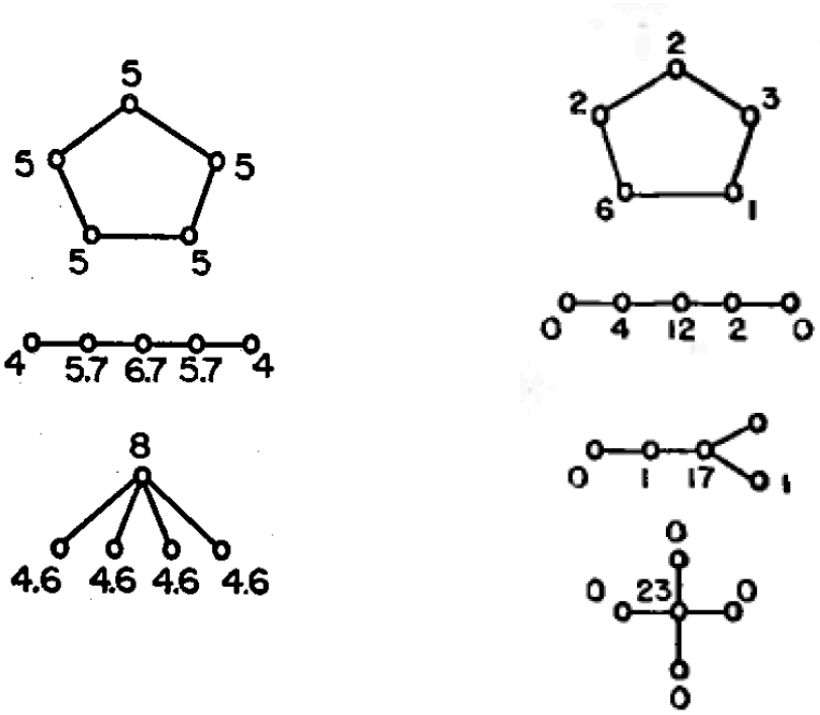


Figure 2.3: Figures from (Bavelas, 1950, pp. 726, 728). The relative centrality is computed as in formula 2.1. (Left.) Theoretical values. (Right.) Experimental values. The original caption reads: "Frequency of occurrence of recognized leaders at the different positions in patterns A, B, C, and D" and cites a work by one of Bavelas' Ph.D. student (Leavitt, 1949).

> [...][11] the findings suggested that the individual occupying the most central position in a pattern was most likely to be recognized as the leader.

Closeness, one of the now-classic measures of centrality, is based on Bavelas' idea of "relative centrality". To our knowledge, Gert Sabidussi gave the first impulse to a mathematical definition of centrality by defining axioms that centrality indices had to verify (Sabidussi, 1966). Sadly, today his work is primarily remembered for his clear definition of closeness centrality, the rest of the article remaining ahead of its time. In 1971, Anthonisse defined betweenness, a way to measure the control a node has on flows across the network as they are assimilated as shortest paths (Anthonisse, 1971). In 1972, Philip Bonacich defined power centrality (also known as eigenvector centrality), an index based on the eigenvectors of the adjacency matrix (Bonacich, 1972). At the time, centrality had chaotically turned into a catch-all concept housing families of measures based on mathematical expressions (Koschützki et al., 2005a). Centrality measures ranked importance by highlighting specific network features

[11] The beginning of the sentence reads "No good theory has been formulated for the differences in number of errors, but [...]". What Bavelas calls "errors" are the mistakes accumulated in the process of transmitting the information among the group participants. This part of Bavelas' work is out of the scope here.

among a wide range of criteria. Later, Linton C. Freeman formulated a clearer definition for betweenness (Freeman, 1977)—the one still in use today—and soon after proposed a review article on the subject of centrality (Freeman, 1978). He chose three measures—degree, closeness and betweenness—that became widely accepted:

> The time has come, it would seem, to stop, take stock and try to make some sense of the concept of centrality and the range and limits of its potential for application. (Freeman, 1978, p. 217)

He introduced centrality as having the characteristic that the maximum centrality of an actor among all possible networks is obtained in the center of a star-graph (see figure 1.1). He also proposed to normalise centrality, because "it might be desirable to have a measure that is independent of network size". Finally, he proposed a global network measure adapted to any centrality index, called centralisation, which shows "the tendency of a single point to be more central than all other points in the network" (Freeman, 1978, p. 227). This was an important step for the then still burgeoning field of social network analysis, but it also imposed in the following years the use of these centrality indices for most of the community. Freeman recalls on the almost immediate acceptance of these centrality measures, including eigenvector centrality:

> Beginning in about 1980 then, the measures based on closeness, degree, betweenness and the first eigenvector became standard in social network analysis. All four were widely used in the field. (Freeman, 2008, p. 3)

Subsequently, over a number of years, research on the concept of centrality focused mostly on statistical properties (Moxley and Moxley, 1974; Donninger, 1986; Mizruchi et al., 1986; Bolland, 1988; Valente et al., 2008; Frantz et al., 2009), while some work followed the suggestion of Gert Sabidussi to define a clear graph-theoretical framework consisting of definitions and a system of axioms. In 1980, G. Kishi would follow his footsteps and obtain some new results for the case of directed graphs (Kishi, 1980).

There are many mathematical definitions of centrality measures, which are usually grouped in families of measures. In fact, there are infinite possibilities, since the concept of centrality is not clearly mathematically defined[12], nor unique to each experimental situation, in most cases. For example, in a recent textbook (Hennig et al., 2012), the authors warn the reader against a blind use of centrality measures, since:

[12] Just as in social sciences, to some extent. In (Borgatti and Everett, 2006), the authors ask "What do centrality measures measure?". Their answer is then only partial.

> The relations between radial[13], medial[14], and feedback[15] centralities are not fully understood. (Hennig et al., 2012)

Measures of centrality have been designed to quantify social or psychological phenomena and to provide insights on experiments in these contexts. Sometimes, a new measure must be defined to answer specific research questions. This can be done by modifying attributes, like with the Google's PageRank algorithm (Page et al., 1999), which is a variation on eigenvector centrality, or by creating a new centrality measure, like in (Tutzauer, 2007), which is based on information theory and is "appropriate for traffic that propagates by transfer and flows along paths".

In spite of the large number of new definitions, sketches of typologies of centrality have been proposed (Ruhnau, 2000; Borgatti, 2005; Koschützki et al., 2005a; Borgatti and Everett, 2006; Koschützki et al., 2005b) and estimates of the efficiency of these indices analysed (Friedkin, 1991; Valente et al., 2008; Boldi and Vigna, 2013).

However, there is no unified framework that allows us to compare the results of one measure with another. There is no axiomatic system, which would allow to explain centrality on the basis of axioms or constraints, and therefore facilitate the comparison of multiple networks. Finally, there is simply no universal theory of centrality[16]. We have to work with the pieces of theory existing here and there, when bringing together the state-of-the-art and some of our ideas. It is also important to note that Noah E. Friedkin (Friedkin, 1991), Mark E. J. Newman (Newman, 2005) and Frank Tutzauer (Tutzauer, 2007), among many others, have all defined new centrality indices which have the particularity of widening the field of possible applications dealing with theoretical questions, and sometimes filling boxes in the typology proposed by (Borgatti, 2005).

[13] "All of the measures [that] assess walks that emanate from or terminate with a given node." (Borgatti and Everett, 2006)

[14] "[...] centrality measures [...] which are based on the number of walks that pass through given node." (Borgatti and Everett, 2006)

[15] For a feedback centrality, the computation of a node centrality depends on its neighbours', and vice versa.

[16] Ulrik Brandes et al. are currently working on a consistent theory of centrality that may provide the answer Gert Sabidussi was looking for fifty years ago (Brandes et al., 2012).

3 Indexing

> A good index is a work of art and science, order and chance, delight and usefulness.
>
> Foreword to *(Bell, 2001)*
> A.S. BYATT

The index of a book is a table containing numeric references—occurrences—to words appearing in the text. These words generally belong to a single category, like concepts, places, names, etc[1]. In the case of an elaborated index, subheadings indicating context are associated with the entries. The index is found in the pages at the back of a book.

In this work, we use an index to extract information from the text without having to consult it. Many indexes of literary works are available. This fact makes our framework reproducible for both stories and sets of stories, such as the *Sagas of Icelanders* or the previously cited *Les Rougon-Macquart*, which share common characters from one story to another. There is no extraction process of occurrences other than the one previously executed by editors and indexers and resulting in the published index. In this chapter, we grant importance to the particular case of indexes of characters. Characters from *Les Confessions* are part of an autobiography, which is at the intersection of biography and fiction[2]. Here, we explore both indexing for biography and fiction (3.2), and indexing for autobiography (3.3).

3.1 Some elements of history

We find proto-indexes in Antiquity already, and even that (Maniez and Maniez, 2009). The origin of their actual form dates back to Middle Ages.

> The first subject indexes were "distinctions," collections of "various figurative or symbolic meanings of a noun found in the scriptures" that "are the earliest of all alphabetical tools aside from dictionaries. [...]" Distinctions were biblical tools designed to assist preachers in writing sermons. (Kilgour, 1998, p. 76)

[1] E.g. (Lemoine, 1985) gives indexes of animal and boat names in the books of George Simenon.
[2] See chapter 4.

These tables were used in religious contexts, and a whole century was needed for them to evolve to their current form:

> By the end of the thirteenth century the practical utility of the subject index is taken for granted by the literate West, no longer solely as an aid for preachers, but also in the disciplines of theology, philosophy, and both kinds of law. [(Goodman, 1990) as cited by (Kilgour, 1998, p. 77)]

The oldest known printed indexes came in the early 1460s (Wellisch, 1986) (see figure 3.1). From that point, they would be spread to larger audiences[3]. The history of indexes is fascinating because closely related to the history of both book and literature; however, a deeper discussion of this specific point would be outside of the scope of this work. Detailed accounts can be found in (Bell, 2001; Maniez and Maniez, 2009). The important point here is that indexes have been present for a very long time: they cover all kinds of works in all kinds of genres and eras.

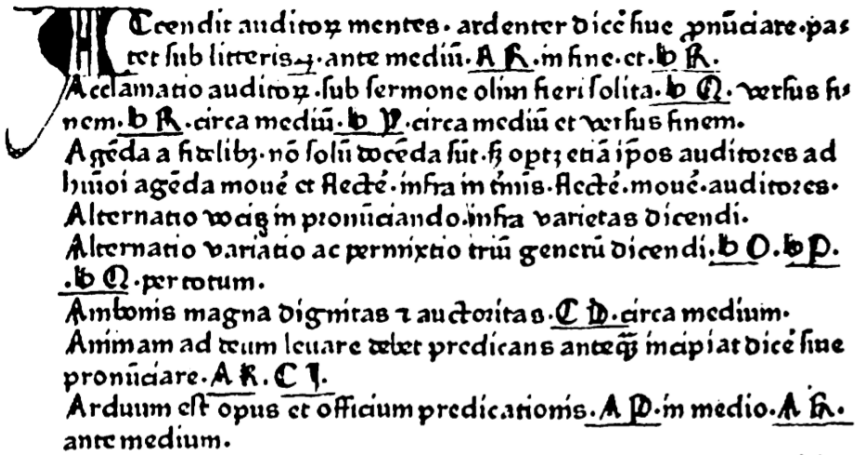

Tcendit auditoꝝ mentes · ardenter dicẽ siue ꝓnũciare · pas
ter sub litteris ꝛ · ante mediũ · A K · in fine · et · b K ·
Acclamatio auditoꝝ · sub sermone olim fieri solita · b Q · versus fi
nem · b K · circa mediũ · b Y · circa mediũ et versus finem ·
Agẽda a fictibꝫ · nõ solũ wcẽda sũt · sꝫ optꝫ etiã ipos auditores ad
huiusmoi agẽda mouẽ et flectẽ · infra in tmis · flectẽ · mouẽ · auditores ·
Alternatio wcis in pronũciando · infra varietas dicendi ·
Alternatio variatio ac permixtio triũ generũ dicendi · b O · b P ·
· b Q · per totum ·
Ambonis magna dignitas ꝛ auctoritas · C D · circa medium ·
Animam ad deum leuare debet predicans anteq̃ incipiat dicẽ siue
pronũciare · A K · C I ·
Arduum est opus et officium predicationis · A D · in medio · A K ·
ante medium ·

Figure 3.1: First lines of oldest known printed index, for St Augustine's *De rate praedicandi*, circa 1466. Figure taken from (Wellisch, 1986).

3.2 Use & types

Most of the time, an index is compiled for a book because research of its content carries interest. Sometimes, a textbook is not meant to be read from first to last page, without skipping any, despite the fact that there is usually a chronological order created by the narrator. An index gives an insight into the content it covers, as demonstrated by Slate collaborator Christopher Beam:

[3] E.g., William Shakespeare cites indexes in the play *Troilus and Cressida*, act I, scene III.

> When Sarah Palin's 413-page autobiography, *Going Rogue: An American Life*, hit stands Tuesday, readers discovered the governor's most mavericky move yet: The book lacks an index. So Slate has compiled its own. Just print out this index, paste it into the back of your copy, and start skipping around![4]

An index is not a summary, but a thematic contingency table proposing an approach on any given corpus from a different angle. It projects the book in another dimensional system. An index also contains common entries; however, precise and rare descriptions (e.g. those having only one occurrence) or abnormally frequent uncommon concepts may give the best insights on the text (Barwick, 2006). Indexing is more present in monographs than in fiction. In nonfiction, the interest in producing an index does not seem to raise questions:

> The necessity to provide indexes for any serious works of nonfiction that are likely to be consulted or reread has long been established. (Bell, 1991)

In this first case, the role of the index is to facilitate the search for names and events. The narrative is about facts, and travelling through the text following an unexpected order does not spoil it; which is not the case with works of fiction. An autobiography is considered at the crossing of biography and fiction (Lejeune, 1975). We will now describe the situation in both cases.

3.2.1 Biography

We frequently find an index at the end of biographies (Sassen, 2012). Most of the time, it consists of mentions of proper nouns—names, and maybe places. The index proposes a thematic summary of the subject's life: we discover which persons other than the subject are present in the composition, and in which places events happened. From there, it is easy to deduce which are the key actors and places based on the number of their occurrences.

Still, there are some problems: before studying the index of a biography or any similar work, the researcher should verify that the indexing method is consistent with his/her expectations and research questions. For example, the indexer may have made drastic choices about the main character.

> Carey denounced long entries for the major character of a biography as 'overloading', and suggested they should be omitted, or restricted to entries which could not go under any other heading, such as his birth, character, honours. (Bell, 1989)

For example, in Pierre Assouline's biography of photographer Henri Cartier-Bresson, the subject is simply omitted, however there remain twelve entries for "Cartier-Bresson": they are for his family. Thus, we can guess he had a large family, and from the pages on which they appear, which ones are older and which ones are younger:

[4] http://www.slate.com/id/2235917/ Accessed on 06/03/2014. For more on this controversy, see (MacGlashan, 2010).

CARTIER-BRESSON André : 30
CARTIER-BRESSON Anne : 409
CARTIER-BRESSON Claude : 190, 207
CARTIER-BRESSON Denise : 31
CARTIER-BRESSON Hortense : 409
etc. (Assouline, 2001, p. 433)

3.2.2 Fiction

In the case of fiction, an exhaustive index reveals information on the story that was supposed to stay hidden. Indexes of fiction are rare and complex to create (Mirabile, 1997; Maniez and Maniez, 2009). Hazel K. Bell sees three types of character indexes in fiction: the "seriously informative dictionary-type", the paratext, and some mixed cases (Bell, 1991). They infer choices by the indexer and are somewhat accused of proposing a second reading. On the other side, as Vladimir Nabokov wrote when "proposing to undertake the preparation of an index to Alexander Pushkin's verse-novel *Eugene Onegin*" (Bell, 1997):

> An index to a work like this should reflect its virtues and shortcomings, its tone and personality (as I have proved in *Pale fire*). It should be an afterglow and not a yawn. [Vladimir Nabokov, cited by (Bell, 1997)]

The index of a work of fiction should not be consulted by the reader whose primary intentions are not those of a scholar:

> Providing an index to a book may attempt two effects: to avoid people's having to read the book through to come across its contents in the order designed by the author; and to reduce the whole to very much less than the sum of its parts, rearranged on some alternative basis. (Bell, 1991)

This fear denotes exactly how our strategy can work on this text: most elements of the organisation of the discourse are available from the index.

Indexes are rarely realised for first editions of novels (Bradley, 1989; Duncan, 2014). In most cases, if not all, the index is the author's initiative. For example, Georges Pérec in *Quel petit vélo à guidon chromé au fond de la cour ?* and Vladimir Nabokov in *Pale fire* use nonconformist indexes. They are parts of the narrative—they act as paratexts. In the first novel, the index reflects the author's objective to create a text that includes a large number of metaphors and paratexts—it seems to induce the text itself... and it remains unfinished[5]! In the second novel, the index is compiled by the protagonist of the story. Therefore, it belongs to the discourse; its narrative function is to highlight the protagonist's madness (Bell, 1997).

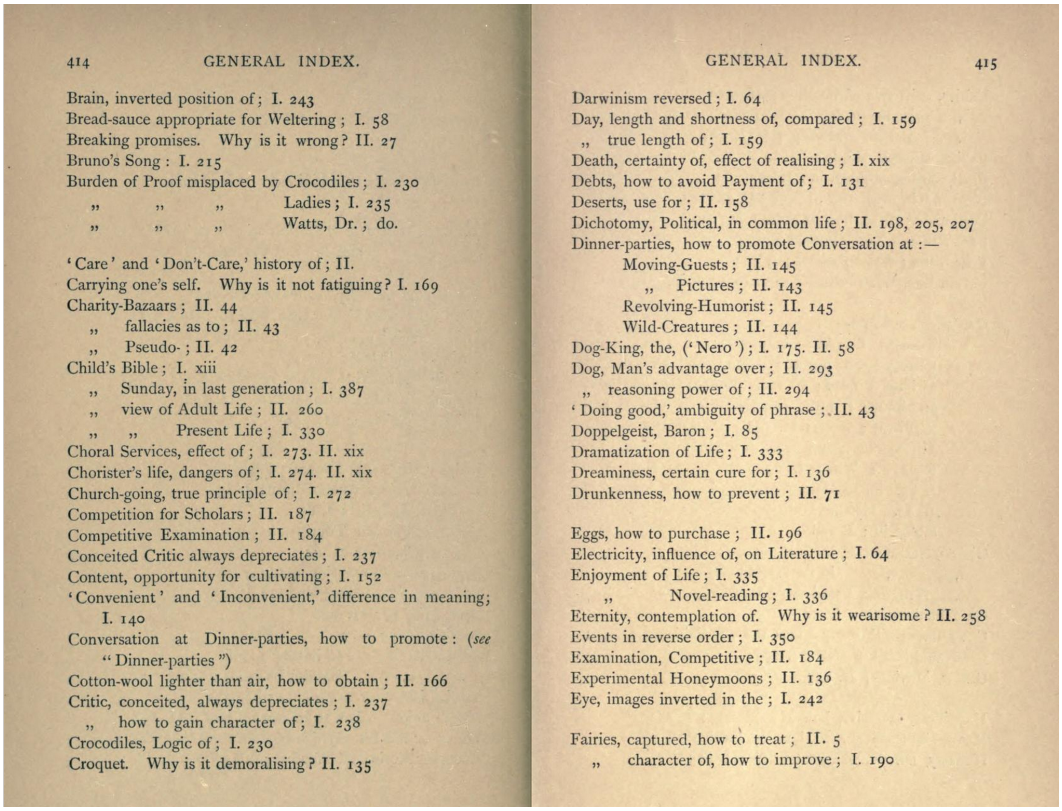
[5]The colourful title of the index reads: "INDEX des fleurs et ornements rhétoriques, et, plus précisément, des métaboles et des parataxes que l'auteur croit avoir identifiées dans le texte qu'on vient de lire." Georges Pérec, *Quel petit vélo à guidon chromé au fond de la cour?*, folio, 2012, pp. 113 - 119.

The use of indexing in fiction reveals interesting questions; Judy Batchelor summarises the index—as paratext—to *Sweet desserts* by Lucy Ellmann:

> On the whole, the index is more entertainment than use [...] We might call this a *para-index* [...] (Batchelor, 1989, p. 94)

In a sense, this remark is also valid for works like *Sylvie and Bruno* from Lewis Carroll who compiled an index with a comic function (see figure 3.2[6]). And finally, the most extreme literary example may be the short story *The Index* by J.G. Ballard; the narrative is simply composed of two paragraphs along with the index of a *lost* biography. The index accounts for nearly 90% of the story. The last entry of the index mentions the indexer himself :

> Zielinski, Bronislaw, suggests autobiography to HRH, 742; commissioned to prepare index, 748; warns of suppression threats, 752; disappears, 761.



414 GENERAL INDEX.

Brain, inverted position of; I. 243
Bread-sauce appropriate for Weltering; I. 58
Breaking promises. Why is it wrong? II. 27
Bruno's Song: I. 215
Burden of Proof misplaced by Crocodiles; I. 230
„ „ „ Ladies; I. 235
„ „ „ Watts, Dr.; do.

'Care' and 'Don't-Care,' history of; II.
Carrying one's self. Why is it not fatiguing? I. 169
Charity-Bazaars; II. 44
„ fallacies as to; II. 43
„ Pseudo-; II. 42
Child's Bible; I. xiii
„ Sunday, in last generation; I. 387
„ view of Adult Life; II. 260
„ „ Present Life; I. 330
Choral Services, effect of; I. 273. II. xix
Chorister's life, dangers of; I. 274. II. xix
Church-going, true principle of; I. 272
Competition for Scholars; II. 187
Competitive Examination; II. 184
Conceited Critic always depreciates; I. 237
Content, opportunity for cultivating; I. 152
'Convenient' and 'Inconvenient,' difference in meaning; I. 140
Conversation at Dinner-parties, how to promote: (*see* "Dinner-parties")
Cotton-wool lighter than air, how to obtain; II. 166
Critic, conceited, always depreciates; I. 237
„ how to gain character of; I. 238
Crocodiles, Logic of; I. 230
Croquet. Why is it demoralising? II. 135

GENERAL INDEX. 415

Darwinism reversed; I. 64
Day, length and shortness of, compared; I. 159
„ true length of; I. 159
Death, certainty of, effect of realising; I. xix
Debts, how to avoid Payment of; I. 131
Deserts, use for; II. 158
Dichotomy, Political, in common life; II. 198, 205, 207
Dinner-parties, how to promote Conversation at:—
Moving-Guests; II. 145
„ Pictures; II. 143
Revolving-Humorist; II. 145
Wild-Creatures; II. 144
Dog-King, the, ('Nero'); I. 175. II. 58
Dog, Man's advantage over; II. 293
„ reasoning power of; II. 294
'Doing good,' ambiguity of phrase; II. 43
Doppelgeist, Baron; I. 85
Dramatization of Life; I. 333
Dreaminess, certain cure for; I. 136
Drunkenness, how to prevent; II. 71

Eggs, how to purchase; II. 196
Electricity, influence of, on Literature; I. 64
Enjoyment of Life; I. 335
„ Novel-reading; I. 336
Eternity, contemplation of. Why is it wearisome? II. 258
Events in reverse order; I. 350
Examination, Competitive; II. 184
Experimental Honeymoons; II. 136
Eye, images inverted in the; I. 242

Fairies, captured, how to treat; II. 5
„ character of, how to improve; I. 190

Figure 3.2: Extract of Lewis Carroll's *Sylvie and Bruno* index, pages 414-415 of MacMillan and co. edition (1893).

Outside of these cases, where an index carries a narrative function, they are usually compiled and printed for scholarly editions of literary works:

[6] Taken from https://archive.org/details/sylviebrunoconcl00carriala, accessed on 23/05/2014.

> The demand for indexes to fiction is small. On the whole they are not wanted by novelists, reviewers, readers or publishers. [...] When support is shown it is for books of merit, especially acknowledged classics, and not for the bulk of fiction which will soon be forgotten. (Bradley, 1989)

On a more global level, there are indexes of fiction that reference in which books characters occur. e.g. (Patterson, 1973; Lemoine, 1985; Smiley, 2001). This is of primary interest[7] when an author has created literary works that overlap outside of the limits of the novel, such as Marcel Proust or Honoré de Balzac (Bradley, 1989, p. 242).

3.3 Autobiography

> "Il ne tiendra qu'à moi de renouer les deux bouts de mon existence, de confondre des époques éloignées, de mêler des illusions d'âges divers..."
>
> *Mémoires d'outre-tombe*
> CHATEAUBRIAND

One of the main differences between indexing biography and autobiography is that the narrator has a strictly different status; in the first case he/she is distant[8], in the second case he/she is the protagonist. Thus, the question of how to index becomes more complicated. Should we ignore all uses of first person pronouns, just as we ignored the subject's name in the biography? What is the status of his/her name in quoted speech? Examples of autobiography indexes are rare, since they are a subfield of fiction, which has a low indexation rate. Still, we should consider the biography rule allowing the non-indexing of the main character.

Usually, we know more about a character in fiction, in terms of intimacy, than that of an individual in a biography[9] (Bell, 1991). The index of fiction gives deeper insights into the characters and their personal stories, while the biography is more scholar-edited, thus giving fewer insights into the perception of the world of the narrator, and more about facts (who meets whom, etc.). In the case of autobiography, flaws from both approaches are inherited, plus a third one; however, they can be turned into advantages. The autobiography is all about its author: his/her life (like in a biography, he/she is omnipresent), his/her point of view on the story (like in fiction, the author tells a story, thus clearly creating a literary work as opposed to the more documentary monographs), and his/her narration (the way of articulating the facts—and perhaps omitting some—, biases the story towards the authors' intentions). The life is described as he/she lived it, his/her point of view is not general (e.g. no interviews of

[7] We consider this as one possible future work.

[8] With the exception of narrator being part of the subject's history, e.g. a political figure's counsellor writing his/her biography.

[9] "Fiction also extends further than normal biography into the sexual lives of its characters, both in vividness of description and in degree of physical intimacy portrayed. 'Sleeps with' was a recurring subheading—perhaps a palely restrained one." (Bell, 1991)

other characters in order to document the story) and its narration is intimate with the subject rather than distant. In addition to the flaws inherited by the biography indexes, familiarity may cause problems (White, 2005).

The indexation of an autobiography raises questions to the point where some autobiography's authors eliminate the potential of having the story spoiled by refusing an index:

> Nicholas Fairbairn explains the lack of an index to his book *Life is too short: autobiography, vol. 1* (Quartet Books, 1987), by saying: "The men, in their vanity, would have turned only to the index, hoping that their names would be in it, and the women, in their terror, would have turned to it hoping that theirs would not!" (Bradley, 1989)

This can be done while the author is alive, and if nobody provides one of his/her own (MacGlashan, 2010). This also illustrates the power of the index on re-telling a story and highlighting outlying facts—and the author's awareness of that. This fact is acknowledged by indexer Hazel K. Bell:

> The indexes may be some indication of the serious status of the books. And to the indexer, the work is fascinating: one reads the text repeatedly and closely; and indexing becomes a form—albeit humble—of literary criticism. (Bell, 1991)

Types of indexes can be split in two categories: with and without subheadings. The first may contain biases, and not only by circumventing the author's absence. The second is more methodic, belonging to Bell's "seriously informative dictionary-type", but omitting the subject.

In the end, the index of an autobiography provides an overview of the people cited by the author, implying a spatial, social or contextual proximity with them. The autobiographical *Les Confessions* runs through a mainly chronological order. Its index provides a summary of the entries and exits of the characters throughout the story: it is a preliminary representation of the plot. Considering not only the entry of a single character in the index, but all the entries concurrently, will allow us to build a relational picture of the plot based on that depiction.

4 *Les Confessions*

> "Il n'y avait pas là de quoi briller pour le pauvre Jean-Jacques. J'eus le bon sens de ne vouloir pas faire le gentil malgré Minerve, et je me tus."
>
> *Les Confessions*
> JEAN-JACQUES ROUSSEAU

Jean-Jacques Rousseau (1712-1778) was a citizen of Geneva. He produced a great variety of work (epistolary novel, music, biology, educational, philosophical and political essays, etc.) which had a lot of influence in his lifetime, and still does to this day. *Les Confessions* is an autobiographical work written in two parts (see table 4.1), which spans fifty years of Rousseau's life; it was published posthumously. In this text, Rousseau tells his story from an intimate point of view. He presents facts along the main story arc, his life, but also reveals personal experiences that had impact on him and gives details on his private life—an uncommon custom at that time. His intended goal was to present his life with total transparency, so that people would be able to understand or to judge him with all the necessary details to do so.

Amongst other things, he was convinced of the existence of a plot to disparage him in public opinion and thus wanted to tell the story of his life in his own words. He was aware of the novelty of revealing a real person's entire life and feelings, and he described this manoeuvre in his introduction to *Les Confessions*[1]:

> Voici le seul portrait d'homme, peint exactement d'après nature et dans toute sa vérité, qui existe et qui probablement existera jamais. Qui que vous soyez, que ma destinée ou ma confiance ont fait l'arbitre du sort de ce cahier, je vous conjure par mes malheurs, par vos entrailles, et au nom de toute l'espèce humaine, de ne pas anéantir un ouvrage unique et utile, lequel peut servir de première pièce de comparaison pour l'étude des hommes, qui certainement est encore à commencer, et de ne pas ôter à l'honneur de ma mémoire le seul monument sûr de mon caractère qui n'ait pas été défiguré par mes ennemis. Enfin, fussiez-vous,

[1] There are three manuscripts of *Les Confessions*. This introduction is present in the Geneva manuscript, but not in the Paris and Neuchâtel manuscripts.

> vous-même, un de ces ennemis implacables, cessez de l'être envers ma cendre, et ne portez pas votre cruelle injustice jusqu'au temps où ni vous ni moi ne vivrons plus, afin que vous puissiez vous rendre au moins une fois le noble témoignage d'avoir été généreux et bon quand vous pouviez être malfaisant et vindicatif : si tant est que le mal qui s'adresse à un homme qui n'en a jamais fait ou voulu faire, puisse porter le nom de vengeance.[2] (Rousseau et al., 2012, p. 65)

This foreword illustrates the difficult situation at the time of writing and some of the reasons—"défiguré par mes ennemis"—motivating his action.

His story brings together hundreds of persons. Transposed in the text, they become characters, and the text a discourse—the narrative image of the story. Even if all of them are based on individuals who lived at the time, the autobiography transforms a story into fiction, and thus Rousseau's writing turns people into fictional entities—characters. By making choices in writing his story, he builds a unique narration conveying his ideas, convictions, feelings and views:

> S'en tenir au portrait du personnage serait perdre de vue la vision plus large mise en place par l'écrivain. C'est bien dans un ensemble qu'il faut le prendre en compte, non seulement celui de la société décrite, mais, au-delà, celui des conceptions esthétiques qui définissent sa manière d'être, d'agir ou de parler.[3] (Bardet, 2007, pp. 20-21)

Behind the pact of transparency promised to the readers, Rousseau controls the words—the style—and more generally the narrative flow. He puts together and sorts what he considers as necessary information, while he also omits some events or makes mistakes (consciously?). His narration does not render time flow linearly, but puts emphasis on the times and events requiring longer development, where he then adds commentaries. As a result, he shares, consciously or unconsciously, an interpretation of the events dictated by his interpretation of transparency and by his narrative choices, as well as the constraints borne by his own writing style.

[2] "Here is the only portrait of a man, precisely painted according to his nature and his whole truth, which exists and will probably ever exist. Whoever you are, that my destiny or my confidence have made mediator of this book, I entreat you by my misfortune, by your guts, and in the name of the whole mankind, not to destroy a unique and useful work, which can serve as a first piece of comparison for the study of men, which certainly must still be initiated, and not to take off from the honour of how I will be remembered the only certain monument of my personality which was not disfigured by my enemies. Finally, were you one of these tenacious enemies, cease to be one to my ashes, and do not hold your cruel injustice up to the time where neither you nor I would be living anymore, thus you could at least once commit the noble testimony of having been generous and good while you could have been harmful and vindictive: whether the damage addressed to a man who never did or wanted to do any may wear the surname of vengeance." [Own translation.]

[3] "Sticking to the portrait of the character would be losing sight of the wider vision set up by the author. It is indeed as a whole that the character must be taken into account, not only the one of the described society, but, farther, to the one of aesthetically conceptions which define its way of being, acting and talking."[Own translation.]

4.1 Jean-Jacques Rousseau

In the following paragraph, we will share a few elements of Jean-Jacques Rousseau's life. The presentation is succinct; the objective of this section is to provide a factual timeline which is useful for the later analysis of *Les Confessions*. The facts presented here are principally adapted from summaries and comments (Rousseau et al., 2009, 2011), as well as our own reading.

Rousseau was born in Geneva in 1712. Like his father, a watchmaker, he was a citizen of Geneva. His mother died a few days after giving birth to him. In addition to his work, his father was interested by politics and community involvement, and he possessed many books on those subjects amongst others (Perrin, 1997). Rousseau had early access to much reading, and he began to consult these books and to become familiar with them. He was sent to study with M. Lambercier and his sister, and after a few chaotic apprenticeships, he left Geneva at the age of sixteen. He wandered through the countryside before arriving in Annecy, where he met Mme de Warens, a woman benefitting from annuities to convert people to catholicism. She hosted him and from that time—and for years—became a mentor to him. Initially, she made him convert, then recommended him to diverse societies in order to help him find work. He stayed attached and lived with her most of that time. During these years, he travelled in Savoy (Annecy, Turin), France (Lyon, Paris, Montpellier), today's Switzerland (Geneva, Lausanne, Fribourg, Neuchâtel, Soleure), and regularly came back to Mme de Warens, whom he intimately called "Maman". At that time, he met Mme Basile in Turin and developed feelings for her, then Mme de Larnage on his way to Montpellier, and lived with her during a short love affair. He learned music but could not yet earn a living from that activity. He also worked as private tutor for the children of M. de Mably, where he began his reflections about education. Mme de Warens and Rousseau would then become lovers. In 1742, he left her to live in Paris. There, he worked again as a private tutor, and tried to distribute—with no success—his own system of musical scores. He would meet Denis Diderot and they would become close friends at that time. From 1743 to 1744, he went to Venice to work as secretary of the French ambassador, the comte de Montaigu. There, he also had a love affair with a courtesan named Zulietta. After less than one year, his relationship with Montaigu was so deteriorated that Rousseau left and came back to Paris. In 1745, he met Thérèse Levasseur, who would become his partner for the rest of his life. In total, they had five children together, all left with public welfare. He met Mme Dupin, an influential woman with a salon, and for whom he also worked. In 1749, he met Friedrich Melchior Grimm who became another close friend; he would introduce him to many of the societies in which he was active. In 1750, he wrote the *Discours sur les sciences et les arts* for an academic contest: it won the first prize. In 1752, one of his musical compositions was played before the King of France. Fleeing any notoriety, he did not present himself before the King, who wanted to meet him and would have allocated him an annuity. He subsequently lived from copying and writing music. In 1755, he wrote the *Discours sur l'origine et les fondements de l'inégalité parmi les hommes*. He was famous and

sought after at the time, and in consequence decided to retire to the countryside to flee the fame. In 1756, his friend Louise d'Épinay offered for him to stay in a house she owned outside of Paris. He went to live there with his partner Thérèse and her mother. In 1757, he quarrelled with Diderot, feeling insulted by one of his writings, though they rapidly reconciled. He then fell in—platonic—love with Louise d'Épinay's sister-in-law, the comtesse d'Houdetot. This was divulged and, in addition to other conflicts, led to a break-up of Rousseau and Louise d'Épinay, as well as with Diderot and Grimm. The hate of Voltaire and d'Alembert towards him also dates from that time. He was then hosted by the duc de Luxembourg and his wife in Montmorency. At the time, he wrote *Julie ou la Nouvelle Héloïse, Émile ou De l'éducation*, and *Du contrat social*. In 1761, the first text was published and became a bestseller. In 1762, the two others were published, then subsequently banned by the Paris Parliament and burnt. In consequence, Rousseau fled to the principality of Neuchâtel, a territory under Prussian sovereignty. There he met Du Peyrou with whom he would become friends, and who would keep some of his files. He also met George Keith, governor of Neuchâtel on the account of Prussia. In 1763, an anonymous pamphlet was published against him in Geneva . In 1764, an another anonymous pamphlet aiming at him was published, this time written by Voltaire. This pamphlet revealed the secret of his abandoned children. In 1765, the inhabitants of Môtiers—the village where he was living—became hostile towards him, and he fled to St-Pierre island, then Biel[4], Strasbourg, Paris (anonymously), and England, advised by the comtesse de Boufflers, and hosted by the philosopher David Hume. There, he started writing *Les Confessions*. In 1766, they had a falling out, and made mutual accusations in public. This was probably caused by the ever growing paranoia of Rousseau and the undermining work of a few persons outside of England. In 1767, he came back to Paris, anonymously, then wandered through the countryside. He did not publish anything, which led to him being tolerated by the authorities. Around years 1770-1771, Rousseau finished *Les Confessions*. He began public readings, but Louise d'Épinay complained about the possibility that parts of her private life would be publicly divulged; the authorities forbade him to do any more readings. For the last years of his life, he lived by copying music and spent time studying nature and writing botanical works. He began writing *Les Rêveries du promeneur solitaire* but died before finishing it, in 1778, in Ermenonville.

4.2 *Les Confessions*

Les Confessions is an account of Rousseau's life. It is influenced by the time of writing and probably by the suspected plot against him. There are elements in the text showing that Rousseau was already used to confessing personal stories to his friends[5]. It is also worth noting that a text focusing on his life had been consistently asked by his editor, who would have allocated an annuity in exchange. Eventually, Rousseau received financial help from

[4] The text of *Les Confessions* stops here.

[5] See (Perrin, 1997, pp. 38-39).

another source, then wrote the novel mainly motivated by his ideal of transparency as an answer to the attacks (Starobinski, 1991). An excerpt of his correspondence confirms this:

> [Mes ennemis] travaillent beaucoup à me faciliter l'entreprise d'écrire ma vie [...] j'ai beaucoup à dire et je dirai tout, je n'omettrai pas une de mes fautes, pas même une de mes mauvaises pensées. Je me peindrai tel que je fus, tel que je suis ; le mal offusquera presque toujours le bien, et malgré cela, j'ai peine à croire qu'aucun de mes lecteurs ose se dire, je suis meilleur que cet homme-là.[6] [Jean-Jacques Rousseau, as cited by (Perrin, 1997, pp. 47-48)]

4.2.1 Modern autobiography

Les Confessions is not the first autobiography ever written (Perrin, 1997). The genre was popular in Rousseau's own time, as well as memoirs which focus less on the life of the author.

> Rousseau a fait la renommée du genre en s'adressant au public des romans et en démontrant que l'autobiographie, jusque-là à audience limitée, pouvait faire un succès de librairie, mais il existe, en tenant compte des domaines anglais et allemand, des dizaines, sinon des centaines d'autobiographies antérieures à la sienne.[7] (Trousson and Eigeldinger, 2012, vol. 1, p. 21)

Nevertheless, the work of Rousseau had the originality of divulging intimacy and explaining, using real life events, how he built his persona and came to express some of his theories. Parts of *Les Confessions* have turned into famous elements of literature and psychological theories, e.g. the excitement caused to Rousseau as a child by Mme Lambercier spanking him, or him lying about a ribbon he had stolen, which lead to the accusation of a young servant, and caused him to feel guilty for the rest of his life. Thus, *Les Confessions* is seen as the origin of the modern autobiography (Lejeune, 1975) :

> C'est ce mécanisme complexe et fondateur qu'on découvre à l'œuvre dans Les Confessions. Rousseau n'a pas inventé, créé l'autobiographie de toutes pièces, mais il en a réalisé les virtualités et élaboré la véritable problématique. Si l'on considère la tradition proprement littéraire du genre, c'est désormais – quel que soit l'apport de ses devanciers – par rapport à lui que se définiront ses successeurs.

[6] "[My enemies] work much in helping me writing my life [...] I have much to say and I will say everything, I will not omit any of my mistakes, not even any of my unpleasant thoughts. I will depict myself as I was, as I am; the evil will almost always offend the good, and despite this, I am scarcely able to believe that none of my readers dares to say: I am better than this man." [Own translation.]

[7] "Rousseau has made well-known the genre by addressing the public of the novels and by demonstrating that autobiography, up until then having a limited audience, could be a success, but there is, when taking into account the english and german areas, tenths or hundreds of autobiographies that are prior to his." [Own translation.]

> Dans le sens et la portée qu'il lui confère, Jean-Jacques disait vrai : son entreprise est sans exemple.[8] (Trousson and Eigeldinger, 2012, p. 34)

4.2.2 The truth

The objective of being honest in his storytelling is indeed difficult to follow. *Les Confessions* is an autobiography; its material is based on historical facts. But from reality to the text, the writer gains total control of the story—on the direction taken by the narration. (Trousson and Eigeldinger, 2012) list some examples (page 56), e.g. how he deleted three paragraphs about remorse at the end of chapter III in order to create a better transition, and concludes :

> Rousseau autobiographe reste bien, selon l'expression de J.-L. Lecercle, écrivain jusqu'au bout des ongles.[9]

In the addition to elements of style, the need for contemporary commentaries—at the time of writing—on decades old facts also biases the resulting discourse. *Les Confessions* can be thought of as a filtered—or equally augmented—version of the story, as seen through the lens of an older man than of all those who went through these events.

> On conçoit aussi que l'autobiographe ne peut déverser les faits en vrac, sans les ordonner ni leur conférer le sens qu'ils prennent avec le recul, c'est-à-dire sans céder à la tentation de transformer une succession d'événements en destin.[10] (Trousson and Eigeldinger, 2012, p. 46)

We take this into account with the angle used in our analysis: the material concerns most of his life, but the story crosses over his alleged paranoid state.

4.2.3 Content

The story of *Les Confessions* is mainly told chronologically, thus we consider that the temporal continuity in the narration is monotonically aligned with the page order. It contains many distinct and consecutive narrative events, and the use of (mostly short) analepses and prolepses are generally made in order to explain the event currently being described. They are generally justified by the narrative context, thus creating links across the novel between characters who never met in real-life. Thus, we do not analyse Rousseau's life, but *Les Confessions*' narration. Told in narratological terms (Genette, 2007), the discourse mostly evolves in parallel with the story. Time flow is monotonous, but not linear. Some events last longer than others, some chapters too, and this is also the case in terms of pages.

[8] "It is this complex and founding mechanism that we discover in Les Confessions. Rousseau did not invent or created the autobiography from scratch, but he realised its potentialities and elaborated set up true aim." [Own translation.]

[9] "Rousseau autobiograph remains, according to J.-L. Lecercle, author up to his fingertips." [Own translation.]

[10] "We conceive that the autobiography can not pour facts higgledy-piggledy, without sorting them or giving them the sense they take with distance, that is to say without giving in to the temptation of transforming the succession of events into fate." [Own translation.]

4.2.3.1 Time

Some parts of Rousseau's life necessitated more text than others, shaping the sizes[11] of the chapters and the durations spanned. In table 4.1, we show the corresponding pages and dates of the chapters in our printed version of the book, as well as the times of redaction. This table is a useful tool for two reasons: we can initiate the analysis from the present introductory chapter with a short statistical description, and later refer back to it, and we also gather information on the narrative process that we will use during the analysis and interpretation of the network data.

Chapter	Pages	Duration	Dates	Redaction
I	67 - 115	16 years		
II	117 - 167	9 months	March - December 1728	1763 (?) -
III	169 - 220	16 months	December 1728 - April 1730	1765
IV	221 - 269	19 months	April 1730 - October 1731	
V	271 - 326	10 years	End of 1731 - Autumn of 1741	1766 -
VI	327 - 380			1768
VII	383 - 471	8 years	Autumn of 1741 - Summer of 1749	
VIII	473 - 538	6.5 years	October 1749 - April 1756	
IX	539 - 637	21 months	April 1756 - December 1757	1769 (oct.) -
X	639 - 704	3 years	December 1757 - December 1760	1771 dec. (?)
XI	705 - 758	2.5 years	December 1760 - June 1762	
XII	759 - 847	3.5 years	June 1762 - October 1765	

Table 4.1: Chapters' pages, duration and time of writing (Trousson and Eigeldinger, 2012, vol. 1, p. 49). The last thirteen years of Rousseau's life are not told in *Les Confessions* (Voisine, 2011, p. xxxvi).

In order to give a preliminary quantified overview of the edition we are working on, we provide bar charts of table 4.1, with focus on various aspects of the chapters' properties. The results appear in figures 4.1, 4.2 and 4.3. These plots are glimpses into the narrative process.

The first one (4.1[12]) considers the durations, and converts them from years to months when needed. Chapter I appears as an outlier, and to some extent chapters V to VIII too, spanning significantly more years than the other chapters, while chapters II to IV and chapter IX correspond to much shorter durations. Indeed, the first chapters (including chapter I

[11] A comparison of the footnote blocks sizes between chapters IV (smallest total of pages) and IX (highest total) do not show any difference of average sizes.

[12] Chapters V and VI are equally divided in two. The transition in between them happens around when Mme de Warens and Rousseau move in the *Charmettes*, but this date is not known (see (Rousseau et al., 2011, p. 258, footnote 1)). Still, this has absolutely no consequences.

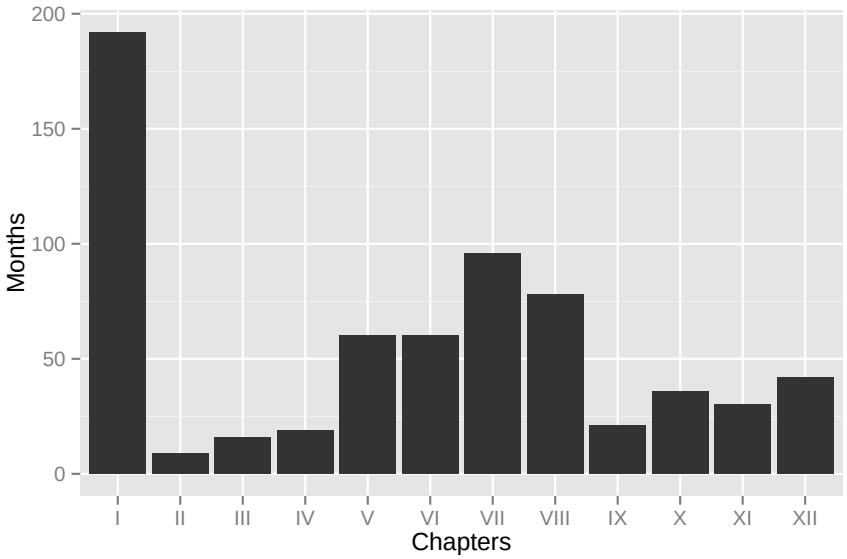


Figure 4.1: Barplot of chapters durations. Values in table 4.1 are converted into months.

to some extent) need more time and boundaries to express the influential events, from his childhood to his life's apprenticeship. They are told because of the importance he gives them into the understanding of his growth as a person, emotionally and intellectually. For example, in chapter II, the religious instruction given to him during a short time spent in a monastery in Turin[13] will have later influence his position towards religion. The other events from these chapters are the encounters with Mme de Warens, before this scene, and Mme Basile and Mme Vercellis, after. It ends with the theft of the ribbon, another fundamental event that Rousseau recalls later in the novel. The importance of these events in understanding his life, from his point of view, and thus the validation of the reason why Rousseau spends deliberately more time describing these moments, is confirmed by the following quote concerning his situation in the entourage of Mme de Vercellis:

> Je crois que j'éprouvai dès lors ce jeu malin des intérêts cachés qui m'a traversé toute ma vie, et qui m'a donné une aversion bien naturelle pour l'ordre apparent qui les produit.[14] (Rousseau et al., 2012, p. 162)

The second chart (4.2) shows the length of each chapter in units of pages. This is simply the subtraction of the first page from the last one, plus one. In this case, chapters VII, IX and XII are significantly longer. The other nine chapters—especially the first six—have similar sizes, showing that there was some sort of editorial regularity that it was necessary to exceed three times.

[13] The stay is described from page 136 to page 148. One quarter of the pages of the chapter.

[14] "I believe I felt from then this sly game of hidden interests which crossed me all my life, it gave me a natural dislike against the apparent order that produces them." [Own translation.]

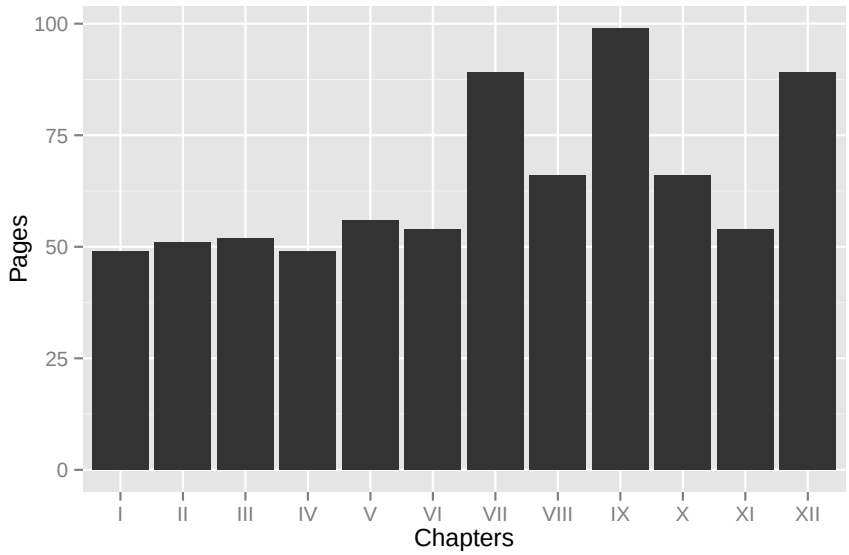


Figure 4.2: Barplot of chapters pages. Values are taken from table 4.1.

The third chart (4.3) is the resultant of plots 4.1 and 4.2. It is the ratio of the number of months by the number of pages, and it gives the speed at which the narration flows. We observe that the start of his life logically needs less time[15] to be told: memories from a very young age are rare for everybody, and Rousseau does not relate many situations he did not live consciously (e.g. the description of his parents' life prior to his birth). On the contrary, the following years—when he experienced striking influential events—take more time to develop. For example, the second chapter has one of the shortest numbers of pages but also spans the shortest duration by far. After this chapter, we await longer and more developed narrative events[16], but which also contain a much lower density of characters. This is an important fact to keep in mind for the time of the analysis. Chapter IX shares similar properties: its speed is low in comparison to the other chapters. It sees Rousseau describing his stay in the countryside, his platonic relationship with the comtesse d'Houdetot, and the end of his

[15] According to the figure, a page corresponds to almost four months.

[16] At the end of chapter IV, which bears similarities of that kind with chapter II, he says:

> "Ces longs détails de ma première jeunesse auront paru bien puérils, et j'en suis fâché : quoique né homme à certains égards, j'ai été longtemps enfant, et je le suis encore à beaucoup d'autres. Je n'ai pas promis d'offrir au public un grand personnage ; j'ai promis de me peindre tel que je suis ; et, pour me connaître dans mon âge avancé, il faut m'avoir bien connu dans ma jeunesse." (Rousseau et al., 2012, 268)
>
> "These long details of my early youth must have appeared trifling, and I am sorry for it: though born a man, in a variety of instances, I was long a child, and am so yet in many particulars. I did not promise the public a great personage: I promised to describe myself as I am, and to know me in my advanced age it was necessary to have known me in my youth."

Translation by S. W. Orson, Privately Printed for the Members of the Aldus Society, London, 1903, see http://en.wikisource.org/wiki/Confessions_(Rousseau), accessed on 23/07/2014.

friendship with Louise d'Épinay, Diderot and Grimm. Since it may be the origin of the plot against him, this chapter can be considered as setting the stage.

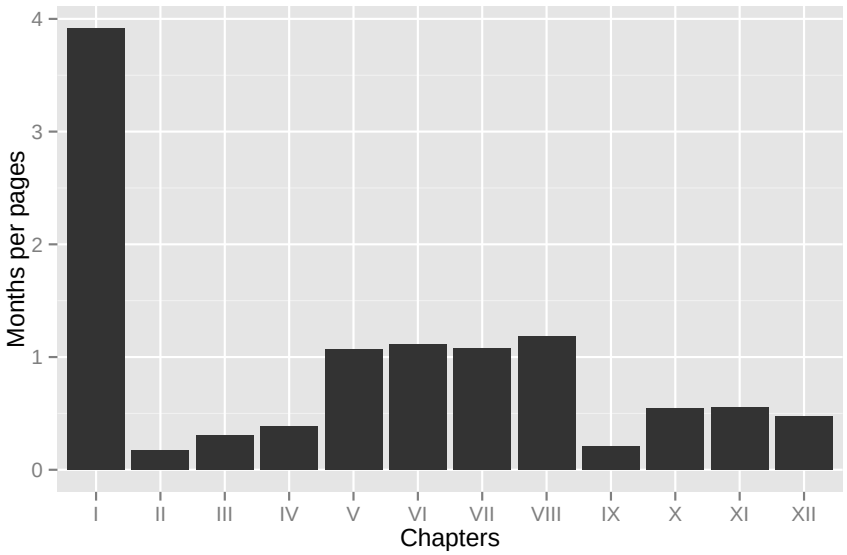


Figure 4.3: Barplot of months per pages ratio for all chapters. Values are taken from table 4.1.

There clearly were constraints and forms of editorial symmetry in the building of the novel's structure, highlighted by the corresponding tones of the two parts:

> L'ensemble révèle un souci architectural : deux parties, de six Livres chacune, couvrant approximativement la même durée, la première contenant l'évocation des temps heureux, la seconde relatant surtout les combats contre la société corrompue et la chute dans le malheur.[17] (Trousson and Eigeldinger, 2012, vol. 1, p. 55)

This is remarkable in the sense that it shows that non-symmetrical features, like the high number of pages of three chapters, or the high speed of a few chapters and very low speed of others, are non-negligible facts.

4.2.3.2 Places

Jean-Jacques Rousseau travelled extensively during his life. In table 4.2, we show the main places where he lived or travelled during each chapter. This table helps us to understand some of the results in later chapters, such as explaining why given people are close in the network, or why some are recurrent over time.

There seems to be no correlation—at least visually—between the number of places visited and the narrative speed (see figure 4.3) of a chapter.

[17] "The whole reveals an attention to the architecture: two parts composed of six chapters each, covering approximately the same duration. The first part contains the evocation of happy times, the second mostly recalls the fights against the corrupted society, and the fall into misfortune." [Own translation.]

Chapter	Places
I	Geneva
II	Annecy, Turin
III	Turin, Annecy, Lyon, Annecy
IV	Annecy, Thônes, Geneva, Nyon, Lausanne, Neuchâtel, Fribourg, Bern, Solothurn, Paris, Lyon, Chambéry
V	Chambéry, Besançon, Chambéry
VI	Chambéry, Geneva, Chambéry, Montpellier, Chambéry, Lyon, Chambéry
VII	Lyon, Paris, Venice, Paris
VIII	Paris, Fontainebleau, Paris, Geneva, Paris, Montmorency (Ermitage)
IX	Montmorency (Ermitage), Montmorency (Mont-Louis)
X	Montmorency (Mont-Louis), Montmorency (Luxembourg)
XI	Montmorency (Luxembourg), Bern
XII	Yverdon, Môtiers, St-Pierre island, Biel/Bienne

Table 4.2: Main places in chapters of *Les Confessions* (in chronological order).

4.3 Index of characters

An index of occurrences of characters comes with the text of the Slatkine edition (Rousseau et al., 2012). This index is part of our corpus, if not, in a sense, the whole corpus itself. We use this index to encode the character-system of the novel, and to build a corresponding character network (see chapter 6). This index of characters contains only names and their occurrences on pages. Even if it had contained subheadings, we would not have needed them for the building and the analysis[18]. The first page of the index[19] is shown in appendix A.

The index is positioned on pages 1035 to 1057 in the printed edition and concerns only the first two volumes, which contain Rousseau's autobiographical works, and therefore *Les Confessions*. The complete edition is composed of twenty-four volumes and around twenty scholars have been working on it (Rousseau et al., 2012, p. XVIII). If the index mentions a scholar or an individual cited by them, the page number is accompanied by an "n", e.g. "ABES

[18] Still, we used the index of characters from (Rousseau et al., 2011) to help in the process of interpretation, because that index possesses subheadings (discussion on subheadings, see chapter 3).

[19] On *Les Confessions* only, and not from the complete works.

Francesco : 163n" in appendix A. In our extraction, we omit the occurrences corresponding to these last two cases.

The compilation of the index has been done by Jacqueline Benoit-Rohde (Rousseau et al., 2012, p. XXIII, vol. 1). In the online version[20], the pages containing the index do not exist. Instead, the index has been made interactive. For this work, we had access to the files used to construct the searchable index, which are based on the printed version, and which have been corrected and updated even after the publication of the printed version. Generally speaking, the scanning and optical-character recognition (OCR) of such indexes is a simple task—when it is not already available.

The index comes in two parts. The first one consists of a table with numerical IDs and their corresponding character names. These were modified for both computational and aesthetic reasons. Each character corresponds to one and only one entry. The second file is a table in which each of the 46540 entries is one occurrence in a page of the text. Two occurrences on a same page still imply only one entry. For each entry, there are four attributes: the ID of the character, the volume and the page in which it appears, and finally a binary indicator stating whether the name appears in a footnote. When restricted to *Les Confessions*, and after excluding meta-occurrences—footnotes and comments—, the number of occurrences is reduced to 2088.

The index includes some of the existing transcriptions—concordances—of certain characters. For example, the nickname "Maman" on page 309 refers to the character Mme de Warens[21]. To our knowledge, no pronouns or indirect mentions are referenced[22]. In the text of *Les Confessions*, Rousseau cites 583 characters[23].

It is important to notice that Rousseau himself is not indexed, despite his first or last name appearing a few times. We discussed that aspect of indexing in section 3.3. There is the notable exception of "Dudding", which is a pseudonym Rousseau used when he met Mme de Larnage, in chapter six. Otherwise, some non-contemporary characters have been indexed (e.g. greek philosophers).

In figure 4.4, we show the distribution of occurrences per page along the whole text of *Les Confessions*. The units on horizontal axes of both plots are equal, showing that the second part is more extended. We remark that chapter two is indeed populated by fewer characters, on average, as announced in section 4.2.3.1. In appendix C, we show the distributions of occurrences for all characters and chapters. It is a useful referential tool which brings a

[20] rousseau.slatkine.com

[21] More exactly : "Françoise-Louise Éléonore de la Tour, baronne de Warens (ou Maman)" as the indexers have recorded it.

[22] For example, on page 71, "ma mère" and "mon père" are not indexed, but they were on the previous pages when explicitly cited.

[23] There are 8014 characters in the index for the twenty-four volumes of the complete works; 4306 are cited by Rousseau, the rest by scholars only, e.g. as bibliographic references.

temporal dimension to the index of characters. It is also a basic quantified representation of character-spaces.

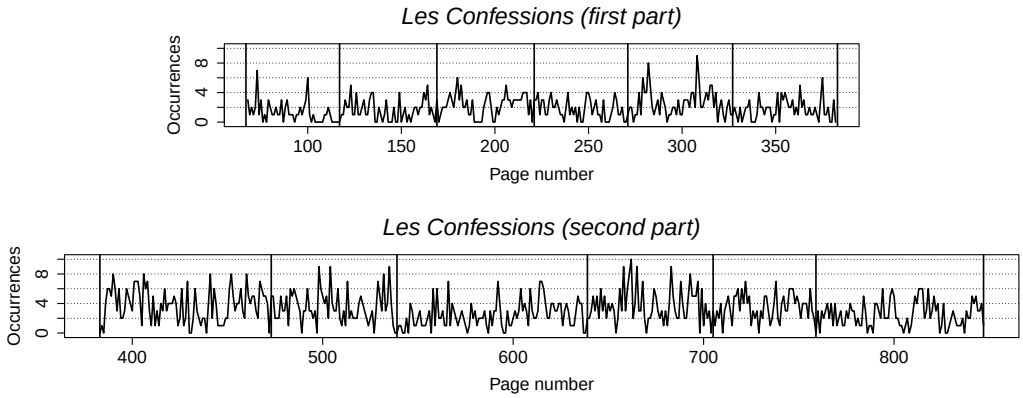


Figure 4.4: Number of occurrences per unique character per page. Page numbering is based on Slatkine Edition (Rousseau et al., 2012). Vertical lines are chapter limitations. One page occupies the same space on horizontal axes of both plots.

5 Character network analysis

[As Chase's nitro-laden hot rod careens towards the Gila Monster.]
Tom [as Colonel Kurtz]: The horror! The horror!
[The hot rod collides with the lizard and explodes.]
Crow: Aw, they killed off the only likeable character![1]

S04E02: The Giant Gila Monster
MYSTERY SCIENCE THEATER 3000

In a novel, characters are positioned all along a narrative, in an order usually dependant on the storyline and their role in the narrative. For example, in detective fiction, we are accustomed to the murder of at least one of characters appearing in the first pages. In Joseph Conrad's *Lord Jim*, the narrator plays an active role in manipulating the narration of the main storyline. He tells the story non-linearly and with frequent interruptions—this modifies the perception of the story. In general, some characters are more prominent than others in two ways:

- The distribution of their occurrences along the text (they may occur only once or on each page).
- Their apparitions along with other characters (they may be clustered inside a few pages, or spread throughout the whole book, along with many characters or none).

All of this originates in the author's act of writing. Determining how the author obtains that arrangement, if possible, is not evident, since he or she may not be conscious of the final relational image shaped by the characters' relations in the narration. We use films as an analogy to demonstrate that phenomenon: according to screenwriter teacher John Truby, relations between characters in film scripts are a component the scriptwriter must manage, citing examples in (Truby, 2007, chap. 4). In this reference, his advice is to define the character-system of the film as a "web":

> The single biggest mistake writers make when creating characters is that they think of the hero and all other characters as separate individuals. Their hero is

[1] As transcribed on en.wikiquote.org/wiki/Mystery_Science_Theater_3000, accessed on 27 Feb. 2014.

> alone, in a vacuum, unconnected to others. The result is not only a weak hero but also cardboard opponents and minor characters who are even weaker. [...] To create great characters, think of all your characters as part of a web in which each helps define the others. To put it another way, a character is often defined by who he is not. (Truby, 2007, chap. 4)

This point of view focuses on how relations build the character descriptions—this web has an influence on the elements of the character-system. In some cases, scriptwriters show "social networks" of characters with complex and obviously non-random design. Such unique networks prove that the authors have perspective on their work and are meticulously organising the "social" relations following a given purpose: e.g. in the case of *Magnolia*[2], written by Paul Thomas Anderson, or *Babel*[3], written by Guillermo Arriaga, both works show highly clustered groups of loosely connected characters forming cycles at the scale of the whole systems (see figure 5.1). Thus, these networks show that there is a structural motivation behind the organi-

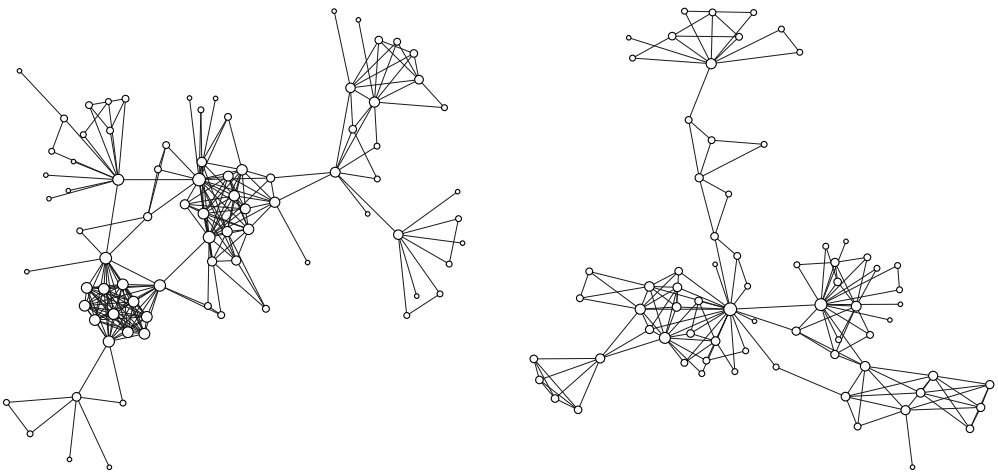


Figure 5.1: Character networks of two movies[4]. (Left.) Magnolia (1999). (Right.) Babel (2006). Both show highly clustered peripheral communities, as well as large non-reducible cycles in their centres.

sation of the narrative. However, a movie may show two or more characters at the same time on screen for the sake of aesthetics or playful reasons, it is not straightforward to determine if there is an equivalent in written stories.

Since in this work we focus on a corpus from the specific genre of autobiography, we leave the question of whether the author knowingly or randomly positions his characters—the process of configuring the character-system, that is to say space of character-spaces—

[2] See moviegalaxies.com/movies/92-Babel, accessed on 20/03/104.
[3] See moviegalaxies.com/movies/523-Magnolia, accessed on 20/03/104.
[4] Data are imported from moviegalaxies.com, see section 1.3.

untouched. The genre to which the work under study belongs also implies that the story is based on real-life facts, therefore slightly limiting the author's freedom in defining the underlying character network.

We identify four categories of network uses in literary studies:

- Networks of characters from a single work, e.g. (Moretti, 2011; Agarwal et al., 2012).
- Networks of recurring characters from a set of interlocking works:
  - depending on the definition of a relation as a co-presence in a narrative event, e.g. (Cuckier, 2013; Grandjean, 2014),
  - depending on a partition by books or chapters, e.g. (Alberich et al., 2002)).
- Networks based on the text and implying narrative entities outside characters, e.g. (Batagelj et al., 2002).
- Networks of literary influences, e.g. (Jockers, 2013; So and Long, 2013).

For this work, we consider the first two cases, which focus on characters and their relations only. We consider two dimensions in these works, which are the extraction process and the analysis.

Firstly, we review the status of the character (5.1), and introduce the notion of "character-system" (5.2). Then, we explore the character network analysis discipline, along with the concepts of character-space and character-system, and how to express them in terms of network analysis concepts (5.3). Eventually, we propose a framework for character network analysis (5.4). This framework is used in chapters 6 and 7, where it links potential research questions with existing works.

5.1 Characters

> "[...] I'm always embarrassed when I see an index an author has made of his own work."
> "Embarrassed?"
> "It's a revealing thing, an author's index of his own work," she informed me.
> "It's a shameless exhibition to the trained eye."
> "She can read character from an index," said her husband.
>
> *Cat's cradle*
> KURT VONNEGUT

There is a philosophical debate about the role of the character in fiction. This notion was first acknowledged in Antiquity, in Aristotle's *Poetics*, and then evolved with time, thanks to works revolutionising that idea—among them such milestones as Cervantes' *Don Quixote*, Mme de La Fayette's *Princesse de Clèves*, Honoré de Balzac's *Comédie humaine*, and *Nouveau roman* authors (Bardet, 2007). We do not use mathematics to investigate what a character is, but rather to describe the entities as they are eventually cited in an index. A discussion

about the concept of character—to determine if it is a simple vector for the action or the one element that must be present in every work of fiction and on which all the rest depends—is outside the scope of this work. The status of the character does not matter here: we study the entities chosen by the indexer.

A definitive position on the role and the importance of the character in narratives is a complex question, and probably an unresolvable one. When trying to define that notion, two opposing approaches are used:

> The status of characters is a matter of long-standing debate: can characters be treated solely as an effect created by recurrent elements in the discourse (Weinsheimer 1979), or are they to be seen as entities created by words but distinguishable from them and calling for knowledge about human beings? (Jannidis, 2012)

Thus, is the character defined by a narrative claim or feelings to convey—thus by content—, or does it exist beforehand? Is it a puppet with the sole purpose of inhabiting an author's ideas, or is it alive, a figure that we feel may exist outside of the story world? These ideas are rather in contradiction, but in the end they summarise the two sides of the narrative process. Thus, the problem to solve can be reduced to a question of narrative process: is the message about a society and a set of events and actions defined on them, or is it about situations for which we use a bunch of puppets, for the sake of illustration? Mudrick labels these two views as "realistic" and "purist" arguments:

> One of the recurring anxieties of literary critics concerns the way in which a character in drama or fiction may be said to exist. The "purist" argument—in the ascendancy nowadays among critics—points out that characters do not exist at all except insofar as they are a part of the images and events which bear and move them, that any effort to extract them from their context and to discuss them as if they are real human beings is a sentimental misunderstanding of the nature of literature. The "realistic" argument—on the defensive nowadays—insists that characters acquire, in the course of an action, a kind of independence from the events in which they live, and that they can be usefully discussed at some distance from their context. (Mudrick, 1961) as cited by (Rimmon-Kenan, 2002, pp. 31-32)

Diverse views about the character are currently being presented and compared. However, in our approach, since we eliminate all words and textual context (our data consists of an index of characters and nothing else), we can rely on a simplified definition, which follows the intuition behind the "realistic" argument. This stance, added to the fact that *Les Confessions* is a novel inspired by real-life persons, motivates us to consider the set of characters from the book as a society bearing many similarities to those in the real-world. The characters may

be envelopes or vectors for ideas; but they also are the core of the story; the pivots of a pure factual narration. As Mieke Bal states:

> Un événement n'est pas possible sans acteur [...].[5] (Bal, 1984, p. 5)

Roles Characters may be manipulated or, on the contrary, be strong willed incarnations; however, in any case, they play a role in the discourse: they follow characterisation archetypes. According to (Rimmon-Kenan, 2002), in his influential study on russian folk tales, Vladimir Propp states that the character roles can be reduced to just a few[6], and a character may personify more than one role. In part, these roles are defined by co-appearing context and thus by the neighbouring characters. The character plays a central role in the narrative environment:

> In fiction the character is used as the structuring element: the objects and the events of fiction exist—in one way or another—because of the character and, in fact, it is only in relation to it that they possess those qualities of coherence and plausibility which make them meaningful and comprehensible. (Ferrara, 1974) cited by (Rimmon-Kenan, 2002, pp. 35)

Despite these positive assertions, Jannidis states in a review article that all theories up to this point have been inconclusive. Character studies have explored many paths, but there seems to be no way out, no alternative to a nearly century old basic theory of characterisation:

> However, most proposals seem to be either too complex or theoretically unsatisfying, so that Forster's classification into flat[7] vs. round[8] characters continues to be widely used. (Jannidis, 2012)

Even so, we go on with the exploration of characterisation. Our case is specific and simplified enough to circumvent the difficulties of assembling a new theory of characterisation. In some

[5] "An event is not possible without an actor." [Own translation.]
These two concepts are defined earlier in her text:

> "Un *événement* est le passage d'un état à un autre. Tout changement, aussi minime qu'il soit, constitue un événement. Les *acteurs* sont les instances qui agissent. Ils ne sont pas nécessairement humains." (Bal, 1984, p. 4)
>
> "An *event* is the passage from a state to another. Every change, as slim as it is, forms an event. The *actors* are the acting instances. They are not necessarily humans." [Own translation.]

[6] "The villain, the donor, the helper, the sought-for-person and her father, the dispatcher, the hero and the false hero." (Rimmon-Kenan, 2002, p. 34)

[7] "In their purest form, they are constructed round a single idea or quality: when there is more than one factor in them, we get the beginning of the curve towards the round. The really flat character can be expressed in one sentence such as "I never will desert Mr. Micawber." There is Mrs. Micawber—she says she won't desert Mr. Micawber, she doesn't, and there she is." (Forster, 1963)

[8] A round character is a character with more dimensions than a flat one. It is a more unpredictable character: his situation changes along the story.

of the main works on the subject, we find hints to the mathematical approach we would like to set up. However, before delving deeper into that idea, we cite a warning from Philippe Hamon when considering a globally quantified analysis of a book. His arguments help to identify the potential pitfalls. He states that:

> [...] il est particulièrement difficile de circonscrire une étude du type: "Les personnages dans l'oeuvre de X ou Y". En effet le "personnage" n'est pas, à la différence du "dialogue", des "dates", de la "description", de "l'histoire", du "titre", des "métaphores", du "récit", de tel ou tel "thème", etc., un champ d'étude facilement et immédiatement identifiable: le personnage n'est pas réductible à la seule apparition textuelle d'un nom propre; il n'est pas dénombrable même [...] et est donc inaccessible aux méthodes quantitatives; il est d'autre part mal localisable en un point précis du texte, ce qui le rend inaccessible aux méthodes purement distributionnelles; [...][9] (Hamon, 1998, pp. 18-19)

We will avoid his warnings in our work by studying what results from the work of the indexer. The task of locating characters is not automatically done, but it is frequent enough to let us consider a large-scale study. This method of relying on the index solves the problems of identification and location. In this way, we study a filtered version of the novel; we study the text through the lens of the indexer. In addition to this argument, recent scientific production contradicts Hamon's somewhat pessimistic position (Agarwal, 2011; Moretti, 2011; Gil et al., 2011; Elson, 2012; Mac Carron and Kenna, 2012; Sack, 2012; Nalisnick and Baird, 2013a; Heyn, 2013; Mac Carron and Kenna, 2013). Vermeule highlights the necessary continuity to consider when studying characterisation. In particular, she studies the character with the reading process in mind:

> Fictional characters come trailing many cognitive puzzles. Some are constant over time while others change. Yet for all their historical complexity, fictional characters stay the same in one respect: they are the greatest practical-reasoning schemes ever invented. We use them to sort out basic moral problems or to practice new emotional situations. We use them to cut throughout masses of ambient cultural information [...] so we use them in place of statistics as told to muddle through. (Vermeule, 2011, p. xii)

[9] "It is particularly difficult to circumscribe a study such as "the characters in the complete works of X or Y". The "character" is not, on the contrary to "dialogue", "dates", "description", "story", "title", "metaphors", "discourse", such and such "theme", etc., an area of research easily and immediately identifiable: the character is not reducible to the sole textual occurrence of a proper noun; it is not countable [...] and is thus unreachable to quantitative methods; it is localisable with difficulty in a single spot in the text, which renders it unreachable to the totally distributive methods; [...]" [Own translation.]

Shlomith Rimmon-Kenan shares this vision of the character aimed at being an interaction between the writer and the reader. She summarises the perception of its role from outside the novel:

> In the text characters are nodes in the verbal design; in the story they are—by definition—non (or pre-) verbal abstractions, constructs. Although these constructs are by no means human beings in the literal sense of the work, they are partly modelled on the reader's conception of people and in this they are person-like.
> Similarly, in the text, characters are inextricable from the rest of the design, whereas in the story they are extracted from their textuality. (Rimmon-Kenan, 2002, p. 33)

Incidentally, her language as well as a previous quote from her show some glimpses of quantitative thinking: "nodes", "abstractions", "constructs"—they all concern a structural approach focusing on representation and placement of these actors in the narrative environment. To add to the semantic, the same author proposes an adaptative solution for character classification:

> In order to avoid reductiveness[10], Ewen ([...]) suggests a classification of characters as points along a continuum rather than according to exhaustive categories. (Rimmon-Kenan, 2002, p. 41)

This distinction should take place along three axes—"complexity, development, penetration into the 'inner life'"—, she adds. This way, the characters are positioned in space, and can therefore be clustered. However, this interesting approach relies on close reading to decide on the attribute values—the positions on the axis. This could be solved by semantic extraction, but once again, this goes outside the scope of our thesis.

Relations in story and discourse After explaining that structuralism had difficulties defining the concept of the character, Rimmon-Kenan suggests—among other arguments—that characterisation should consider all characters occurring in the corpus with no exception; it should study their function in the narration by taking into consideration all narrative elements:

> [...] do not even the minimal depersonalised characters of some modern fiction "deserve" a non-reductive theory which will adequately account for their place and functioning within the narrative network? (Rimmon-Kenan, 2002, p. 31)

Characters are defined at their introduction, by a physical or mental description, quoted speech, the situation, or by receiving no description. Recognising a character in the text means that there is an identification process (Jannidis, 2012). Amongest other things, characters

[10] The "flat" versus "round" characterisation from (Forster, 1963).

co-appearing help define one another. For example, Sherlock Holmes and his assistant Watson show the division of a principal character into two: one to bring the story forward, raise the confidence of the reader, and also to generate suspense by leaving him/her ignorant; the other by being an average person, asking the questions the reader has in mind. Each character's existence in the story depends on the other's. This is expressed by Rimmon-Kenan with more generality:

> When two characters are presented in similar circumstances, the similarity of contrast between their behaviour emphasises traits characteristic of both. (Rimmon-Kenan, 2002, p. 70)

5.2 Character-system

> "As to the mere physical possibility we all know that some speeches in Parliament have taken nearer six than three hours in delivery; whereas all that part of the book which is Marlow's narrative can be read through aloud, I should say, in less than three hours. Besides–though I have kept strictly all such insignificant details out of the tale–we may presume that there must have been refreshments on that night, a glass of mineral water of some sort to help the narrator on."
>
> Foreword to *Lord Jim*
> JOSEPH CONRAD

Despite many initiatives to generate and to study character networks with mathematical methods, only few of them have a real basis in literary theory rather than computer science/statistical physics[11]. Thus, we begin the discussion here with two pamphlets by Franco Moretti, presenting the ties between literary theory and network analysis (Moretti, 2011, 2013b). We consider them as good starting points to discuss the foundations of character network analysis.

The first pamphlet is a preliminary exploration of using network analysis concepts to understand a literary work—here William Shakespeare's *Hamlet*. The second is a subsequent contextualisation of the first text—it deals with the "operational approach":

> Operationalizing means building a bridge from concepts to measurement, and then to the world. In our case: from the concepts of literary theory, through some form of quantification, to literary texts. (Moretti, 2013b, p. 1)

The article's intentions are "pamphletary" in the sense that it was written without too much rigour, but rather with the will to convince that measurement can bring something new to literary study. To establish that position, Moretti chooses the notions of character-space by Alex Woloch (Woloch, 2003) as the "concept of literary theory", and social network analysis as the "form of quantification". Woloch's concepts of character-system and character-space

[11] A state of the art focused on character network analysis only is in section 5.4.

bring together many elements related to mathematical intuition. His work does not contain any histograms or scatter plots, nor a graph or a single figure; nonetheless, the developed ideas regularly refer to quantified concepts[12], inviting such an interpretation. Here is how Alex Woloch defines character-spaces (a node and its neighbourhood), and the subsuming character-system (the "social" network):

> My interpretive method rests above all in the combination of two new narratological categories which I will formulate and continually return to: the *character-space* (that particular and charged encounter between an individual human personality and a determined space and position within the narrative as a whole) and the *character-system* (the arrangement of multiple and differentiated character-spaces—differentiated configuration and manipulations of the human figure—into a unified narrative structure). (Woloch, 2003, p. 14)

Using that, Moretti uses some plays (*Phèdre*, *Hamlet*, *Macbeth*, etc.) to show how quantification induces other points of view on existing analysis of literary work. On this occasion, he interprets the notion of character-space in two manners: the first as the unidimensional "word-space"—the "number of words allocated to a particular character"—; the second as a network analogy. His analysis allows him to find some discrepancies when computing both spaces: in some cases the perceived structure is what results from the analysis, while in others irregularities suggest that the focus be on the reasons for that. For example, in *Phèdre*, the character with that name has the largest word-space, while Thésée is at the centre of the play's character network. Thus, his two measurements of "character-spaces" lead to different conclusions if taken separately:

> And it's not that one is right and the other is wrong; rather, they capture different features of dramatic networks: the number of links tells us how connected a character is [...]; the number of words tells us how much meaning the character brings into the play [...]. (Moretti, 2013b, p. 5)

Then, his reasoning reaches a more final conclusion[13], going back to the topic of operationalization:

> The protagonist is a utensil; character-space, is an instrument. The protagonist is a utensil because it belongs to the world of readerly common sense, and doesn't go beyond it. Character-space is an instrument, because it's the realization of a theory that wants to understand something "that does not fall under the domain

[12] For example: "space", "set", "amount", "duration", "arrangement", "intersection", "socionarrative matrix", "interconnected", "refracted", "delimit", "distort", "assymetry", "permutation", "networks/web of characters", "biography matrix", etc.

[13] The first part of the pamphlet discusses how to import "quantification-ready" theories like Woloch's. The second part deals with exportation of operationalisation to qualitative concepts.

> of our senses": instead of individual characters, the relations among characters. That's why, in the end, its operationalization produced more than the refinement of already-existing knowledge: not the protagonist, improved, but an altogether new set of categories. (Moretti, 2013b, p. 9)

We see Moretti's approach, as well as the consistency of Woloch's characterisation[14], as the theoretical foundations of the mathematical framework supporting our analysis. This is unveiled later, in section 5.3. In addition, we discuss the echoes and possible interpretations of concepts like character-system and character-spaces in topology and graph theory. Eventually, in section 7.5, we show how centrality measures undermine dimensions of narrative roles, starting with Woloch's reasoning for the major–minor distinction between characters.

5.3 Definition

> How much greater a thing it is to compile than to analyse—almost, we might say, to create than to destroy.
>
> *Indexing fiction: a story of complexity*
> HAZEL K. BELL

We consider Alex Woloch's theory of the character-system (Woloch, 2003) as the starting point from which to sketch a definition of character network analysis. We borrow ideas from Franco Moretti, who explored this approach in (Moretti, 2011, 2013b). The goal of character network analysis is to propose a contextualisation of graph theory to represent characters within a plot. The basic element is the relation we infer between characters—its nature is defined by the research questions and the corpus. Character network analysis describes, in a quantified manner, characters' roles based on their positions: character-spaces. It provides a means for classification of the positions of a whole set of characters—is a given character a protagonist? a minor character? minor but key? minor minor? We use the positions and ignore the textual contexts. When studying a literary work as extended as *Les Confessions*, this approach is useful in obtaining a global overview (6.2), and to highlight local narrative events (6.3) and key characters (7.4 and 7.5).

The relational approach to plots was praised even before Woloch and Moretti. In his character study of the twenty novels forming Émile Zola's *Les Rougon-Macquart* in 1983, Philippe Hamon showed that they have differently structured character-systems:

> Polyfocalisation du système, enfin sur plusieurs héros — plutôt que défocalisation —, qui partagent en alternance le ou les "point(s)-héros" (les "foyers") du système, polyfocalisation dont *Pot-Bouille*, *La Bête Humaine* et *La Débâcle* sont les meilleurs exemples, procédé de la nébuleuse, ou du réseau à "noeuds" marqués

[14] Defined by himself as: "the literary representation of imagined human beings" (Woloch, 2003, p. 14).

> et à tissus interstitiel lâche, qui s'éloigne donc de la hiérarchie fixe "en pyramide" (un héros, des personnages secondaires, des personnages de troisième plan, etc. selon une échelle non modulable) des récits classiques.[15] (Hamon, 1998, p. 320)

The stories of these three books in particular[16] are told in such a way that the character-systems are not strictly hierarchical. For example, *Pot-Bouille* tells the story of middle-class inhabitants from different flats in a Parisian building. The focus changes from one chapter to another and from one flat and its inhabitants to another; it is difficult to identify protagonists[17]. Émile Zola consciously orchestrates these variations of focus in order to soften the boundaries between the characters and to suggest that all of them hide secrets and bad manners in the privacy of their households. A character network analysis of this plot may show many characters having equivalently central positions, implying that they are presented in the narrative in similar manners.

Another manifestation of interest in relations between narrative entities in the text occurred when Shlomith Rimmon-Kenan wrote an influential book on narratology called "Narrative Fiction" (Rimmon-Kenan, 2002). Narratology studies the transformation of a story (the plot in the writer's mind) into a discourse (the text in the hands of the reader); it makes use of many tangible concepts. We cited her work earlier; and she observes that this discipline does not take the character much into consideration—an unavoidable category among the narrative entities (Bal, 1984)—as an object of study. However, she does recognise the existence of underlying relations among the characters as the narration moves forward:

> Far from seeing story as raw, undifferentiated material, this study stresses its structured character, its being made of separable components, and hence having the potential of forming network internal relations. (Rimmon-Kenan, 2002, pp. 6-7)

Character network analysis should aim to represent ideas such as Hamon's and Rimmon-Kenan's. It must allow us to:

- Categorise patterns at local and global levels.
- Reveal points of interest—underlying motifs—, like identifying narrative events or specific narrative roles for characters (the network acts as the exploratory variable).
- Obtain explanations about structural phenomena (the network acts as the explanatory variable).

[15] "Polyfocalisation of the system on a few heroes—rather than unfocalisation—, which alternately shares the "hero spots" of the system, polyfocalisation of which *Pot-Bouille*, *La Bête Humaine* and *La Débâcle* are the best examples, processes issued from a network made of marked "nodes" and interstitial light layers, which take distance from a fixed "pyramid-like" hierarchy (a hero, secondary and marginal characters, etc. according to a non-adjustable scale) of classic works." [Own translation.]

[16] To which we can add *La fortune des Rougon*.

[17] Except maybe Octave Mouret, who is a pretext to visit all households, and will be one of the two protagonists in the following novel, *Au bonheur des dames*.

- Produce and work on graphical representations.

Alex Woloch's theory offers a narratological environment to accompany character network analysis. In particular, we translate the notions of character-space and character-system (defined in section 5.2) into graph theory objects. Mathematically, the character-system is similar to a topological space. The character-spaces are open sets—thus an open set always contains at least a character—and the union of any number of character-spaces is still a character-space. The case of the intersection is somewhat problematic, since the topology must be designed so that the intersection of two open sets always contains a character. Alex Woloch raises a list of questions[18] that illustrate how close his objectives are to ours:

> My study addresses and connects a series of questions that have never been conceptually formulated but that are provoked by, and often essential to, any number of narratives. (1) What is the purpose or significance of a particularly marginal character? (2) How much access are we given to a certain character's thoughts, and (3) how does the partial enactment of this perspective or point of view fit into the narrative as a whole? (4) Why and how are certain narratives divided between two or three central characters? (5) How often, at what point, and for what duration does a character appear in the text? (6) How does she enter and exit specific scenes? (7) How does her delimited position intersect with the achieved representation of her speech, actions, or physiognomy? (8) How are her appearances positioned in relation to other characters and to the thematic and structural totality of the narrative? (9) Why does a particular character suddenly disappear from the narrative or abruptly begin to gain more narrative attention? (10) How does the text organise a large number of different characters within a unified symbolic and structural system? (Woloch, 2003, p. 14)

We clearly see how the motivations underlying his theory are compatible with and invite a quantified approach, since all his questions concern measurement of aspects of the narration ("how much", "how often", etc.) (Moretti, 2013b). In particular, questions 1, 4, 5, 6, 7, 8 and 10 can be totally or partially addressed using a network approach and distribution of characters occurrences[19]. For example, the purpose in question 1 can be studied by exploring the neighbourhood of that "particularly marginal character[20]", along with its occurrences in the text. It would be highly significant—as "marginal"—if it were not to appear with any protagonists, for example. Then, question 8 essentially announces character network analysis:

[18] The number were not in the original quotation. We added them to improve readability.

[19] Most, if not all, of the known methods building networks either require the positions of the characters in the text, or obtain them via an extraction process. Thus, occurrences may be considered as an attribute coming automatically with the network.

[20] First, it would be necessary to agree on a definition of that description.

Woloch describes the network interpretation of his notions of character-space and character-system, respectively.

5.3.1 Node neighbourhood as character-space

In Moretti's theory, a character-space is a "narratological category" describing—or even defining—a character:

> The character-space marks the *intersection* of an implied human personality [...] with the definitely circumscribed form of a narrative [...] In this perspective the implied person behind any character is never directly reflected in the literary text but only partially inflected: each individual portrait has a radically contingent position within the story as a whole; our sense of the human figure (*as* implied person) is inseparable from the space that he or she occupies within the narrative totality. (Woloch, 2003, p. 13)

The description of a character-space rests upon the narrative elements near the character. A network approach extracts the "social" dimension concerning a character in the narration, as well as all the information that it induces, such as the social environment to which it belongs or the narrative events that can be inferred from particular patterns. Character-spaces gather structural information locally, and in the case of a character network analysis, delegate the psychological portrait of the character to a close reading approach. They are defined by the node—or vertex—representing the character, a set of neighbouring vertices, and the relations induced by them (see definition 17, p. 24). In the network, characters' neighbours are the ones with whom it shares narrative events, dialogue or text proximity, depending on the nature of the relations. The open neighbourhood[21] of a character permits us to analyse how they are connected in the absence of the character, what structural importance it has in its part of the character-system, and how the other characters are linked together and rely on its presence. Depending on the research, the quantified character-space may optionally include further nodes by aggregating the character-spaces of the node's neighbours, and so on.

5.3.2 Network as character-system

In summary, the character-space corresponds to the plot structure at the character-level; on a wider scale, the character-system is associated with the structure of the plot at the level of the whole plot. It is the union of all character-spaces, and it spans the characters and their environment through the entire narration:

> [...] all character-spaces inevitably point us toward the character-system, since the emplacement of a character within the narrative form is largely comprised *by* his or her relative position vis-à-vis other characters. (Woloch, 2003, p. 18)

[21] The neighbourhood of a node without the node itself.

Woloch implies that the writing of character spaces is a necessary step in analysing the narrative role of the characters. He also insists that the character-system is the overview of the narrative behaviours of the characters, and that it represents the whole set of characters by referring to their "relative positions". He then gives a more precise description, emphasising from a narratological point of view how character-spaces must be considered not separately but jointly, which leads to the character network. The resulting representation allows us to describe and to explain the role a character adopts in the narrative:

> A novel's character-system consists not merely in the interlocking of a large group of distinct fictional individuals but also in the combination of different kinds of character-spaces, [...][22] Each fictional individual emerges only within a larger narrative framework, shaped by the particular space he or she occupies within a complicated structure. This space is formed through the dynamic interaction, or jostling, among numerous characters who share a limited, and unevenly distributed, amount of narrative attention. The implied person is transformed in this process: whether pushed to the side or fondly rendered, a character's reactive position within the totality flows into his or her specific representation as an individual. (Woloch, 2003, p. 177)

This is stated like a qualitative approach, but it is relevant in a quantified way as well. In character network analysis, the entire network is a representation of the character-system (see table 5.1), thus a model of how the characters appear, interact and evolve. Either character network analysis is what Woloch describes, or his quantitative and our qualitative approaches are so close that describing one essentially describes the other.

	Woloch's theory	**Character network analysis**
Character-system	Social network of the plot	Network
Character-space	Textual environment	Neighbourhood
Character	Name	Node

Table 5.1: Corresponding concepts in qualitative and quantitative approaches.

5.3.3 Asymmetry and centrality

Character-spaces have different ranges depending on the narrative importance of the characters. In general, a protagonist appears along with many other characters in events spanning the whole narrative. A minor character will have shorter appearances. However, there are exceptions, e.g.

[22] "... the various modes through which specific human figures are inflected into the narrative."

> *Great expectations* features a weak protagonist, overwhelmed on all sides by various kinds of minor characters. [...] Dickens subtly and intricately links this structure of characterization to the modern economic and social relations that extinguish Pip's expectations. (Woloch, 2003, p. 35)

In this example, we observe how the character-system and its corresponding character network highlight the intention of the author to use the environment to reflect the trajectory of the protagonist. This contrast between character-spaces is observable throughout the character-system: Woloch calls it, in general, an

> asymmetric structure of characterization—in which many are represented but attention flows toward a delimited centre. (Woloch, 2003, pp. 30-31)

The narration process positions the characters as a function of action and attention. The final static picture that is the whole narration emerges from local phenomena, from the descriptions of diverse sorts of characters, reinforcing one another or, on the contrary, bringing some down while moving others up. These narrative interactions are recorded by co-occurrences in the text, in the narration or in the dialogues.

> In novels like *Madame Bovary* or *Moby Dick*—or *Crime and Punishment*, or *Anna Karenina*—we need both to coordinate the large list of characters and to consider how each individual character-space is combined and differentially refracted through the narrative structure. The realist character-system is always oriented in two directions: toward each uniquely delineated character-space (and the implied human figure that it amplifies or obscures) *and* toward the unified structure, the symbolic or thematic edifice, the interconnected plot that is being constructed through—and often helping to delimit or distort—these character-spaces. (Woloch, 2003, p. 33)

We need a method to measure the asymmetry in the character network. It must quantify how close a character is to the centre, or how far from it, as well as measure its narrative importance and allow for comparison with others. For that purpose, we use the concept of centrality we introduced in section 2.3. It is defined to do a similar task on societies, so we use it in this work to measure the variations of narrative importance of *Les Confessions*' characters (section 5.4 and chapter 7).

5.4 Framework

> "Since the data files were prepared by hand, they are subject to human error. [...] In particular, I recently learned that I forgot to include any connection between Fantine and her infant daughter Cosette, [...] However, I shall never update the file jean.dat, because it is "correct by definition.""
>
> *The Stanford GraphBase*
> DONALD E. KNUTH

Here, we initiate the swing from theory to application. We propose a framework going from the identification of the corpus to a model highlighting the roles of the characters it contains. All along the exposé, we cite numerous examples to provide a consistent overview on the subject, while we focus on character network studies. We consider the following partition:

1. the obtaining of a network via the extraction of characters and links,
2. the analysis of
    - a single network,
    - more than one network.

This outline implies that to conduct a full character network study, one must start with a method of extracting the characters and their underlying relations, thus identify their nature, before moving on to the analysis of the obtained network, which leads to hypotheses or inferences on the corpus and its character-system. Optionally, if the corpus contains several works, a larger-scale approach is necessary. It may bring together all works from an author or that happen inside some boundaries (Mac Carron and Kenna, 2013), or it may compare structural properties between comparable literary works (Gil et al., 2011).

Typical studies are divided into two orientations: statistical physics on one hand, and literary theory on the other hand. The applications in physics began at the end of the twentieth century[23] with the need to test hypotheses, such as small world or preferential attachment, on numerous types of network data sets. It led to the compilation or re-use of networks based on literary works as well as social sciences (Alberich et al., 2002; Stiller et al., 2003; Newman and Girvan, 2004; Stiller and Hudson, 2005; Oliveira and Barabási, 2005). The interpretation of the network in these cases was not the primary interest. This became a concern in a new era starting around 2010—it comprises most of our references. In the last five years, we have seen the popularity of character network analysis grow—this time mostly among humanities researchers, or by scientists borrowing their motivations[24]. This may be caused by

[23] Indeed, this is a caricatural description: computer science and life sciences also have been strongly involved in this subject. However, the seed articles originate from physicists (Watts and Strogatz, 1998; Barabási and Albert, 1999).

[24] There are also examples where one would compile a character network to test visualisation and communication of network results, this time (Favre, 2011; Elson, 2012; Ercolessi et al., 2012; Nalisnick and Baird, 2013b; Park et al.,

the emerging field of digital humanities, aiming at bringing together computational methods of network analysis and humanities, in particular literature studies.

5.4.1 Network building

The first step is to circumscribe the corpus: is it a novel or a play? a chapter? an act? a set of interlocking novels or plays? a movie script? The choice of the source—which may also not be textual, e.g. automatic name recognition in audiobooks or face recognition in movies (Salway et al., 2007; Weng et al., 2009)—is of great importance, as it sets the boundaries of the character-system. The choice is decisive for the identification of characters and relations, as well as setting the research questions.

5.4.1.1 Nodes/characters extraction

We want to recognise the forms a character can adopt in the text, and index them, in order to record the relations and build a network. It is necessary to obtain a list—e.g. extract it from the casting of a film—either manually, but that is costly, or automatically (Elsner, 2012). An intermediary solution—the one we use later—is to use indexes[25]. The words used to depict a character are diverse and provide important preliminary informations about it—e.g. they can be a proper name, or a nickname. Thus, these words also bear information about their corresponding character-spaces: with a semantic extraction process, we may obtain information such as social status, gender, etc. However, in scenes where a character is recurrent, style dictates the use of pronouns in order to eliminate repetition. Pronouns are usually close to the corresponding proper noun, following them, but still, they contain less information and may not be discernible. Pronouns are usually not indexed, but they can be automatically linked to the corresponding character (Kazantseva and Szpakowicz, 2010).

5.4.1.2 Edges/relations extraction

At this stage, we have the list of characters as well as, optionally, their positions in the text. As with the character extraction, there are two available options: the manual or the automatic extraction of relations. In addition, it is also possible to consider the context surrounding the relation, and to measure or describe it.

One possibility is to generate relations from co-occurrences in the same sentence, paragraph, page, scene or event, or from some notion of distance, such as words appearing in the text within a given interval. A definitive choice must be made in order to build a coherent model. For example, in the context of theatre, if we reduce the plot to the co-apparitions of characters, then it suffices to report, for each scene, which characters are present on stage at

2013). This is a popular subject for student projects (Gil et al., 2011; Heyn, 2013; Djurica et al., 2013; Stensson, 2014).

[25] For example, we find interactive visualisations of character networks from *The Silmarillion*, *The Hobbit* or *The Lord of the Rings* from J. R. R. Tolkien at: http://lotrproject.com/statistics/books/cooccurrences, accessed on 8/04/2014.

the same time (Stiller et al., 2003). Some infer relations between characters on the basis of their conversations in the text (Elson et al., 2010), co-occurrences in subtitles (Tan et al., 2014), in scripts parts of the scripts (Gil et al., 2011), or co-appearances in video scenes (Weng et al., 2009).

That work, done for every pairing of two characters, results in a set of relations: two characters significantly co-occurrent induce a relation between them. Otherwise, the link is not created. These links define the states of every dyad and lead to the building of the network. Nonetheless, the result is a flat model of the character structure in the narrative—flat in the sense that dynamics are lost in the extraction process, while the global summing structure emerges from it.

There are works that base their approach and analysis on the semantic context in which the interactions happen (Batagelj et al., 2002; Agarwal, 2011; Hutchinson et al., 2012; Sudhahar et al., 2013; Oelke et al., 2013). For example, Graham A. Sack measures structural balance in character networks—the census of triplets with at least two edges, to see if they also possess the third one—which represent a form of equilibrium, the verification of the idiom "les amis de mes amis sont mes amis" (Sack, 2011). More generally, some works also attempt to model relations through the whole narrative, including actions, and thus to interpret the plot automatically (Kazantseva and Szpakowicz, 2010; Mani, 2013).

5.4.2 Analysis of a single network

The second step is the analysis of one or many corpuses. The framework provides local and global indicators to describe the character-spaces and character-system. It utilises these measures as well as the meaning of the links and nodes for the interpretation of the results.

In a narrative, what defines a character is his or her name, along with other hidden or apparent traits. However, it is not only these, but also whom they meet and how they meet them that is of use when describing them, their importance and their roles alongside the other characters. In the world of the network, we translated the character-space of Woloch to the neighbourhood of the node which represents the character. We measure the node's importance with node centrality, that is any measure associating a value to each node, with computation based on the use of the network's edges, which are the "social", or more precisely "narrative" relations of another character to the one under study. For example, the degree centrality value of a node is the sum of its incident edges, while betweenness centrality takes into consideration the shortest paths across the network and across character-spaces. Put into practice, centrality measures are defined to represent forms of importance among the nodes in the network. In chapter 7, we develop and import methods requiring node centrality to describe, highlight, determine, or differentiate the roles that characters play in our model.

In addition to centrality, the vitality of a node is used to measure how the structure is modified, and how it may collapse, when nodes are removed from the network, that is to say

characters from the plot. We explore this family of indicators based on centrality in section 7.3. We also deal with another method derived from centrality in section 7.4, where we show how a principal component analysis can discriminate the inputs of centrality measures in the computation of importance of characters. Principal component analysis is frequently used for network studies, and it was used, for example, to describe the roles of characters (Hutchinson, 2013) or to study characters in parts of the narration (Cutting et al., 2013).

At the higher level of the character-system and the whole network structure, the indicators are global and unique: an appropriate selection of them form a *signature* of the network. For example, the density computes the ratio of existing edges to the theoretical maximum. It gives insights into how dense the interactions of characters are, thus how spread out the characters are in the narration.

The concept of centrality can be extended dually to the edges of the network. Edge centralities value the importance of a connection on the basis of its situation in the network. It takes into account which edges it is adjacent to, what kind of nodes it is incident to, what importance an edge has in comparison to its direct and local neighbourhoods, as well as the global structure. In the context of our work, exploring this question is of the highest importance, since it shows how strong the relations between characters are as a function of surrounding and/or distant relations. For example, a relation with high edge centrality situated in a neighbourhood of edges with low centralities and empty dyads is of much interest, since it can be a relation that the author wants us to perceive as significant, intentionally or not, via the way the narrative is built. We illustrate edge centrality in the particular case of community detection, where the edges with high betweenness act as separators.

The potential applications and methods that could still be extended to character network analysis is long. For example, the study of average path length and clustering in parallel would allow us to test the small-world hypothesis on a character network: a structural property implying that distance between any two nodes is considered as small, while the nodes neighbourhood are dense, thus providing robust relays (Stiller and Hudson, 2005).

5.4.3 Analysis of several networks

We only outline the previously explored ways of analysing data sets made up of more than one narrative work, for example the study of a given author (Stiller et al., 2003), or the classification of works having common properties (Bearman and Stovel, 2000; Rydberg-Cox, 2011). Studies on large corpus are difficult to sort, since the conditions for creating a relation may be based on narrative units located far away from the dialogue. For example, researchers have compiled character networks of comic characters from Marvel universe. In this case, two characters are linked if they appear in the same publication, but that does not imply that they physically met in the story-world of that episode (Alberich et al., 2002; Gleiser, 2007). The reason to compare some narratives is sometimes motivated by an analysis of the genre, for example

the comparison of the plot structures from a few mythological texts (Mac Carron and Kenna, 2012), or of a universe, based on narration (Mac Carron and Kenna, 2013).

6 The network

"He'll never marry her."
"Why not?"
"I've said all I'm going to say," she said.
"I'm gratified to meet an indexer who respects the privacy of others."
"Never index your own book," she stated.

Cat's cradle
KURT VONNEGUT

In this chapter, we extract a character network of *Les Confessions*. Its links depend on page-proximity of characters. The building of this network is based on co-occurrences of characters on the pages of the edition of *Les Confessions* described in section 4.2. We only use the index of characters (see section 4.3) compiled by scholars for this edition, and not the text. The result is a perspective focusing on the arrangements of narrative events of the novel through the distribution of its characters, as well as the structure and patterns they induce. To some extent, this approach leads to the generation of a somewhat ego-centered *social* network of Jean-Jacques Rousseau over the more than fifty years of his life covered in the book. The expression *social* network and the relevance of using it here are discussed in this chapter: knowing exactly what a link between two characters means is critical to the interpretation of that network. Having this goal in view, we present in detail the method used to build the network, and what its potential uses and interpretations are.

We explain the methodology behind the extraction of the network, with some of the problems encountered (6.1) and the solution proposed (6.2). After this, we present the macro-properties of the network (6.3). The main points of that work, which are attributing roles to characters and adapting or defining centrality measures for that purpose, are outlined in this chapter but more fully developed in the following one (see chapter 7).

6.1 Naive approach

Basically, the index of occurrences detailed in section 4.3 is comprised of a table of two columns, giving each row the name of a character and one page number on which that name occurs.

For each character, we create as many rows as pages on which that character is mentioned at least once. More than one occurrence on a page does not create rows in addition to the one previously created. The distribution of such occurrences per page is given in figure 4.4. It is summarised by the distribution of numbers of characters per page as shown in table 6.1.

Occurrences	0	1	2	3	4	5	6	7	8	9	10
Frequency	101	151	156	150	91	54	45	18	9	7	1

Table 6.1: Number of pages containing a given number of occurrences. For example, in the whole edition, 45 pages contain exactly 6 occurrences. The total number of pages containing text from *Les Confessions* is 774.

As we can see from this table, there are 101 pages containing no names. This implies that there are gaps when studying how narrative events are organised along a temporal scale. However, these pages must be considered as well, as pages containing no names illustrate the intention of Rousseau to have nobody related to the events present at this moment of the novel.

We also need to state that we do not consider the last page of each chapter different to the others. They contain less text and therefore it is less probable that they will have characters appearing on them, but this is also the case on many other pages in the book, when they contain voluminous footnotes[1]. Please note that mentions of many names also implies many footnote entries, since they are used for giving context about most new characters. One solution, independent of the edition, would be to measure the distance between names in terms of words, but this is outside the scope of this work that intends to explore the possibilities offered by character indexes.

The mode and the median are equal to two occurrences per page (it is the case of 156 pages), and then it decreases monotonously, reaching a maximum of ten names once, on page 662 (see figures 6.1 and 6.3).

The distribution of empty pages per chapter across the book is given in table 6.2 and figure 6.2. We observe that there is more sparsity across names in the first chapters (contained in the first volume). A one-sided Mann-Whitney test comparing percentages for the two volumes confirms that observation ($p < 0.05$).

[1] For example, on page 68.

compter pour une de mes voisines de campagne, depuis qu'ils s'étaient fait un établissement à Clichy, où j'allais quelquefois passer un jour ou deux, et où j'aurais été davantage, si M^me^ Dupin et M^me^ de Chenonceaux avaient vécu de meilleure intelligence. Mais la difficulté de se partager dans la même maison, entre deux femmes qui ne sympathisaient pas, me rendait Clichy trop gênant. Attaché à M^me^ de Chenonceaux d'une amitié plus égale et plus familière, j'avais le plaisir de la voir plus à mon aise à Deuil, presque à ma porte, où elle avait loué une petite maison, et même chez moi, où elle me venait voir assez souvent.

J'avais M^me^ de Créqui, qui, s'étant jetée dans la haute dévotion, avait cessé de voir les D'Alembert, les Marmontel, et la plupart des gens de lettres, excepté, je crois, l'abbé Trublet, manière alors de demi-cafard, dont elle était même assez ennuyée. Pour moi, qu'elle avait recherché, je ne perdis ni sa bienveillance ni sa correspondance. Elle m'envoya des poulardes du Mans aux étrennes[1], et sa partie[2] était faite pour venir me voir l'année suivante, quand un voyage de M^me^ de Luxembourg croisa le sien. Je lui dois ici une place à part ; elle en aura toujours une distinguée dans mes souvenirs.

J'avais un homme qu'excepté Roguin, j'aurais dû mettre le premier en compte : mon ancien confrère et ami de Carrio, ci-devant secrétaire titulaire de l'ambassade d'Espagne à Venise, puis en Suède, où il fut, par sa cour, chargé des affaires, et enfin nommé réellement secrétaire d'ambassade à Paris. Il me vint surprendre à Montmorency, lorsque je m'y attendais le moins. Il était décoré d'un ordre d'Espagne dont j'ai oublié le nom, avec une belle croix en pierreries. Il avait été obligé, dans ses preuves[3], d'ajouter une lettre à son nom de Carrio, et portait celui de chevalier de Carrion. Je le trouvai toujours le même, le même excellent cœur, l'esprit de jour en jour plus aimable. J'aurais repris avec lui la même intimité qu'auparavant, si Coindet, s'interposant entre nous à son ordinaire, n'eût profité de mon

[1] Rousseau l'en remercia le 15 janvier 1759, mi-figue, mi-raisin : « Si vous m'aviez fait donner de vos nouvelles sans rien m'envoyer de plus, que vous m'auriez fait riche et reconnaissant ! Au lieu qu'à présent que les poulardes sont mangées, tout ce que je puis faire de mieux, c'est de les oublier : n'en parlons donc plus. Voilà ce qu'on gagne à me faire des présents » (*L* 374).

[2] « Divertissement, projet de divertissement » (*Académie*, 1762).

[3] « Titres qui établissent la noblesse » (*Académie*, 1762).

Figure 6.1: Screenshot of page 662 containing occurrences of 10 different characters, from rousseau.slatkine.com. The characters's names are shown in figure 6.3.

Chapter	Nonempty	Empty	Empty (%)
I	35	14	28.6
II	39	12	23.5
III	43	9	17.3
IV	40	9	18.4
V	52	4	7.1
VI	44	10	18.5
VII	84	5	5.6
VIII	62	4	6.1
IX	89	10	10.1
X	62	4	6.1
XI	52	2	3.7
XII	79	10	11.2

Table 6.2: Number of pages containing occurrences of characters, or containing none, per chapter. The third column gives the percentage of pages containing no occurrences out of the number of pages per chapter. In general, the first volume (chapters I to VI) contains more "empty" pages than the second volume.

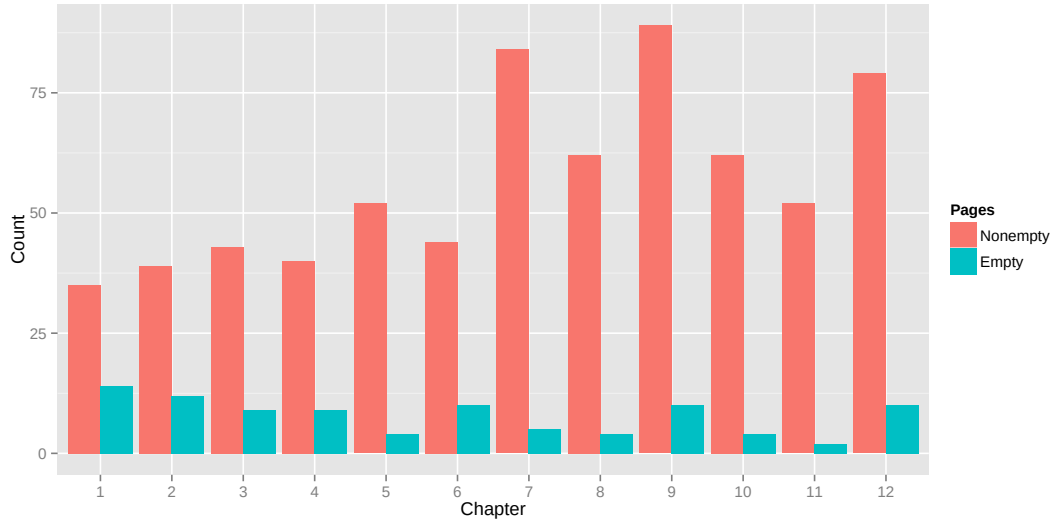


Figure 6.2: Nonempty and empty pages.

6.1.1 First attempt

A first and naive approach to building and highlighting relations between characters, based on proximity, is to use link creation as a basis to isolate the co-occurrences situated on a same page. In the previous example taken from page 662, all ten characters would be linked together, which induces a sub-network of the global network of the book, composed of ten characters and forty-nine edges, all undirected since there is no notion of order taken into account[2].

[2]For example, considering which name appears first can lead to using directed edges.

Doing this exercise with each page of the book leads to many sub-networks of various sizes. They are all complete graphs. For example, figure 6.3 shows the sub-networks created using pages 661 and 662 respectively.

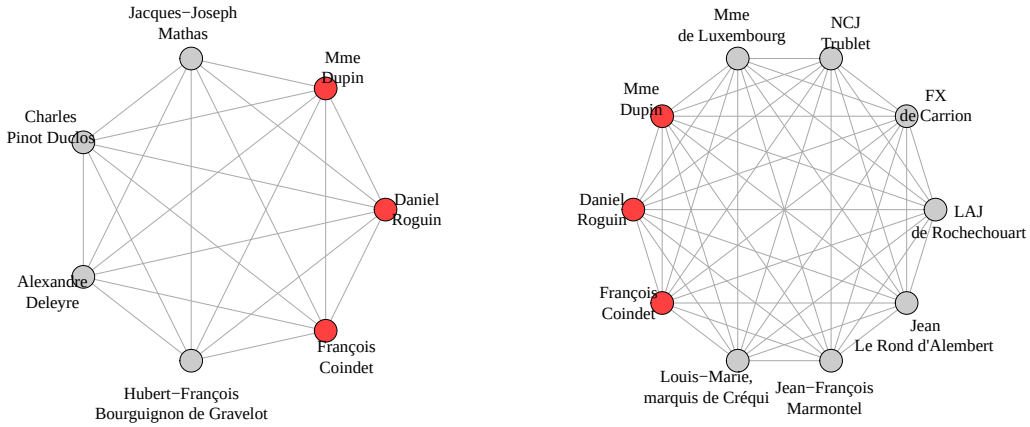


Figure 6.3: The complete networks induced by co-occurrences on pages 661 (left) and 662 (right). Mme Dupin, François Coindet and Daniel Roguin appear in both networks.

These sub-networks share some character names. Based on these co-appearing characters, we build a global network of the book by merging all occurrences of each character over the sub-networks, while incorporating the edges and summing them, as a way of measuring the intensity of the co-occurrence. We illustrate this step with the example given in figure 6.4.

As we verify in figure 6.3, the common occurrences are Mme Dupin, François Coindet and Daniel Roguin. The connections between them exist in both sub-networks, which means that their intensity will be double, while the intensity of other connections remains single. This way, the network is simplified.

This principle is applied to all the pages at the same time. All the characters appearing in the index are considered. We create a link between two characters if they both appear on the same page. We count the number of times it happens, and record that in an edge attribute that we call intensity. The final result is given in figure 6.5, a network with 583 nodes and 2415 edges that we call $\mathcal{G}_0$.

In this visualisation, we observe that there is a giant component[3] containing 547 characters, that is to say 93.8% of them. The 36 other characters belong to 22 much smaller components, 14 of them being isolates[4]. They appear in the periphery of the visualisation due to the choice of the layout algorithm. For example, the isolated Mme Clot appears alone on page 77, and there is nobody on either pages 76 nor 78. The reason, in this case, is that Rousseau

[3]The majority of nodes belong to a single and large connected part of the network.
[4]An isolate is a connected component comprised of a single element (character/vertex).

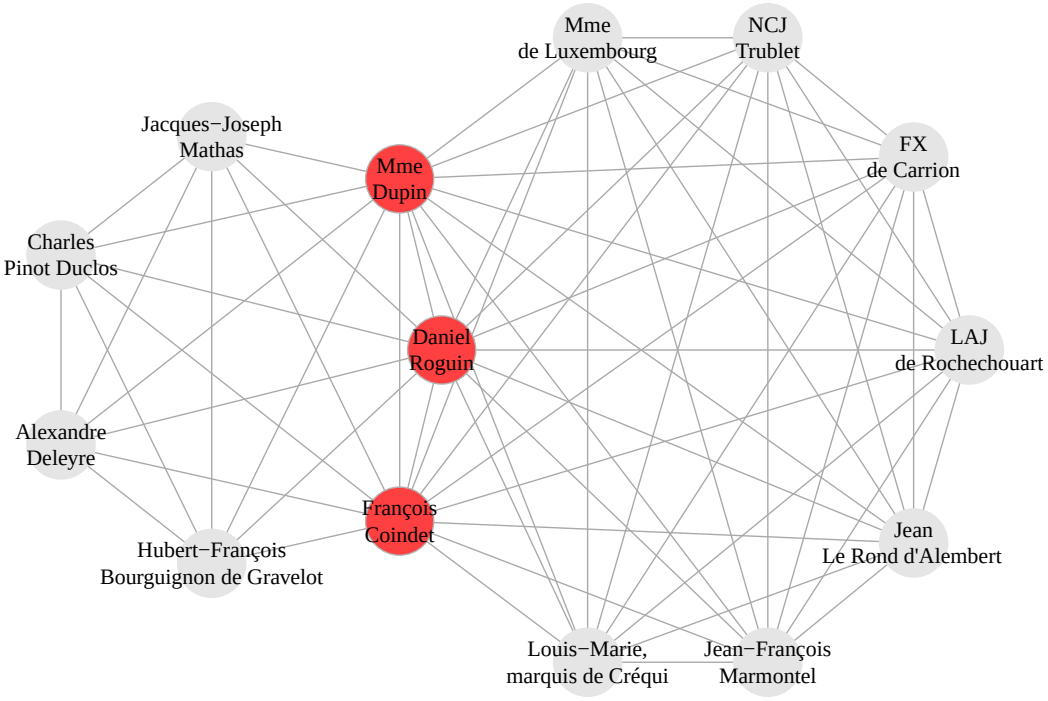


Figure 6.4: The network resulting from a merge of the sub-networks of pages 661 and 662. The three characters appearing in both sub-networks are drawn darker.

explains one of the silly things he did to this neighbour when he was a child. It is a different narrative event. But being an isolate doesn't mean that one does not appear in any narrative events shared with other characters. For example, on page 267, Don Antoine Petitti appears in a quoted sentence of Mme de Warens, while being present in the same place as her and Rousseau:

> J'arrive enfin, je la revois. Elle n'était pas seule. M. l'Intendant général était chez elle au moment que j'entrai. Sans me parler, elle me prend par la main, et me présente à lui avec cette grâce qui lui ouvrait tous les cœurs : « Le voilà, monsieur, ce pauvre jeune homme ; daignez le protéger aussi longtemps qu'il le méritera, je ne suis plus en peine de lui pour le reste de sa vie. » Puis, m'adressant la parole : « Mon enfant, me dit-elle, vous appartenez au roi ; remerciez M. l'Intendant qui vous donne du pain.[5] » [...]

[5] "At length I arrived at Madam de Warrens; she was not alone, the intendant-general was with her. Without speaking a word to me, she caught my hand, and presenting me to him with that natural grace which charmed all hearts, said: "This, sir, is the poor young man I mentioned; deign to protect him as long as he deserves it, and I shall feel no concern for the remainder of his life." Then added, addressing herself to me, "Child, you now belong to the king, thank Monsieur the Intendant, who furnishes you with the means of existence." [...]"
Translation by S. W. Orson, Privately Printed for the Members of the Aldus Society, London, 1903, see http://en.wikisource.org/wiki/Confessions_(Rousseau), accessed on 15/07/2014.

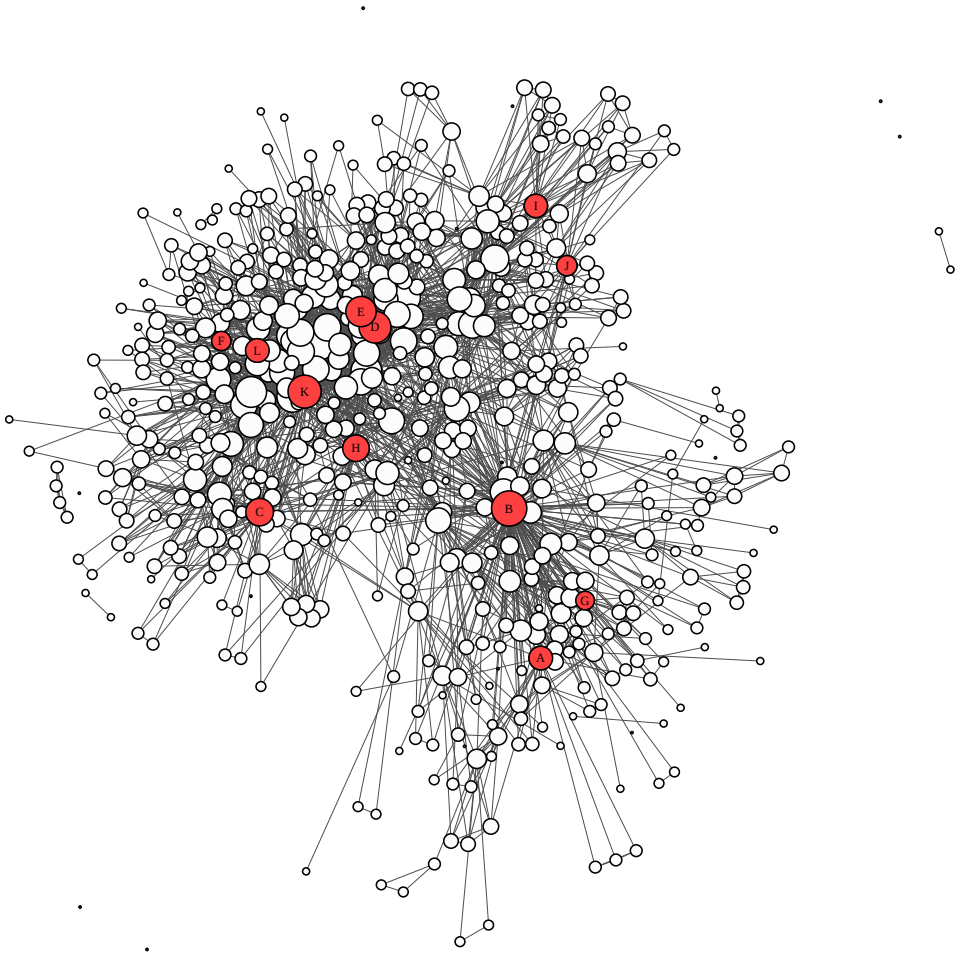


Figure 6.5: Network of co-occurrences per page, $\mathcal{G}_0$. The size of the nodes depends on their degrees. The underscored nodes show some characters playing a significant intermediary role. In order of appearance : (A) Victor-Amédée II, Roi de Sardaigne, (B) Mme de Warens, (C) George Keith, (D), Denis Diderot, (E) Louise-Marie-Madeleine Dupin, (F) François-Charles de Vintimille comte Du Luc, (G) Daniel-François comte de Gelas de Voisins d'Ambres, (H) Voltaire, (I) comte de Montaigu, (J) Abbé de Binis, (K) Jean-Jacques Amelot de Chaillou, (L) Thérèse Levasseur and (M) Jean Néaulme.

Mme de Warens is cited on the next page in the index. They clearly appear together in the same place, speak with each other, and the place is unpopulated. But they are not related. This is a consequence of the work done, not creating an entry in the index if a character appears only via a pronoun. This is also partly due to how the beginning and end of pages of this edition are arranged. Therefore we needed to think about how we wanted to construct the network.

A weight attribute is given to the links, counting the number of co-occurrences of two names on the same page across all the pages of the book. It gives a way to measure the intensity of the relation between any two related characters. At the same time, it highlights one of the main problems of this method, which is the rigidity of the page format, which we needed to overcome.

6.1.2 The page

This method is too likely to cut scenes that are spread across more than one physical page. Think of a page containing a single name but a majority of footnotes, for example. This name will probably be isolated if that person isn't mentioned anywhere else in the book. As we have previously said, there are fourteen such cases[6], and a few others where some names appear on the same page but not anywhere else.

As we are currently describing the smaller connected components, we still need to focus on the two largest components coming after the giant one. In the first one, Rousseau lists thinkers that have influenced him but he did not meet in person; they are Nicolas de Malebranche, John Locke, René Descartes and Gottfried Wilhelm Leibniz, on page 341, chapter 6. In the second case, he lists some of the people to whom he reads *Les Confessions*, an event that happens predictably on the penultimate page of the book. The people named are Mme la comtesse d'Egmont, M. le prince Pignatelli, Mme la marquise de Mesme and M. le marquis de Juigné. In both cases, no links are created with the outside. Those events have their importance in the book, and can be easily categorised, as we have done. But they remain anecdotal when compared to those appearing in the various story arcs that are central to Rousseau's project, such as the ones concerning his youth, transparency, the plot against him, etc. In fact, they mention characters that didn't play a direct role in his life during the interval of time covered by *Les Confessions*, therefore their side-lining.

The choice of delimiting link creation by the format of the page prevents us from using the weight attribute of measuring intensity in order to define a network representing only the significant ties, by calling for intensity to be larger than a given threshold. The observed proximity, that is already sometimes not taken into account when two characters are respectively mentioned, one at the bottom of a page, and the other one at the top of the following one, becomes even more twisted if we require the weight attribute to reach a certain threshold in order to capture that said intensity. We miss some information, and randomly give too much importance to some other. In addition, even if that question were to be resolved, how should we set the threshold? A minimal value allows the maximality of information, but a higher value would only highlight the main characters and the main narrative events or contexts involving relations between them.

[6] For example, on page 103, "M. Verrat" is the only character on the page, and he is not referenced anywhere else in the book.

As a result of this discussion, the network we have just built has two main flaws that we need to correct :

- The system of co-occurrence per page is too strict, and deals with page proximity poorly. It creates disjointed sets of relations that do not take into account the spread of narrative events on many consecutive pages. We need a method overlapping pages, and they should be considered in large sets of consecutive pages, like couples or even more.
- The meaning of a link is too vague. For now, it is simply the meeting of two names, once on one page. We need a stronger condition, and we need to be able to interpret the link, that is to say to characterise the narrative relation of two characters in the book by looking at their dyad, and if there is one, at its intensity. Blindly, a single co-occurrence of two names on more than 700 pages doesn't prove that those two characters are related.

6.2 Valid approach

In this section, we describe a method that corrects the two flaws previously identified, namely the lack of entanglement between occurrences situated on different but close pages, and the need for defining a threshold for significance. In the first case, we define a system that takes into account co-occurrences on consecutive pages in order to create links. The domain is not the set of pages anymore, it is now a set of overlapping couples of pages. A simple computation allows us to weight the links, based on the previous co-occurrence method itself (see section 6.1.1). To deal with the second flaw, we define a method that requires that two characters appear together at least twice, closely or not, in order to register the link in the network. That condition of a minimum intensity is translated into a value that serves as a threshold that determines if the link must be created or not.

6.2.1 Overlapping

Firstly, we define the set of occurrences of any given character on couple of pages instead of single pages (see section 6.1.1). We formalise it here: let A be the name of a character. If A appears on page i, we consider that A occurs on both couples of pages $\{i-1,i\}$ and $\{i,i+1\}$. A belongs to the narrative events ending and starting at page i. Let

$$o : \{\text{Character names}\} \longmapsto \{\text{Couples of pages}\}$$

be the function registering occurrences. In this case, if A appears only on page i, then $o(A) = \Big\{\{i-1,i\},\{i,i+1\}\Big\}$. The number of occurrences of the same character on a single page is not taken into account. This is forced by the structure of an index of characters. The apparition of a character on two consecutive pages lead us to count them twice, that is to say: if B appears on pages i and $i+1$, then

$$o(B) = \Big\{\{i-1,i\},\{i,i+1\},\{i+1,i+2\}\Big\}.$$

The character is already represented in all couples covering this neighbourhood of pages, and the goal is to relate who that character appears with instead of its frequency distribution.

Then, we define the method dealing with co-occurrence. The idea is to take advantage of the intersection of respective occurrence sets. Let A and B be characters in the index.

Definition 29. *A and B are* weakly co-occurrent *if* $|o(A) \cap o(B)| > 0$.

Example 1. *Pietro Basile and the comtesse de Vercellis are* weakly co-occurrent. *The first one appears on pages 157 and 158, and the second one on pages 159 to 164, 169, 171 and 179. Therefore, we obtain that*

$$\Big|o(\text{Pietro Basile}) \cap o(\text{comtesse de Vercellis})\Big| = \Big|\Big\{\{158,159\}\Big\}\Big| = 1 > 0.$$

Definition 30. *More generally, for $k > 0$, A and B are* k-co-occurrent *if* $|o(A) \cap o(B)| \geq k$.

Example 2. *Dominique Vitali and Jean-Charles Patizel are* 2-co-occurrent. *The first one appears on pages 418 and 426, and the second one on pages 423, 425, 426, 427 and 434. Therefore, we obtain that*

$$\Big|o(\text{Dominique Vitali}) \cap o(\text{Jean-Charles Patizel})\Big| = \Big|\Big\{\{425,426\},\{426,427\}\Big\}\Big| \geq 2.$$

Note that A and B are *weakly co-occurrent* if and only if they are *1-co-occurrent*.

Theorem 1. *Two characters on consecutive pages are* weakly co-occurrent.

Proof. $\Big|o(A) \cap o(B)\Big| = \Big|\Big\{\{i-1,i\},\{i,i+1\}\Big\} \cap \Big\{\{i,i+1\},\{i+1,i+2\}\Big\}\Big| = \Big|\Big\{\{i,i+1\}\Big\}\Big| = 1 > 0.$ □

Theorem 2. *For $k > 0$, if characters A and B are k+1-co-occurrent, then they are* k-co-occurrent.

Proof. $\Big|o(A) \cap o(B)\Big| \geq k+1 \geq k.$ □

Theorem 3. *Two characters on the same page are* 2-co-occurrent, *since they appear together in two couples of pages.*

Proof. If A and B both appear on page i, then

$$o(A) \cap o(B) \supset \Big\{\{i-1,i\},\{i,i+1\}\Big\},$$

which implies that

$$\Big|o(A) \cap o(B)\Big| \geq \Big|\Big\{\{i-1,i\},\{i,i+1\}\Big\}\Big| = 2$$

□

Definition 31. *We call the quantity* $|o(A) \cap o(B)|$ *the* intensity *of the couple formed by characters A and B.*

The properties of intensity and *k-co-occurrence* of a couple are related in the following way.

Theorem 4. *Let A and B be characters. The intensity of A and B is equal to k if and only if A and B are* k-co-occurrent *but not* (k+1)-co-occurrent.

Proof. $\left|o(A) \cap o(B)\right| = k \geq k \not\geq k+1$ □

Remark 1. *The co-occurrence of two characters on the same page weighs double as consecutive pages. This is coherent with the fact that the word distance of two positions on the same page is on average half the word distance of two positions on consecutive pages.*

Remark 2. *The definition of* k-co-occurrence *can be generalised to hypergraphs by defining k-co-occurence on the intersection of more than two occurrence sets, but this is out of the scope of this work.*

6.2.2 Significance threshold

The first flaw has found an answer: overlapping co-occurrences allows us to consider proximity as smooth and distributed. However, we have added a lot of information to the resulting network, grabbing connections from one paragraph to another, sometimes hundreds of words later. In this section, we deal with that. We choose and discuss thresholds that we can interpret as significant. Then, we look for the one that is the minimum with such properties, and we use it to build the network of interactions of characters in narratives.

For the sake of comparison, simply applying the previous method to the index increases the size of the network from 2415 to 4919 edges (see figure 6.6, left). The order stays the same, with 583 nodes.

We have regained much of the information that was lost but have generated "noise" at the same time. For example, on the first pages of chapter IV, character Venture de Villeneuve, an adventurer, appears on page 221, and Jean-François de Dortan, a cantor, appears on page 222. They are mentioned in other places of the book, both before and after[7], but this is the only time they appear close one with the other. From reading the text, we can assert that they do not meet, or play a role in a common scene. Therefore, such a weak relation is not necessary for the final network that we are going to study, like many similar ones scattered throughout the book. They must be separated. Thus, from now on, we do not consider *1-co-occurrence* as being a sufficient condition.

[7] Venture de Villeneuve on pages 212, 222, 229, 230, 238, 239, 240, 290, 305, 306, 308 and 535, Jean-François de Dortan on pages 217, 221 and 282.

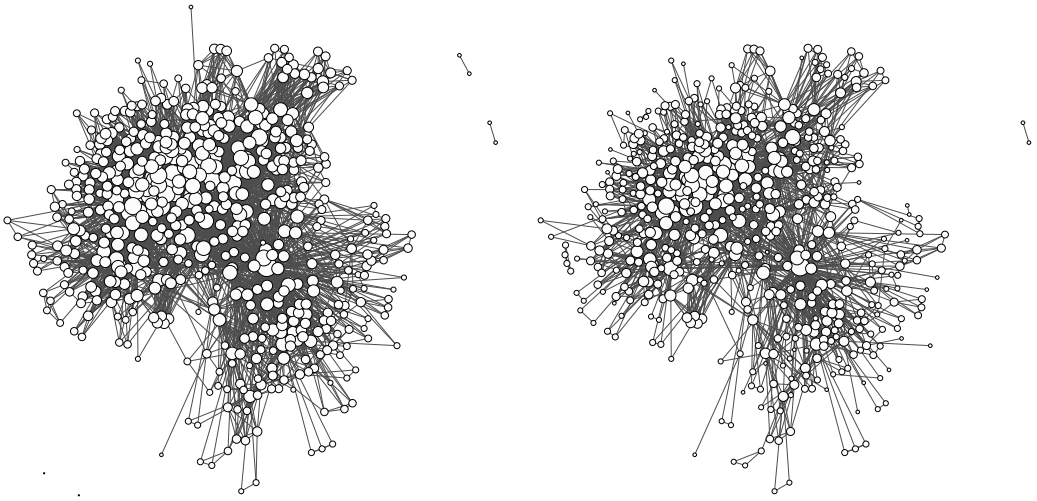


Figure 6.6: (Left.) The network based on *1-co-occurrence*, $\mathcal{G}_1$. It is composed of 583 nodes and 4919 edges. It is disconnected. 99% of the nodes belong to the giant component (577 out of 583). (Right.) The network based on *2-co-occurrence*, $\mathcal{G}_2$. It is composed of 574 nodes and 2675 edges. It is disconnected. 96.9% of the nodes belong to the giant component (556 out of 574). In both cases, the size of a node depends on its degree.

The network obtained from a restriction to *2-co-occurrences* has 575 nodes and 2686 edges (see figure 6.6, right). Like the network built from *1-co-occurrence*, it is disconnected. The next step is to observe the limit-cases where intensity is equal to 2. That happens if two names occur on the same page, and only on this page, or if they appear two times on two couples of consecutive pages. For example, on page 487, Zulietta and Suzanne Rousseau are mentioned together. The first one is a prostitute from Venice that is central at this moment of the story, but the second is his aunt, mentioned in a parallel narrative event: a flashback about his urinary disorders dating from youth. This is not a significant link, even if it means that they are related in some way. As an example of the second case, Thérèse Levasseur appears on pages 460 and 462, and Rebel on page 461. This is the only time they appear closely together, therefore intensity between them is equal to 2. The context shows no significant relation between them: Rousseau talks about her in the context of his everyday life, and mentions him because he is directing an orchestra playing one of his musical works. As we can see, couples of characters verifying *2-co-occurrence*, but not *3-co-occurrence*, can be subject to some sort of randomness, with names unrelated or very weakly related to a common narrative event. Therefore, we decided to exclude edges with intensity equal to 2. The network built from *3-co-occurrences* has 226 nodes (see figure 6.7, left). We call it $\mathcal{G}_3$. It is composed of 38.8% of all characters, and 613 edges remain. It is simple, undirected, weighted and disconnected.

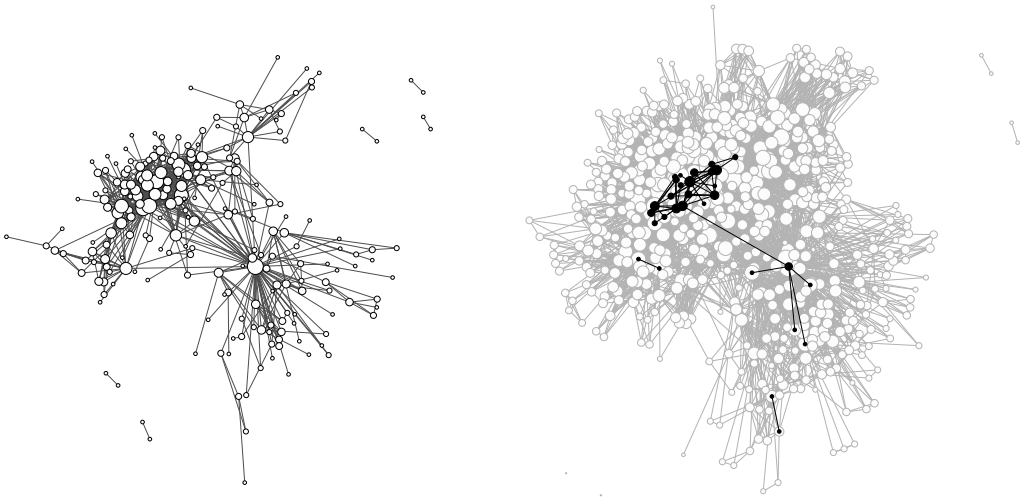


Figure 6.7: [Left] The network[8]based on *3-co-occurrence*, $\mathscr{G}_3$. It is composed of 226 nodes and 613 edges. It is disconnected. 95.6% of the nodes belong to the giant component (216 out of 226). The size of a node depends on its degree. [Right] The network based on *10-co-occurrence*, $\mathscr{G}_{10}$. It is composed of 30 nodes and 53 edges. It is disconnected. 86.7% of the nodes belong to the giant component (26 out of 30). The size of a node depends on its degree.

As the intensity threshold grows, the order and sizes diminish fast, while the sizes of the consecutive giant components remain high (see table 6.3). The sequenced networks don't dislocate (see figure 6.8). The connected components disjointed from the giant component represent narrative events getting separated from the more intense or related ones.

Min. intensity	Order	Size	Giant comp. (%)
1	583	4919	0.99
2	574	2675	0.97
3	226	613	0.96
4	180	394	0.91
5	101	215	0.85
6	77	147	0.81
7	60	116	0.87
8	47	83	0.74
9	38	68	0.84
10	30	53	0.87

Table 6.3: Order and size of sub-networks induced by a minimum edge intensity.

[8]To increase readability, the layout of network $\mathscr{G}_3$ as well as all forthcoming subnetworks of $\mathscr{G}_1$ are coherent with layout of $\mathscr{G}_1$. For example, all nodes in $\mathscr{G}_3$ have the same position in $\mathscr{G}_1$.

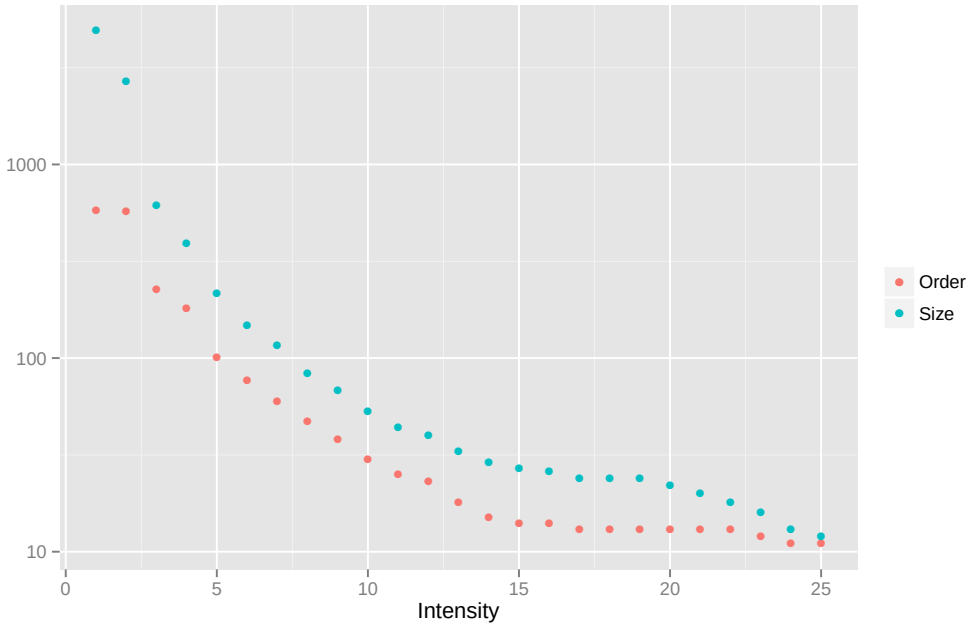


Figure 6.8: Order and size of sub-networks induced by a minimum edge intensity.

This method brings out the successive and underlying structures of the book, from the most global one, encompassing all *weakly co-occurrent* relations, to more sensible ones, highlighting the narrative core (see figure 6.7, right).

6.2.3 *3-co-occurrences*

At this point of our work, we make the hypothesis that, for the relation between characters A and B to be meaningful, they need to be *3-co-occurrent*. Cases with more demanding conditions ($k > 3$) are relevant, and some need a description, but the condition we choose here is a good compromise between too much information and too little, and also a depiction of significant relations between characters, following the distribution of narrative events. The network on which we focus most of our attention from now on is illustrated in figure 6.7, left.

To illustrate what this choice implies, we enumerate the possible cases where a couple of characters are *3-co-occurrent* and not *4-co-occurrent*. Then, we show a few examples of these cases, finally describing the network on which we work for the rest of this chapter. The interpretation we give of the results, the implication it has in the following, and the validity of the network analysis methods we use are discussed in section (6.3).

Determining a threshold and building our whole analysis on its induced network makes a large project depending strongly on a simple assumption. Here, this is not a question of what was Jean-Jacques Rousseau's real social network or, more precisely, the social network derived from *Les Confessions*. It is a choice, that we justify by enumerating the possible cases that are

considered the minimum condition for a link to exist, each time with an example. We want to study appearances and co-appearances of characters through the narrative events of the book. We want to study relational narratives. Of course, there is the possibility of lowering the threshold, in order to capture every situation of two nearby characters (no need to illustrate this case, as it would be illegible). Or to heighten the threshold, so as to highlight the presence (or not) of the most frequent characters, and the strongest relations between those that are present. It is an approach that determines the underlying structure of the relations in the book.

There are three cases in which a relation between two characters is *3-co-occurrent* but not *4-co-occurrent*: when two characters are once co-occurrent on the same page, and once on a disjoint couple of consecutive pages, or when they are both co-occurrent on two consecutive pages, and finally when two characters are three times co-occurrent, but never on the same page. We develop this with the help of a few examples.

Case 1 : *all successive co-occurrences.* If two characters are strongly intricated, appearing together on two consecutive pages for example, their intensity will be equal to 3 if that is their only co-occurrence. This is the most compact case, meaning that at least one of the two characters' names is positioned in between the other name. The probability that they belong to the same scene or narrative event, or have physically met, is high.

For example, Mme de Warens and the abbé Blanchard respectively appear, among others, on pages 304 and 306, and on pages 305 and 306. Which means that

$$\Big|o\Big(\text{Warens}\Big) \cap o\Big(\text{Blanchard}\Big)\Big| = \Big|\Big\{\{304,305\},\{305,306\},\{306,307\}\Big\}\Big| = 3$$

In the text, Blanchard is someone recommended to Rousseau for music lessons. He has to ask the permission to leave the house to Mme de Warens, and she helps him financially on this occasion:

> [...] Je me mis en tête d'aller à Besançon prendre leçon de l'abbé Blanchard, et cette idée me parut si raisonnable, que je parvins à la faire trouver telle à Maman. La voilà travaillant à mon petit équipage, et cela avec la profusion qu'elle mettait à toute chose. Ainsi, toujours avec le projet de prévenir une banqueroute et de réparer dans l'avenir l'ouvrage de sa dissipation, je commençai dans le moment même par lui causer une dépense de huit cents francs : j'accélérais sa ruine pour me mettre en état d'y remédier. [...][9] [Page 306]

[9] "I therefore determined to go to Besancon, and take some lessons from the Abbe Blanchard, and the idea appeared so rational to me, that I soon made Madam de Warrens of the same opinion, who immediately set about the preparations for my journey, in the same style of profusion with which all her plans were executed. Thus this project for preventing a bankruptcy, and repairing in future the waste of dissipation, began by causing her to expend eight hundred livres; her ruin being accelerated that I might be put in a condition to prevent it."

Case 2 : *mixed co-occurrences.* An interesting consequence of our method is that two characters co-occurrent once on the same page, and once earlier or after in the book, will have intensity equal to 3 and enter this category. In summary, there is a strong connection and a weak one, and they are spread on close or distant pages. We interpret that last fact as a high probability that they play a significant role in the same narrative event.

For example, Jacques-Armand Dupin de Chenonceaux and Louise-Marie-Madeleine Dupin (also known as Mme Dupin) respectively appear, among others, on pages 404 and 491, and on pages 404 and 490. Which means that

$$\Big|o\Big(\text{Chenonceaux}\Big) \cap o\Big(\text{Dupin}\Big)\Big| = \Big|\Big\{\{403,404\},\{404,405\},\{490,491\}\Big\}\Big| = 3$$

Jacques-Armand Dupin de Chenonceaux is the son of Mme Dupin. The strong co-occurrence on page 404 happens when Rousseau is hired as his tutor.

> Je passai ces huit jours dans un supplice que le plaisir d'obéir à Mme Dupin pouvait seul me rendre souffrable ; car le pauvre Chenonceaux avait dès lors cette mauvaise tête qui a failli déshonorer sa famille, et qui l'a fait mourir à l'île de Bourbon. Pendant que je fus auprès de lui, je l'empêchai de faire du mal à lui-même ou à d'autres et voilà tout : encore ne fut-ce pas une médiocre peine ; je ne m'en serais pas chargé huit autres jours de plus, quand Mme Dupin se serait donnée à moi pour récompense.[10] [Page 404]

The weaker co-occurrence on pages 490 and 491 happens in a context of large footnotes covering respectively about 40% and 75% of these pages, which means that this weak co-occurrence would be a stronger one in another edition. In this example, Rousseau is leaving his job of secretary of M. de Francueil, a close friend of Mme Dupin, who then hires an old tutor of Mme Dupin's son as a new secretary.

Case 3 : *all disjoint co-occurrences.* This is the least frequent case, where two characters share three weak co-occurrences. They meet three times, on consecutive pages, but never on the same. The probability is high that these pages are close, for example situated in the same chapter. We interpret these co-occurrences as weak ones, but the fact that they are recurrent tends to induce that they are related in some way.

Translation by S. W. Orson, Privately Printed for the Members of the Aldus Society, London, 1903, see http://en.wikisource.org/wiki/Confessions_(Rousseau), accessed on 15/07/2014.

[10] "I passed eight days in such torments as nothing but the pleasure of obeying Madam Dupin could render supportable. *Poor Chenonceaux was so sulky he could have brought shame upon his family, something that later provoked his death on the Isle of Bourbon. While I was by his side, I prevented him from hurting himself or others, and that is all: [...].* I would not have undertaken to pass eight other days like them had Madam Dupin given me herself for the recompense."

Translation by S. W. Orson, Privately Printed for the Members of the Aldus Society, London, 1903, see http://en.wikisource.org/wiki/Confessions_(Rousseau), accessed on 15/07/2014. In italic: own translation.

For example, Jean Le Rond d'Alembert and Louise d'Épinay respectively appear, among others, on pages 536, 645 and 785, and on pages 535, 646, 784. Which means that

$$\left|o\Big(\text{d'Alembert}\Big) \cap o\Big(\text{Épinay}\Big)\right| = \left|\Big\{\{535,536\},\{645,646\},\{784,785\}\Big\}\right| = 3$$

Jean Le Rond d'Alembert and Louise d'Épinay are exactly three times *weakly co-occurrent*, and never *strongly co-occurrent*. It is important to notice that these happen on three couples of pages, each time separated by more than hundreds of pages. These gaps are equivalent to a few years each time, during which these characters are not explicitly cited together, but are mentioned in similar situations.

In the first case, Louise d'Épinay is mentioned because she offers Rousseau a retreat to a house she possesses in the countryside. But before leaving, Rousseau is told, as well as d'Alembert, that an author is being excluded from his Academy. This is definitely a very weak link. The second one concerns the article *Genève* that d'Alembert is writing while having in mind his intention to use it in order to highlight the lack of theatres and cultural life in Geneva. Rousseau gets annoyed by this fact and responses to that article of the Encyclopaedia with the *Lettre à d'Alembert sur les Spectacles*, in which he depicts some people of his time, including Louise d'Épinay. A stronger link, but with characters on various levels: one of the two is active, the other is only quoted. In the third case, Rousseau realises that some of the letters he had conserved had disappeared, and among them some from Louise d'Épinay. When trying to find the culprit, he ends up suspecting d'Alembert.

There is a bond we can infer between these two characters: at least two of the three—distant—weak co-occurrences imply the same shared narrative events. Such cases of relations with intensity equal to three may be difficult to interpret. However, as shown in table 6.4, these cases are rare. One could be a false positive error as the result of the many occurrences of two characters throughout the book, which increase the probability that they appear somewhere together. This is not the case here[11], since Louise d'Épinay appears on 69 pages out of 673 (10.3%), but d'Alembert appears only on 13 pages out of 673 (1.9%).

3 successive	2 successive	0 successive
81	129	7

Table 6.4: Distribution of cases with edge intensity equal to three.

In table 6.4, we present the repartition of relations with intensity equal to three. There are three possible cases. The configuration of two names appearing three times on pages side by side is the rarest. The two other cases are similar in their significance, but slightly not in their

[11] The six other cases are "Voltaire — Grimm", "Abbé de Saint-Pierre — Louise d'Épinay", "Abraham de Pury — Thérèse Levasseur", Saint-Lambert — baron d'Holbach", "Charles Pinot Duclos — Mariel Fel".

meaning: two characters appearing together on three consecutive pages are strongly related through a common narrative event. That is not necessarily the case with the last possibility, in which, most of the time, the co-occurrence that is not on successive pages is close to them (the example above bringing together Jacques-Armand Dupin de Chenonceaux and Mme Dupin acts as a counter-example).

The final version of the network, previously shown in figure 6.7, is given in figure 6.9. We have subtracted the five disconnected couples of nodes from $\mathscr{G}_3$. The number of nodes is then 216 instead instead of 226. We call that network $\mathscr{G}$.

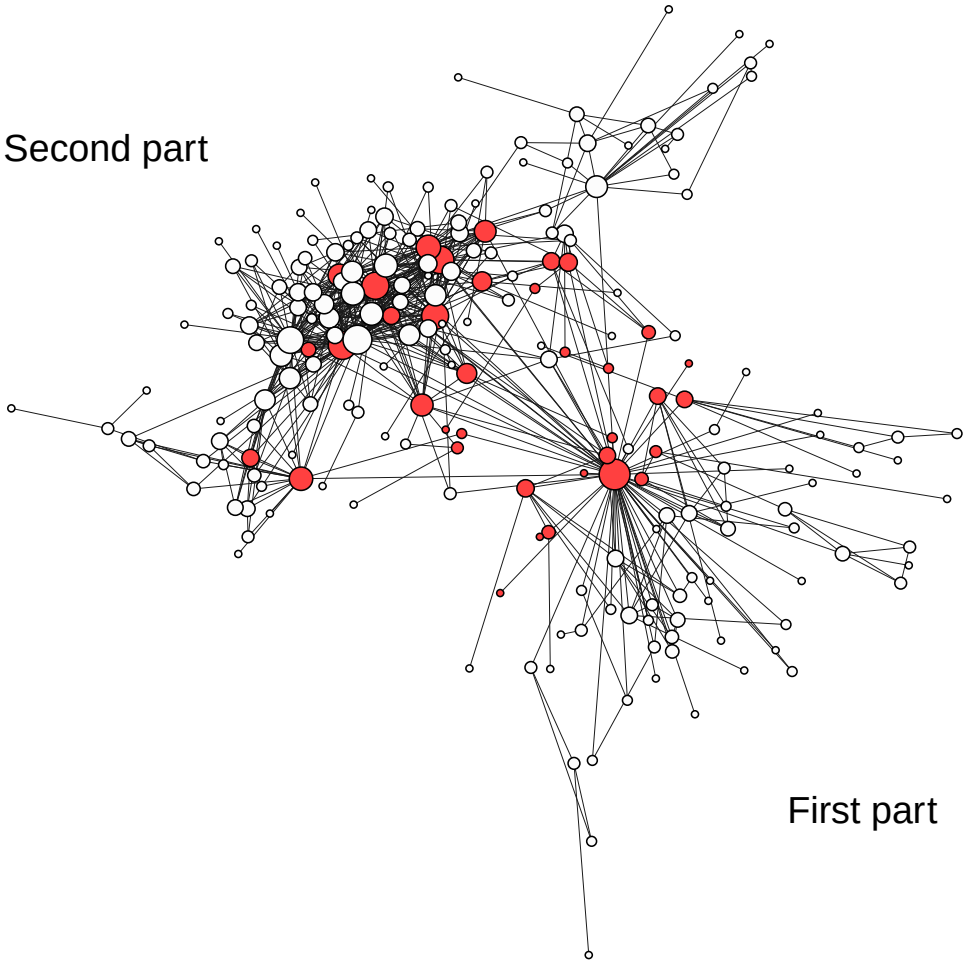


Figure 6.9: Definitive network $\mathscr{G}$ of co-occurrences per page. Colored nodes are characters occurring in both parts of *Les Confessions*.

This network will be the one most under study in the rest of this thesis. We consider it as the best compromise between all of the possible co-occurrences systems that can be taken

into account, and the need to identify the most important thread among the numerous and complex narrative events of *Les Confessions*. For the sake of clarity, we have to mention that when saying "network", we consider the mathematical object made of nodes, edges and a few attributes, and not the visualisation proposed here. In such a case of a non-regular network[12], we use, in order to determine the positions of nodes, a layout that stochastically minimises some conditions and, after a given number of iterations, produces a suitable layout. In this case, the one used is given by (Fruchterman and Reingold, 1991). Some description can be made out of this visual output, but the analysis has to be conducted with methods issued from network analysis. Visualisation just does not suffice.

At this step of the analysis, we make the decision not to normalise the values of intensity[13], but to keep them as they are, beginning from three. Our choice is to keep coherent attributes throughout this work, allowing us to compare intensities between this network and the network including every link (*1-co-occurrence*, meaning that intensities start at 1), or networks with stronger constraints (*k-co-occurrence*, with $k > 3$). We deal with the computing side of this choice in the chapter 7, since the algorithms behind measures computed over weights on the edges of the network necessitate having transformed values as inputs, usually starting at one.

6.3 Interpretation

An important step in any network study is the act of describing and interpreting what the existence of a link between two entities means and implies. The nodes represent characters, thus people from the real life. *Les Confessions* is an autobiographical work; these characters are people who were in contact with Rousseau at one moment in time, something he may explicitly say, or people who were related to him through a shared acquaintance or event, in which case he may not have physically met them. In any case, Rousseau gave their existence a particular attention, since he mentioned a wide spectra of these contemporary individuals by name[14]. Indifferently, they range from aristocrats to servants and other common people[15]. Being cited by Rousseau is the common property of all the characters in the index. It defines the characters of *Les Confessions*' narrative. The construction of relations—the links in the network—is based on Rousseau's narration as much as our interpretation of the narration—whom they bring together, and how they interact.

[12] In this context, we call *regular networks* networks that are subject to automorphisms. Circle or complete networks are regular networks.

[13] Two common ways to normalise them would be to substract 2 to all values, in order to bring the minimum from 3 to 1, or to normalise them in between 0 and 1.

[14] Otherwise, the character is not referred to in the index.

[15] Rousseau presented an opera to the king, who was then in favour of attributing him an annuity. Nonetheless, Rousseau mentions him by "le Roi", "Sa Majesté", or "un si grand monarque" (Rousseau et al., 2012, pp. 511-512), but never by his name. Probably because of that, the indexer did not index the king, thus he does not appear in the character network.

In this section, we show how to read this network. For this purpose, we use the framework for character network analysis defined in section 5.4.

6.3.1 Eccentricity, diameter and average path length

In a graph, the eccentricity of a node is defined as the largest distance from that node to all others. The minimum value of the eccentricity is the radius (the distance from the "centre" to a peripheral point), and the maximum is the diameter. It is a rough measure. This is similar to the geometry of the circle or of the sphere, with the difference being that the diameter of a network is the length of the longest shortest paths of the network, while the diameter of a circle is the segment binding two corresponding points on the periphery. Thus, the diameter of the graph is the radius of the circle. In our case, Mme de Warens is the only centre of the graph, with eccentricity equal to four; five nodes are peripheral, with eccentricity equal to eight (see table 6.5).

Eccentricity	4	5	6	7	8
Nodes	1	64	125	21	5

Table 6.5: Eccentricity. *Mme de Warens* is the unique node at the centre of the network. The periphery is filled by five nodes (see figure 6.10).

This shows a not too compact nor centralised network. The diameter is equal to 8, it is the maximal value of the eccentricity. The sixteen possible[16] diameters are given in figure 6.10. The diameter allows us to highlight the longest geodesics, which are the longest possible paths that link any two characters. The diameter is seen as an upper boundary to the cost of transfer within the network. It also strongly depends on the structure of the network: if it is disposed as lightly connected clusters, then the diameter will be high, signifying that the information would be distributed with difficulty, like in a medieval society with no mail service. A low diameter with respect to the order and size of the network signifies that all other geodesics are even shorter than that value, and that, globally, information spreads easily[17]. In the case of network $\mathscr{G}$, the diameter is small, which means that even for two characters with different profiles and appearing at different epochs, there are overlapping characters and narrative events that link them. There are many possible paths with length equal to the diameter. Most of the minor characters are linked to some of the primary characters that overlay the book. These play the roles of *hubs* in terms of network theory. That is to say: they have a high number of connections, which situate them on many of the shortest paths, therefore at strategic places like the centres of clusters or places linking them. These hubs bring all the characters closer together, with the exception of the longest geodesics, who find paths through only two such hubs, otherwise through cascading secondary characters.

[16] Same characters are used multiple times.

[17] This still depends on what the links represent, and if there is a weight attribute that must be taken into account.

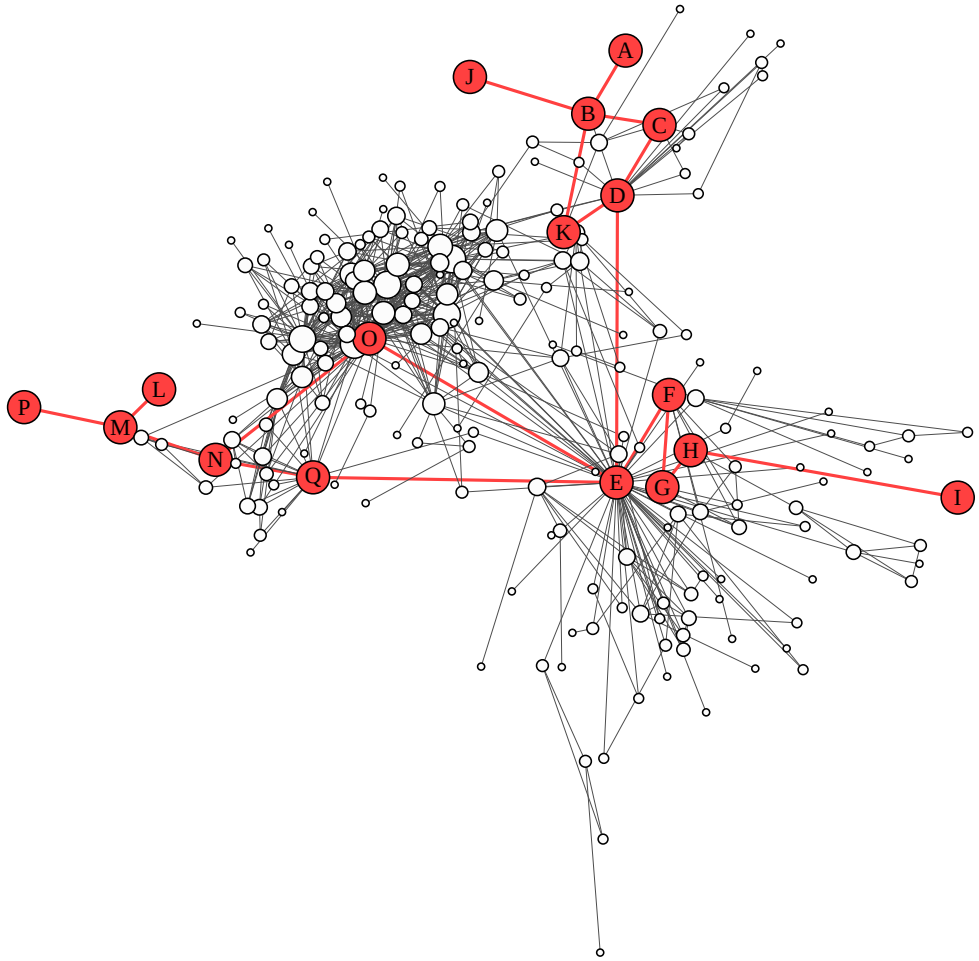


Figure 6.10: The sixteen possible diameters. They are obtained by combinations of the peripheral nodes. In any case, they transit through Mme de Warens, thus she is considered as one of the potential centres of the network. (A) Zanetto Nani. (B) de Altuna y Portu. (C) Abbé de Binis. (D) comte de Montaigu. (E) Mme de Warens. (F) comte de Gouvon. (G) Pauline-Gabrielle de Breil. (H) Mme Basile. (I) Pietro Basile. (J) Mme La Selle. (K) FX de Carrion. (L) Seigneur de Cromessière. (M) Mathieu Buttafoco. (N) KE von Graffenried. (O) Thérèse Levasseur. (P) Pasquale Paoli. (Q) George Keith.

The average path length is equal to 3.208, which means that any two characters are on average separated by 3.208 edges, or 2.208 vertices. This implies that any two characters may be very close together. Since these characters relations or interactions take place on a given set of pages, which correspond to one or more narrative events in total, this well connected network implies that the network of narrative events, where links are shared characters, is tight too. Thus, there is continuity and recurrent characters along the main story line.

6.3.2 Narrative levels

The weight attribute on the edges—called "intensity" (def. 31)—was used to generate networks $\mathcal{G}$ and $\mathcal{G}_3$ in section 6.2.3. This step was necessary to target the relations having an interpretable meaning, and to keep only these in the final network.

In addition to generating the network we study in the rest of this thesis, intensities have the potential to show the underlying narrative structure, floor by floor. This attribute shows how strong a relation between two characters is, in terms of co-occurrences, thus proximity of the characters in the narration. Subsequently, we generate a network by raising the minimum intensity required: we obtain an interlocked sequence of networks, peeling off the narrative levels of the character network, up to the most important ones. We show the networks obtained for minimum intensity, varying from one to nine in figure 6.11.

In this figure, we see how the different parts of the character network are dismantled. We knew of $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$, and the next ones are more readable, thus instructive. As between $\mathcal{G}_1$ and $\mathcal{G}_2$, the gap between $\mathcal{G}_3$ and $\mathcal{G}_4$ is not tremendous. The vertices become detectable when reaching $\mathcal{G}_5$. In this case, we see an important cycle in the middle, containing brothers Gabriel Bonnot de Mably and Etienne Bonnot de Condillac. This cycle implies that some narrative events are related via diverse manners through their incident characters. From $\mathcal{G}_5$ to $\mathcal{G}_9$, we see that the agglomeration of characters from the second part is strong: they not only occurred many times but were also often co-occurrent. The small edge on the side, but related to them, links Rousseau's late friends George Keith and Du Peyrou. We observe that the representation of Mme de Warens has links intense enough to have her stay. The remaining disconnected edge on her side is composed of M. and Mme Lambercier. Via the strongest links of the network, we observe a representation of the narrative core of the novel in terms of characters. Of course, there is no deterministic threshold perfectly suited for this part of the analysis. In each case, it is important to generate the interlocking sequence of networks, then describe them depending on the size of the network and the interpretation we make of the links.

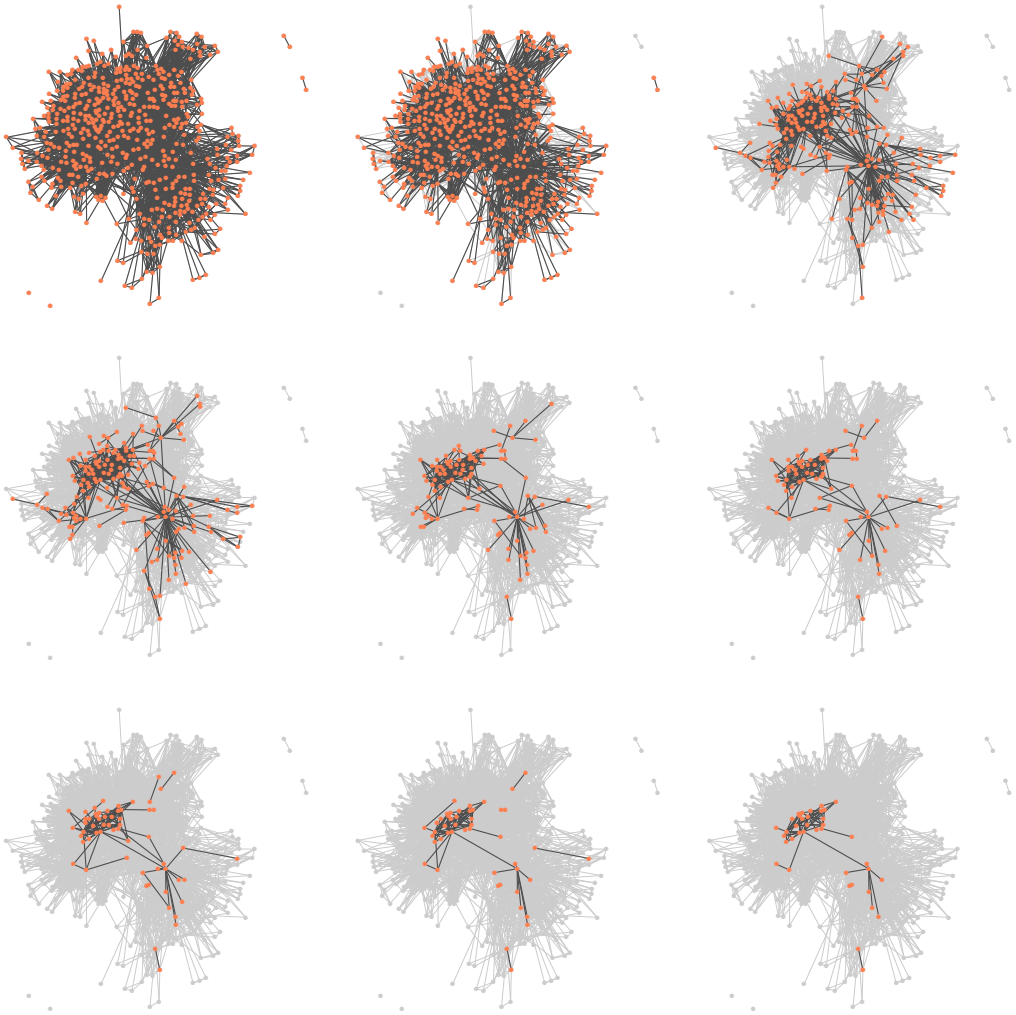


Figure 6.11: Narrative levels of the character network, induced by minimum values of intensity going from one to nine. These raw figures are meant to sketch the layers contained in the character network.

7 Centrality

Nous avouerons que notre héros était fort peu héros en ce moment.

La chartreuse de Parme
STENDHAL

We propose an approach to centrality that fits into the environment of character networks. The characters not only reinforce their roles by appearing side-by-side, on multiple occasions, but they play one of many possible narrative roles in the story as well and, by extension, in the discourse. Among the possible roles are the protagonists, on which the attention is focused, or minor characters, that play secondary roles[1]. Usually, these would outnumber the protagonists, and their distribution would follow a "long tail" (Sack, 2011; Moretti, 2011). Still, they are potentially of some importance at a given time in the narrative, but lost in the midst of a large number of characters at the global level. Or they are simply present to help a transition from one story arc to another, or by means of loyalty to the story. To describe the narrative roles of the characters who occur most is a simple task, which only requires us to draw conclusions from the observed measures of centrality. Analyses of novels document the roles of such characters, since they are intricated into narratives which cannot be understood without their presence. We propose a classification of all characters, including minor ones, by following the steps of (Woloch, 2003). Centrality measures will help us to analyse all characters in a book's character network and to define a typology of narrative roles. Such an angle of attack can potentially lead to analyses at larger scale.

[1] Alex Woloch even defines *minor minor characters* :

> "If Lydia and Wickham are minor characters who, through their very minorness, are central to the novel's semantic field, how can we account for most of the minor characters in *Pride and Prejudice*, who are not thematically significant? It is an error to concentrate on how Wickham and Lydia come to *represent* multiplicity in contrast to Darcy's and Elizabeth's singularity and forget that this opposition is effective only because of the larger structure of multiplicity in which all four characters are embedded. Lydia and Wickham disappear into a fragmented world evoked in the narrative itself, through the many flattening descriptions of more periphera minor characters." (Woloch, 2003, p.116)

Anthropologist Robin Dunbar has estimated the number of relations to others a human being can manage in our social environment at 150 (Dunbar, 1992). When reading a novel, we can suppose the existence of such a boundary with the number of characters we can keep in mind. Some characters appear throughout the novel, while others appear only on a few pages but make a strong impression. They may stay on the reader's mind until the end of the narrative without even reappearing. In most cases, it is difficult to remember all of the characters appearing in the text, as well as who they were with and when it happened[2]. In the text, the characters are organised around the other constitutive units that are the events, the places and the durations (Bal, 1984). The narrative proximity of two characters is deduced from their positions in the text, which itself depends on their simultaneous apparitions in same scenes. The network represents this narrative proximity. By searching for clustered groups of vertices, we can find similar characters occurring in a given event, or side-by-side along the main story arc. Nodes having similar properties can also lead the analysis to highlight a shared narrative role.

The importance of a character is usually computed by deductions based on the nearest neighbours or all neighbours. Eigenvector centrality is the closest of the canonical centrality indices to quantify the cohesive role of a character in terms of narration. Another approach for measuring narrative cohesion is the quantification of the omission of a character. Finding the minimum set of characters whose disappearance makes the network collapse brings forward the narrative backbone of the novel.

In this chapter, we discuss the importation of centrality from social networks to character network analysis (7.1). Then, we apply centrality measures and study them independently, illustrating their particular properties and making adjustments relative to the substitution of social networks by character networks. (7.2). We discuss vitality, a versatile method seldom used in social network analysis, but whose suitability with character network analysis was demonstrated by (Moretti, 2011). It leads us to mock up the narrative importance of a character by measuring the result of its disappearance (7.3). The diverse centrality distributions are based on nodes' information, but when they are known in their entirety, they comment on the network itself. The indices are also used to identify the narrative roles of the characters. We sketch a typology of these and illustrate it with characters from *Les Confessions* (7.4). Eventually, we present an appraisal of characterisation on the basis of character networks, and of using a network approach to build character-spaces and their character-system (7.5).

[2] This argument was used by Cole Heyn to discuss student projects at University of Wisconsin Green Bay in 2013 (Heyn, 2013):

> "[...] when I am reading a lengthy novel, even closely, my brain can only retain so much information. I may not recall, when on page 436, which character interacted with another early in the text, or how much over the course of the text. In this way the character network is a visual annotation of the text's interactions, and can show the level of connection between characters."

7.1 From social sciences to literary theory

In this chapter, we apply the concept of centrality to character networks, or more generally literary networks. To begin with, we discuss its use in social sciences, then the way we can incorporate it into our analytic framework.

7.1.1 As social phenomena

In the context of incomplete definitions and understandings of centrality (see section 2), we base our analysis on mathematical definitions first, and then on the usual interpretations existing in social network analysis. We want to start from the most abstract rather than using social-oriented ideas. We would also like to highlight differences in centrality measures rankings, or an abnormal absence of these differences, or even completely normal and expected results. Centrality is useful in detecting curious configurations in the network, as well as for describing such configurations when brought forward by other methods.

At this stage, we pay heed not to introduce circular reasoning by extracting interesting patterns with centrality and then explaining them with the centrality concept used in the detection. Instead, we must explain them with other independent properties of centrality, or adopt a narratological approach. If we need to explain a given local structure with centrality, then it must be accomplished on characters or groups of characters that bear a narrative interest. We justify the need for and use of centrality in the analysis in this way.

7.1.2 The use of centrality on character networks

In section 5.4, we showed examples of character networks that are difficult to extract with character network analysis because of the lack of documentation on the creation process. This warning applies to identifying and understanding the network before the analysis, and more generally to integrating how centrality works in the context of character network analysis.

The objects being studied are different in subtle but significant ways. A social network is composed of individuals moving freely in a world whose boundaries are defined by the observer. They are auto-organised: the network structure is the result of micro-level emergence. On the other hand, a literary network is composed of (fictional) characters moving along a narrative written by the author in a world whose boundaries are also defined by the observer[3]. They all are projections from the author. In the writing process, we do not know if he consciously organised them as they appear in the character network, or if this is an emerging result the author could not foresee. Still, the author has control of the story-world and is responsible for the whole narrative. The network structure results from a macro-level process.

Extracting the meaning of a network's constitutive elements in social science or literary analysis may differ, but the mathematical formulas do not. This is a stable feature: the proper-

[3] Usually the novel itself, but it can be a chapter, a series of novels, subset of context-defined scenes, etc.

ties from graph theory that are depicted by centrality are common to both cases. Formulas are computed on the graphs, which are the abstract mathematical objects: the networks without any context. Therefore, the results are to be read the same, but the interpretation depends on the discipline's context.

Before going any further, we must understand how to the interpret centrality. In social network analysis, centrality measures the importance of a node according to social phenomena like popularity, influence, control, or integration. In literary networks, we highlight the narrative functions of the character. It is about how they appear in the narration/discourse. In which events do they appear ? Based on the importance, or centrality, of the events, this gives hints on the character's role. In which places are they located? This steers the analysis on the geographical context, which also plays its part in interpretative contents. At what time, or how long are they present in a given scene ? And so on. They help to define the character under study by showing similarities or contrasts. Their proximity in the story-world also highlights shared or, on the contrary, different social and geographical properties. For example, in figure 6.9, we perceive the community structure of the character network of *Les Confessions*. It separates the Venetian, provincial French, Swiss and Parisian societies.

We have used the convention of assimilating centrality and vertex centrality. It is important to consider edge centrality as well, which is the computation of centrality on edges. This is realised by building the dual graph of the graph in use and computing vertex centrality. Edge centrality evidences the key relations in the character network. Though we do not want to consider a complete dual network analysis, we use the measures obtained on edges to improve the quantified knowledge on their ends.

7.2 Measures & interpretations

We have anticipated the stakes of a detailed centrality analysis, and we now present the different steps we used to analyse the results. We began by considering every index separately. We consider the four canonical centrality indices — degree, betweenness, closeness, eigenvector — since they measure various dimensions of nodes structural properties (Freeman, 1978; Koschützki et al., 2005b; Borgatti and Everett, 2006; Brandes et al., 2012). In each case, we present the underlying mathematics as well as the narrative meaning, and we comment on selections of values from common cases as well as outliers. Then, the subsequent sections cover the crossing use of the centrality indices and extensions of this concept and its implications.

7.2.1 Degree

Degree centrality is a locally-based index. Let $G = (V, E)$ be a non-directed network with $|V| = n$ and adjacency matrix $a_{ij}\}_{1\leq i,j\leq n}$. Degree centrality of node $x_i \in V$ is defined by :

$$c_D(x_i) = \sum_j a_{ij}. \tag{7.1}$$

It is the sum of the incident edges of a node, or equivalently the number of its neighbours. The computation does not take into account elements at a distance greater than one. The measure reports the size of node's neighbourhood. It is the most simple centrality measure[4], the mandatory introduction of a centrality analysis. According to (Degenne and Forsé, 1994, p. 156), degree centrality measures

> [...] l'activité ou la capacité de communication ou d'échange de chaque individu au sein du réseau, en ne tenant pas compte de sa capacité à contrôler ces communications.[5]

Additionally, the degree centrality is related to the centre versus periphery structure of the network:

> An actor with a high centrality level, as measured by its degree, is "here the action is" in the network. [...] An actor with a large degree is in direct contact or is adjacent to many other actors. This actor should then begin to be recognized by others as a major channel of relational information, [...] actors with low degrees are clearly peripheral in the network. Such actors are not active in the relational process. (Wasserman and Faust, 1994, p.179)

7.2.1.1 Distribution

In figure 7.1, the cumulative frequency, density (left) and frequency (right) show the degree distribution of network $\mathscr{G}$. Most of the nodes have low values of degree centrality.

We plot the distribution on log-log scales in order to obtain the alignment of the points. The distribution can be fitted by a power-law[6], meaning that a small subset of nodes has an important number of relations (max = 64), while a large subset of nodes has few relations (min = 1). According to the structural properties of power-law networks, some nodes play the roles of hubs (Barabási and Bonabeau, 2003). Mme de Warens is one of them. She is connected to a large number of characters, and this implies that her narrative role is the one of a revolving character. It is interesting to observe that Mme de Warens appears in a perceptible way in the network visualisation as well as in the distribution.

We know from the properties of a power-law distribution that if some nodes are to be considered as hubs, having high degree, most of the characters have small degree. Therefore they do not appear side-by-side with many characters in the network. In fact, the network adopts a shape in which nodes with large degrees are slightly anti-correlated with nodes

[4] For the anecdote, the simplest centrality measure would trivially be one giving 1 to the node under study and 0 to the others. Since centrality has no admitted mathematical definition, it allows this.

[5] "[...] the activity, or the capability of each individual to communicate or exchange inside the network, while not taking into account his or her capability to control these communications."

[6] The method used to fit the distribution is defined in (Clauset et al., 2009) and implemented in *igraph*, a R package (Csárdi and Nepusz, 2006). This is an ambiguous result since the sample is not large.

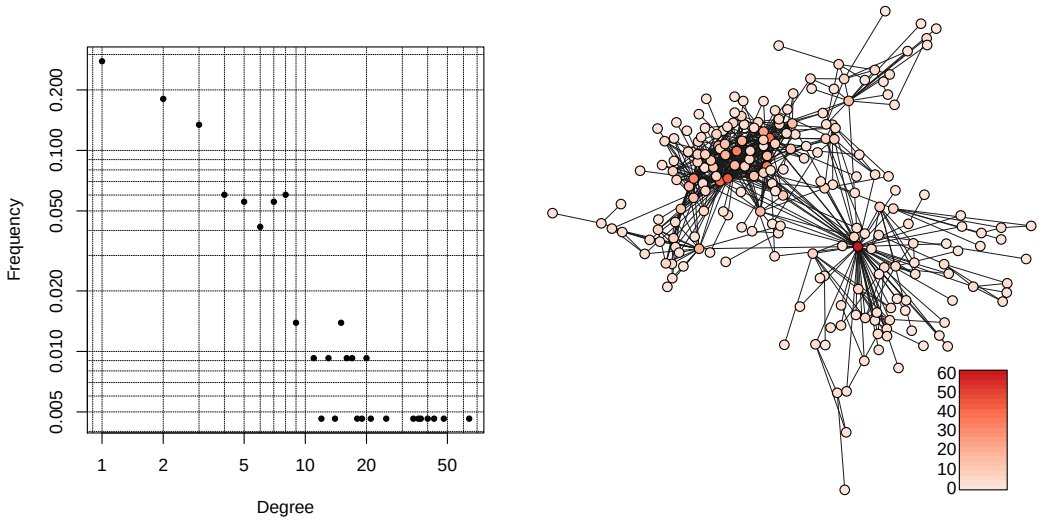


Figure 7.1: (Left.) Degree distribution on a log-log scale. Its tail ($x_{min} = 6$) follows a power-law with exponent of the fitted distribution $\alpha = 2.43$. The case $x_{min} = 6$ maximises the likelihood ($p = 0.841 > 0.1$). The tail can be fitted less accurately, but on a larger interval, by a power-law from $x = 3$ ($\alpha = 2.05$, $p = 0.227 > 0.1$). (Right.) The network with degree centrality visualised.

having small degrees. This is expressed by a measure called degree assortativity, which takes up a value between 1 (assortative, or correlated) and -1 (disassortative, or anti-correlated), with case non-correlated when the value is close to 0. In the case of the network model $\mathscr{G}$, it is equal to -0.173.[7] We can therefore conclude from that discussion about degree distribution and assortativity that, in the narrative structure of the characters of *Les Confessions*, minor characters in the narration have the tendency to be linked with nodes having large numbers of connections. This is explained by the few characters that accompany Rousseau over long periods of time (Mme de Warens, Thérèse Levasseur, Grimm, Diderot, etc.), playing the narrative role of consolidating the coexisting presence of characters in the discourse, as told and maybe desired by Rousseau. In addition to that, nodes with high degree are not necessarily clustered together, as we have already observed with narrative levels (see section 6.3.2), showing that there is not a single dense group of hubs, or highly connected nodes, but rather hubs distributed throughout the network.

7.2.1.2 Results

In table 7.1, we show the characters with highest degree centrality in the network.

In character network analysis, the interpretation of a relation implies that the two characters at its extremities have been co-occurrent in at least one narrative event. A large measure of

[7] -0.07 for $\mathscr{G}_0$ and $-.103$ for $\mathscr{G}_1$. Our method increases assortativity negatively.

	Degree
Mme de Warens	64
Thérèse Levasseur	48
Mme de Luxembourg	43
Denis Diderot	40
Grimm	37
Louise d'Épinay	36
Duc de Luxembourg	34
Mme Dupin	25
George Keith	21
Comtesse d'Houdetot	20
Jean Le Rond d'Alembert	20
M. et Mme Levasseur	19
Guillaume de Lamoignon	18
Voltaire	17
Comtesse de Boufflers	17
CL Dupin de Francueil	16
Comte de Montaigu	16
LAJ de Rochechouart	15
Charles Pinot Duclos	15
NB Duchesne	15

Table 7.1: The characters with the twenty highest degree centrality values.

degree centrality generally implies that the character took part in many events, or in populated ones. For example, Mme de Warens, the character with the highest degree centrality, appears regularly and frequently in chapters II to VI, alongside many different characters and in many disjointed events. On the other hand, characters with small degree centrality usually appear in one or two events at most. Jean-Baptiste Salomon appears on pages 334 and 335 and has a degree equal to one, since he is related only to Mme de Warens. The character is a doctor, who takes care of Rousseau's and Mme de Warens' health while they are retired in the countryside. This character is part of a narrative event. The boundaries of this event coincide with the occurrences of that character.

In general, nodes with high degree centrality play the role of *hubs*. In a literary network, there are many possible interpretations of that concept. For example, the hub can represent the narrator, if the narrator and the protagonist are related, and narration tells a story in which he or she is present. It is what would happen in our case if each mention of Rousseau via the first-person pronoun was recorded. There can be more than one hub in a network. For example, a few central characters around which the story is told, like Hamlet and Claudius in William Shakespeare's play *Hamlet* (Moretti, 2011). On the other hand, examples with no hubs exist. Trivial situations happen in cases where the network is complete or very dense, or when the network's shape is circular. The *huis clos Twelve Angry Men* by Reginal Rose is an example of narrative with no hub, and no character more central than the others. A method based on presence gives a complete network and methods based on spoken interactions are

rather well distributed. It would be interesting to analyse if the reticent juror, the one without whom there is no story, has different network characteristics.

We know from chapter 6 that the two parts of *Les Confessions* have slightly different structures. Degree assortativity behaves differently in each case. By going down to that level, we omit the edges contained inside the other part, as well as the edges that were interlinking parts one and two. The resulting construction gives rather significant returns, showing the strong differentiated structures of both induced networks. The network reduced to the characters of part one has a degree assortativity of -0.248 while degree assortativity for part two is equal to -0.145. The central role of Mme de Warens in part one, linked to 58 characters out of 94, shows how the narration itself is tremendously centralised.

7.2.1.3 Occurrences

One's having a high number of occurrences does not necessarily obtain high degree centrality in the network. The study of degree centrality on the whole corpus implies that some properties of the narration are flattened, like occurrences of characters. The number of occurrences measures the importance Rousseau gives to the character or to the scene that contains it. Occurrences distribution show the integration of a character in the narration, without bearing relations to the other characters. Degree centrality also measures the inclusion of a character, but in the society formed by all characters. If a character always appears alone, it is always left aside by the algorithm defining the final network (see sections 6.2.1 and 6.2.2). For example, Charlotte de Vulson appears on six pages, alongside nine characters[8]. But none of these relations is significant enough to be retained and to make this character be part of the final network. It has to be kept in mind that the object under study is the narrative society in the novel, and that key characters in the narration that are not related to key characters in that society are not part of the analysis.

Occurrences are used to deduce the co-occurrences, but characters with identical numbers of occurrences do not necessarily have same degrees, and vice versa. We plot occurrences against degree centrality in figure 7.2. We highlight outliers in the plot, that is to say characters occurrent on many pages and having few connections, or the contrary. In general, their degree and occurrences are related, as is shown by the diagonal shape scatter plot. Still, we need to understand which are the borderline cases.

For example, Louise-Françoise-Pauline, duchesse de Montmorency-Fosseux appears on only two pages, 675 and 683, and is connected to Madeleine-Angélique, duchesse de Luxembourg, the comtesse de Boufflers-Rouverel and Charles-François-Frédéric de Montmorency-Luxembourg. Therefore, her degree centrality is equal to three. On the other side, Jean-Jacques

[8] That is according to the method developed in section 6.2.1. She appears alongside four characters when using the "naive" method (see section 6.1).

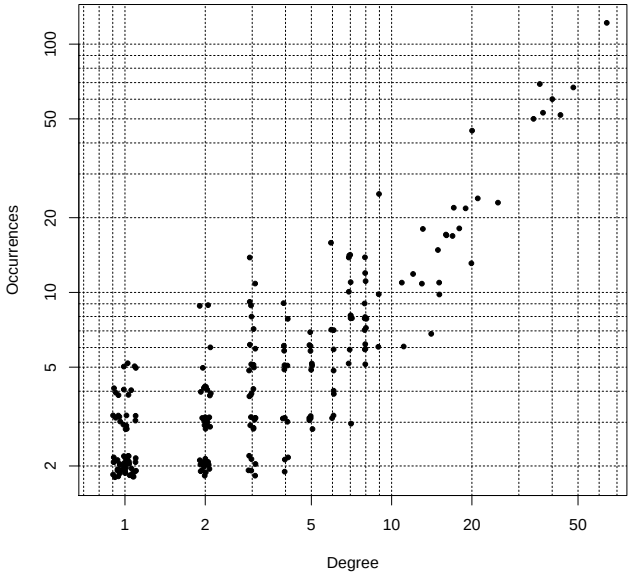


Figure 7.2: Degree and occurrence distributions (logarithmic scales).

Lambercier has degree centrality equal to three, too. Nevertheless[9], he appears on fourteen pages, most of them in the first chapter, and is connected to Gabriel Bernard, Abraham Bernard and Gabrielle Lambercier. This indicates that at the start of the novel, the network density is lower.

In many cases, a large number of occurrences leads to a large number of acquaintances[10]. However, as we have seen, if at one point the concentration of characters is less dense in the narration — e.g. encounters with reduced crowd and lasting many pages —, then a popular character according to the number of his appearances in the narration is not necessarily a central character. The network approach reveals aspects of the hidden internal structure in those types of observations.

7.2.2 Betweenness

Betweenness centrality measures the frequency of appearance of a node on the shortest paths of the network. In social as well as communication networks, this index is useful in

[9] Before the reduction in the network building method (section 6.2.1), she is connected to 16 characters in $\mathscr{G}_1$, and he is connected to 22 characters in network $\mathscr{G}_1$. We conclude that there is is no influence of the method in this case.

[10] Pearson correlation coefficient between degree centrality and occurrences is equal to 0.938 and Spearman correlation coefficient is equal to 0.772. In the first case, the high value is in some way biased by the large proportion of nodes having both low degree and low number of occurrences. In the second case, the use of ranks on that same large mass introduces variability and lower the value of the coefficient. Both coefficient are therefore imprecise, but generally show a significant measure of correlation.

highlighting actors and objects who possess control on transit in the network. It is a concept motivated by the understanding of flows in the network. It is given by

$$c_B(x_i) = \sum \sum \frac{g_{jk}(i)}{g_{jk}}, \tag{7.2}$$

with g_{jk}, the number of shortest paths from node x_j to node x_k, and $g_{jk}(i)$ the number of shortest paths from node x_j to node x_k that comprise node x_i. The double sum is computed on all the pairs (j,k) such that $j \neq i \neq k$ and $j < k$. Given a node and the set of all shortest paths between two other nodes, the measure of betweenness of the node under study increases by one if it appears on all shortest paths (there can be only one such). If there are two shortest paths and it belong to only one of them, the measure is increased by $\frac{1}{2}$. This definition is only valid on connected networks.

This index offers another dimension to the analysis of the roles of the nodes in that it highlights key positions in the narrative flow:

> Betweenness centrality differs from the other centrality measures [...] in being not principally a measure of how well-connected a vertex is. Instead it measures how much a vertex falls "between" others. (Newman, 2010, p.188)

Newman illustrates the use of betweenness on a network of co-occurrences of actors: two actors are related if they appear in the same film according to the *Internet Movie Database*[11]. Under these conditions, Fernando Rey, a frequent collaborator of Luis Buñuel, appears to be the most central actor of that network according to betweenness, ahead of prolific and long-standing Christopher Lee. He proposes a qualitative explanation for that result, an example of how to interpret betweenness in this real-life social network:

> It is perhaps no coincidence that the highest betweenness belongs to an actor who appeared in both European and American films, played roles in several different languages, and worked extensively in both film and television, as well as on stage. Rey was the archetypal "broker", with a career that made him a central figure in several different arms of the entertainment business that otherwise overlap relatively little. (Newman, 2010, p.189)

7.2.2.1 Distribution

We show the betweenness centrality distribution in figure 7.3. The shape is similar as the degree distribution's, with the majority of values regrouped and a few nodes getting significantly higher results.

The median is situated at 6.833 and 92 characters (42.6% of the vertices) have betweenness equal to zero, meaning that they are connected to only one node[12], or simply not situated on

[11] www.imdb.com
[12] This is the case for sixty of them.

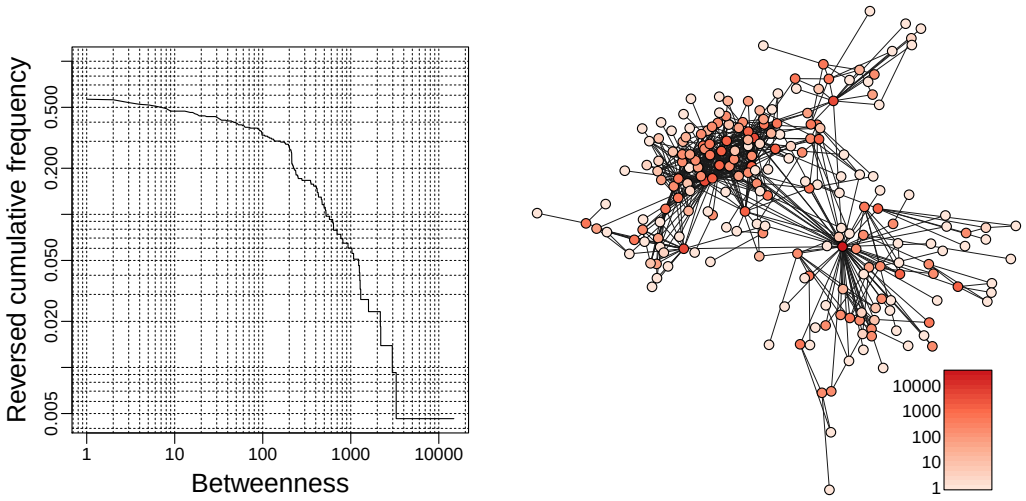


Figure 7.3: (Left.) Betweenness centrality distribution. The double log scales highlight a two-parts power-law. We plot the cumulative frequency of the reversed distribution. The segment on the right is an artefact caused by the one measure excessively larger than the following ones. (Right.) Betweenness visualised on the network itself.

any shortest path. For example, the node representing the Venetian noble Zanetto Nani is a leaf, that is to say it has only one connection, and is thus not crossed by any shortest path. Its only neighbour is Manoel Ignazio de Altuna y Portu, whose presence is close in the narrative. Anyway, they probably never met, and this is not what their relation in the network signifies. Altuna y Portu is the node that gets Zanetto Nani connected to the rest of the network. A relation can represent people knowing each other, but more generally we consider that it is a narrative bond. And in this case, it exists because they occur on related pages, non-randomly, therefore in shared narrative events.

Narratively, we presume that nodes having low or null betweenness have their character-space fully contained in small subsets of narratively more prominent character-spaces. On the contrary, a character with high betweenness acts as a narrative bridge, linking groups of characters together, and thus the events or groups of events they inhabit. Schematically, let's call three characters A, B and C. If A is linked to B and C, which are in two different and consecutive events, not co-occurrent, and A has high betweenness, then the chances are that $A_{cs} \cap B_{cs} \neq \emptyset$ and $A_{cs} \cap C_{cs} \neq \emptyset$, with A_{cs}, B_{cs} and C_{cs} their corresponding character-spaces. For example, David Hume has degree equal to four and betweenness equal to zero. Among his four neighbours are George Keith and the comtesse de Boufflers, who are central characters according to most of the centrality measures. David Hume appears to be narratively depending on their character-spaces.

7.2.2.2 Results

The highest betweenness values are shown in table 7.2. We saw that many characters don't play intermediary roles in the network. Here, we observe that some characters have very high values compared to the mean of the distribution equal to 261.2. Only 37 characters out of 216 (17.1%) possess a higher measure. These characters with high betweenness form a skeleton of narrative entities with pivotal role.

	Betweenness
Mme de Warens	14824.21
Comte de Montaigu	3280.87
Mme de Luxembourg	2962.41
George Keith	2189.27
Denis Diderot	2166.44
Grimm	1589.35
Mme Dupin	1296.00
Louise d'Épinay	1277.33
Mme de Larnage	1266.00
Voltaire	1251.54
Thérèse Levasseur	1228.63
Comtesse d'Houdetot	1076.30
CE La Roque	1031.33
Jean Bonnot de Mably	930.55
Abbé de Gouvon	812.97
M. et Mme Levasseur	770.11
Daniel Roguin	698.87
Gabriel Bernard	638.00
KE von Graffenried	636.00
Claudine de Galley	622.67

Table 7.2: The characters with the twenty highest betweenness centrality values.

Mme de Warens is the node with significantly higher centrality than others. This is due to the fact that she plays a central intermediary role inside the first part as well as on the whole narrative level. At the first part level, she is linked to two thirds of the characters in $\mathscr{G}_0$ thanks to her social skills and will to help Rousseau in his life. At the whole narrative level, she plays the hub relating characters in the second part to the first one. Despite the fact that she is not the only one in such a position—appearing in the two parts of the book—, her high degree increases her influence. A measure of betweenness centrality based on random walks (Newman, 2005) reduces the importance of degree centrality in the computation.

A node with high betweenness does not always have high degree, as figure 7.4 shows. For example, Mme de Larnage, at the 9th position, is only at the 43th in terms of degree. Here we show that such a case is possible. In addition, while Mme and M. de Luxembourg were close in function of degree centrality (3rd and 7rd), there is an important difference in betweenness. The discussion about their roles, with interpretation of betweenness, as well as the one of Mme de Larnage, who has an especially high betweenness, happens in section 7.4 in the

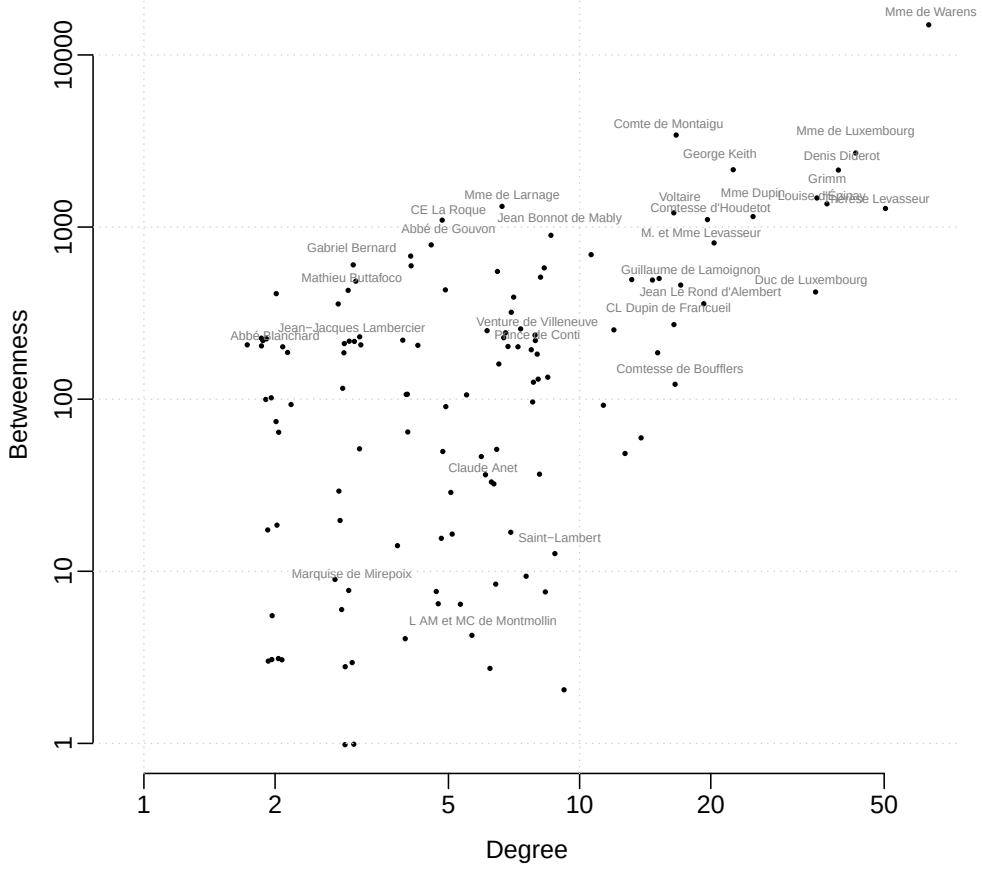


Figure 7.4: Degree and betweenness distributions.

context of a typology based on centrality of the different possible narrative roles. The high betweenness values of comte de Montaigu and George Keith are also explained there and in the forthcoming section 7.2.2.3.

7.2.2.3 Community detection

A subset of nodes inside which links are more frequent than to the nodes outside is called a *community* (or *cluster*). A network possesses a *community* structure if a partition in such subsets exists. We measure the efficiency of that partition via a measure called modularity (Newman and Girvan, 2004). Networks extracted from real-world have the tendency to show a community structure (Porter et al., 2009).

In this section, we show that there is a community structure in the character network of *Les Confessions*. It runs over the simple and highly intuitive partition that would divide the two parts of the book, or three if including the subset of nodes appearing in both[13]. The algorithm is hierarchical, meaning that we create a partition of two, then of three, etc. until the partition is composed only of singletons[14]. Modularity estimates the quality of the cut. The partition that maximises modularity is considered optimal. In our case, there are clearly defined partitions, and more than three. They appear to be representative of a selection of story events. The use of betweenness centrality defined on edges completes its vertex counterpart in highlighting narrative changeovers.

The algorithm At the time when Michelle Girvan and Mark Newman defined their original algorithm of community detection (Girvan and Newman, 2002), network analysis was being rediscovered by physicists; a situation that led to numerous publications in famous journals. Literature on the subject was exploding (Fortunato, 2010), and in this context, community detection became its own new research field. Indeed it is a challenging problem, since there is no exact solution: researchers have to develop an algorithm, an optimal partition criteria and take care of computational efficiency. Since the first algorithm, most of the focus has been on large-scale networks (Clauset et al., 2004; Pons and Latapy, 2005; Raghavan et al., 2007; Blondel et al., 2008; Ovelgönne and Geyer-Schulz, 2012).

In our case, the network is small; therefore, we use the original algorithm of Girvan and Newman. It is based on a measure of betweenness centrality defined on edges. Edge betweenness centrality formula is nearly identical to formula 7.2. It is defined by:

$$c_{B'}(e_i) = \sum\sum \frac{g_{jk}(i)}{g_{jk}}, \tag{7.3}$$

with g_{jk}, the number of shortest paths from node x_j to node x_k, and $g_{jk}(i)$ the number of shortest paths from node x_j to node x_k that comprise edge e_i. The double sum is computed on all the pairs (j,k) such that $j < k$. The results from the computation of that index are interesting independent of the community detection process[15]. The idea behind the algorithm is to find the link with the highest capacity — basically the number of shortest paths it belongs to — and to remove it. We then start over again on the network with the edge removed. We stop when there are no edges left, or if a sufficient quantity of clusters has been generated. In the end, a measure called modularity estimates the quality of the cut (Newman and Girvan, 2004). We keep the cut maximising modularity. Modularity is

[13] In fact imposing a partition of three clusters brings the previously shared nodes into the second part cluster, the third cluster being Venice.

[14] A singleton is a node with no incident edge.

[15] Surprisingly, edge betweenness centrality is not correlated with edge weights: Pearson correlation gives −0.047 and Spearman correlation gives −0.008.

> [...] the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random. (Newman, 2006)

We choose the version defined in that last reference. Mathematically, it is expressed by:

$$Q = \frac{1}{2m} \sum_{ij} \left(a_{ij} - \frac{c_D(x_i) c_D(x_j)}{2m} \right) \delta(C_i, C_j) \,, \tag{7.4}$$

where m is the number of edges in the network, $\{a_{ij}\}_{1 \leq i,j \leq n}$ the adjacency matrix and C_i the label of the cluster node x_i belongs to. Due to the Kronecker delta[16], the sum is actually computed on all pairs of vertices belonging to a same cluster. In the parenthesis, the fraction measures the probability for a relation to exist in the random case. Therefore, the denser the links inside every cluster, the higher the value of the parenthesis, thus the sum. (Brandes et al., 2006) proved that finding the partition that maximises modularity is an NP-complete problem, meaning that even for a relatively modest size, we would need years or more to find the best partition.

Application We have seen that there are many variants of community detection algorithms. Here, we use the classic one of (Girvan and Newman, 2002) for which the reasoning is intuitive and explainable in a narrative sense. At this stage, we still have two things to discuss: shall we consider the weights of the edges, and shall we stop when modularity indicates to do so.

Concerning the first point, we tested the two methods[17]: weighted and unweighted. Weights are used in the computation of the shortest paths for edge betweenness. In this case, the length of a path is the sum of the weights of the edges, instead of the number of edges. The shortest path is the one minimising the total distance between two nodes. Since our weights show the intensity of a relation, the higher the intensity, the easier it must be to "circulate". Thus, we use the inverted values of the weights in the algorithm.

The comparison clearly goes in favour of the weighted version. The highest value reached by modularity in this case is equal to 0.5345, for twelve communities, while in the unweighted case it is equal to 0.4527, for fourteen communities. The weighted algorithm makes the first cut between what will be eight and four different communities, while the first cut of the unweighted algorithm identifies a community that is then left untouched, as with the second one; it excludes groups from the start, while in the weighted case the divisions were more uniformly distributed on the whole network. It seems to consider the global network structure less. In addition to that, the unweighted algorithm seems to come unstuck with the densely connected characters found in the second part of the novel. It cannot find a cut, nor drop

[16] $\delta(x, y) = 1$ if and only $x = y$, otherwise $\delta(x, y) = 0$.

[17] The algorithm we used from igraph (Csárdi and Nepusz, 2006) cites as reference (Brandes, 2001). The code is available at https://github.com/igraph/igraph/blob/master/src/centrality.c from line 1945. (Accessed on 01/03/2014.)

them untouched, hence it gets a giant community of 88 characters, which contains 40.7% of all the characters. With the weighted algorithm, the previous problematic group is separated, and the biggest community is now made up of 52 characters. In conclusion, the intensity measured on the links guided the algorithm towards a more efficient partition.

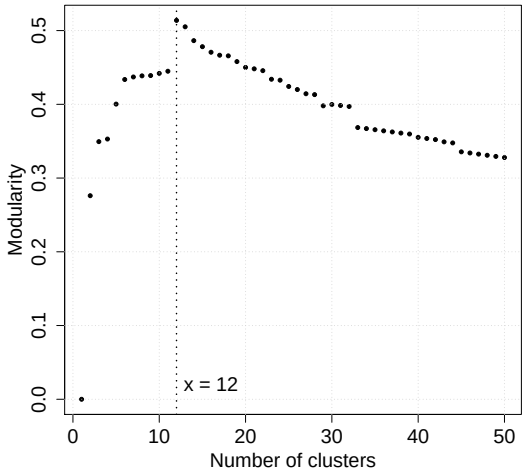


Figure 7.5: Modularity scores of the fifty first community partitions obtained with the weighted edge betweenness community detection algorithm.

Regarding the stop criteria, figure 7.5 shows that in our case modularity is at its maximum for twelve communities, and close to it in thirteen. The value is significantly far from zero (the maximum is one). It also shows that having a different number of cuts is not really relevant in any direction — smaller or higher number of partitions — since there is no other local extremum. Therefore, in the rest of this section, we keep the twelve community partition.

The resulting twelve communities are shown in figure 7.6. The hierarchical cuts are shown in figure 7.7. Nodes of a same colour belong to the same community.

The clusters in the partition are related to fractions of the narrative. It can be a single story event, which is so strictly delimited that it appears aside, e.g. the small community composed of Jacques-François Deluc, Jean Jallabert and Jean Perdriau, whom Rousseau met in Geneva and who are all connected thanks to their narrative proximity in chapter VIII. In other cases, the characters in the cluster are grouped together because they share a series of events linked via constitutive elements of the narration like time or place, e.g. the Venetian cluster led by the infamous comte de Montaigu.

Interpretation His Venetian lover, Zulietta, belongs to the community centred around a previous lover, named Mme de Larnage. We cannot and we do not want to say that the algorithm

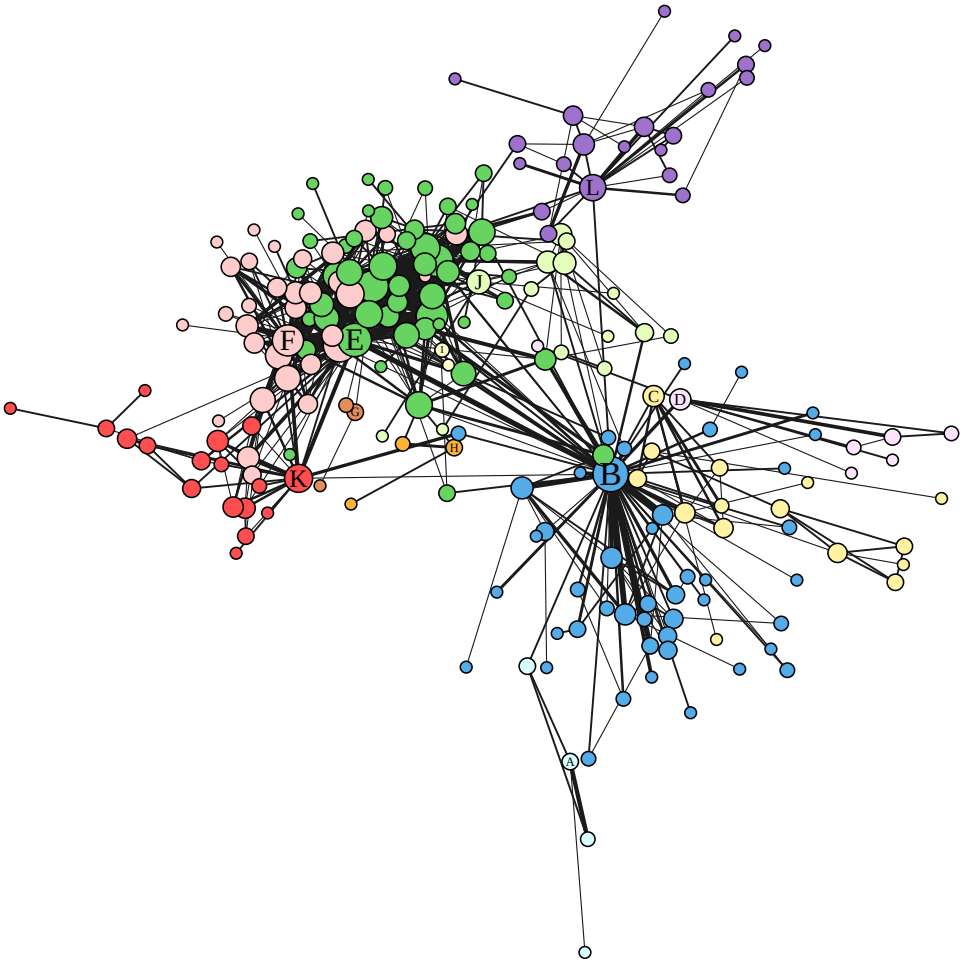


Figure 7.6: Communities in the character network of *Les Confessions*. The composition of the communities is given by a dendrogram in figure 7.7, in which colours correspond with those of this plot. One prominent character is given in order to identify each community: A. Jean-Jacques Lambercier. B. Mme de Warens. C. Comte de Gouvon. D. Mme de Larnage. E. Thérèse Levasseur. F. Duc de Luxembourg. G. Jean-Françoise Deluc. H. François-Robert Mussard. I. M. de Cury. J. Daniel Roguin. K. George Keith. L. Comte de Montaigu.

detected the common sentimental background. Nonetheless, there is an explanation as to why they are so close, and why Zulietta does not belong to the Venetian crowd in the network. In his narration, Rousseau compares the two women. However, Mme de Larnage travels with a wide group of persons at that moment of the story, while Zulietta barely seems accompanied. At least, this is the result of what Rousseau wrote: the consequence of the chosen narration

Figure 7.7: Dendrogram of hierarchical cuts, corresponding to figure 7.6.

at this point is that Zulietta appears lonely. Interestingly, we observe that the community detection algorithm separates the characters on the basis of time and places[18].

7.2.3 Harmonic closeness

One of the canonical centrality measures in social sciences since (Freeman, 1978), closeness centrality was used in the premises of what would become social network analysis (Bavelas, 1950). It is similar to the Wiener index in chemical sciences (Wiener, 1947). Closeness centrality is the sum of the distances from a node to all others which then considers the inverse of the result in order to give the largest value to the most central node:

$$c_C(x_i) = \frac{1}{\sum_{j \neq i} \text{dist}(x_i, x_j)}. \tag{7.5}$$

It quantifies the effort needed to reach all other nodes. In terms of literary analysis, closeness centrality represents the narrative proximity of a character to all others. This information gets more useful if we consider that a character-space bears with it contexts of narrative units such as places, events and time. In this case, closeness centrality ranks characters with respect to how close they are to all the others.

Closeness highlights the centre of the network. It is intuitive, but it has serious drawbacks (Rochat, 2009). In particular, the variance of closeness centrality is usually low, which makes it necessary to use ranks in order to discriminate the data. Thus, we use harmonic centrality, a similar index but with particular properties. It is not a trifling adaptation of closeness centrality, even if they are strongly correlated in most cases. (Boldi and Vigna, 2013) define a list of axioms with the goal of describing centrality. They reach the conclusion that harmonic centrality is the only index verifying them all[19].

Harmonic centrality considers the harmonic mean of distances, that is to say the sum of reciprocated distances instead of the inversion of the sum of all distances in closeness case. It is defined by :

$$c_H(x_i) = \sum_{i \neq j} \frac{1}{\text{dist}(x_i, x_j)}, \tag{7.6}$$

where $\text{dist}(x_i, x_j)$ is the distance from vertex x_i to vertex x_j. Harmonic closeness measures are correlated with closeness centrality in most cases. Results are more spread out, while nodes at great distances play roles equivalent as when using closeness. Unlike closeness, harmonic closeness deals with disconnected networks. It is defined in a similar way to eigenvector centrality (see section 7.2.4) since one's measure depends on the connectivity of their direct neighbours and also to a lesser extent on the next neighbours. In comparison, degree is

[18] This is increased by the weighted algorithm: a link that was hazardous because of happenstance is considered less than links reflecting frequent co-occurrences of two characters.

[19] They define three axioms, one on size, one on density and one on monotonicity.

restricted to direct neighbours, and a node can get high betweenness centrality despite poor local connectivity, by standing between communities.

7.2.3.1 Distribution

We observe from figure 7.8 that the distribution of harmonic centrality is quite symmetric. Most of the nodes have a value close to the mean. A few in the centre of the network have high centrality and a few on the periphery have low centrality. This measure highlights a hierarchy in the structure of the network. The nodes with the lowest harmonic centrality values are situated on the periphery, therefore distributed in various parts of the network[20].

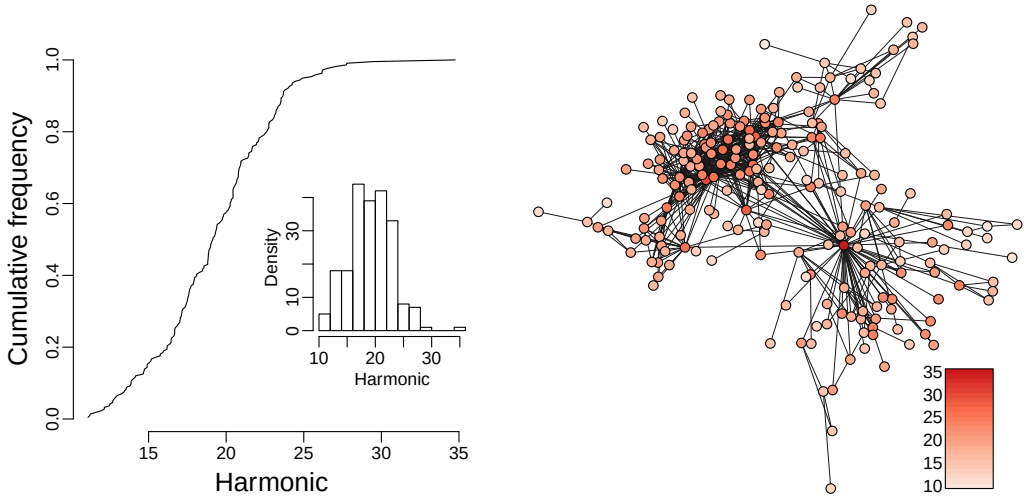


Figure 7.8: (Left.) Harmonic centrality distribution, cumulative frequency and histogram. (Right.) The network with harmonic centrality visualised.

7.2.3.2 Results

Harmonic centrality does not simply give high values to nodes situated at the centre, being there only thanks to a link with a node truly situated in the centre. To get a high value, a node must be situated at the junction of many clusters of characters, but also being connected to a number of them, since without that condition the sum in formula 7.6 decreases rapidly. Results are indicated in table 7.3.

There are many previously encountered names on the top, but also characters that are not considered highly central under the other centrality dimensions. We saw in figure 7.8 that their values of harmonic centrality are close, therefore this is not strong enough to imply a whole new characterisation of their role. However, there is an importance to these characters

[20] This is coherent with a computation of closeness centrality we did for control.

	Harmonic
Mme de Warens	34.757
Mme de Luxembourg	29.539
Denis Diderot	27.771
Voltaire	27.746
Mme Dupin	27.040
Grimm	26.563
George Keith	26.227
Thérèse Levasseur	26.183
Comtesse d'Houdetot	26.182
M. de La Poplinière	25.733
Louise d'Épinay	25.513
Marquise de Verdelin	24.901
Jean Bonnot de Mably	24.775
Guillaume de Lamoignon	24.427
NB Duchesne	24.257
CP de Montenoy	24.246
JA de Condillac	24.065
VC de Gauffecourt	23.989
François Coindet	23.794
Duc de Richelieu	23.694

Table 7.3: The characters with the twenty highest harmonic centrality values.

that are not protagonists, but also not *minor minor characters*. For example, the marquise de Verdelin is someone Rousseau met and corresponded with, but with whom he went from a friendship to distancing himself. She visited him in Switzerland and put him in contact with David Hume in England. Clearly, she is an important minor character. The next character in the list, Jean Bonnot de Mably, was an influential man in Lyon. He hires Rousseau to take care of the education of his children, then present him to other influential people. He appears on only eleven pages and has nine neighbours in the network. Again, he is also an important minor character.

In this case study of a character network analysis on *Les Confessions*, harmonic centrality shows the property of detecting narratively influential characters and putting them in evidence. They are not protagonists but at the centre of non-anecdotic ramifications. The interpretation of the results from computing harmonic centrality is not as straightforward as for the other measures: here, harmonic centrality is a symmetrical distribution that aligns with a hierarchical structure going from the unique center to the periphery. It is especially relevant later, in section 7.4, where we use it along with the other measures in order to classify different types of narrative roles. In this occasion, harmonic centrality allows for example to differentiate two characters having same degree or betweenness by determining if they are in the middle of the character network or closer to the outside.

7.2.4 Eigenvector

Eigenvector centrality[21] attributes a value to a node based on the centrality values of its neighbours. This is dealt with simultaneously for all nodes in the network (Bonacich, 1972, 1987). It belongs to the type of measures called *feedback centrality*, for which every value relies on neighbouring values. It is the solution of

$$c = A_w c, \tag{7.7}$$

where A_w is the weighted adjacency matrix and c the vector of centrality. Thus, solving this problem leads to the resolution of a system of n linear equations with n unknowns, and the use of linear algebra. This is equivalent to the problem of finding the eigenvalues and their corresponding eigenvectors expressed by

$$\lambda c = Ac. \tag{7.8}$$

We know from the Perron-Frobenius theorem that, since the adjacency matrix is non-negative, then there is always an eigenvalue[22] such that the corresponding eigenvector is composed only of positive entries (Newman et al., 2006; Hogben, 2013). This eigenvector, c_E, is our vector of centrality.

The eigenvector centrality index is sensitive to piles of nodes, since they reinforce the importance of one another position. In the context of a structural analysis of the character, the central nodes are the ones connected to the other central nodes. In particular, there is a reinforcement of eigenvector centrality in the centres of the communities, which are densely connected. Still, we find that more importance is given to one community on top of the others, rather than to the various centres of communities (Ilyas and Radha, 2010). We use the adjacency matrix in its weighted form[23], meaning that the elements of the matrix are replaced by the corresponding weights. The resulting centrality values are differentiated on the basis of the link intensities. A higher intensity implies better communication through the observed link, meaning that two nodes strongly connected will significantly increase each other's values. This choice helps to enhance the local structures.

7.2.4.1 Distribution

We show the distribution of eigenvector centrality in figure 7.9 (left). The resulting distribution is long-tailed as degree and betweenness. The maximal value is 1, reached by Louise d'Épinay, then come a few others with high scores. Finally, most of the values are situated between 0 and 0.1, with some very low values among the extremely peripheral vertices, as shown in figure

[21] Or power centrality, or Bonacich centrality from the name of its author.
[22] Positive and the largest one.
[23] Weights were defined in section 6.1.2

7.9 (right). The minimum is attained by character Pietro Basile whose position is situated at the end of a chain of non-structurally prominent nodes along which centrality regularly decreases[24].

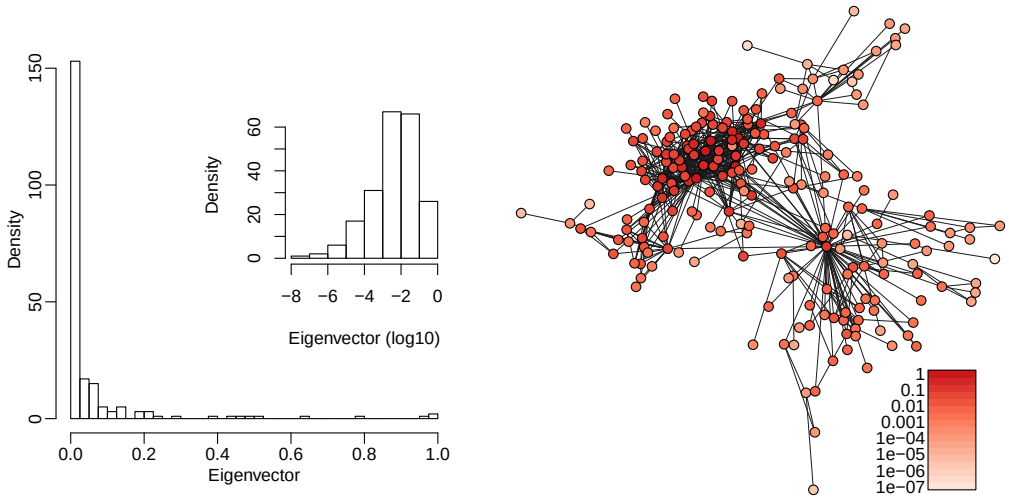


Figure 7.9: Eigenvector centrality distribution. (Left.) Histograms with linear and logarithmic scales on the horizontal axis. (Right.) In the network, vertices color depict eigenvector centrality.

7.2.4.2 Results

Table 7.4 shows the twenty highest eigenvector centrality values. We see that Mme de Warens is not in the first position, differing from all the other centrality measures. Instead, characters from the second part of *Les Confessions* monopolise the highest ranks, with a head trio of three characters that were among the closest acquaintances of Rousseau, and who had the particularity of turning from friends into enemies.

The next in line are also to be found amongst Rousseau's closest friends, or family. The number of important characters is irregular at certain moments of his life, and therefore across the chapters of *Les Confessions*. In particular, chapters VIII to XI are overrepresented here because of characters such as Louise d'Épinay, Grimm and Diderot whose occurrences are recurrent and most often co-occurrent with other protagonists (see appendix C). This shows a fracture, after the spread out first part chapters, where quite possibly the only common trait behind two randomly chosen characters would be Mme de Warens, and the volatility of

[24] Mme de Warens (0.190) → comte de Gouvon ($4.31e^{-3}$) → Pauline-Gabrielle de Breil ($1.83e^{-4}$) → Mme Basile ($5.24e^{-6}$) → Pietro Basile ($8.27e^{-8}$).

	Eigenvector
Louise d'Épinay	1.000
Grimm	0.978
Denis Diderot	0.964
Thérèse Levasseur	0.793
Comtesse d'Houdetot	0.637
M. et Mme Levasseur	0.511
Mme de Luxembourg	0.475
Duc de Luxembourg	0.469
Baron d'Holbach	0.426
Saint-Lambert	0.389
Mme Dupin	0.277
Charles Pinot Duclos	0.246
CL Dupin de Francueil	0.224
Jean Le Rond d'Alembert	0.210
Comtesse de Boufflers	0.203
LAJ de Rochechouart	0.199
Comte de Friesen	0.196
Mme de Warens	0.190
VC de Gauffecourt	0.147
Guillaume de Lamoignon	0.144

Table 7.4: The characters with the twenty highest eigenvector centrality values.

chapter VII (see figure 7.10), with many names emerging from the effervescence of his rich social life in Paris at the time.

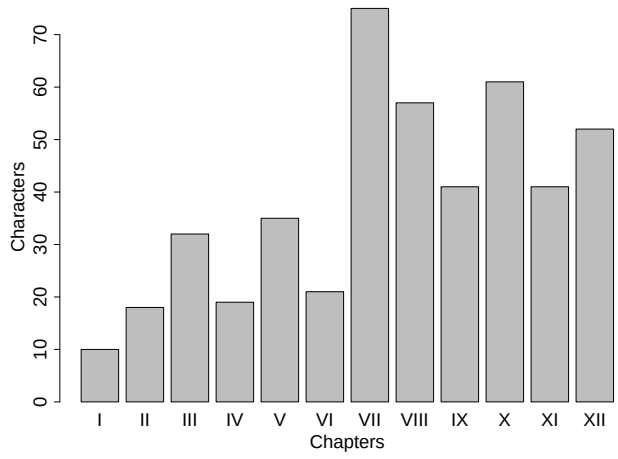


Figure 7.10: Occurrences per chapter. Chapter VII contains 34.7% of all characters of $\mathscr{G}_0$.

7.3 Vitality

Vitality is a concept measuring how some structural properties of the network depend on a node by evaluating the impact of its deletion (Koschützki et al., 2005a). Equivalently, it is a way to measure the importance of a node. This idea is present in the literary analysis works:

> L'importance du personnage pourrait se mesurer aux effet de son absence. Sans lui, comment raconter des histoires, les résumer, les juger, en parler, s'en souvenir ?[25] (Reuter, 1988)

For a function f defined at the graph level and a graph $G = (V, E)$, vitality of a node $x \in V$ is

$$c_v(x, f) = f(G \backslash \{x\}) - f(G). \tag{7.9}$$

The result can be either positive or negative. We do not recommend the use of absolute value, because the sign carries much signification. For example, deleting a hub in a highly centralised network will probably increase its diameter, while deleting one of the peripheral vertices may diminish the diameter.

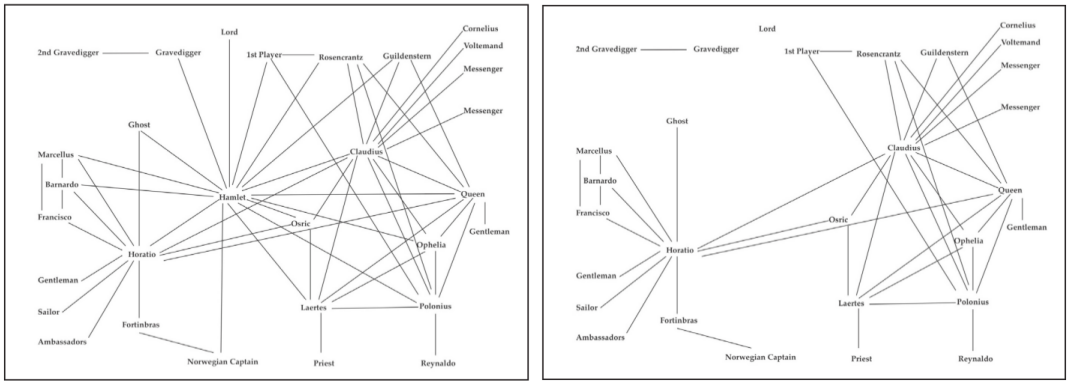


Figure 7.11: Figures from (Moretti, 2011). (Left.) Figure 10: character network of *Hamlet*. (Right.) Figure 11: character network after removing Hamlet.

To our knowledge, Franco Moretti was the first to use it in character network analysis. In a case study on *Hamlet*, he demonstrates the cohesive narrative role of a central figure in the character network of that play:

> Take the protagonist again. [...we] remove Hamlet, to see what happens [original diagrams: see figure 7.11]. And what happens is that the network almost splits in half: between the court on the right, and the region that includes the Ghost and Fortinbras on the left all that remains are the three edges linking Horatio to Claudius, Gertrude, and Osric: a few dozen words. (Moretti, 2011, p.5)

[25] "A character's importance could be measured by the impact of its absence. Without it, how do one tells, summarises, appreciates, talks about or remembers of stories ?" [Own translation.]

Franco Moretti shows the narrative importance of Hamlet by removing him. Once a character has disappeared, all the links that were incident with him disappear as well. This makes the network look scarce (see figure 7.11, right), acknowledging the key narrative role of Hamlet and confirming him, obviously, as a protagonist. He then goes deeper in the analysis by showing how vitality highlights levels in the structure of the character network. His example is based on the sum of all the distances in the network.

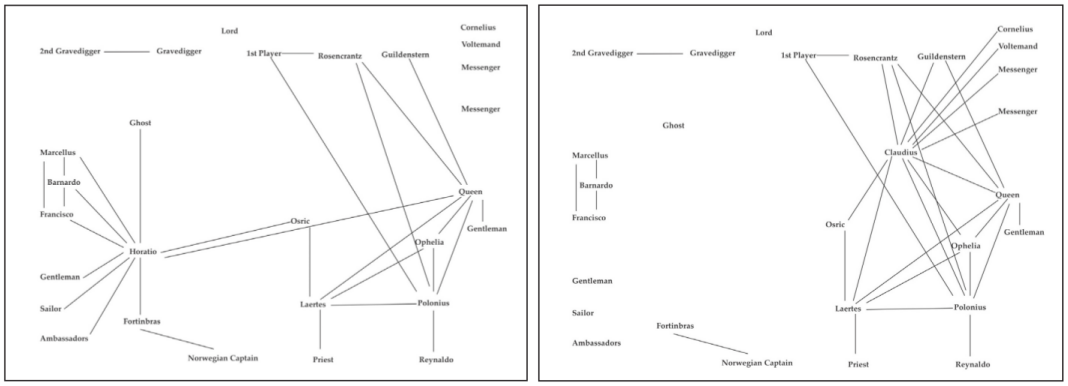


Figure 7.12: Figures from (Moretti, 2011). (Left.) Figure 18: character network after removing Hamlet and Claudius. (Right.) Figure 21: character network after removing Hamlet and Horatio.

He considers the three most central characters using closeness centrality. They are, in order, Hamlet, Claudius and Horatio. Omitting Hamlet from the network has a strong impact on its cohesion, almost creating two large connected components, but then deleting Claudius does not have a significant impact. Right next to that, he shows the interest of the vitality approach by deleting Hamlet and the less central Horatio. At this point, the network collapses. Deleting Hamlet and Claudius has no such impact, despite Claudius being more central than Horatio. In fact, it leads to the same result as when Hamlet was the only one left out. There are many potential applications: we can dualise all known centrality dimensions, or other network measures, and complete the previously obtained conclusions, or find new ones.

Vitality is not a centrality measure in itself, but an operator associated with a measure defined at the level of the network globally. For example, vitality associated with the operator of centralisation (Freeman, 1978), itself associated with measures of centrality, demonstrates how deleting a node causes a centralised structure to emerge or to disappear. Franco Moretti associates vitality and closeness centrality[26] to show that a given measure of centrality does not necessarily predict which are the most interesting central characters. In the case of vitality used with centralisation, a negative measure means that the network without the picked node is less centralised than with it, thus the node prompts an effect of centralisation. Conversely, a

[26]Technically, he divides the sum of all distances by the number of nodes in the network.

positive value signifies that the presence of the node implies decentralisation, or equivalently that its absence increases centralisation.

Returning to our corpus, we compute some vitality measures on all the nodes of the network. We use centrality and connectedness. In some cases, we show and discuss which nodes have the largest negative or positive values. The results from the computation for the four centrality measures are shown in figure 7.13. In all cases, the majority of nodes are all on one side, and close to zero (but none reach it), and a minority of nodes follow the inverted tendency, with larger results in absolute value. The case of eigenvector centrality is opposite to those associated with the other centrality indices: in this case, for most of the nodes, if we take one of them away, centralisation will slightly increase, meaning that by way of chain reaction, having any node disappear diminishes the impact of the most central nodes. In the three other cases of degree, betweenness and harmonic centrality, and for most of the nodes, centralisation slightly decreases. A brief explanation is that taking a node away increases the influence of the central nodes. In the examples of betweenness and harmonic vitalities, this is clearly related to shortest paths disappearing, therefore the redirected ones will pass through hubs more than other nodes.

Extremal characters As expected, Mme de Warens is an extremum in all cases, being the extreme positive value in degree, betweenness and harmonic cases, and the extreme negative one in the eigenvector situation. We can observe all these distant exceptions in figure 7.13. What may look surprising is that she is the most central node according to a lot of criteria, and still her absence increases centralisation in the network in the three first cases. This is caused by the fact that in the network she is an articulation node[27], therefore her own contribution to centralisation is compensated by the absence of 29 characters now disconnected from the giant component of 187 characters[28].

Dealing with the question of hubs and articulation points, vitality handles the problem of detecting nodes in such positions. We associate it with a function counting the number of connected components (which is 1 in the case of $\mathscr{G}$). For some characters, they are the inevitable gate to the total network. For example, the next articulation points with high consequences after Mme de Warens are Mme de Larnage and the comte de Montaigu, whose deletions create five connected components in each case. Rousseau met her during a journey to cure his illness and would spend a lot of time with and accompany her on legs of her own journey. The betweenness value of the comte de Montaigu is high, as shown in table 7.2.2. This

[27] A node such that if we remove it, the graph becomes disconnected.

[28] Interestingly, it is only in the eigenvector vitality case that that logic is respected, her presence being more centralising than those of Diderot (2nd) or Grimm (3rd). This is a counter-intuitive result when compared to section 7.2.4. In fact, the use of vitality with a quantitative approach is problematic when the network is not 2-vertex-connected.

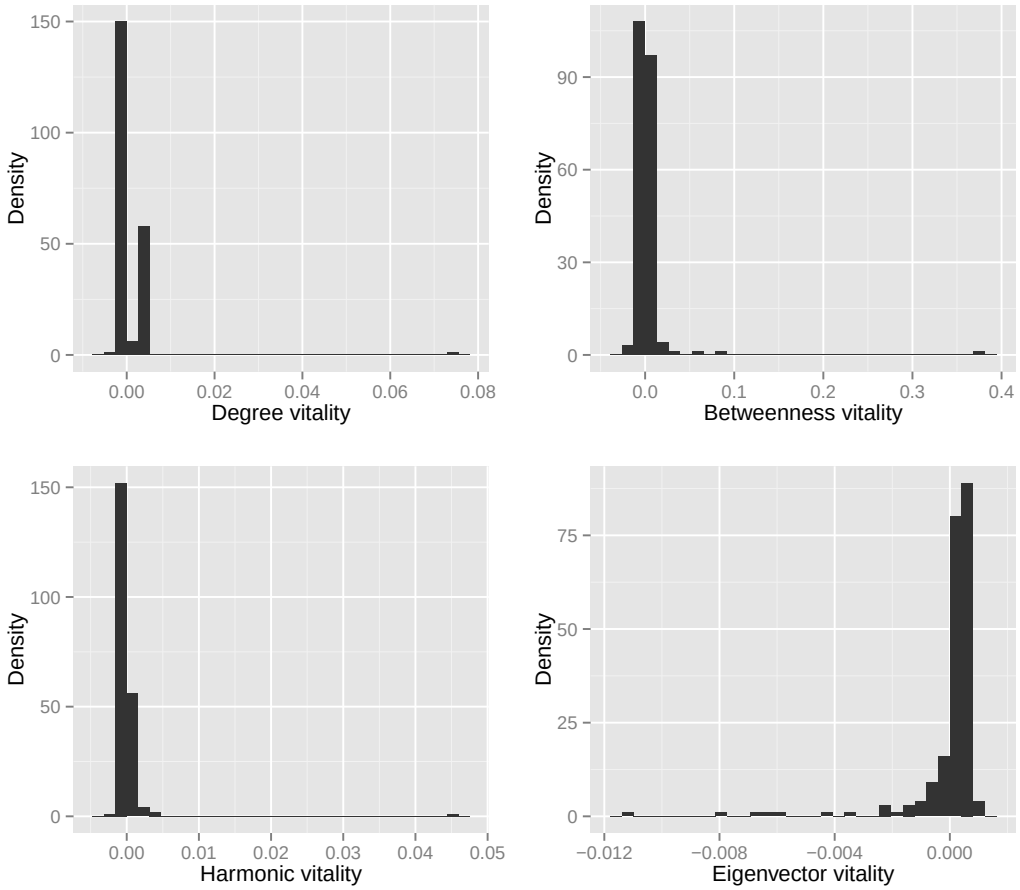


Figure 7.13: Histograms of centralisation-based vitality distributions for the four canonical centrality measures. Extremal values are rare.

confirms the role they play between communities, creating narrative paths between groups of characters, or with distant lonely characters.

The more meaningful cases are when focusing on the characters whose demise diminishes centralisation, meaning that they play a central role not necessarily detected by classic centrality values. In the degree and harmonic vitality cases, with the exception of Mme de Warens, values are too close to one another to allow interpretation. This is not the case for the two other instances. In the betweenness case, for example, Denis Diderot owns the highest negative value with -0.024, while the narratively[29] close Grimm has the sixth highest positive value, with 0.014. They have most of their neighbours in common, so what can explain this contrast? Possibly, we highlighted some internal and undetectable property of their local environment, according to dendrogram in figure 7.7 in which they belong to the same community, but only loosely. As for the case of eigenvector vitality, it is more spread out. With the exception of Mme

[29] And intermediary close, see table 7.2.

de Warens, as we have explained earlier, the other high negative cases show the same actors that had high eigenvector centrality values (see table 7.5). In this occasion, vitality scores maintain similar properties to their corresponding centrality index.

	Eigenvector vitality
1	-0.01132
2	-0.00781
3	-0.00661
4	-0.00620
5	-0.00594
6	-0.00412
7	-0.00356
8	-0.00227
9	-0.00222
10	-0.00205

Table 7.5: The characters with the ten highest eigenvector vitality values.

7.4 Discussion

The goal of this section is to retrieve, gather and analyse results from sections 7.1 to 7.3. Earlier, we considered centrality measures separately. Here, we describe characters with the help of all these measures and vitality. We also present a principal component analysis which explores the narrative dimensions induced by the four chosen centrality measures. We then describe the narrative roles of some characters based on their properties in the character network. A typology is proposed in the next—and last—section of this chapter based on the ideas developed here (section 7.5).

7.4.1 Summary

In table 7.6, we show the characters that are one of the twenty most central nodes on the whole network at least once for one of the four indices. Many characters are several times among the twenty highest, therefore the table contains only 39 characters. Some of them are significantly central according to all measures (Thérèse Levasseur, Mme de Luxembourg, Denis Diderot, Grimm, etc.), while some others get contrasted results, with some high and low values (Mme de Warens, duc de Luxembourg, George Keith, Voltaire, etc.). Eventually, some characters are central according to one of the centrality measures, significantly more so relative to the others (Mme de Larnage, M. de la Poplinière, comte de Montaigu, baron d'Holbach, etc.).

In the table, we immediately remark that the ranks are not conserved from one measure to another. Mme de Warens has the highest degree centrality but is only the 18th with eigenvector. The duc de Luxembourg has the 7th measure of degree centrality but only the 44th with harmonic centrality. Finally the comte de Montaigu and George Keith have very high betweenness values (2nd and 4th), but poor measures of eigenvector centrality (122th and

	Degree		Betweenness		Harmonic		Eigenvector	
		Rank		Rank		Rank		Rank
Mme de Warens	64	(1)	14824.2	(1)	34.76	(1)	0.18977	(18)
Thérèse Levasseur	48	(2)	1228.6	(11)	26.18	(8)	0.79273	(4)
Mme de Luxembourg	43	(3)	2962.4	(3)	29.54	(2)	0.47534	(7)
Denis Diderot	40	(4)	2166.4	(5)	27.77	(3)	0.96440	(3)
Grimm	37	(5)	1589.4	(6)	26.56	(6)	0.97783	(2)
Louise d'Épinay	36	(6)	1277.3	(8)	25.51	(11)	1.00000	(1)
Duc de Luxembourg	34	(7)	405.1	(33)	22.63	(44)	0.46873	(8)
Mme Dupin	25	(8)	1296.0	(7)	27.04	(5)	0.27725	(11)
George Keith	21	(9)	2189.3	(4)	26.23	(7)	0.08072	(31)
Comtesse d'Houdetot	20	(10.5)	1076.3	(12)	26.18	(9)	0.63721	(5)
Jean Le Rond d'Alembert	20	(10.5)	352.3	(35)	23.32	(30)	0.20969	(14)
M. et Mme Levasseur	19	(12)	770.1	(16)	23.53	(26)	0.51091	(6)
Guillaume de Lamoignon	18	(13)	462.2	(28)	24.43	(14)	0.14381	(20)
Voltaire	17	(14.5)	1251.5	(10)	27.75	(4)	0.11745	(24)
Comtesse de Boufflers	17	(14.5)	125.8	(68)	22.67	(43)	0.20304	(15)
CL Dupin de Francueil	16	(16.5)	278.9	(37)	23.69	(21)	0.22353	(13)
Comte de Montaigu	16	(16.5)	3280.9	(2)	23.64	(23)	0.00456	(122)
LAJ de Rochechouart	15	(19)	177.7	(64)	22.83	(38)	0.19871	(16)
Charles Pinot Duclos	15	(19)	513.6	(24)	22.53	(45)	0.24620	(12)
NB Duchesne	15	(19)	496.8	(25)	24.26	(15)	0.12633	(23)
Comte de Friesen	14	(21)	62.8	(81)	21.95	(52)	0.19644	(17)
Baron d'Holbach	13	(22.5)	46.8	(87)	20.68	(74)	0.42611	(9)
Marquise de Verdelin	13	(22.5)	472.8	(26)	24.90	(12)	0.05180	(46)
VC de Gauffecourt	12	(24)	249.7	(39)	23.99	(18)	0.14674	(19)
Daniel Roguin	11	(25.5)	698.9	(17)	23.14	(34)	0.06712	(36)
Jean Bonnot de Mably	9	(28)	930.5	(14)	24.78	(13)	0.03291	(56)
Saint-Lambert	9	(28)	13.2	(102)	18.47	(128)	0.38946	(10)
M. de La Poplinière	8	(36)	135.6	(66)	25.73	(10)	0.05287	(44)
JA de Condillac	8	(36)	237.3	(41)	24.06	(17)	0.05499	(42)
François Coindet	8	(36)	118.2	(69)	23.79	(19)	0.05327	(43)
Mme de Larnage	7	(48.5)	1266.0	(9)	16.90	(160)	0.00701	(105)
CP de Montenoy	7	(48.5)	201.2	(59)	24.25	(16)	0.05667	(41)
Duc de Richelieu	7	(48.5)	349.6	(36)	23.69	(20)	0.02358	(64)
Claude Anet	6	(59)	33.5	(90)	17.34	(150)	0.02782	(62)
CE La Roque	5	(69.5)	1031.3	(13)	23.51	(27)	0.00300	(142)
Abbé de Gouvon	5	(69.5)	813.0	(15)	21.67	(54)	0.00423	(127)
KE von Graffenried	4	(82)	636.0	(19)	20.26	(88)	0.02236	(67)
Claudine de Galley	4	(82)	622.7	(20)	23.46	(28)	0.00321	(139)
Gabriel Bernard	3	(103)	638.0	(18)	20.85	(66)	0.00400	(129)

Table 7.6: Centrality measures of the twenty most central nodes according to each index. They are reunited in a single table and ordered according to their degrees. Between brackets: their ranks related to the corresponding index.

31st). This is a first step towards the confirmation that in an analysis of character networks, and by extension of literary networks, the four chosen centrality measures reveal different narrative properties.

The network in figure 7.14 shows the apparent disparity between the structural properties these four classic centrality indexes measure. The example is adapted from (Brandes et al., 2012), where they pretend that

> [Diese] Beispiel [...] zeigt den kleinsten bekannten Graphen, in dem der jeweils zentralste Knoten bezüglich der vier vorgestellten Masse ein anderer ist.[30]

Correlations In order to get an overview of how correlated the measures are to one another, we compute the Spearman correlation coefficient[31] on all of them, since our goal is to compare, or highlight which are the most or least central nodes (see figure 7.15). The values of centrality matter less than the rankings, e.g. betweenness and eigenvector centrality values of comte de Montaigu. The resulting matrix of paired distributions is useful to detect outliers, like in table 7.6.

Logarithmic scales are necessary in the plots because of the long-tail shapes of most of the distributions. Correlation coefficients are there to help figure out these results. At first, we see that all coefficients imply rather significant correlations. This was expected, as even if a node is central according to one measure and not another, most of the nodes bear similar properties, and equalise the correlation towards a positive value.

Nonetheless, there are different levels of correlation, and, as expected from some previous examples like the centrality measures of comte de Montaigu, Mme de Larnage or George Keith, betweenness and eigenvector centrality share the smallest correlation coefficient. They measure different types of importance, different dimensions of influence emerging from the node positions. Betweenness and harmonic centrality are also less correlated than degree and betweenness. Betweenness gives importance to nodes at the border of different communities, while harmonic centrality gives more importance to nodes at the centre of populated communities. Computed correlations involving degree centrality trivially indicate that increasing the number of incident links to a node also increases its centrality under any form.

Covariances In order to explore the relations between the results of our centrality distribution, we perform a principal component analysis on the four indices. Principal component analysis seeks a linear transformation so that the first axis of the new basis is the one minimising the sum of the orthogonal projections. Then, the second axis follows the same constraints, but in a hyperplane orthogonal to the first axis, etc. The resolution requires us to compute the eigenvectors of the covariance matrix. Their order is found as the part of variance they describe decreases.

[30] "This example shows the smallest known graph in which the node with highest centrality is different for each of the four centrality indexes presented."

[31] For two distributions, their Spearman correlation coefficient is the result of the computation of the Pearson correlation coefficient — the classic correlation coefficient — on their ranks instead of their values.

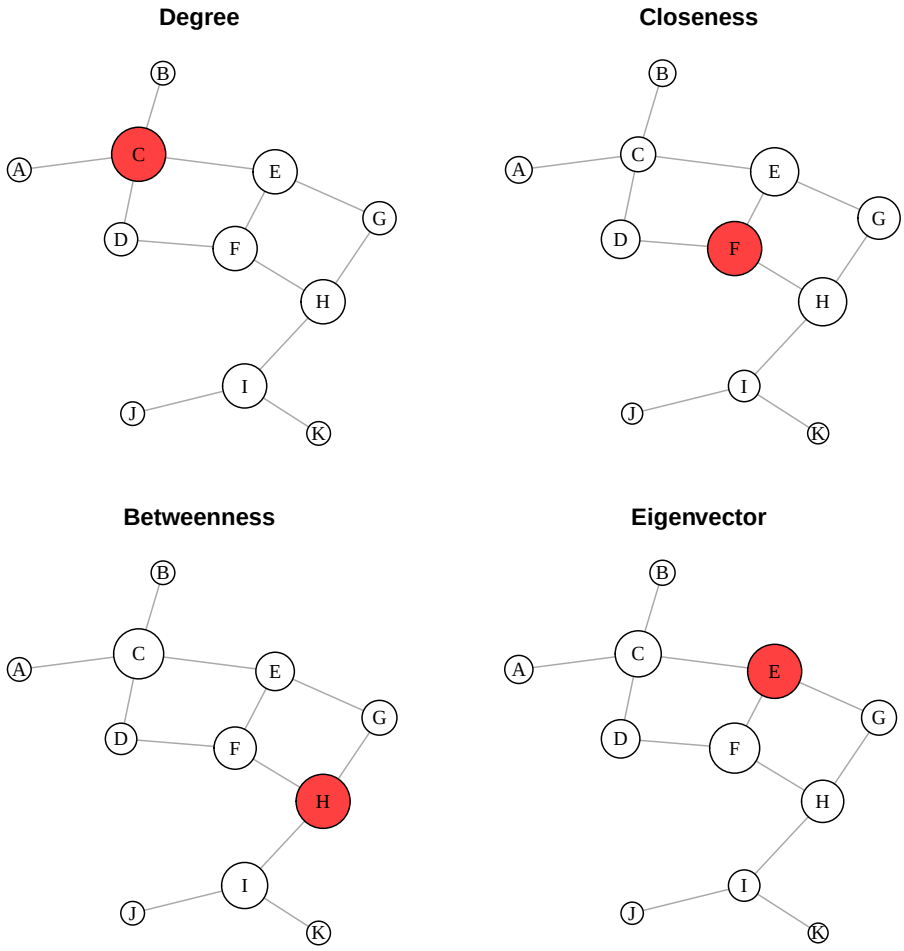


Figure 7.14: In this graph, the actors obtaining the highest centrality scores are different each time, depending on the index chosen. Size of nodes is proportionally normalised on the centrality scores. (**Degree**) Node C has the highest degree centrality, with degree equal to 4 since it is connected to nodes A, B, D, and E. (**Closeness**) Node F has the highest closeness centrality. The sum of the distances from that node to all other nodes is 21. For comparison, in the case of node K, the sum is 36. (**Betweenness**) Node H has the highest betweenness centrality. It appears on $\frac{21.83}{45} = 48.5\%$ of geodesics, with respect to the quantity of geodesics between each couple of nodes. (**Eigenvector**) Node E has the highest eigenvector centrality, since its corresponding value in the eigenvector is 1.

In our case, we keep the two first components out of four. The first one is a composition of the four centrality indices and explains 67.72% of the variance ($s_d = 1.646$). The second one opposes the betweenness and the eigenvector centrality, and it accounts for 18.16% of the variance ($s_d = 0.852$). Table 7.7 shows the contribution of each index to the components.

[32] "Small loadings are conventionally not printed (replaced by spaces), to draw the eye to the pattern of the larger loadings." (R Core Team, 2013, *stats::loadings*) (Manual of package *stats* (version 3.0.2) accessed on 02/03/2014.)

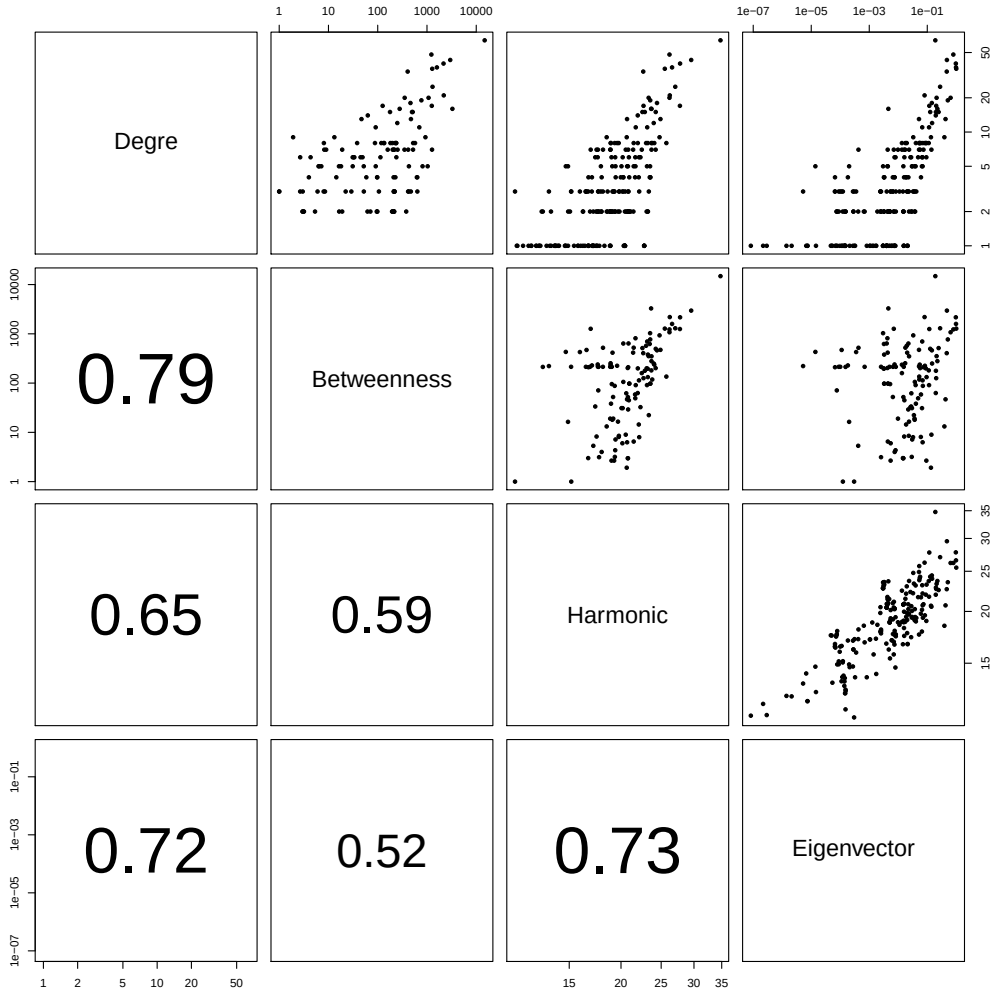


Figure 7.15: Scatterplot matrix of centrality measures. Scales are logarithmic. Correlations (Spearman) are shown in the lower panel.

	Comp.1	Comp.2	Comp.3	Comp.4
Degree	-0.59		-0.23	0.78
Betweenness	-0.44	-0.75	-0.31	-0.38
Harmonic	-0.48		0.87	-0.11
Eigenvector	-0.48	0.66	-0.30	-0.49

Table 7.7: Loadings[32] of the centrality measures PCA.

Figure 7.16 shows the placement of the characters in accordance with the two first components. Vertically, i.e. along the betweenness-eigenvector axis, we see how Mme de Warens, the comte de Montaigu, George Keith and Mme de Larnage, who have high betweenness and low eigenvector centrality, are opposed to the likes of Louise d'Épinay, Grimm, Denis Diderot and

Thérèse Levasseur, who have low betweenness but high eigenvector centrality. Between them, and at the edge of the crowd lies Mme de Luxembourg.

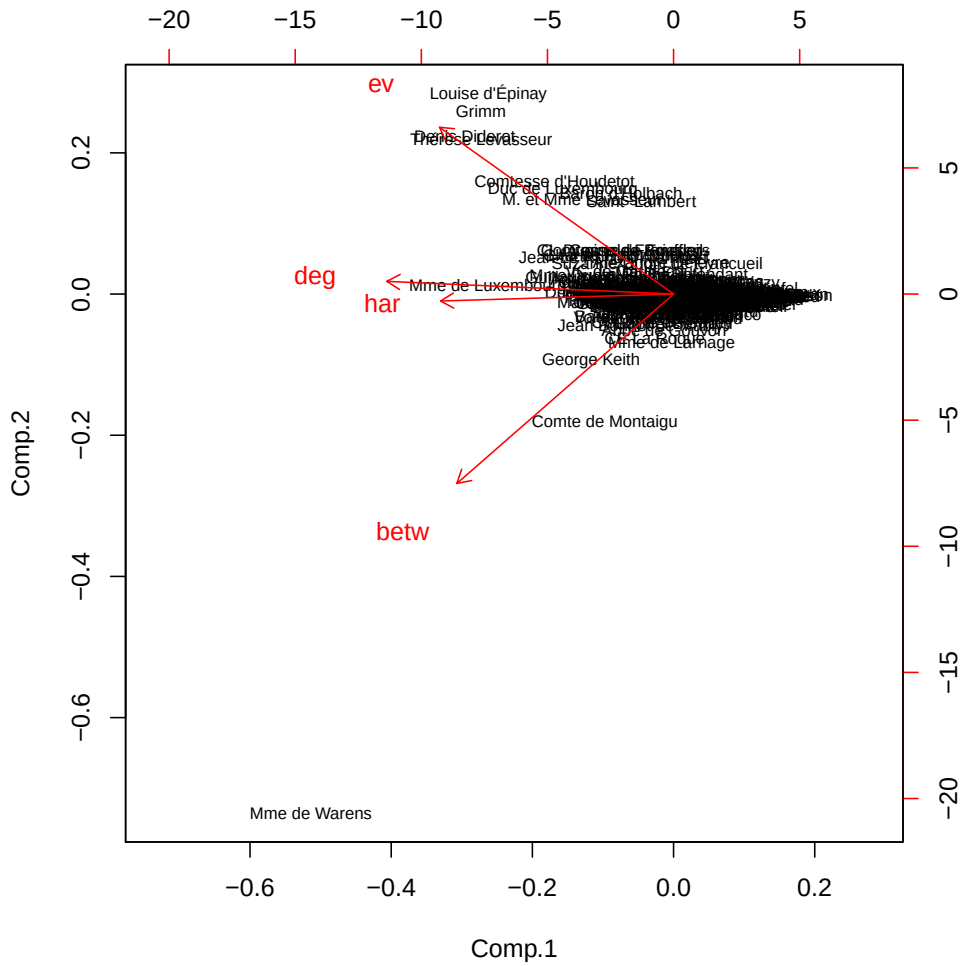


Figure 7.16: Principal component analysis of the four dimensions represented by degree, betweenness, harmonic and eigenvector centrality measures. Normalised betweenness gives the same result, with Mme de Warens sent very far away.

On this aspect of centrality, our work of character network analysis is coherent with the results related to social network analysis. John Bolland, in a study where he compared degree, betweenness, closeness and eigenvector centrality, concluded that[33]

> Additionally, these results suggest c_B as a useful companion to c_E. The assumptions underlying the two models are quite different: one is based on the ability

[33] Original symbols C_B and C_{CF} (CF for "continuing flow") are hereby replaced by ours.

> to facilitate network flow, while the other is based on the ability to hinder flow. Empirically, they are the least correlated of the models, and the strengths of each complement the weaknesses of the other. Used together, they should yield a useful description of network centrality. (Bolland, 1988, p.252)

Therefore, we found that all centrality measures house unique properties despite being somehow correlated. In particular, betweenness and eigenvector centrality highlights almost disjointed structural attributes.

Following up this centrality analysis on our character network of *Les Confessions*, in the coming subsections we describe the narrative roles of a few characters.

7.4.2 Mme de Warens

Despite being mentioned in five of the six chapters in the second part, Mme de Warens occurs in 87% of the cases in the first part of *Les Confessions* (see appendix C). Her narrative role is the one of a central figure, which is proved by her high degree and harmonic centrality, and an overwhelming measure of betweenness, as shown in figure 7.16. Vitality also demonstrated her key position as an articulation node: in terms of narration, she is essential, appearing together with more than half of the characters in the first part, many of them being connected to the rest of the network only through her. Her positioning is shown in figure 7.17: even if the theory advises us not to draw conclusions from network visualisation — and this one is based on a heuristic layout algorithm —, in this case it clearly appears that she is central based on most of the definitions we can imagine for that term.

However, there is one exception in her centrality portrait: in spite of being connected to some of the influential characters in the second part and to many characters throughout the whole network, her eigenvector centrality score is low in comparison to the other scores. Her absence in the most socially dense periods of Rousseau's life leaves her aside according to this particular criteria. Her intermediary role between Rousseau and some small disjointed societies (his first journey in Italy, Annecy, Lyon, etc.), and his trips away from her, followed by his returns, create this constellation of small groups which she dominates in the network, and which have little influence at the level of the whole network.

7.4.3 M. et Mme de Luxembourg

Husband and wife, the duc and Mme de Luxembourg hosted Rousseau after he left the house that Louise d'Épinay was putting at his disposal. They are two of the most central nodes, obviously often cited together (see appendix C). They appear as such in the network (see figure 7.18) and their close eigenvector centrality values reflects that (see table 7.6). Their central position is partly due to the fact that they appear along with other central nodes[34]. They are

[34] For example, intensities of edges between :
Mme de Luxembourg & comtesse de Boufflers : 21.

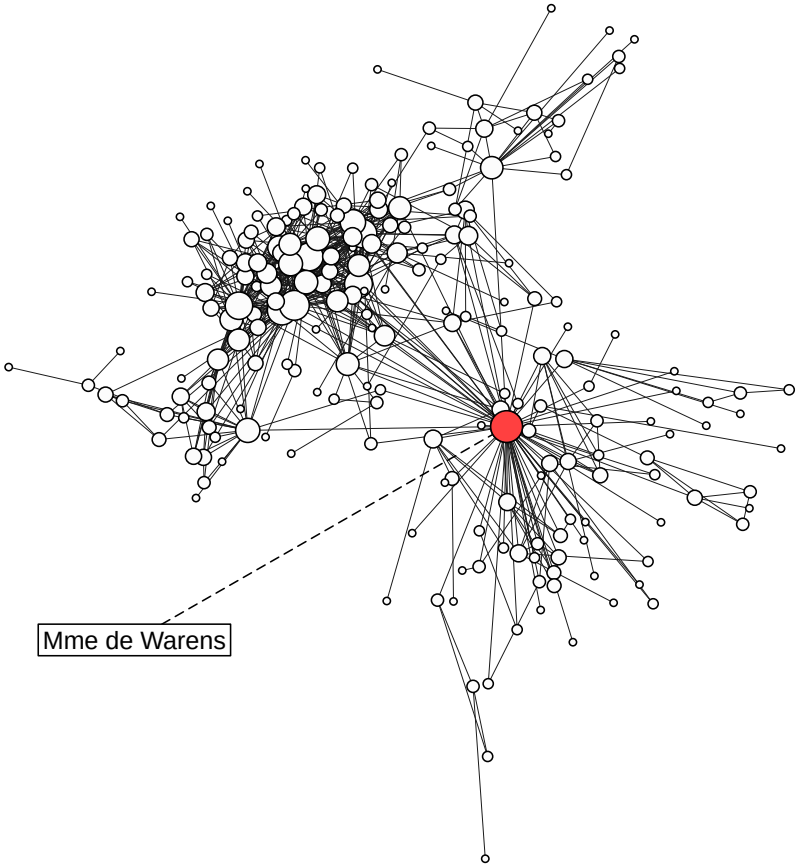


Figure 7.17: Network $\mathcal{G}$ with Mme de Warens in evidence.

well connected and omnipresent in chapter XI, which is populated above the average (see figure 7.10). This situation shows how the continuous presence of a character in the narration causes it to gain relations and thus increase its centrality.

However, in this case, there is a split between them in terms of betweenness and harmonic centrality. When two characters are close one to another, one generally gets the upper hand over the other via a more strategic positioning on shortest paths. This accounts for the high betweenness value of Mme de Luxembourg: she "steals the flow". She does that by being connected to Mme de Warens, whereas her husband is not. The intensity of her edge incident to Mme de Warens is equal to three, the minimum weight required. They are co-occurrent on

Mme de Luxembourg & Thérèse Levasseur : 20.
Mme de Luxembourg & d'Alembert : 13.

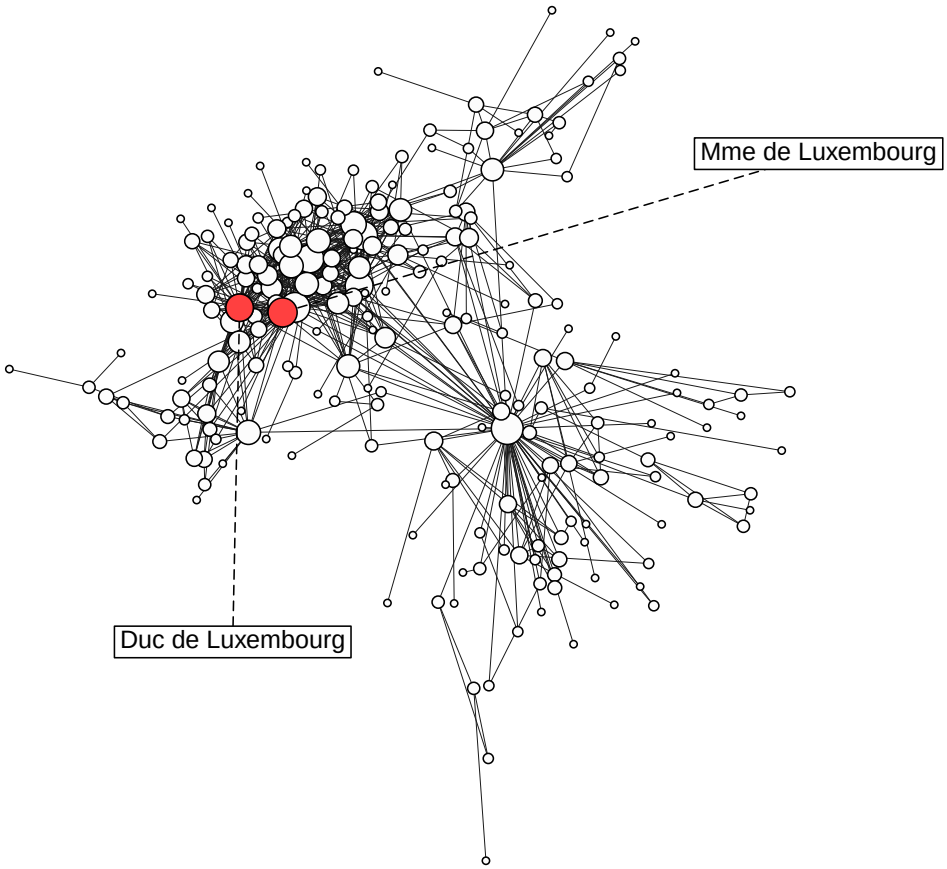


Figure 7.18: Network $\mathscr{G}$ with M. and Mme de Luxembourg.

page 277 and weakly co-occurrent on pages 449-450. On the first co-occurrence, Rousseau tells an anecdote about a small cabin:

> [...] je passais une partie de mon temps à l'orner et à y préparer à Maman quelque surprise agréable lorsqu'elle s'y venait promener. Je la quittais pour venir m'occuper d'elle, pour y penser avec plus de plaisir ; autre caprice que je n'excuse ni n'explique, mais que j'avoue que la chose était ainsi. Je me souviens qu'une fois Mme de Luxembourg me parlait en raillant d'un homme qui quittait sa maîtresse pour lui écrire. Je lui dis que j'aurais bien été cet homme-là, et j'aurais pu ajouter que je l'avais été quelque fois.[35]

[35] "[...] spending part of my time in ornamenting it during the absence of Madam de Warrens, that I might surprise her the more agreeably on her return. Sometimes I quitted this dear friend, that I might enjoy the

This co-occurrence reflects their narrative proximity in this anecdote. There is a facet of his personality and life story that is echoed in his friendship with Mme de Luxembourg. The weak co-occurrence relies on his meeting with Thérèse Levasseur. She is "un successeur à Maman" on page 449. He then describes her role in his life and her personality on page 450 and explains that he made a

> dictionnaire de ses phrases pour amuser Mme de Luxembourg, et ses quiproquos sont devenus célèbres dans les sociétés où j'ai vécu.[36]

Pages 449-450 contain a unique narrative event, which focuses on their encounter. This explains how Mme de Luxembourg has such high betweenness and harmonic centrality. Indeed, an important number of shortest paths joining characters in the first and second parts go through Mme de Warens and then through her. Moreover her harmonic centrality benefits from the massive neighbourhood of Mme de Warens. To conclude, this narrative relation with her rather than with the duc de Luxembourg is not coincidence: she has nine more neighbours than he has.

7.4.4 Thérèse Levasseur, comtesse d'Houdetot

Thérèse Levasseur was Rousseau's partner in the second part of *Les Confessions*, while the comtesse d'Houdetot is a character he fell in love with during that period. The later often occurs in chapters IX (see appendix C). They bear very similar centrality rankings (see table 7.6), with the exception of degree, which is high in both cases. We show their central placement in figure 7.19. In particular, the comtesse d'Houdetot is a character with many occurrences (45) in comparison to degree (20)[37].

Thérèse Levasseur and the comtesse d'Houdetot symbolise both sides of Rousseau's love life: the loyal companion on one side, and the platonic and intense but unattainable love on the other. In the character network, they obtain very close measures, implying similar narrative relational properties. This does not mean that they are interchangeable. On the contrary, Rousseau places them together in his narration, so they share the same environment and supplement one another. Rousseau gives them similar importance in the character-system of *Les Confessions*.

---

uninterrupted pleasure of thinking on her; this was a caprice I can neither excuse nor fully explain, I only know this really was the case, and therefore I avow it. I remember Madam de Luxembourg told me one day in raillery, of a man who used to leave his mistress that he might enjoy the satisfaction of writing to her; I answered, I could have been this man; I might have added, That I had done the very same."

Translation by S. W. Orson, Privately Printed for the Members of the Aldus Society, London, 1903, see http://en.wikisource.org/wiki/Confessions_(Rousseau), accessed on 15/07/2014.

[36] "I formerly made a dictionary of her phrases, to amuse M. [!] de Luxembourg, and her 'qui pro quos' often became celebrated among those with whom I was most intimate."

Translation by S. W. Orson, Privately Printed for the Members of the Aldus Society, London, 1903, see http://en.wikisource.org/wiki/Confessions_(Rousseau), accessed on 15/07/2014.

[37] She is an outlier on figure 7.2.

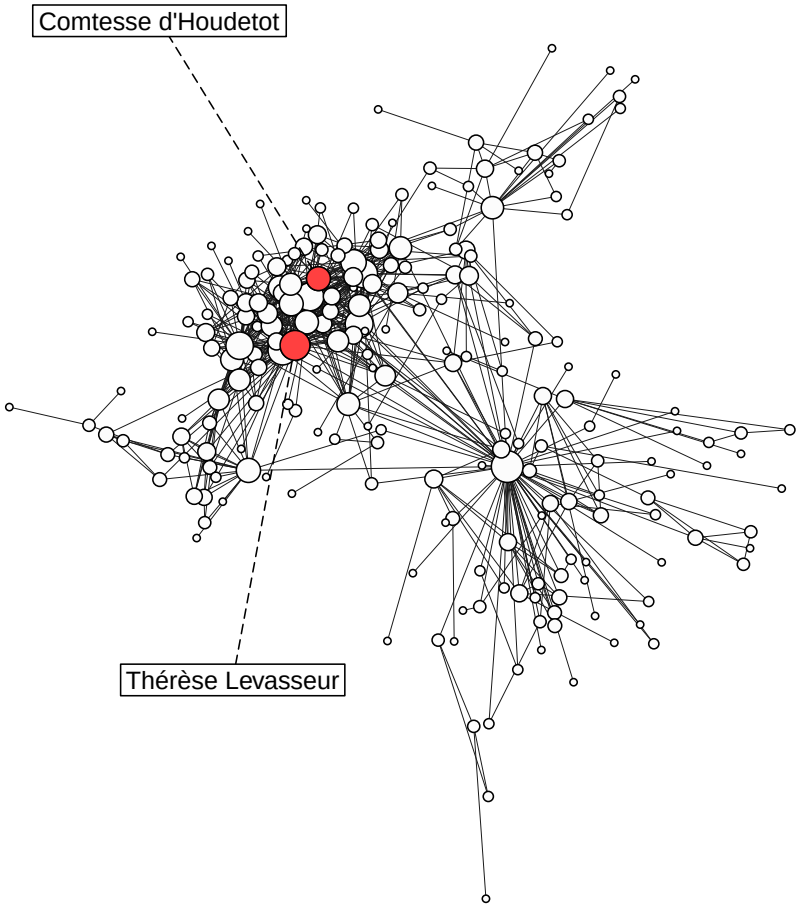


Figure 7.19: Network $\mathscr{G}$ with two characters having high degree centrality in contrast with other centrality measures.

7.4.5 Comte de Montaigu, George Keith, Mme de Larnage

These characters are the gates between clusters of characters. They play key roles as narrative intermediaries in the detection of communities (see section 7.2.2.3). In figure 7.20, we show characters that have high betweenness ranks in comparison with the other centrality measures. We can clearly see their strategic positions in sparse areas between clusters, which causes the reduction of their harmonic and eigenvector centrality.

7.4.6 The Plot against Rousseau

As we have seen, centrality rankings show the domination of Mme de Warens on important parts of the narrative. This was expected, since she is the character with the most occurrences.

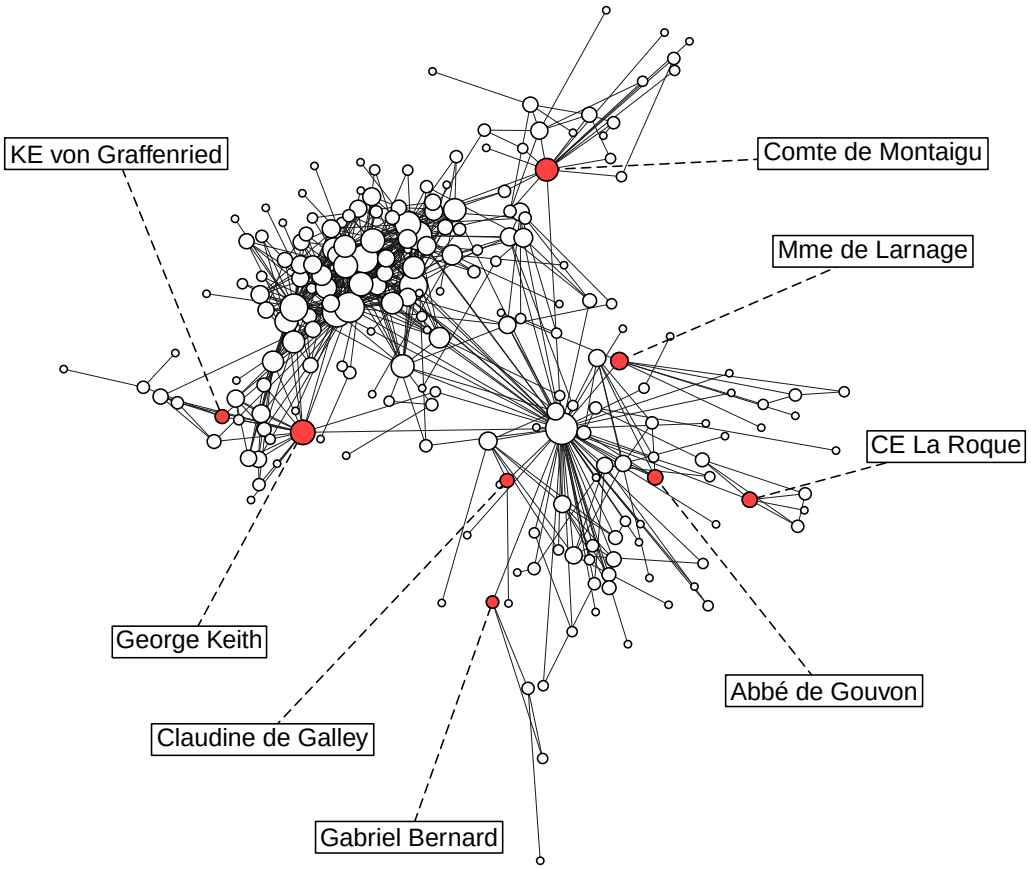


Figure 7.20: Network $\mathcal{G}$ with some characters having high betweenness centrality contrasting to other centrality measures.

The central characters of the second part are all situated after Mme de Warens in the rankings (see table 7.6), with the notable exception of eigenvector centrality. A cluster of otherwise quite central nodes would significantly increase that coefficient, and this is the case here. Eigenvector centrality clearly identifies a group of central nodes, which happen to be, for some of them, the most important characters in the plot fomented against Rousseau in his suspicions. The *schemers* behind the plot were attacking his writings and his public image. We can see in the network analysis how Rousseau transcribed that in his narration.

Among nodes with the highest eigenvector centrality measures, we exclude obvious characters like Thérèse Levasseur, who was his companion and later his wife, the comtesse d'Houdetot, with whom Rousseau was in love (a fact that the *schemers* used to attack him),

or the couple Luxembourg, who were his friends at the time and helped him escape France to Neuchâtel. We find a cluster of characters with shared structural properties, although not necessarily individually central under the other centrality measures. This group[38] is composed of Louise d'Épinay (1st), Friedrich Melchior Grimm (2nd), Denis Diderot (3rd), the baron d'Holbach (9th) Jean Le Rond d'Alembert (14th) and the comtesse de Boufflers (15th).

In the narration, they constitute a meta-character made up of interchangeable figures: a group of characters sharing enough narrative similarities to be perceived and detected as one narrative element. We justify[39] this metaphor by the fact that deleting one of them leaves the structural properties of the group intact. Despite their high eigenvalue centrality, none of them is truly "central". Characters recurrently associated one with another are, in practice, interchangeable from a narrative point of view. They act like a crowd, and their character-spaces have similar properties. This lack of individuality is correlated with fear and suspicion, as Rousseau identifies them on several occasions as faceless enemies. This assumption is verified by the analysis of the network with a contraction of these characters into a single node. In this operation[40], we replace the six corresponding vertices with a unique vertex gathering all the connections they had individually (see figure 7.21). We call this network $\mathscr{G}_s$. In table 7.8, we show how the centrality measures behave when we combine multiple character-spaces to a single one.

	Degree	Rank	Betw.	Rank	Harm.	Rank	Eigenv.	Rank
Schemers	74	(1)	5910.5	(2)	0.137	(2)	1.000	(1)
Mme de Warens	62	(2)	13631.6	(1)	0.156	(1)	0.212	(12)
Thérèse Levasseur	44	(3)	1534.9	(6)	0.121	(5)	0.778	(2)
Mme de Luxembourg	40	(4)	3249.5	(3)	0.136	(3)	0.683	(3)
Duc de Luxembourg	31	(5)	432.6	(27)	0.106	(29)	0.667	(4)
Comte de Montaigu	16	(11.5)	2924.9	(4)	0.111	(18)	0.007	(121)
George Keith	21	(7)	2049.9	(5)	0.119	(8)	0.118	(25)
Comtesse d'Houdetot	17	(9)	1160.3	(9)	0.121	(6)	0.494	(5)
Mme Dupin	22	(6)	1353.7	(7)	0.124	(4)	0.285	(9)

Table 7.8: Measures and ranks of the most central nodes according to centrality. *Schemers* is the contraction of six characters to one node.

[38] The number in brackets is the eigenvector centrality rank.

[39] Woloch takes a similar step in his study of Jane Austen's *Pride and Prejudice*:

> "Ironic and omniscient, the narrator configures the five sisters' central situation in to an asymmetrical structure, developing Elizabeth Bennet into a strong, central protagonist and making Lydia, Catherine, Mary, and Jane into different kinds of minor characters. Elizabeth's centrality emerges only in dynamic interaction with the development of these (and other) minor characters, so that the narrative price of her achieved interiority is the distortion of many other human figures." (Woloch, 2003, p. 34)

[40] The nodes are not simply replaced by a single one, but the network extraction is done again from the list of occurrences. This prevents us from drawing hasty conclusions caused by artefacts at the level of the network extraction.

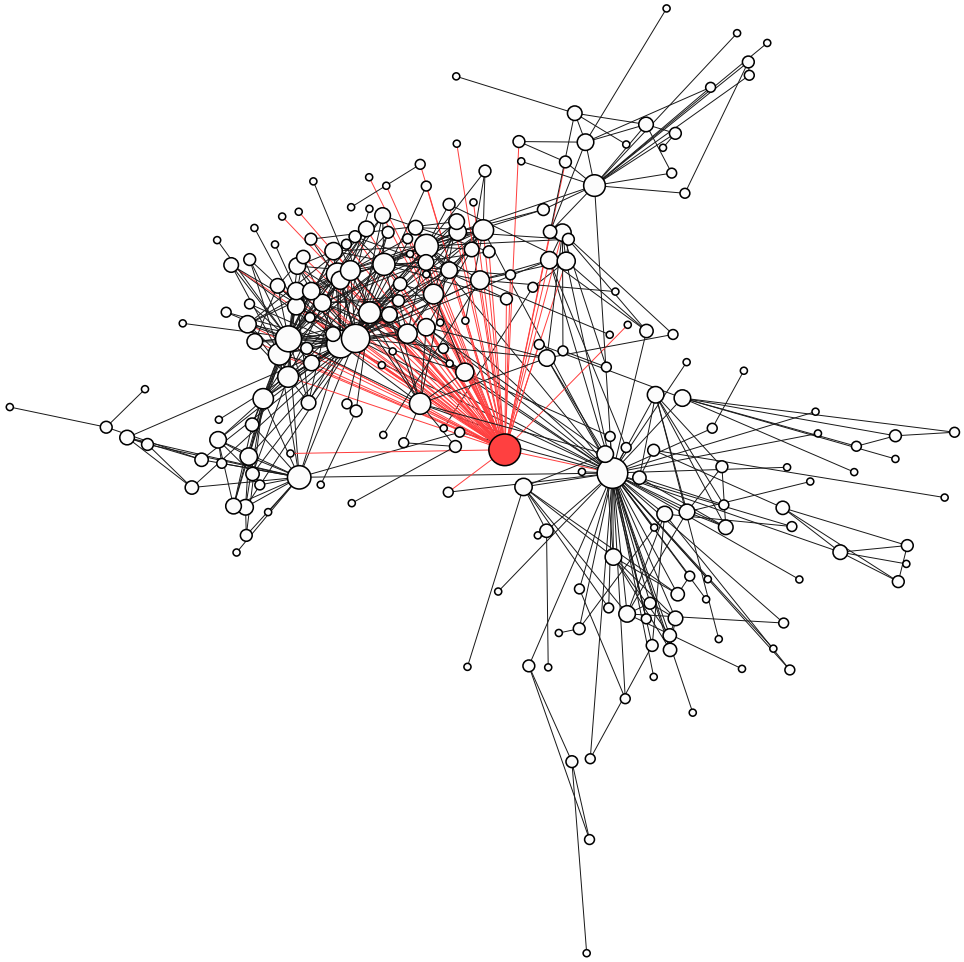


Figure 7.21: Network $\mathscr{G}_s$, highlighting the *schemers* (in red).

The resulting measures for *schemers* and Mme de Warens are close, with the exception of eigenvector centrality. They have the same narrative importance in the second part that she had in the first. The first part of *Les Confessions* tells the story of life through learning and discovering the world, gaining experience and developing an adult persona. This is symbolised by the central figure of Mme de Warens. The second part is more complicated. Rousseau had bad experiences as price of success, for instance in Venice or in Paris, leading him to withdraw from a restless social life and settle in the countryside. The *schemers* meta-character is the shadowy reflection of Mme de Warens in the second part. It is the expression of Rousseau's

unease and paranoia in terms of network language and character-spaces. The ubiquitous position of the *schemers* in one half of the network is shown in figure 7.22.[41]

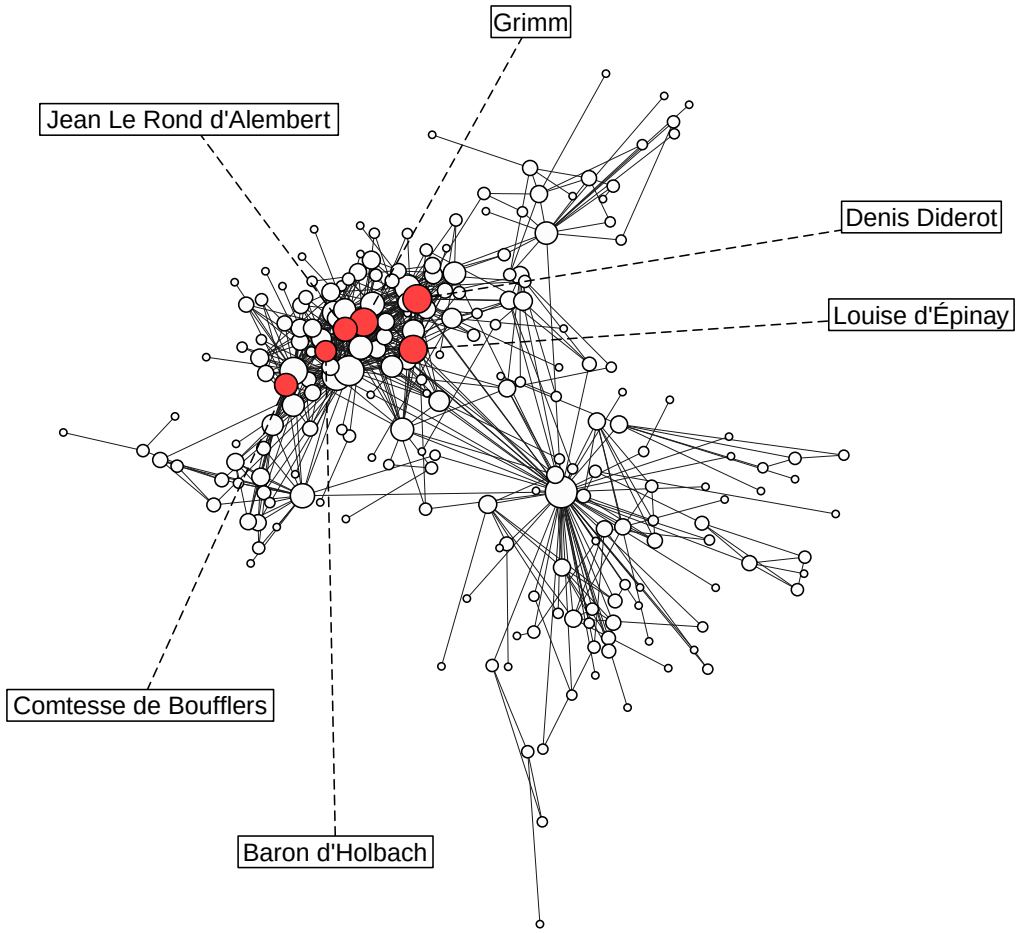


Figure 7.22: Plot of network $\mathscr{G}$ with characters having high eigenvector centrality and that Rousseau incidentally suspected to plot against him.

7.5 Character-system

In the conclusion of this chapter, we interpret the results to illustrate the narrative role of the characters on the level of the whole plot. We quantify the "configurations" from Woloch's theory:

[41] These ideas and results in this section appeared previously in (Rochat and Kaplan, 2014). An online version is available here: http://www.txdhc.org/character-networks-and-centrality/, accessed on 20/11/2014.

> If any character-system contains a constellation of intersecting and simultaneously unfolding character-spaces (minorness, after all, is partially developed *as* the attention rests on other figures), the character-system itself unfolds against the horizon of other possible configurations—not just configurations hypothetically of implicitly suggested within the narrative but achieved configurations forming the intertextual and generic conventions that any particular structural presentation of character is set against. (Woloch, 2003, p. 36)

We describe a typology of characters solely based on centrality analysis, or other derived structural methods, and not on close reading of the text. We follow Alex Woloch's steps by considering *protagonists*[42] and *minor characters*[43]. We then divide the "minor" category into *great minor characters*[44] and *minor minor characters*[45], following his proposed expansion. We have previously shown that some characters fall into the "minor" category because of a lack of occurrences or long-term presence (see section 7.2.3); nonetheless, some of them appear in key moments, play central roles over a short duration, or put Rousseau in contact with characters central to the narration. We call *great minor characters* those that fall into this category. They are characters that have high structural importance but in some way have not been brought forward by the narration.

We develop the ideas behind these concepts and show how they are illustrated by structural properties. The typology we present here is purely based on the network features; the characterisation is therefore about the structural component of their narrative roles. In each case, we illustrate the category with examples, some of them taken from other sections of the chapter in which we were doing the "spadework" for the subject. We must keep in mind that the characterisation done here is on an autobiography which, on top of many conceptual differences, is not structured like any work of fiction. In this case, the discourse is mostly chronological; this is not always true for autobiography, but in proportion it is more often the case than in fiction, due to its goal.

7.5.1 Protagonists

In this case, the chronological property borne by the narration implies that *Les Confessions*' protagonists and minor characters are not necessarily as such because of the same quantified properties, but rather for specific and non-exclusive reasons. For example, Mme de Luxembourg is cited in a concise part of the text—chapters X and XI[46]—; however, she is clearly a

[42] "[...] the protagonist, whose identity rests on a narrative centrality that always threatens to take the form of wrath (erasing or absorbing all the other persons who surround him) [...]" (Woloch, 2003, p. 7).

[43] "[...] the minor characters who, simply through their subordinated multiplicity, hover vulnerably on the borderline between name and number." (Woloch, 2003, p. 7).

[44] We called them *major minor characters* before discovering the concept had already been defined in (Woloch, 2003, p. 367, n.34)—for the best.

[45] (Woloch, 2003, p. 117).

[46] See appendix C.

central character as evidenced by high quantities for each centrality index, in addition to a high number of occurrences. A protagonist in this particular autobiography—and probably in most of them—must be seen as a protagonist in a significant story arc: in this case, the comtesse d'Houdetot plays this role starting with Rousseau's falling out with Louise d'Épinay, Grimm and Diderot, to his departure from France. She hosted him and helped during that time—they met on a regular basis and she was part of his life. The other characters in this case, being *local* protagonists, would be: comtesse d'Houdetot, the duc de Luxembourg, Louise d'Épinay.

Another type of protagonist, Thérèse Levasseur occurs frequently in more than one or two chapters. She is a recurring character, like Voltaire or Venture de Villeneuve, but unlike them, she is attributed high centrality values which reflect a more central narrative role. She is not an omnipresent character over a short period of time, but rather a discreet and constant one, often situated in key positions. This defines the *macro protagonist*, a category to which characters such as Diderot and Grimm belong.

Mme de Warens is clearly a dominant protagonist over all others, being the only centre of the structure in the first part, thus obtaining high local values and an extended recurring presence.

7.5.2 Minor characters

Conventionally, characters that are not protagonists are obviously minor characters; these are complementary notions. We have mentioned subcategories of protagonists; and here we propose similar subcategories for minor characters. Rather than having equally important characters quantified by different methods, we use a more classic hierarchy for minor characters. The minor characters of Woloch are extracted from fiction, and as such they clearly play roles directly defined by story and narration. In our case, it is the story—the real world story—which primarily defines their roles; before the narrator describes them and inserts them into a situation.

We follow Woloch on the definition of *minor minor characters*. In our case, they can trivially be considered as "out of network" characters. Our analysis is based on relations, and the results must be considered in this scope. Thus, *minor minor characters* in our framework are characters bearing no relation to any other character intense enough that it is retained in the final network. This concerns 357 characters if we consider network $\mathscr{G}_3$, and 367 if we consider the connected network $\mathscr{G}$.

The detection of *great minor characters* necessitates an aforementioned adaptation of the context of both the autobiography and a relational framework. It can be argued that some of them are protagonists, in the sense that at one moment in the story[47] they are the most

[47] For example a chapter or a group of successive pages concerning a place, such as Venice at chapter VII, and Neuchâtel at chapter XII.

important character along with Rousseau. This does not concur with Woloch's words about the *great minor characters*: they are "unlikely to survive [...] in the reader's affective memory of the novel". This is where both autobiography and network analysis provoke a bifurcation with his work. In character networks, *great minor characters* will be minor characters with outlying centrality. They are characters perceived differently in the narration, being prominent at one point or playing a non-anecdotic narrative role. The comte de Montaigu and George Keith are such *great minor characters* in *Les Confessions*, with key roles as narrative intermediaries. The members of the plot as highlighted in section 7.4.6, and after the exclusion of the ones considered as protagonists, are *great minor characters*, having influential narrative roles. This concerns the baron d'Holbach, Jean Le Rond d'Alembert and the comtesse de Boufflers.

8 Conclusion

"Diderot a fait de grands compliments à Richardson sur la prodigieuse variété de ses tableaux et sur la multitude de ses personnages. Richardson a, en effet, le mérite de les avoir tous bien caractérisés : mais, quant à leur nombre, il a cela de commun avec les plus insipides romanciers, qui suppléent à la stérilité de leurs idées à force de personnages et d'aventures."

Les Confessions, Livre XI
JEAN-JACQUES ROUSSEAU

J'ai dit la vérité. Si quelqu'un sait des choses contraires à ce que je viens d'exposer, fussent-elles mille fois prouvées, il sait des mensonges et des impostures [...] quiconque, même sans avoir lu mes écrits, examinera par ses propres yeux mon naturel, mon caractère, mes moeurs, mes penchants, mes plaisirs, mes habitudes, et pourra me croire un malhonnête homme, est lui-même un homme à étouffer.

Les Confessions, Livre XII
JEAN-JACQUES ROUSSEAU

8.1 Character network analysis

Character networks model plots using the relations derived from the positions of the characters in the text. They have the following functions:

- They reduce the complexity of the plot, from the text and the various sets of narrative entities to the category of characters; they do not rely on the presence of narrative entities such as places, times of action, or events[1].
- They reveal the structure of the character-system: Is the shape of its structure simple or complex? Symmetric or asymmetric? What can we infer about the plot?

[1] At least not directly.

- They allow analysis of relations between characters, considered in this case as the backbone of the plot, and simultaneously give a *signature* of their roles in the narration.

Borrowing the words of Franco Moretti[2]: we *quantified the plot*[3] by using a network representation as well as methods from network analysis to study the plot and characters' roles in *Les Confessions*. We worked on the dimension of occurrences, which gave us a description of how the characters are distributed throughout the chapters. Then, we included a higher dimension, the co-occurrences, which resulted from the systematic comparison of these distributions of characters. We extracted the relations that model the plot using these observations. Thus, in contrast to the distant reading application on literature that is *culturomics*, we needed other methods than n-grams to "read the book without opening it". We used an index (many are available, at least for classic works) which provided the tables of occurrences from which we deduced weighted co-occurrences. We developed a method for extracting them that is applicable to any index, with or without index subheadings. The set of co-occurrences was subsequently considered as the set of network edges, vertices being the characters. At this stage, we had to decide on how to conduct a literary analysis on an object now conceptually far from a book; the character network was a mathematical representation of the plot.

Our work relies both on close and distant reading. Manual indexing, if included in the workflow, and parts of the interpretation of the character networks, are illustrations of close reading. Automatic indexing, an option that was not considered in our work, the transformation of the index into a network, and the network analysis methods, are illustrations of distant reading. Building on this work, the collection of existing indexes will allow us to build an *atlas* of character networks depicting novels, and will thus potentially uncover *signatures* for different authors, movements, or epochs.

Our approach was strongly inspired by the work of Alex Woloch, who provided a structuralist study of characters (Woloch, 2003). We imported his notions of character-system and character-space to describe *Les Confessions*' characters and their relations. They imply that a character is defined by its narrative neighbourhood[4], and that the society in the text (the society of the story world) is described by the reuniting of these neighbourhoods. In the network, we considered a neighbourhood as the sum of adjacent characters, their relations one to another and with the character under study, and the attributes of both characters and relations. To some extent, character-spaces are not bound by their direct neighbours, since network measures take into account edges outside of the close neighbourhood, with the exception of degree. Thus, a character-space was not necessarily restricted to the co-occurrent

[2] (Moretti, 2013a).

[3] This wording is unfortunate, since a character network is more a topological than a quantified representation of the plot. In a character network, the absolute position of a character does not matter, but the mutual positioning of characters matters. However, a network can loosely be seen as a quantifier, by considering its adjacency matrix, which is an array consisting of non-negative integers...

[4] Close characters as well as elements of the description.

characters: if character A was not linked to B but to C and D, themselves linked to B, then B also intervened in the quantification of A's role, for example when using the eigenvector centrality.

The quantification of plot and narrative roles allows us to describe:

> [...][5] the importance of distributed attention not just to the realist novel but to the vexed problem of characterisation itself: a problem that lies at the heart of contemporary literary theory. (Woloch, 2003, p. 14)

We defined our framework by adapting methods from social network analysis to character network analysis. In a sense, these disciplines are comparable, since in both cases nodes represent human beings, or groups of human beings. When applying the methods of social network analysis to literature, the definition of the existence of relations becomes a necessary piece of information for the narrative interpretation. Similar to previous works which compared networks of different origins (Barabási, 2002), we have suggested that for the measures and concepts we used, network analysis enables a comparison of narratives and societies.

We focused on centrality for the characterisation of the roles; this concept quantifies the importance of nodes as a function of the structure of the network. We found that the distributions of the classical measures of centrality are the same as in most social networks: degree, betweenness and eigenvector follow long-tailed distributions, meaning that in each case a few characters have the central roles, while harmonic closeness follows a gaussian distribution, in particular highlighting a centre and a periphery. The centrality measures describe particular roles in the narrative. We used them in several ways:

- To describe the narrative roles of dominating characters, e.g. Mme de Warens or Denis Diderot.
- To detect and highlight key characters in the narrative, e.g. the comte de Montaigu, Mme de Luxembourg or Mme de Larnage, then to explain the roles they play in the narration.
- To show how the *schemers* are represented as a meta-character with interchangeable network positions, providing a contrasting figure in the second part to the one of Mme de Warens in the first part.
- To get a hierarchical partition of all characters into narrative groups.
- To measure the importance of characters by their absence.

Studies on character networks were conducted before ours, most of them borrowing methods from social network analysis. Among them, the marriage of Alex Woloch's character theory with networks was first proposed by Franco Moretti (Moretti, 2011, 2013b). We have

[5] "In developing these categories [character-space and character-system] through a series of nineteenth-century novels, I want to demonstrate [...]"

brought together the knowledge related to that topic, and outlined the theoretical premises of character network analysis—the analysis of character-systems via a relational approach. In the following section, we discuss what we consider to be interesting paths to follow in the future.

Moreover, we have dissected aspects of the narration of *Les Confessions*. Our position is similar to that of Gérard Genette defining narratology with *À la recherche du temps perdu*[6]. We showed how characters are distributed throughout the novel, how they may be lost in the crowd, on the top of it, positioned at the meeting of different narrative scenes, or isolated. We discussed these results with numbers and illustrated them by labelling the corresponding characters in network visualisations. Thus, *Les Confessions* provided a rich corpus as well as a suitable data set to conduct this analysis. We do not claim that we have brought to light new interpretations of Rousseau's story or even life. Instead, we provide a different view on *Les Confessions*'s narration.

8.2 Future research

At this stage, we would like to discuss two topics that, in our opinion, require further exploration, beyond our current character network analysis framework. They are: the dynamics supporting the character-system (8.2.1) and the study of plots on a larger scale (8.2.2).

The first topic introduces time as a variable—and thus dynamical character networks along with a dedicated methodology—in order to capture how the character-spaces and the character-system are created and formed as the narrative is unfolded. In particular, we define *moving windows* of pages to show the evolution of the structures in series of fixed intervals. We also discuss the extension of centrality measures to make them able to deal with time—we call this *temporal centrality*.

The second topic prepares the character network analysis to a much larger scale, the natural step to take after this work: how do we expand the present framework in order to allow such an analysis to be conducted on a corpus of two (e.g. comparing two novels), twenty (e.g. studying the novels composing *Les Rougon-Macquart*), hundreds (e.g. the moviegalaxies.com database) or even millions of novels as well as scripts, plays and all other kinds of stories?

[6] "On a pu déjà observer que ni le titre ni le sous-titre de cette étude ne mentionnent ce que je viens de désigner comme son objet spécifique. Ce n'est ni par coquetterie ni par inflation délibérée du sujet. Le fait est que bien souvent, et d'une manière peut-être exaspérante pour certains, le récit proustien semblera ici oublié au profit de considérations plus générales : ou, comme on dit aujourd'hui, la critique s'effacer devant la "théorie littéraire", et plus précisément ici la théorie du récit ou narratologie." (Genette, 2007, p. 10)

"Readers may already have observed that neither the title nor the subtitle of this book mentions what I have just designated as its specific subject. The reason is neither coyness nor deliberate inflation of the subject. The fact is that quite often, and in a way that may exasperate some readers, Proustian narrative will seem neglected in favor or more general considerations; or, as they say nowadays, criticism will seem pushed aside by "literary theory," and more precisely by the theory of narrative or *narratology*."

Translation by Jane E. Lewin, published by Cornell University Press, 1980,
see https://archive.org/details/NarrativeDiscourseAnEssayInMethod, p. 22, accessed on 22/07/2014.

8.2.1 Temporality

> "[Le personnage] n'a pas d'autonomie par rapport au texte. [...] Sa "vie" se voit rythmée par la progression de la narration, par les liens qu'il entretient avec son environnement humain, géographique."
>
> *Le personnage de roman*
> JEAN BARDET

The network $\mathscr{G}$ is a representation of the whole narration. It summarises the plot with a significant selection of characters. However, its 216 characters corresponded to 608 edges, while the entire corpus contains 583 characters and 4919 weak co-occurrences (see section 6.1). Our approach did not capture and analyse this overload of information. We set conditions that extracted the most meaningful information and defined a suitable network for the study. The focus on a smaller part necessitated the extraction of the names and relations for that portion of the global network $\mathscr{G}_1$, as this network contains the raw information, so we began from that point in order to build new network models (see appendix B).

A study of the dynamics behind the narration also requires us to go back to that network. It should not be considered as a "wrong" network, but as a preliminary stage to many applications. We compile a family of networks that catch the temporal changes in the shapes of character-spaces as the narration goes forward. At this step, a sentence and the events it contains are found in a context that includes the narration's past. The accumulated information on previous pages helps us contextualise what we read. Character network analysis has the ability to capture the temporality of narrative relations, and to give a representation of the structure they have shaped in the recent past.

In this section, we outline how to extract and analyse sequences of networks built from overlapping page intervals and the corresponding index subsets. They capture local phenomena better than a single global network, or discrete parts of it. Our goal is to understand the dynamic of the narration: how a node or a relation appears, disappears, or reappears, as pages go by. We need to understand how these interact within the more global structure. For example, some characters may appear and disappear very episodically, meaning that they are not central; however they could play a key role in dramatic situations, or could be used as a means for the author to combine disjointed narrative story arcs. On the other hand, a character that is frequently present during an important timeframe, but never appearing in the narrative again, would play a different role. This role is not necessarily less influential and, in this case, may even highlight the leader of a precise story arc. Alex Woloch suggests *temporal centrality* when analysing Herman Melville's *Benito Cereno* when he writes:

> The remarkable tale is structured around the three characters' continual shifts in *centrality* (as potential protagonists within the discourse) and *power* (as potential captains within the story). (Woloch, 2003, p. 347, n.30)[7]

[7] This is not the same *centrality* concept as in chapter 7.

Even though the focus here is on the interlocking ties between discourse and story, he shows interest in the narrative shifts in the importance of the character.

In studies about network dynamics, social network analysts usually get snapshots of the society in focus, then compare the changes between them. They obtain a set of discrete network states. Here, we want the smallest level of definition possible—in our case of extraction from an index, it is the page. For this purpose, we build and analyse subnetworks based on intervals that overlap: e.g. the character network of pages one to fifteen, then two to sixteen, and so on, for a temporal window of fifteen pages (see figure 8.1). We do not consider the approach of taking a few snapshots along the narrative, because by doing so we would lose the continuity and the exhaustivity of the corpus.

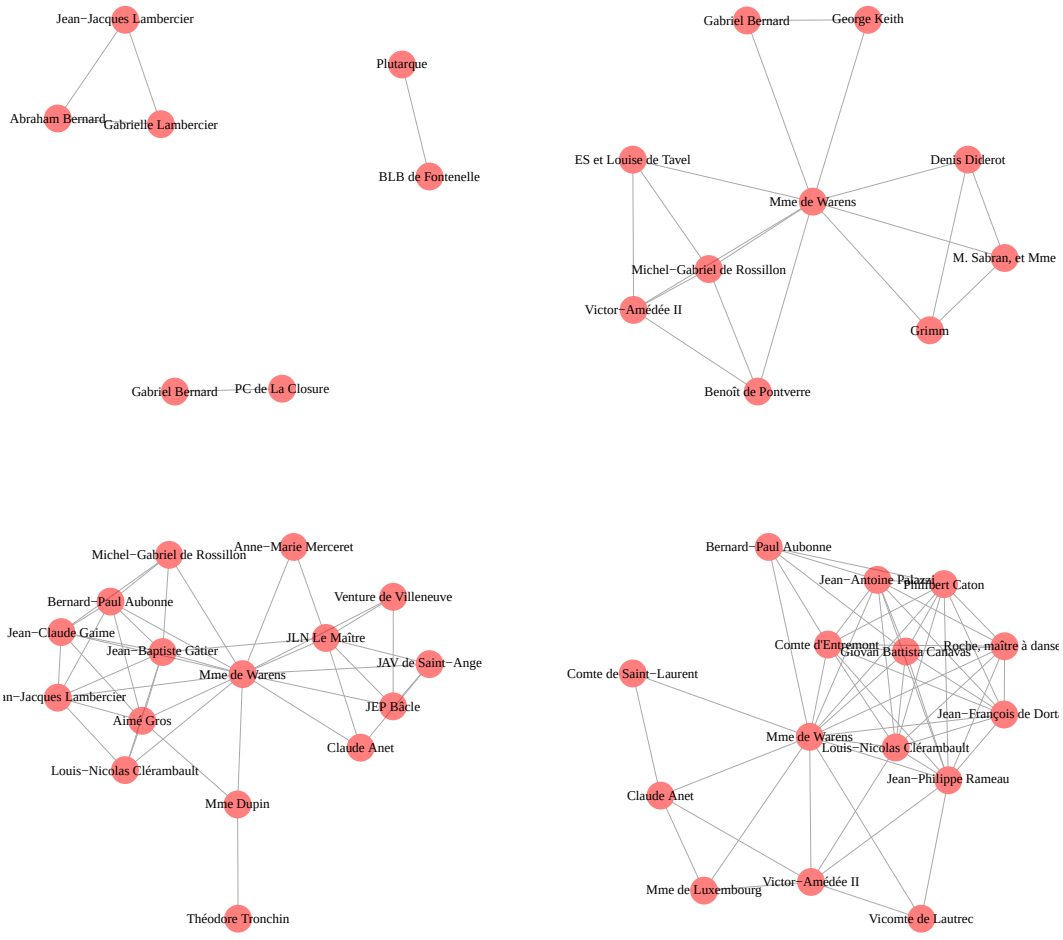


Figure 8.1: Temporal subnetworks compiled from fixed intervals of fifteen pages, respectively (from left to right, top to bottom) 67-82, 121-136, 200-215 and 272-287. For the same number of pages, we immediately see variations in order, size, density centralisation and clustering.

The ideal number of pages to keep in order to define the temporal window remains an open question. It depends on the size of the corpus, and editorial choices like the quantity of text on each page, to capture a significant number of characters and relations. It also depends on the research questions, which may ask for less extended memory.

A sequence of networks allows us to capture local phenomena. If we consider two consecutive networks from that sequence, say G_i for pages x to $x+t$ and G_{i+1} for pages $x+1$ to $x+t+1$, the slip from one network to the other leaves the occurrences and co-occurrences of page x and includes those of page $x+t+1$. This is what we presented as continuity[8]: a small modification—a character disappearing from a narrative event—does not cause a significant modification of the networks, while a complete change of focus from one place to another, with the substitution of all the characters, is observable and significantly modifies the shape of the sequence at that point, thus any temporal indicators as well. Sequences also offer visual opportunities: we compiled a short movie[9] showing the evolution of the plot, with the help of visual effects highlighting the appearance and departure of characters in the plot, as well as a clustering phenomenon and radical changes of focus. These changes imply that the whole set of present characters has been replaced and suggest a transition to another narrative context.

We consider that analytic methods on temporal networks extend the measures we used in our work. For example, temporal degree centrality is the result of the computation of degree centrality on each network of a sequence (see figure 8.2)[10]. It shows the central position of a character according to degree centrality and to the structure of the character-system at that narrative interval.

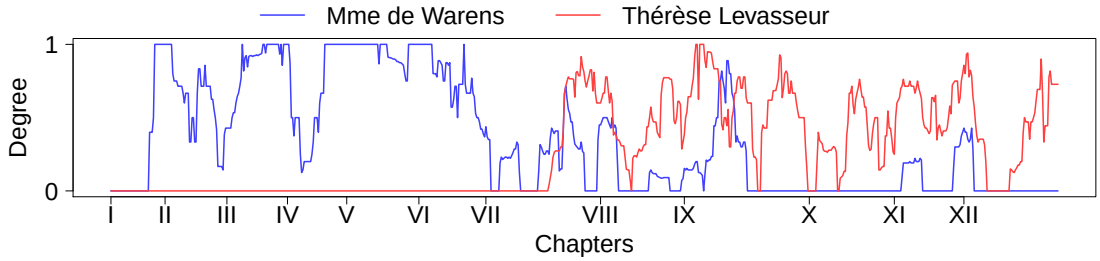


Figure 8.2: Temporal degree centrality of Mme de Warens and Thérèse Levasseur. It is normalised by the order of the network minus one at the given time.

In further research, we suggest characterising the distributions of these measures and developing a narratological framework in order to use this type of data for the definition of characters' roles over time.

[8] It is interlocked at a level of precision such that we consider it continuous rather than discrete.

[9] Presented at the DH 2013 conference in Lincoln, Nebraska. See the video here: http://infoscience.epfl.ch/record/200020, accessed on 22/07/2014.

[10] For literature on temporality and temporal centrality, see (Braha and Bar-Yam, 2006; Tang et al., 2010; Pan and Saramäki, 2011; Kim and Anderson, 2012; Holme and Saramäki, 2012, 2013).

8.2.2 Large-scale

> "Ce n'était pas une petite tâche que de peindre les deux ou trois mille figures saillantes d'une époque, car telle est, en définitive, la somme des types que présente chaque génération que *La Comédie humaine* comportera. ce nombre de figures, de caractères, cette multitude d'existences exigeaient des cadres, et, qu'on me pardonne cette expression, des galeries. De là, les divisions si naturelles, déjà connues, de mon ouvrage en *Scènes de la vie privée, de province, parisienne, politique, militaire* et *de campagne*."
>
> Foreword to *La comédie humaine*
> HONORÉ DE BALZAC

Our method builds character networks automatically and is accessible as long as the index has been obtained in a digitised form[11]. Theoretically, it allows the compilation of a database whose size is the number of all existing indexes. Many indexes are not readily available, as most of the time they are omitted during the transformation of a printed text into a digitised form. However, they are available for many classics and in different languages (Bradley, 1989): an automatic scanning process, along with optical character recognition methods, can help to gather hundreds or even thousands of these potential representations of plots. This is an illustration of the "large-scale gathering of data" Franco Moretti is looking for in (Moretti, 2013a, p. 211).

At a micro level, we saw that narrative roles of characters can be characterised by network measures. The question of exhaustivity was not answered, however, nor the definition or existence of a complete typology. For now, we have shown the relevance of indicators, such as centrality measures (and derived measures for vitality and centralisation), to describe and explain the importance of each character in the plot based on the underlying structure of relations between characters. Nonetheless, the indicators were robust in our typology. We believe that it is possible to use network measures to attribute a *signature* to a narrative role, and interpret narrative relations without the help of the text (here, discourse and style).

At the macro level, measures like average path length, diameter, centralisation or density[12], as well as methods for community detection, allow us to describe the entire structure and to make assumptions on the repartition of the roles, and thus on the structure of the plot. For example, it is possible to provide a preliminary description of the subplots corresponding to each chapter of *Les Confessions* by studying the networks in figure B.1. We can compare them,

[11] The use of automatic indexation was discussed in chapter 5, where we cited previous works having optimistic results.

[12] In a recently submitted abstract in which we conduct such a macro character network analysis on Zola's *Les Rougon-Macquart*, we worked with the notion of *k-core* (Seidman, 1983) to differentiate for each character network the characters positioned in the center from the others, close to it or situated in the periphery. We showed how the *cores* of some novels were sometimes made of one or two characters, while at other times they were composed of up to a dozen characters with comparable structural properties, revealing how the shapes of Zola's narratives may rapidly vary when going from a theme to an unrelated other.

apply the distributions of order, size, and other indicators, then recognise if the network is compact or divided into communities of characters, or linked by shared events or characters. Thus, we can extract a representation of the narrative's rhythm, and highlight the centres of attention.

In conclusion, based on these three points, we consider that a large-scale character network analysis can provide relevant results when willing to refine the definitions of *signatures* of roles and of networks. For example, the comparison of other autobiographical works with *Les Confessions* would help to highlight additional features unique to this text. A second example would be the study of other novels from the same epoch, which would help better characterise, contextualise and explain how the story is told and how its character-system is organised. Furthermore, outside of the Rousseau scope, we could study cases of series of novels with interlocking characters, such as *Les Rougon-Macquart* by Émile Zola, *La Comédie humaine* by Honoré de Balzac, or the *Sagas of Icelanders*, where character network analysis would show the higher systems of all reunited character-systems. Finally, at the highest level, indexes exist for thousands of classic texts: from these, by conducting the systematic work of gathering, constructing and analysing which we presented in this thesis, we can expect to detect *signatures* for genres, authors, literary tendencies, epochs, etc., and assemble an *atlas* of novels based on the properties of their character networks.

A Index excerpt

In the following figure, we show the first page of the index used as the basis of this work (see section 4.3). Each entry has a name associated with one or more numbers for occurrences on pages. In the index, we find the names cited in the novel, as well as in the comments and footnotes. In these two last cases, the page number is accompanied by a "n". The names correspond to fictitious characters and real persons, among those people from Rousseau's world as well as those related to the editors themselves.

INDEX DES NOMS CITÉS

Figure A.1: First page of the index of characters from (Rousseau et al., 2012, Vol. 1-2).

B Subnetworks

When we deal with questions on the structure adopted by the set of characters and their interactions in the text, we may have to provide a different network for each question. For example, to answer the question "What is the character-system of a given chapter?", we need a network based on that chapter: the choice of a subnetwork of $\mathscr{G}$ must be dismissed because it will possess links from other parts of the text. It is a better solution to compile a network from the original data set, perhaps with a different threshold, and take that into account in the analysis. As additional examples: answering the "How do the character-spaces from two non-adjacent chapters taking place in the same towns interact?" question requires the union of the networks induced by these chapters, and "What are the characters at the core of the narrative, on a relational basis?" necessitates that we extract a character network with a higher threshold.

The model is not necessarily based on the original network—it may be compiled differently from the same data set. Every question or context implies and induces a method that may be new. This is what we described in sections 6.1 and 6.2 when wanting to understand the global structure of a character network based on *Les Confessions*. At first, we chose the network with a threshold equal to one, but we had to deal with many potential artefacts, like links that correspond to weak co-occurrences. To focus on the significant ones, we levelled the threshold up to three and obtained a more suitable network. In a character network analysis, the research questions define the methods, and thus define the network itself. We may have chosen one that we consider as the representation of the society from the novel, $\mathscr{G}$, but this does not help when defining other networks to model other narrative properties of the characters.

If the focus is on finite parts of the novel, we build a network based on co-occurences situated in a given interval of pages, and choose the appropriate method. This is what we propose in this section of the appendix: to compare the character-systems in chapters I and VII. We build networks based on the pages dedicated to these chapters, with the page adjusting method, but without defining a threshold, since the occurrences are much fewer in chapter I. This is an important point, letting us compare networks of chapters I and VII, but we are not

allowed to compare these networks with the global one defined earlier, $\mathscr{G}$. Hereby, we propose a network analysis in a smaller scale. We show how the structures of two character networks differ across the different chapters of the novel. The comparison of these networks is a work that cannot be done simply on the basis of the visuals.

To extract the co-occurrences from a given chapter, we delimit the occurrences to consider the pages of that chapter only[1]. This was done in table 4.1. Thus, for a given chapter, say chapter I, we consider only the co-occurrences for that chapter. Then, we consider all the *1-co-occurrences*, and generate the network. The reason for considering *1-co-occurrences* instead of *3-co-occurrences* is the size of the corpus. Moreover, having a threshold too high—e.g. equal to three—means losing important parts of that local character-system[2]. It implies that the structures we are studying here are less robust that the one of network $\mathscr{G}$, for example. The chosen method—*1-co-occurrences*—prevents us from comparing the character networks with the global definitive network $\mathscr{G}$, since links account for different things. For example, in the case of the restriction of network $\mathscr{G}$ to the characters occurring in chapter I, the links are the ones we retained at a higher and global level. Using this method also means taking into account co-occurrences outside of the chapter: if nodes A and B both appear in chapter I and are related in $\mathscr{G}$, this does not imply that they are related in chapter I. Taking a subnetwork of the global network constructed on *1-co-occurrences*, $\mathscr{G}_0$ would lead to the same difficulties. In table B.1, we give the order, size and density of the resulting subnetworks for *1-*, *2-* and *3-co-occurrences*. We observe the same drop as in section 6.3 when passing from *2-* to *3-co-occurrences*. These last character networks are comparable with $\mathscr{G}$ and $\mathscr{G}_3$. Nonetheless, here we choose *1-co-occurrence*, which reunites all of the characters that appear in the index.

The twelve resulting networks (see figure B.1) have various orders and sizes. All of the characters co-occurring with at least one other character are in these networks, even if it occurred only once in the narrative. Table B.2 shows these networks' statistics, along with their connectedness. The order—the number of nodes—largely varies from one chapter to the other. Chapters from the first part (I to VI) have fewer different characters than chapters from

[1] When we extracted the whole network, there were no cases of co-occurrences above the page delimitating any two chapters: we did not have to deal with that special case. We wanted to clarify that point: we remarked that Mme de Warens was frequently mentioned on the last and first pages containing names, of successive chapters (chapters III and IV, IV and V, V and VI). Mme de Warens frames the story in these chapters from part one. Denis Diderot appears in such a case between chapters VII and VIII, Louise d'Épinay between chapters IX and X, the comtesse d'Houdetot between chapters X an XI, Daniel Roguin between chapters XI and XII. We have considered only the last and first pages, and despite this strong constraint, in part two like part one, some names assure the transition. However, in part two, they are all different.

[2] If the corpus considered is the chapter, then it is a character-system. If it is part of a larger corpus, then the object of study is the union of character-space. In a sense, when the focus changes inside a study, exceptionally a character-space and a character-system may be confounded.

Chapters	Order			Size			Density		
	1	2	3	1	2	3	1	2	3
I	33	28	8	90	54	6	0.17	0.14	0.21
II	37	35	14	91	60	19	0.14	0.10	0.21
III	50	49	20	217	121	29	0.18	0.10	0.15
IV	40	40	15	135	76	13	0.17	0.10	0.12
V	66	66	20	365	196	36	0.17	0.09	0.19
VI	38	37	15	118	72	15	0.17	0.11	0.14
VII	173	170	47	1389	690	77	0.09	0.05	0.07
VIII	104	104	29	795	421	75	0.15	0.08	0.18
IX	56	56	20	331	211	55	0.21	0.14	0.29
X	98	98	38	783	436	101	0.16	0.09	0.14
XI	69	69	22	482	259	66	0.21	0.11	0.29
XII	96	89	31	550	282	56	0.12	0.07	0.12

Table B.1: Order, size and density of subnetworks induced by *1-*, *2-* and *3-co-occurrences* in all twelve chapters.

the second part, with the exception of chapters V[3] and IX[4]. The minimum (33 characters) is reached in chapter I, explaining the beginning of his life. The maximum (173 characters) is obtained in chapter VII, where Rousseau travels to Lyon, Paris, meets Denis Diderot and Mme de Francueil there, then works in Venice as the ambassador's secretary, then finally comes back to Paris, citing many characters from his stays in these cities, all appearing in this unbalanced chapter.

The number of edges of the networks—the size—varies regularly in function of the order. Usually, more characters bring more relations. This regularity between order and size, probably due to the extraction method, is not present with *3-co-occurrences* (see table B.1). Moreover, the ranks of order and size of the twelve chapters are perfectly correlated[5], meaning that if network A has more nodes than network B, it will also have more edges. We show later that this is due to a constant effect of cohesion in the way Rousseau's narrative deals with co-occurrences of characters.

Density measures the ratio of edges in the network with the theoretical maximal number of edges, that is the number of edges in a complete network with same order. Density varies with the number of nodes: the low value of chapter VII is caused by the high number of nodes, not by a lower mean degree[6]. Instead, a low density implies that the network is less compact, probably wrapped around a few communities. We studied that phenomenon in general on

[3]The high number may be explained by the birth of his passion for playing and teaching music, which results in some social success. He meets many influential people for him during his times in Annecy with Mme de Warens.

[4]The low number can be explained by the isolation of Rousseau outside of Paris, and meetings with recurring characters only, among which are the main ones behind the plot.

[5]This is the Spearman correlation coefficient.

[6]The average degree is the highest in this chapter (16.06)

Figure B.1: Subnetworks defined by chapters, one for each of the twelve chapters, to be read from left to right, and then from top to bottom. In each plot, lighter in the back, we show the global graph with threshold equal to one in order to give a topological context, since some nodes do not appear in the main graph, having a threshold equal to three and omitting some of these characters. Note that the positions of nodes in the twelve subnetworks depend on theirs in the global network.

Chapter	Order	Size	Density	Connected
I	33	90	0.17	NO
II	37	91	0.14	NO
III	50	217	0.18	YES
IV	40	135	0.17	YES
V	66	365	0.17	YES
VI	38	118	0.17	YES
VII	173	1389	**0.09**	NO
VIII	104	795	0.15	YES
IX	56	331	**0.21**	YES
X	98	783	0.16	YES
XI	69	482	**0.21**	YES
XII	96	550	0.12	NO

Table B.2: Order, size, density and connectedness of subnetworks induced by *1-co-occurrences*, in all twelve chapters.

the whole network in section 7.2.2.3. The visualisation of chapter VII in figure B.1 is based on a layout common to all twelve chapters.

Connectedness of the subnetworks may appear irrelevant, since the sizes of the giant components of the unconnected networks remain quite high (94.6%, 99.4%, 97.9% for chapters II, VII and XII respectively), with the exception of chapter I (81.8%), a case once again explained by fewer characters[7]. However, this shows some kind of continuity inside the chapters, with recurring characters, or short transitions from an event to the next one, thus keeping characters that help the networks from getting disconnected close together, or at least disconnected in more than two components.

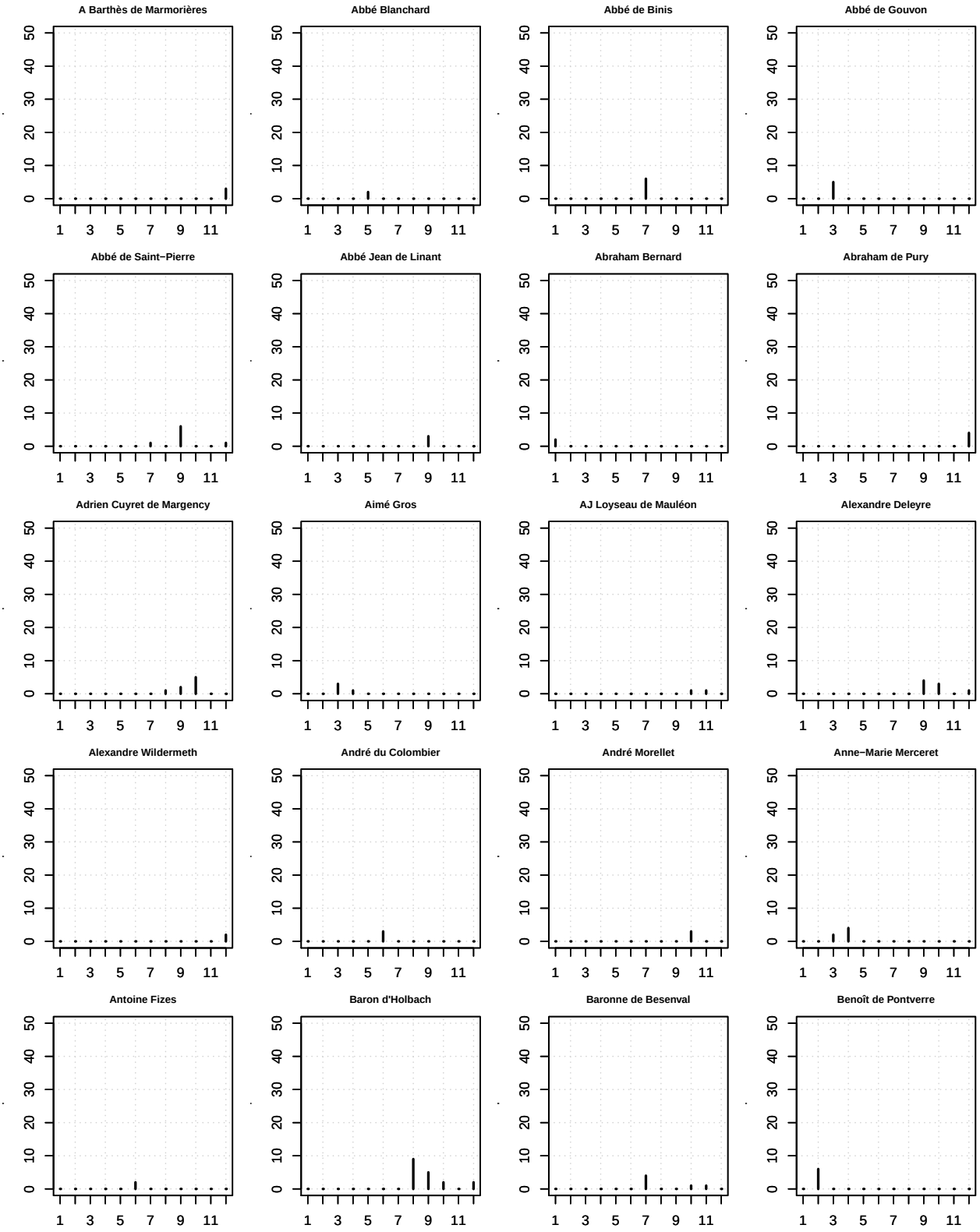
The variation of order and size can be observed in the visualisations contained in figure B.1. The networks are plotted on top of the naive network $\mathscr{G}_0$. This creates some distortion, if two characters are structurally close in one of the chapter networks, and not in $\mathscr{G}_0$, or the contrary. These distortions can be seen with most of the networks corresponding to chapters in the first part of the book. But the reason why we show these visualisations is to give a first insight into the models we are dealing with. A graph visualisation cannot be the main object under study, but it remains a useful descriptive tool.

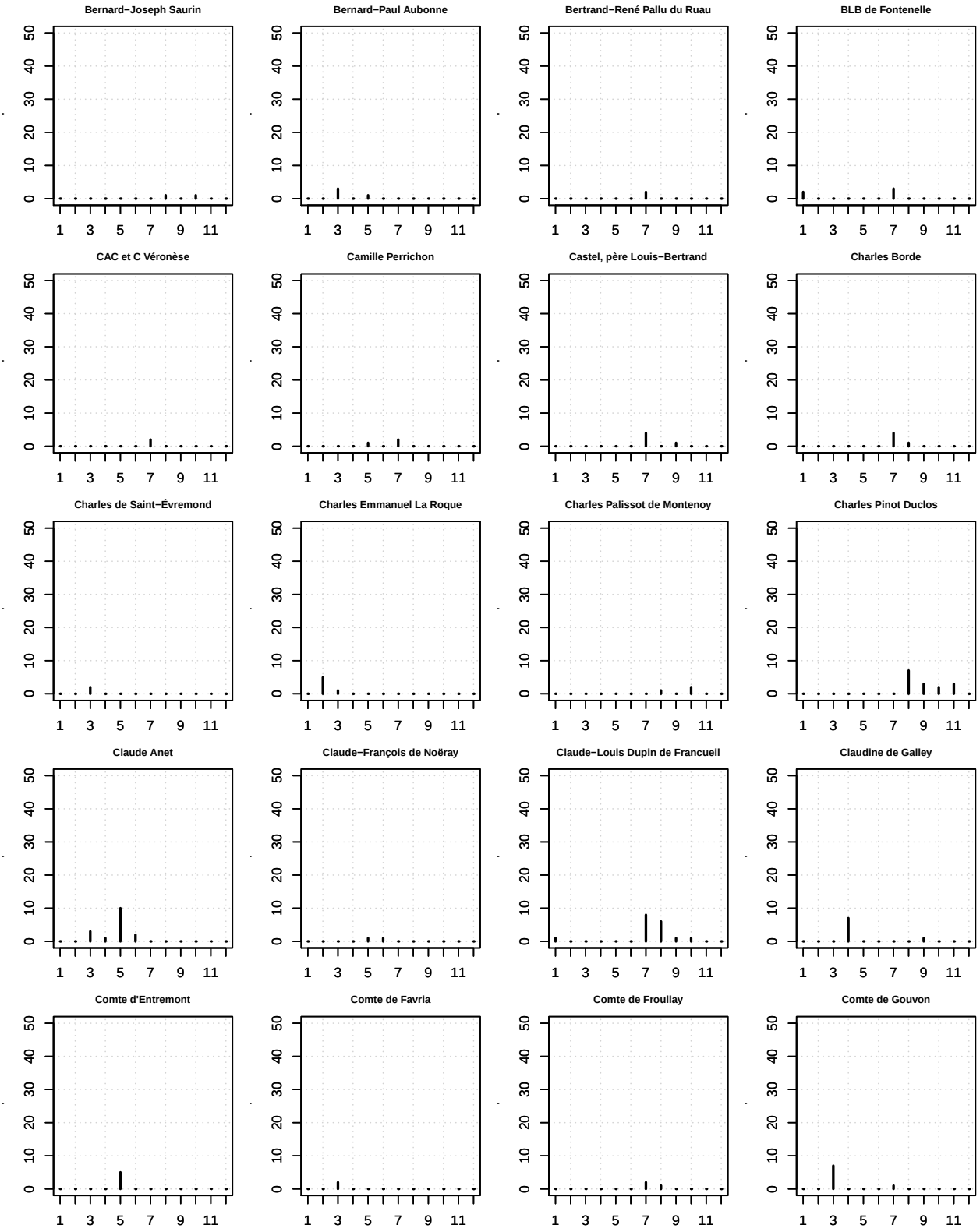
[7] When explored in detail, the connected components show three scenarios: in chapter VII, the disconnected component from the giant one appears at the beginning of the chapter, being situated on the border. In chapters II and XII, the disconnected components (only one in each case, composed of a couple of characters) appear in the middle of the chapter: Dame Lorenza and Henri IV for chapter II (respectively at the top and the bottom of the page, there is no story interpretation of a link between them), Jacob Favre and Jean-Robert Tronchin for chapter XII (both are state men, in an event where Rousseau describes how he defends himself against the public opinion. This social similarity, and the context linked to it, explains why appear in the same place). Finally, chapter I is a mixed case: it has a disconnected component in its middle (p. 77, Mme Clot, someone Rousseau . . . disturbed.) and two other disconnected components at the end of the chapter.

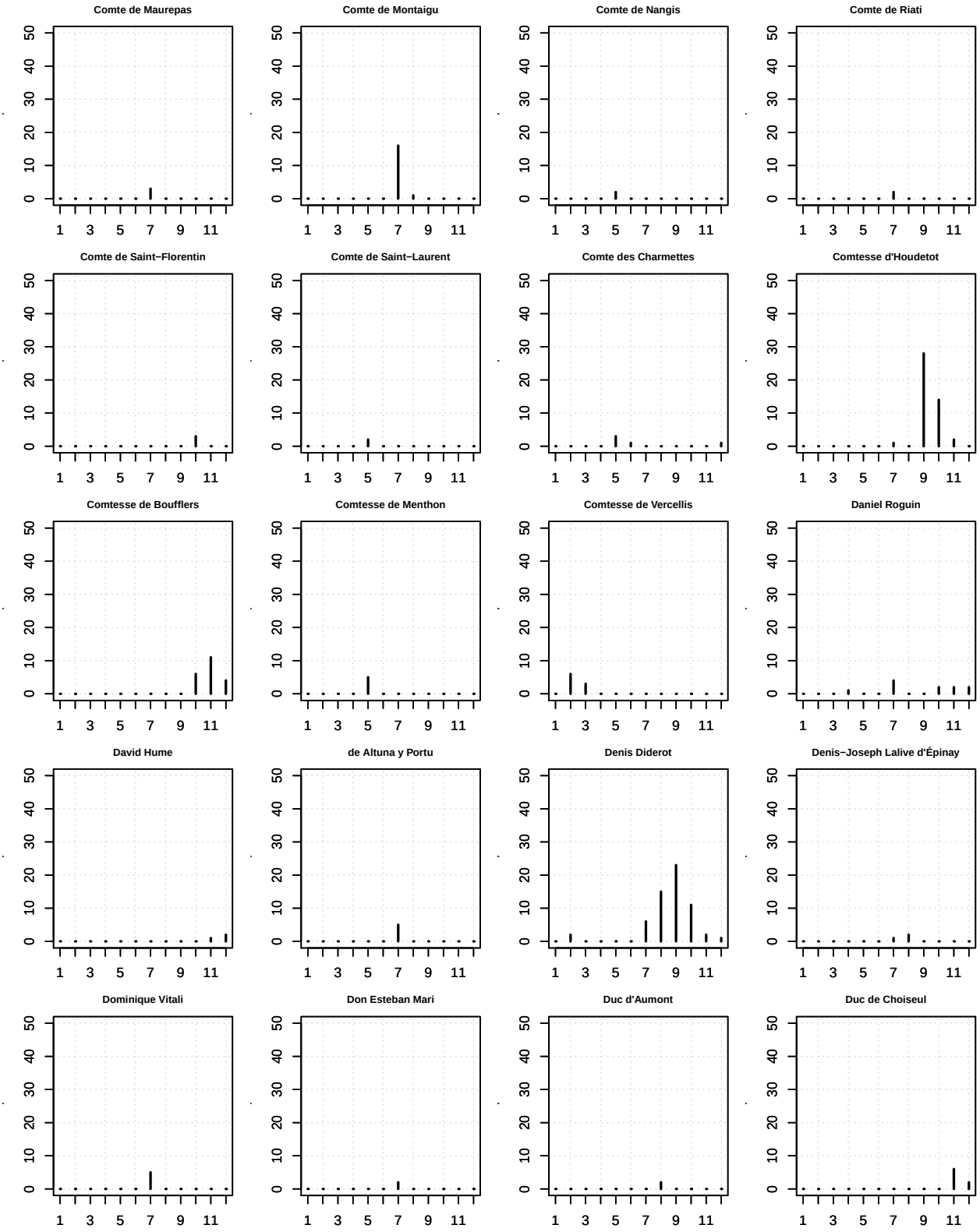
To describe the network visualisations in figure B.1, we first have to remember that the positions of the nodes, which form the *graph layout*, are stochastically computed on the basis of the existence or not of a link between each couple of nodes in the network. Briefly, the algorithm used here (Fruchterman and Reingold, 1991) is a forced-directed algorithm. It brings closer nodes that are related together, considering edges as springs, while a repulsive force is emitted by each node, like an electric field. The algorithm starts from a random position, and usually converges to a stable position. Thus, in the visualisations, two nodes situated close together are often related, but they more probably share the same neighbourhood. Basically, the best way we can describe it is that the more nodes the network of a chapter has, the closer its nodes will be to the center of the bigger plot.

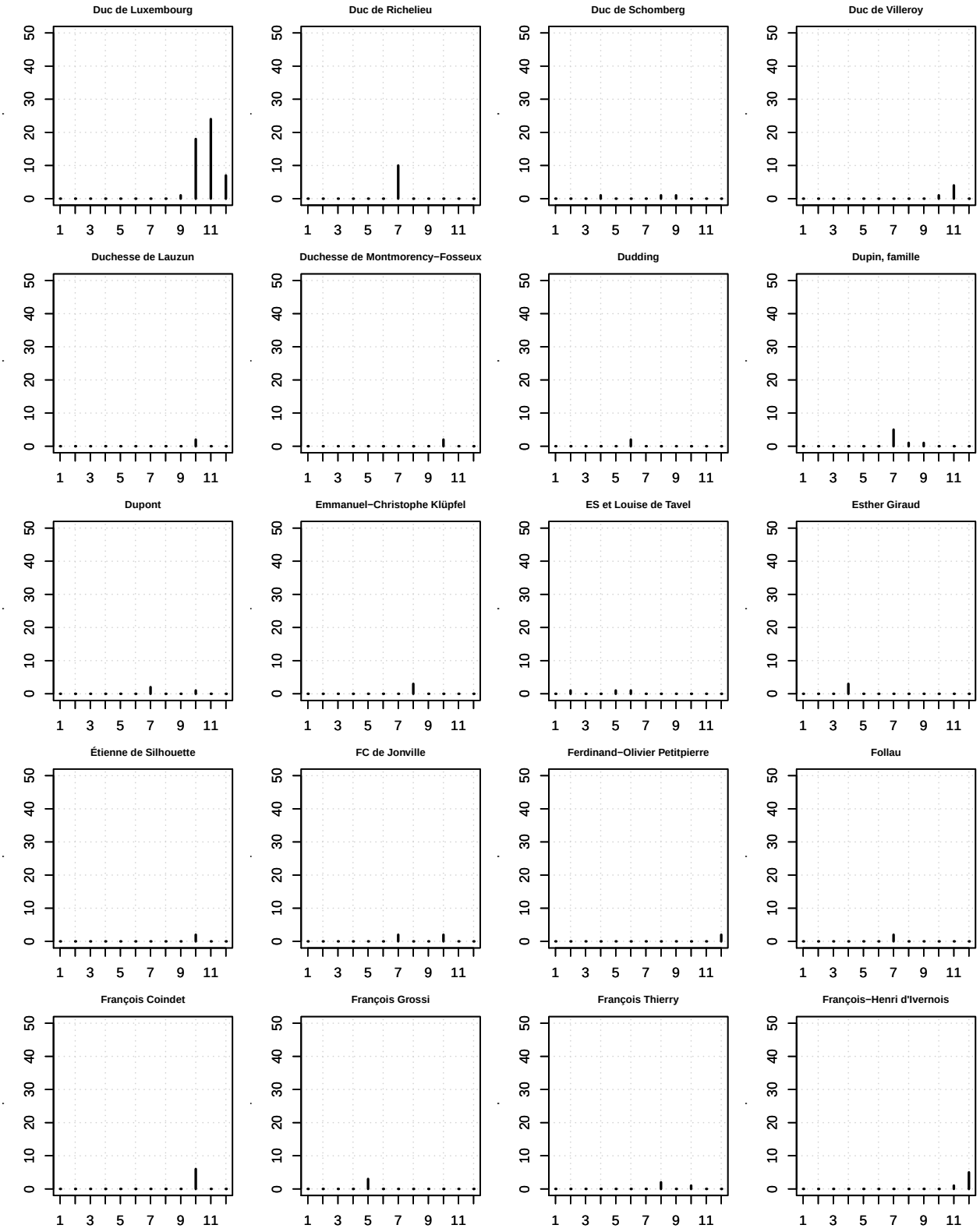
C Occurrences of all characters

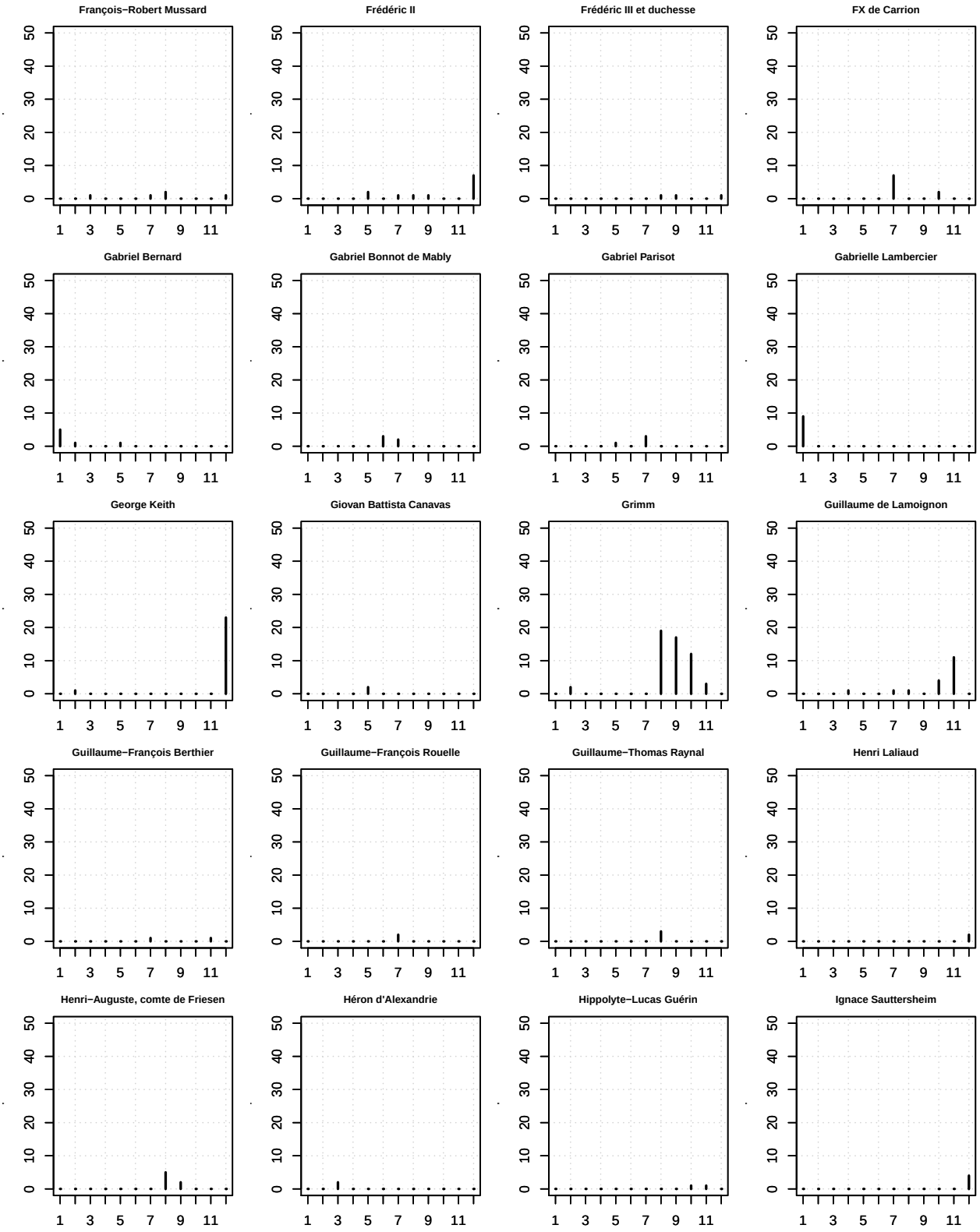
In the following figures, we show the occurrences of the 216 characters in $\mathscr{G}$ regarding the twelve chapters of *Les Confessions*. They are listed in a rough alphabetical order. These plots were helpful during the analysis. It can be seen as a preliminary step to the study of co-occurrences.

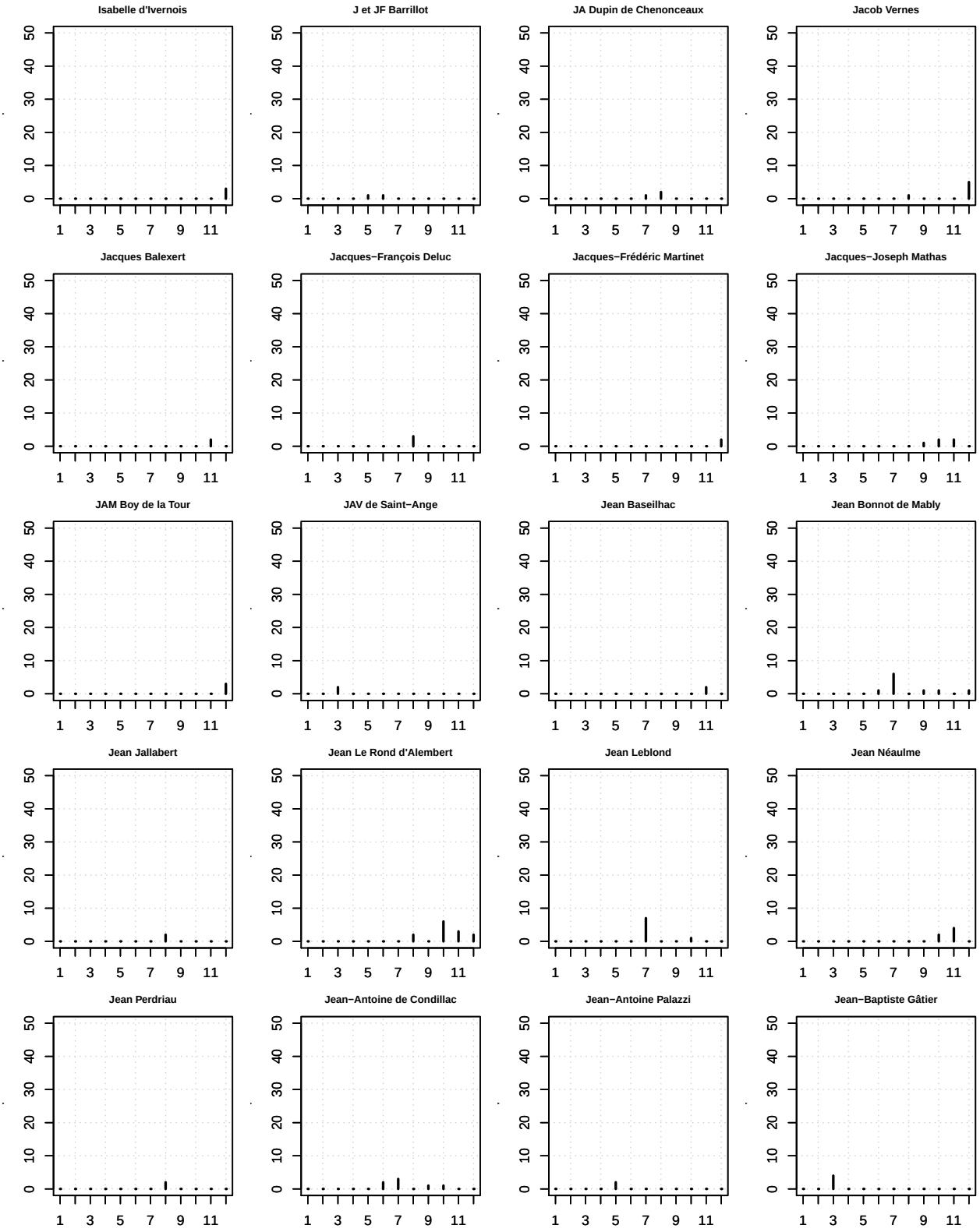

A Barthès de Marmorières
Abbé Blanchard
Abbé de Binis
Abbé de Gouvon
Abbé de Saint−Pierre
Abbé Jean de Linant
Abraham Bernard
Abraham de Pury
Adrien Cuyret de Margency
Aimé Gros
AJ Loyseau de Mauléon
Alexandre Deleyre
Alexandre Wildermeth
André du Colombier
André Morellet
Anne−Marie Merceret
Antoine Fizes
Baron d'Holbach
Baronne de Besenval
Benoît de Pontverre
0 10 20 30 40 50
1 3 5 7 9 11

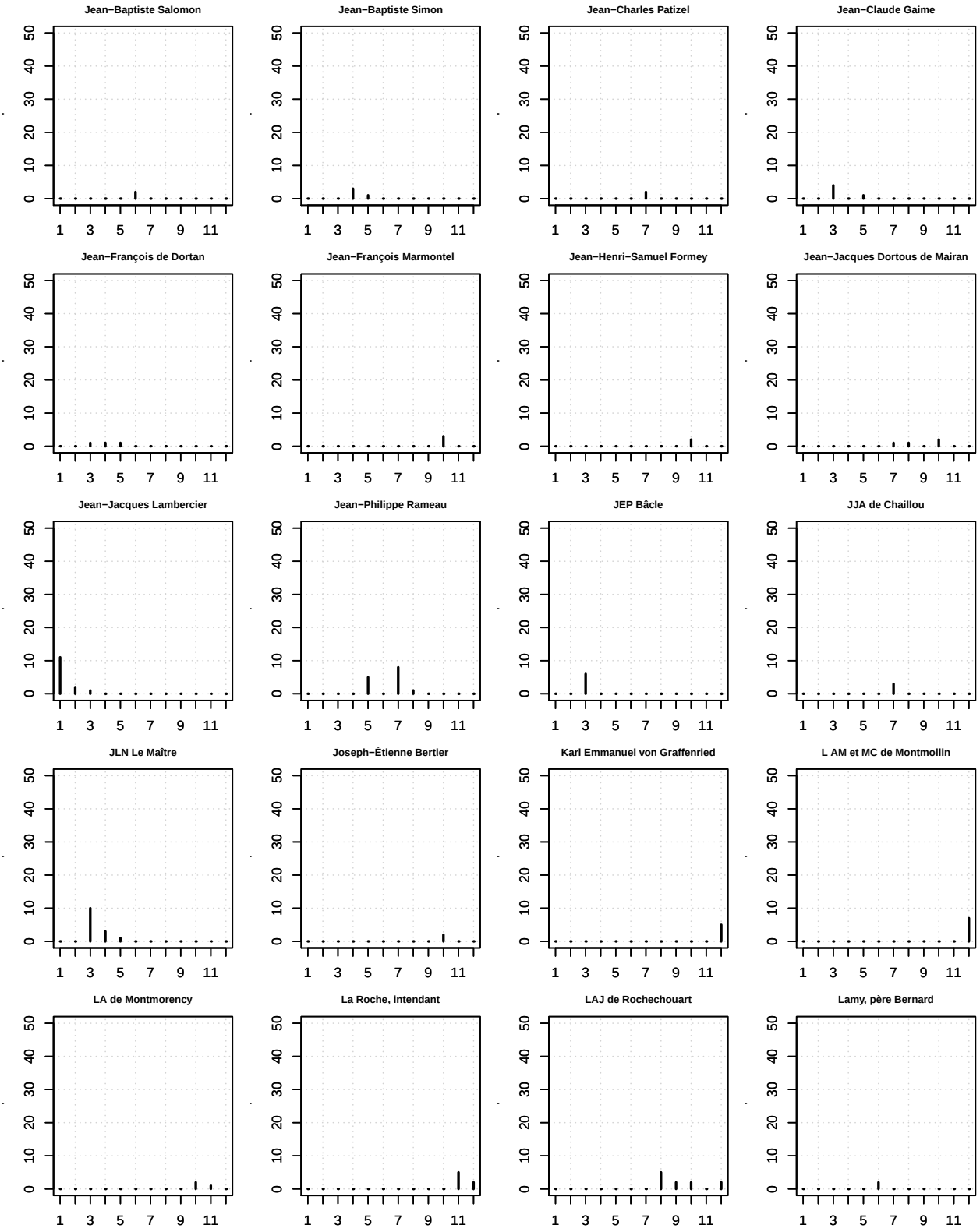

Bernard−Joseph Saurin
Bernard−Paul Aubonne
Bertrand−René Pallu du Ruau
BLB de Fontenelle
CAC et C Véronèse
Camille Perrichon
Castel, père Louis−Bertrand
Charles Borde
Charles de Saint−Évremond
Charles Emmanuel La Roque
Charles Palissot de Montenoy
Charles Pinot Duclos
Claude Anet
Claude−François de Noëray
Claude−Louis Dupin de Francueil
Claudine de Galley
Comte d'Entremont
Comte de Favria
Comte de Froullay
Comte de Gouvon
0 10 20 30 40 50
1 3 5 7 9 11

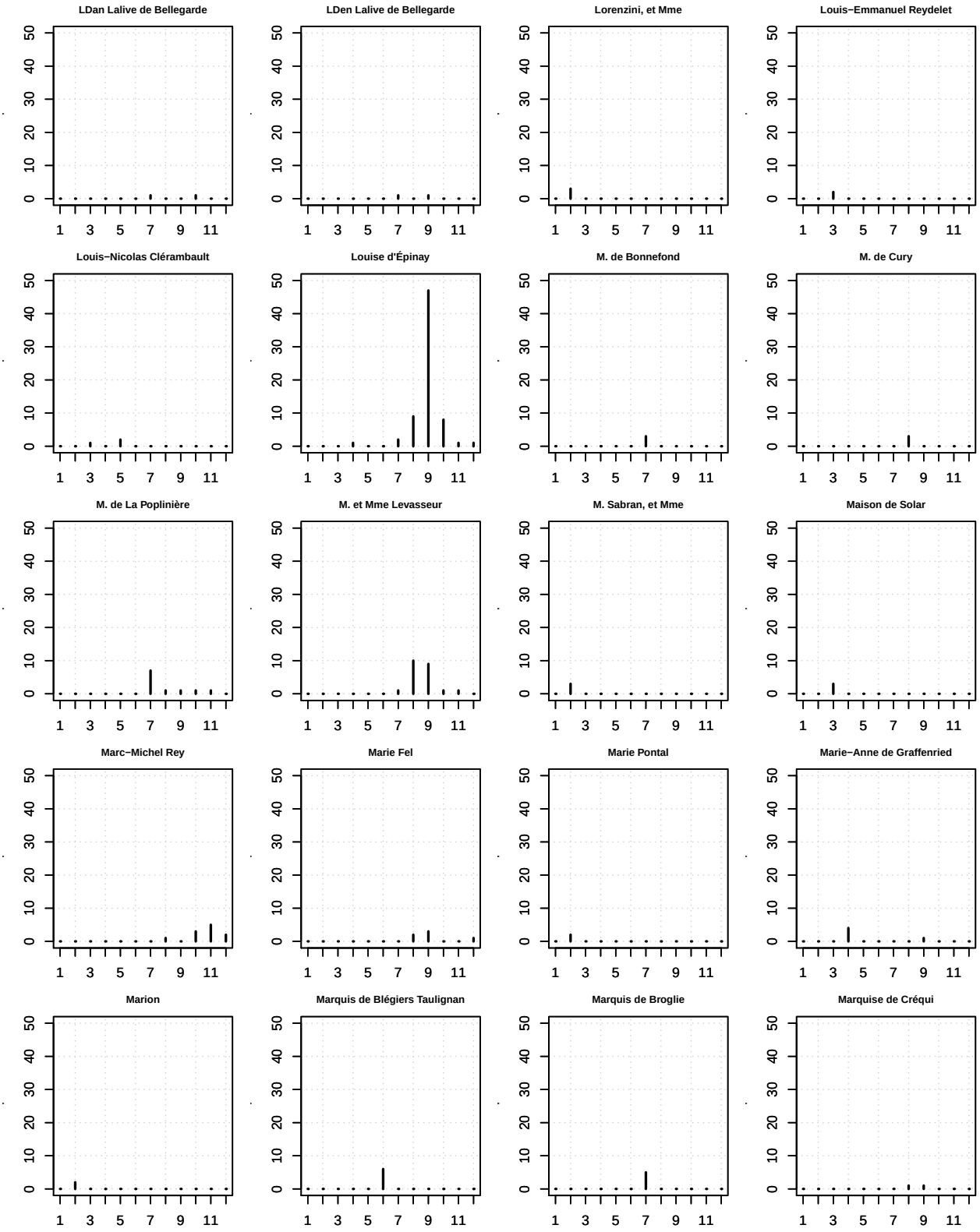

Comte de Maurepas
Comte de Montaigu
Comte de Nangis
Comte de Riati
Comte de Saint−Florentin
Comte de Saint−Laurent
Comte des Charmettes
Comtesse d'Houdetot
Comtesse de Boufflers
Comtesse de Menthon
Comtesse de Vercellis
Daniel Roguin
David Hume
de Altuna y Portu
Denis Diderot
Denis−Joseph Lalive d'Épinay
Dominique Vitali
Don Esteban Mari
Duc d'Aumont
Duc de Choiseul
0 10 20 30 40 50
1 3 5 7 9 11

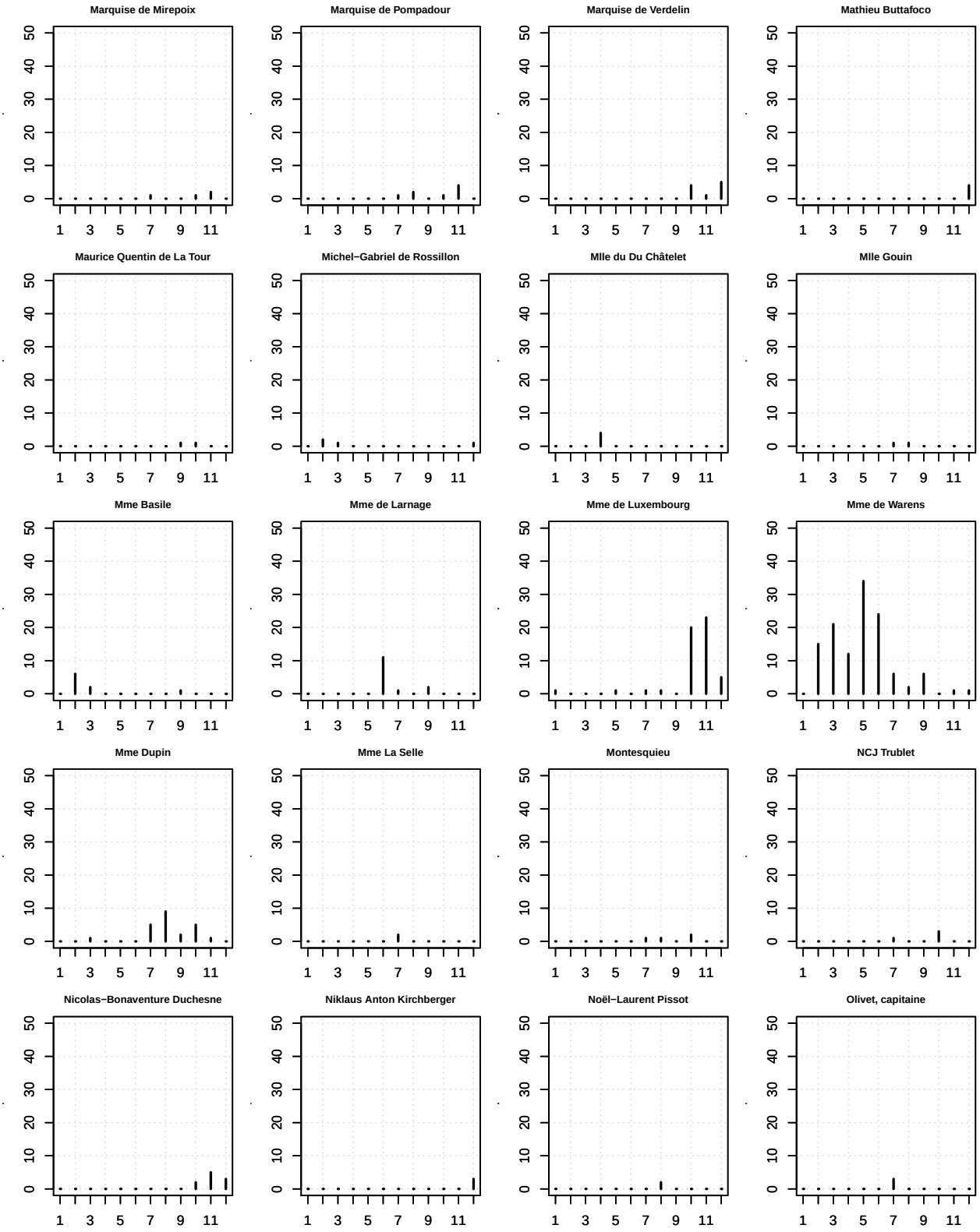

Duc de Luxembourg
Duc de Richelieu
Duc de Schomberg
Duc de Villeroy
Duchesse de Lauzun
Duchesse de Montmorency−Fosseux
Dudding
Dupin, famille
Dupont
Emmanuel−Christophe Klüpfel
ES et Louise de Tavel
Esther Giraud
Étienne de Silhouette
FC de Jonville
Ferdinand−Olivier Petitpierre
Follau
François Coindet
François Grossi
François Thierry
François−Henri d'Ivernois
0 10 20 30 40 50
1 3 5 7 9 11

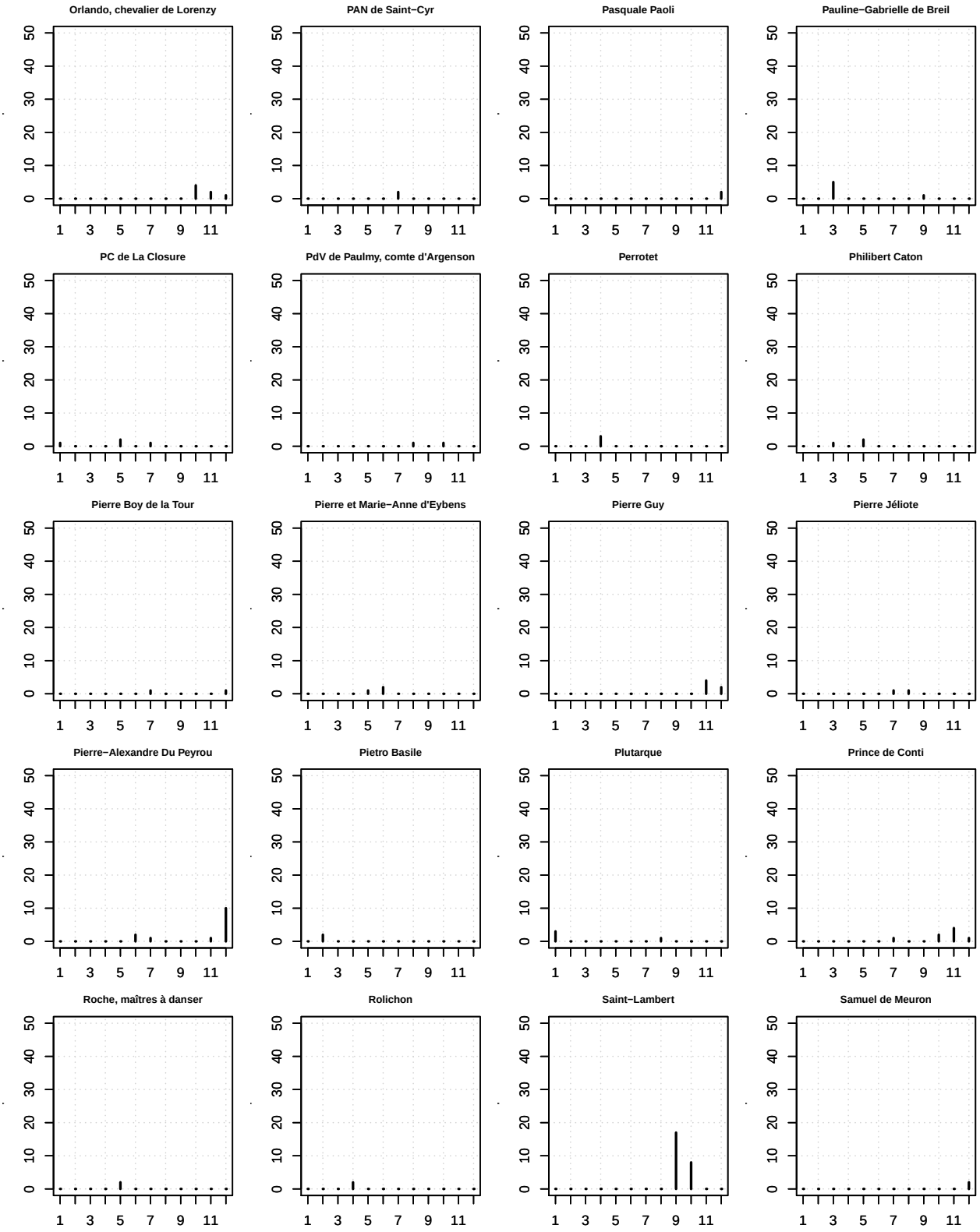

François−Robert Mussard
Frédéric II
Frédéric III et duchesse
FX de Carrion
Gabriel Bernard
Gabriel Bonnot de Mably
Gabriel Parisot
Gabrielle Lambercier
George Keith
Giovan Battista Canavas
Grimm
Guillaume de Lamoignon
Guillaume−François Berthier
Guillaume−François Rouelle
Guillaume−Thomas Raynal
Henri Laliaud
Henri−Auguste, comte de Friesen
Héron d'Alexandrie
Hippolyte−Lucas Guérin
Ignace Sauttersheim
0 10 20 30 40 50
1 3 5 7 9 11

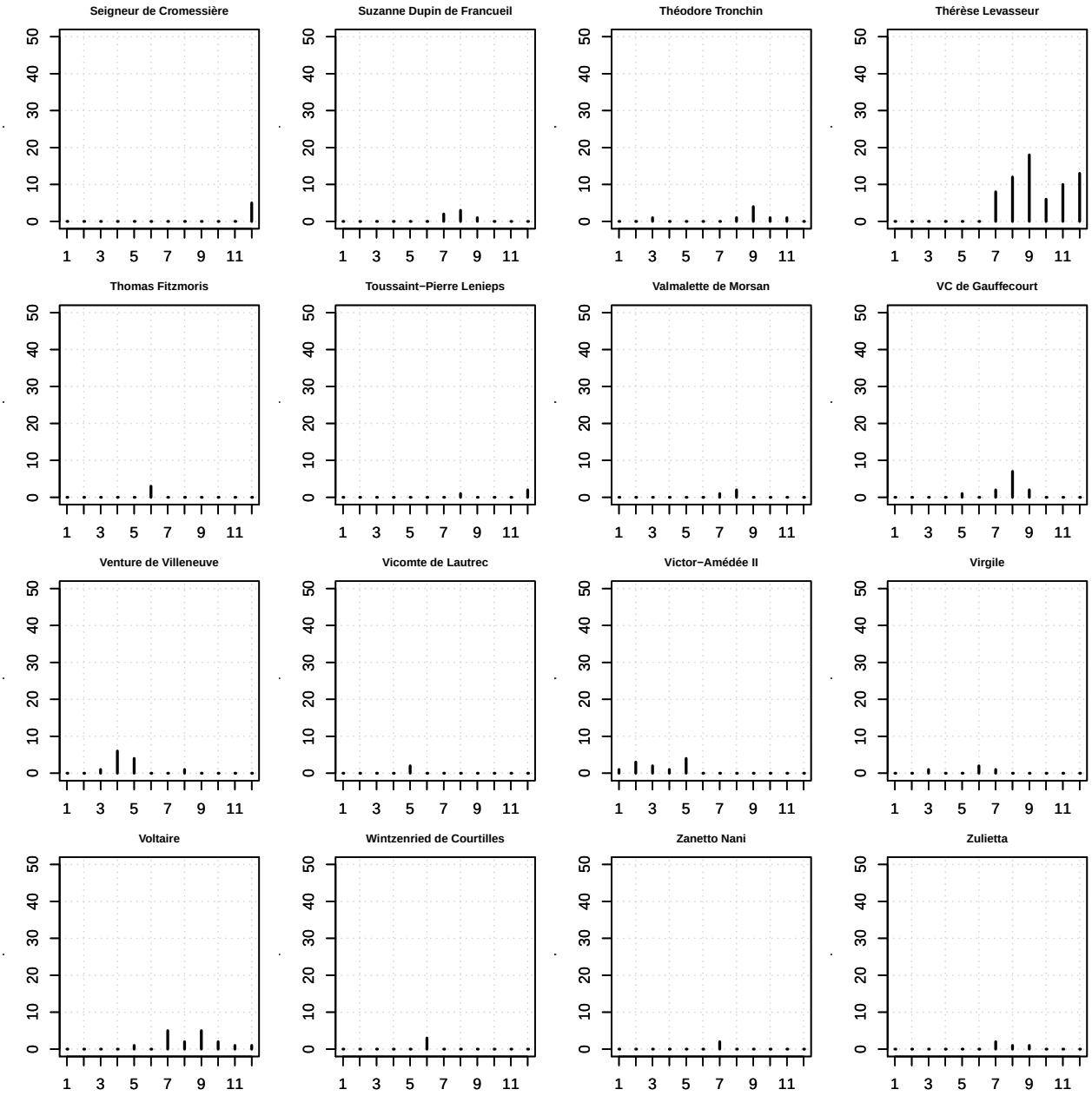

Isabelle d'Ivernois
J et JF Barrillot
JA Dupin de Chenonceaux
Jacob Vernes
Jacques Balexert
Jacques−François Deluc
Jacques−Frédéric Martinet
Jacques−Joseph Mathas
JAM Boy de la Tour
JAV de Saint−Ange
Jean Baseilhac
Jean Bonnot de Mably
Jean Jallabert
Jean Le Rond d'Alembert
Jean Leblond
Jean Néaulme
Jean Perdriau
Jean−Antoine de Condillac
Jean−Antoine Palazzi
Jean−Baptiste Gâtier
0 10 20 30 40 50
1 3 5 7 9 11


Jean−Baptiste Salomon
Jean−Baptiste Simon
Jean−Charles Patizel
Jean−Claude Gaime
Jean−François de Dortan
Jean−François Marmontel
Jean−Henri−Samuel Formey
Jean−Jacques Dortous de Mairan
Jean−Jacques Lambercier
Jean−Philippe Rameau
JEP Bâcle
JJA de Chaillou
JLN Le Maître
Joseph−Étienne Bertier
Karl Emmanuel von Graffenried
L AM et MC de Montmollin
LA de Montmorency
La Roche, intendant
LAJ de Rochechouart
Lamy, père Bernard
0 10 20 30 40 50
1 3 5 7 9 11


LDan Lalive de Bellegarde
LDen Lalive de Bellegarde
Lorenzini, et Mme
Louis−Emmanuel Reydelet
Louis−Nicolas Clérambault
Louise d'Épinay
M. de Bonnefond
M. de Cury
M. de La Poplinière
M. et Mme Levasseur
M. Sabran, et Mme
Maison de Solar
Marc−Michel Rey
Marie Fel
Marie Pontal
Marie−Anne de Graffenried
Marion
Marquis de Blégiers Taulignan
Marquis de Broglie
Marquise de Créqui
0 10 20 30 40 50
1 3 5 7 9 11


Marquise de Mirepoix
Marquise de Pompadour
Marquise de Verdelin
Mathieu Buttafoco
Maurice Quentin de La Tour
Michel−Gabriel de Rossillon
Mlle du Du Châtelet
Mlle Gouin
Mme Basile
Mme de Larnage
Mme de Luxembourg
Mme de Warens
Mme Dupin
Mme La Selle
Montesquieu
NCJ Trublet
Nicolas−Bonaventure Duchesne
Niklaus Anton Kirchberger
Noël−Laurent Pissot
Olivet, capitaine
0 10 20 30 40 50
1 3 5 7 9 11


Orlando, chevalier de Lorenzy
PAN de Saint−Cyr
Pasquale Paoli
Pauline−Gabrielle de Breil
PC de La Closure
PdV de Paulmy, comte d'Argenson
Perrotet
Philibert Caton
Pierre Boy de la Tour
Pierre et Marie−Anne d'Eybens
Pierre Guy
Pierre Jéliote
Pierre−Alexandre Du Peyrou
Pietro Basile
Plutarque
Prince de Conti
Roche, maîtres à danser
Rolichon
Saint−Lambert
Samuel de Meuron
0 10 20 30 40 50
1 3 5 7 9 11


Seigneur de Cromessière
Suzanne Dupin de Francueil
Théodore Tronchin
Thérèse Levasseur
Thomas Fitzmoris
Toussaint−Pierre Lenieps
Valmalette de Morsan
VC de Gauffecourt
Venture de Villeneuve
Vicomte de Lautrec
Victor−Amédée II
Virgile
Voltaire
Wintzenried de Courtilles
Zanetto Nani
Zulietta
0 10 20 30 40 50
1 3 5 7 9 11

List of Figures

List of Tables

List of Symbols

Bibliography

Apoorv Agarwal. Social network extraction from texts: A thesis proposal. In *ACL (Student Session)*, pages 111–116, 2011. 60, 72

Apoorv Agarwal, Augusto Corvalan, Jacob Jensen, and Owen Rambow. Social network analysis of Alice in Wonderland. *NAACL-HLT 2012*, pages 88–96, 2012. 17, 57

Ricardo Alberich, José Miro-Julia, and Francesc Rossello. Marvel universe looks almost like a real social network. *arXiv:cond-mat/0202174*, February 2002. 57, 70, 73

Jac. M. Anthonisse. The rush in a directed graph. Technical report, Stichting Mathematisch Centrum, Amsterdam, 1971. 29

Pierre Assouline. *Henri Cartier-Bresson: l'oeil du siècle*. Gallimard, Paris, 2001. 36

Mieke Bal. *Narratologie. Essais sur la signification narrative dans quatre romans modernes*. HES Publishers, Utrecht, 1984. 14, 59, 65, 100

Albert-László Barabási. *Linked: How everything is connected to everything else and what it means*. Plume, 2002. 147

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. 21, 70

Albert-László Barabási and Eric Bonabeau. Scale-free networks. *Scientific American*, 288: 50–59, 2003. 103

Jean Bardet. *Le personnage de roman*. Gallimard, 2007. 42, 57

John. A Barnes. Class and committees in a norwegian island parish. *Human Relations*, 7:39–58, 1954. 25

John A. Barnes and Frank Harary. Graph theory in network analysis. *Social Networks*, 5(2): 235–244, June 1983. 25

Laura Barwick. *The Opinions of Literature Faculty and Students on Back-of-the-Book Indexes in Fiction*. Master thesis, University of North Carolina, Chapel Hill, 2006. 35

Vladimir Batagelj, Andrej Mrvar, and Matjaž Zaveršnik. Network analysis of texts. In *Third Language Technologies Conference: Preprint Series*, volume 40, Ljubljana, Slovenia, 2002. 57, 72

Judy Batchelor. Para-index and anti-index. *The Indexer*, 16(3):194, 1989. 37

Alex Bavelas. A mathematical model for group structures. *Applied Anthropology*, 7(3):16–30, 1948. 28

Alex Bavelas. Communication patterns in task-oriented groups. *Journal of the Acoustical Society of America*, 22:725–730, 1950. 28, 29, 117, 175

Peter S. Bearman and Katherine Stovel. Becoming a nazi: A model for narrative networks. *Poetics*, 27(2–3):69–90, March 2000. 73

Hazel K. Bell. Indexing biographies: lives do bring their problems. *The Indexer*, 16(3):168–172, 1989. 35

Hazel K. Bell. Indexing fiction: a story of complexity. *The Indexer*, 17(4):251–256, 1991. 35, 36, 38, 39

Hazel K. Bell. Indexes as fiction and fiction as paper-chase. *The Indexer*, 20(4):209, 1997. 36

Hazel K. Bell. *Indexers and Indexes in Fact and Fiction*. University of Toronto Press, 2001. 33, 34

Claude Berge. *Théorie des graphes et ses applications*. Dunot, Paris, 1958. 27

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. 112

Paolo Boldi and Sebastiano Vigna. Axioms for centrality. *arXiv:1308.2140 [physics]*, August 2013. 27, 31, 117

John M. Bolland. Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. *Social Networks*, 10(3):233–253, September 1988. 30, 133

Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 2(1):113–120, 1972. 29, 120

Phillip Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92:1170–1182, 1987. 120

Stephen P. Borgatti. Centrality and network flow. *Social Networks*, 27(1):55 – 71, 2005. 31

Stephen P. Borgatti and Martin G. Everett. A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466 – 484, 2006. 30, 31, 102

Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, February 2009. 21, 25, 26, 27

Philip Bradley. Indexes to works of fiction: the views of producers and users on the need for them. *The Indexer*, 16(4):239–248, 1989. 36, 38, 39, 152

Dan Braha and Yaneer Bar-Yam. From centrality to temporary fame: Dynamic centrality in complex networks. *Complexity*, 12(2):59–63, November 2006. 151

Ulrik Brandes. A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology*, 25(2):163–177, 2001. 113

Ulrik Brandes and Thomas Erlebach. Introduction. In *Network Analysis*, volume 3418, pages 1–6. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005. 12

Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Goerke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. Maximizing modularity is hard. *arXiv:physics/0608255*, August 2006. 113

Ulrik Brandes, Sven Kosub, and Bobo Nick. Was messen zentralitätsindizes ? In *Die Integration von Theorie und Methode in der Netzwerkforschung*, pages 33–52. Springer, 2012. 31, 102, 129

Ulrik Brandes, Garry Robins, Ann McCranie, and Stanley Wasserman. What is network science? *Network Science*, 1(01):1–15, 2013. 13, 21, 26

Ronald S. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, MA, 1992. 26

Aaron Clauset, Mark E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6), December 2004. arXiv:cond-mat/0408187. 112

Aaron Clauset, Cosma Rohilla Shalizi, and Mark E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, November 2009. 103

Gábor Csárdi and Tamás Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, http://igraph.org, 2006. 19, 103, 113

Jérôme Cuckier. Making the game of thrones visualization, 2013. URL http://www.jeromecukier.net/blog/2013/05/13/making-the-game-of-thrones-visualization/. Accessed on 21/03/2014. 57

James E. Cutting, Catalina Iricinschi, and Kaitlin L. Brunick. Mapping narrative space in hollywood film. *Projections*, 7(2):64–91, December 2013. 73

Alain Degenne and Michel Forsé. *Les réseaux sociaux*. Armand Colin, Paris, 1994. 103

Keith Devlin. The hidden math behind Alice in Wonderland, March 2010. URL https://www.maa.org/external_archive/devlin/devlin_03_10.html. Accessed on 30/03/2014. 9

Aleksandra Djurica, Felix Oech, and Horia Radu. Measuring and visualizing the increase in narrative complexity in TV series. DH101 final report, EPFL, Lausanne, 2013. 71

Christian Donninger. The distribution of centrality in social networks. *Social Networks*, 8(2):191–203, June 1986. 30

Robin I. M. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493, June 1992. 100

Dennis Duncan. 'As if we were reading a good novel': fiction and the index from Richardson to Ballard. *The Indexer*, 32(1):2–11, March 2014. 36

Micha Elsner. Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 634–644, 2012. 71

David K. Elson. *Modeling narrative discourse*. PhD thesis, Columbia University, New York City, 2012. v, 60, 70

David K. Elson, Nicholas Dames, and Kathleen R. McKeown. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147. Association for Computational Linguistics, 2010. 72

Philippe Ercolessi, Christine Sénac, Hervé Bredin, and Sandrine Mouysset. Vers un résumé automatique de séries télévisées basé sur une recherche multimodale d'histoires. *Document numérique*, 15(2):41–66, 2012. 70

Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Opera Omnia*, 7:128–140, 1736. 22

Sarah Favre. *Social network analysis for automatic role recognition*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2011. 70

Fernando Ferrara. Theory and model for the structural analysis of fiction. *New Literary History*, 5(2):245–268, 1974. 59

Edward M. Forster. *Aspects of the novel*. Harmondsworth, Middlesex: Penguin Books Ltd, 1963. Originally published in 1927. 59, 61

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5):75–174, February 2010. 112

Terrill L. Frantz, Marcelo Cataldo, and Kathleen M. Carley. Robustness of centrality measures under uncertainty: Examining the role of network topology. *Computational and Mathematical Organization Theory*, 15(4):303–328, December 2009. 30

Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1): 35–41, 1977. 30

Linton C. Freeman. Centrality in social networks: conceptual clarification. *Social Networks*, 1 (3):215 – 239, 1978. 30, 102, 117, 124

Linton C. Freeman. *The development of social network analysis: a study in the sociology of science*. Empirical Press, BookSurge, 2004. 25

Linton C. Freeman. Going the wrong way on a one-way street: Centrality in physics and biology. *Journal of Social Structure*, 9, 2008. 30

Noah E. Friedkin. Theoretical foundations for centrality measures. *American Journal of Sociology*, 96(6):1478–1504, 1991. 31

Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991. 93, 162

Gérard Genette. *Discours du récit*. Seuils, 2007. 14, 46, 148

Sebastian Gil, Laney Kuenzel, and Caroline Suen. Extraction and analysis of character interaction networks from plays and movies. CS 224W Final Project Report, Stanford, 2011. 60, 70, 71, 72

Michelle Girvan and Mark E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. 112, 113

Pablo M. Gleiser. How to become a superhero. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(09):P09020, September 2007. 73

Lenn E. Goodman. The translation of Greek materials into Arabic. *The Cambridge History of Arabic Literature: Religion, Learning and Science in the 'Abbasid Period*, pages 477–94, 1990. 34

Martin Grandjean. Comparing the relational structure of the gospels. network analysis as a tool for biblical sciences. Harvard, 2014. Center for Hellenic Studies. Submitted. 57

Mark Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380, 1973. 26

Philippe Hamon. *Le personnel du roman: le système des personnages dans les Rougon-Macquart d'Émile Zola*. Librairie Droz, 1998. 18, 60, 65

Heather A. Harrington, Mariano Beguerisse Díaz, M. Puck Rombach, Laura M. Keating, and Mason A. Porter. Teach network science to teenagers. *arXiv:1302.6567*, 2013. 15

Marina Hennig, Ulrik Brandes, Jürgen Pfeffer, and Ines Mergel. *Studying social networks: a guide to empirical research*. Campus Verlag, September 2012. 21, 25, 30, 31

Cole Heyn. Dracula and the tangled web he weaves: character networks as close reading tools, 2013. URL http://www.hastac.org/blogs/coleheyn/2013/10/15/dracula-and-tangled-web-he-weaves-character-networks-close-reading-tools. Accessed on 11/02/2014. 60, 71, 100

Leslie Hogben, editor. *Handbook of linear algebra*. Chapman and Hall/CRC, 2013. 120

Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, October 2012. 151

Petter Holme and Jari Saramäki. *Temporal networks*. Springer, Berlin, Heidelberg, 2013. 151

Sterling Hutchinson, Vivek Datla, and Max M. Louwerse. Social networks are encoded in language. In *Proceedings of the 34th annual conference of the cognitive science society*, pages 491–496, 2012. 72

Sterling Chelsea Hutchinson. *Language statistics encodes social network information*. PhD thesis, The University of Memphis, 2013. 73

Muhammad Usman Ilyas and Hayder Radha. A KLT-inspired node centrality for identifying influential neighborhoods in graphs. In *44th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–7. IEEE, 2010. 120

Fotis Jannidis. Character. In Peter Hühn, Jan Christoph Meister, John Pier, and Wold Schmid, editors, *The Living Handbook of Narratology*. Hamburg University, Hamburg, 2012. 58, 59, 61

Matthew L. Jockers. *Macroanalysis. Digital methods and literary history*. University of Illinois Press, March 2013. 10, 57

Frédéric Kaplan and Stylianos Stylianou. Semantic text mining in 4 millions press articles, June 2013. URL http://ic.epfl.ch/page-95510-en.html. Accessed on 12/01/2014. 18

Anna Kazantseva and Stan Szpakowicz. Summarizing short stories. *Computational Linguistics*, 36(1):71–109, 2010. 71, 72

Frederick G Kilgour. *The evolution of the book*. Oxford University Press, New York and Oxford, 1998. 33, 34

Hyoungshick Kim and Ross Anderson. Temporal node centrality in complex networks. *Physical Review E*, 85(2), February 2012. 151

G. Kishi. On centrality functions of a graph. In *Graph Theory and Algorithms*, pages 45–52. Springer, 1980. 30

Eric D. Kolaczyk and Gábor Csárdi. *Statistical analysis of network data with R*. Springer New York, 2014. 19

Dirk Koschützki, Katharina Anna Lehmann, Leon Peeters, Stefan Richter, Dagmar Tenfelde-Podehl, and Oliver Zlotowski. Centrality indices. In Ulrik Brandes and Thomas Erlebach, editors, *Network Analysis*, volume 3418, pages 16–61. Springer, Berlin, Heidelberg, 2005a. 29, 31, 123

Dirk Koschützki, Katharina Anna Lehmann, Dagmar Tenfelde-Podehl, and Oliver Zlotowski. Advanced centrality concepts. In Ulrik Brandes and Thomas Erlebach, editors, *Network Analysis*, volume 3418, pages 83–111. Springer, Berlin, Heidelberg, 2005b. 31, 102

David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science*, 323(5915):721, 2009. 21

Harold J. Leavitt. *Some Effects of Certain Communication Patterns on Group Performance*. PhD thesis, Cambridge, MA: Massachussetts Institute of Technology, 1949. 29

Harold J. Leavitt. Some effects of communication patterns on group performance. *Journal of Abnormal and Social Psychology*, 46:38–50, 1951. 28

Philippe Lejeune. *Le pacte autobiographique: Signes de vie*, volume 1. Seuil, 1975. 35, 45

Michel Lemoine. *Index des personnages de Georges Simenon*. Éd. Labor, 1985. 33, 38

Pádraig Mac Carron and Ralph Kenna. Universal properties of mythological networks. *Europhysics Letters*, 99(2):28002, July 2012. 17, 18, 60, 74

Pádraig Mac Carron and Ralph Kenna. Network analysis of the Íslendinga sögur – the sagas of icelanders. *The European Physical Journal B*, 86(10):1–9, October 2013. 17, 18, 60, 70, 74

Maureen MacGlashan. Wot, no index? or the death of the 'Washington read'. *The Indexer*, 28 (1):18–19, 2010. 35, 39

Inderjeet Mani. *Computational modeling of narrative*. Morgan & Claypool, San Rafael, Calif., 2013. 72

Jacques Maniez and Dominique Maniez. *Concevoir l'index d'un livre: histoire, actualité, perspectives*. ADBS, 2009. 33, 34, 36

Pierre Mercklé. *Sociologie des réseaux sociaux*. La Découverte, Paris, 2011. 25

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, January 2011. 10

Stanley Milgram. The small-world problem. *Psychology today*, 1(1):61–67, 1967. 26

Lisa Mirabile. *Origins and objectives of an index to Michael Ondaatje's The English Patient*. Final Project for L.S. 419, Indexing for Information Retrieval, Simmons GSLIS, 1997. URL http://www.birchile.com/index_report.html. Accessed on 06/03/2014. 36

Mark S. Mizruchi, Peter Mariolis, Michael Schwartz, and Beth Mintz. Techniques for disaggregating centrality scores in social networks. *Sociological Methodology*, 16:26–48, 1986. 30

Franco Moretti. Conjectures on world literature. *New Left Review*, 1:54–68, 2000. 10

Franco Moretti. *Graphs, maps, trees: abstract models for a literary history*. Verso, 2005. 10

Franco Moretti. Network theory, plot analysis. *New Left Review*, 68:80–102, April 2011. v, 10, 11, 15, 17, 57, 60, 62, 64, 99, 100, 105, 123, 124, 147, 176

Franco Moretti. *Distant reading*. Verso, 2013a. 18, 146, 152

Franco Moretti. 'Operationalizing'. *New Left Review*, 84:103–119, December 2013b. 62, 63, 64, 66, 147

Robert L. Moxley and Nancy F. Moxley. Determining point-centrality in uncontrived social networks. *Sociometry*, 37(1):122–130, March 1974. 30

Marvin Mudrick. Character and event in fiction. *Yale Review*, 50(1):1, 1961. 58

Eric T. Nalisnick and Henry S. Baird. Character-to-Character Sentiment Analysis in Shakespeare's Plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 2013a. 60

Eric T. Nalisnick and Henry S. Baird. Extracting Sentiment Networks from Shakespeare's Plays. In *12th International Conference on Document Analysis and Recognition*, pages 758–762, 2013b. 70

Mark E. J. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54, January 2005. 31, 110

Mark E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, May 2006. 113

Mark E. J. Newman. *Networks : An Introduction*. Oxford University Press, 2010. 27, 108

Mark E. J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004. 70, 111, 112

Mark E. J. Newman, Albert-László Barabási, and Duncan J. Watts. *The structure and dynamics of networks*. Princeton University Press, April 2006. 120

D. Oelke, D. Kokkinakis, and D. A. Keim. Fingerprint matrices: Uncovering the dynamics of social networks in prose literature. In *Eurographics Conference on Visualization (EuroVis)*, volume 32. Blackwell Publishing, 2013. 72

João Gama Oliveira and Albert-László Barabási. Human dynamics: Darwin and Einstein correspondence patterns. *Nature*, 437(7063):1251–1251, October 2005. 21, 70

Oulipo. *La littérature potentielle*. Gallimard, Paris, May 1988. 9

Michael Ovelgönne and Andreas Geyer-Schulz. An ensemble learning strategy for graph clustering. In *Graph Partitioning and Graph Clustering*, pages 187–206, 2012. 112

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web, November 1999. URL http://ilpubs.stanford.edu:8090/422/. 31

Raj Kumar Pan and Jari Saramäki. Path lengths, correlations, and centrality in temporal networks. *Physical Review E*, 84(1), July 2011. 151

Gyeong-Mi Park, Sung-Hwan Kim, and Hwan-Gue Cho. Structural analysis on social network constructed from characters in literature texts. *Journal of Computers*, 8(9), September 2013. 70

J. G. Patterson. *A Zola Dictionary: The Characters of the Rougon-Macquart Novels of Emile Zola*. Georg Olms Verlag, January 1973. 38

Jean-François Perrin. *Les Confessions de Jean-Jacques Rousseau*. Gallimard, Paris, 1997. 43, 44, 45

Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *arXiv:physics/0512106*, December 2005. 112

Mason A. Porter, Jukka-Pekka Onnela, and Peter J. Mucha. Communities in networks. *Notices of the AMS*, 56(9):1082–1097, 2009. 111

R Core Team. R: A language and environment for statistical computing, 2013. URL http://www.R-project.org/. 19, 130

Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106, 2007. 112

Yves Reuter. L'importance du personnage. *Pratiques*, 60:3–22, 1988. 123

Shlomith Rimmon-Kenan. *Narrative fiction: Contemporary poetics*. Psychology Press, 2002. 14, 58, 59, 61, 62, 65

Yannick Rochat. Closeness centrality extended to unconnected graphs : the harmonic centrality index. In *Applied Social Network Analysis*, Zürich, August 2009. 117

Yannick Rochat and Frédéric Kaplan. Character networks in les confessions from jean-jacques rousseau. In *Texas Digital Humanities Conference*, Houston, Texas, USA, 2014. http://www.txdhc.org/character-networks-and-centrality/ Accessed on 25/11/2014. Winner of graduate travel bursary and best paper award. 141

Jean-Jacques Rousseau, Bernard Gagnebin, and Marcel Raymond. *Les Confessions*. Gallimard, Paris, 2009. 43

Jean-Jacques Rousseau, Jacques Voisine, Jacques Berchtold, and Yannick Séité. *Les Confessions*. Classiques Garnier, Paris, 2011. 14, 43, 47, 51

Jean-Jacques Rousseau, Raymond Trousson, and Frédéric S. Eigeldinger. *Oeuvres complètes en 24 volumes*. Slatkine, Genève, Paris, May 2012. v, 14, 19, 42, 48, 49, 51, 52, 53, 93, 156, 176

Britta Ruhnau. Eigenvector-centrality — a node-centrality? *Social Networks*, 22(4):357–365, October 2000. 31

Jeff Rydberg-Cox. Social networks and the language of greek tragedy. In *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, volume 1, 2011. 73

Gert Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, December 1966. 27, 29

Graham A. Sack. Simulating plot: Towards a generative model of narrative structure. In *2011 AAAI Fall Symposium Series*, November 2011. 72, 99

Graham A. Sack. Character networks for narrative generation. In *Eighth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2012. 60

Graham A. Sack. Character networks for narrative generation: Structural balance theory and the emergence of proto-narratives. In *2013 Workshop on Computational Models of Narrative*, volume 32, pages 183–197, 2013. v

Andrew Salway, Bart Lehane, and Noel E. O'Connor. Associating characters with events in films. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 510–517. ACM, 2007. 71

Catherine Sassen. Biography indexes reviewed. *The Indexer*, 30(3):136–140, 2012. 35

Stephen B Seidman. Network structure and minimum degree. *Social networks*, 5(3):269–287, 1983. 152

Jane Smiley, editor. *The Sagas of Icelanders*. Penguin Books, 2001. 38

Richard Jean So and Hoyt Long. Network analysis and the sociology of modernism. *boundary 2*, 40(2):147–182, June 2013. 57

Jean Starobinski. *Jean-Jacques Rousseau: la transparence et l'obstacle suivi de sept essais sur Rousseau*. Gallimard, 1991. 45

Einar Stensson. Network assignment. Technical report, Department of Sociology, Stockholm University, February 2014. URL http://einarstensson.se/files/GoTNetPaper.pdf. Accessed on 20/03/2014. 71

James Stiller and Mathew Hudson. Weak links and scene cliques within the small world of Shakespeare. *Journal of Cultural and Evolutionary Psychology*, 3(1):57–73, 2005. 70, 73

James Stiller, Daniel Nettle, and Robin I. M. Dunbar. The small world of Shakespeare's plays. *Human Nature*, 14(4):397–408, 2003. 70, 72, 73

Saatviga Sudhahar, Gianluca De Fazio, Roberto Franzosi, and Nello Cristianini. Network analysis of narrative content in large corpora. *Natural Language Engineering*, pages 1–32, September 2013. 72

M. S. A. Tan, E. A. Ujum, and K. Ratnavelu. A character network study of two sci-fi TV series. In *AIP Conference Proceedings*, pages 246–251, 2014. 72

John Tang, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Vincenzo Nicosia. Analysing information flows and key mediators through temporal centrality metrics. In *Proceedings of the 3rd Workshop on Social Network Systems*, page 3. ACM, 2010. 151

Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969. 26

Raymond Trousson and Frédéric S. Eigeldinger. Introduction générale. In *Oeuvres complètes en 24 volumes*, volume 1. Slatkine, Genève, Paris, May 2012. 45, 46, 47, 50

John Truby. *The Anatomy of Story*. New York: Faber and Faber, 2007. 55, 56

Frank Tutzauer. Entropy as a measure of centrality in networks characterized by path-transfer flow. *Social Networks*, 29(2):249–265, May 2007. 31

Thomas Valente, Kathryn Coronges, Cynthia Lakon, and Elizabeth Costenbader. How correlated are network centrality measures? *Connections*, 28(1):16–26, 2008. 30, 31

Blakey Vermeule. *Why Do We Care about Literary Characters?* JHU Press, February 2011. 60

Jean-Jacques Voisine. Introduction. In *Les Confessions*. Garnier, Paris, 2011. 47

Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, November 1994. 103

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, June 1998. 21, 70

Hans H. Wellisch. The oldest printed indexes. *The Indexer*, 15(2):73–82, 1986. 34

Chung-Yi Weng, Wei-Ta Chu, and Ja-Ling Wu. RoleNet: Movie analysis from the perspective of social networks. *IEEE Transactions on Multimedia*, 11(2):256–271, February 2009. 71, 72

Martin L. White. Indexing lives. In *Index it right: Advice from the experts, vol.1*, pages 31–47. Information Today, Inc., 2005. 39

Harry Wiener. Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69(1):17–20, 1947. 117

Alex Woloch. *The One Vs. the Many: Minor Characters and the Space of the Protagonist in the Novel*. Princeton University Press, 2003. v, 9, 10, 62, 63, 64, 66, 67, 68, 69, 99, 139, 142, 146, 147, 149

Émile Zola. *La fortune des Rougon*. Librairie générale française, Paris, 1985. 8

List of publications

Journal Papers

Yannick Rochat and Frédéric Kaplan, Analyse des réseaux de personnages dans Les Confessions de Jean-Jacques Rousseau. *Les Cahiers du Numérique*, vol. 10, num. 3, p. 109-133, 2014.

Jorge Peña and Yannick Rochat, Bipartite Graphs as Models of Population Structures in Evolutionary Multiplayer Games. *Plos One*, 2012.

Conference Papers

Yannick Rochat, Mélanie Fournier, Andrea Mazzei and Frédéric Kaplan, A Network Analysis Approach of the Venetian Incanto System. *Digital Humanities*, Lausanne, Switzerland, July 7-12, 2014.

Mélanie Fournier and Yannick Rochat, Modeling Venice's maritime network - End 13th to Mid. 15th centuries. *International Workshop ERC World Seastems - Maritime Networks in Space and Time*, Paris, France, June 16-18, 2014.

Yannick Rochat and Frédéric Kaplan, Character networks in Les Confessions from Jean-Jacques Rousseau. *Texas Digital Humanities Conference*, Houston, Texas, USA, April 10-12, 2014.

Yannick Rochat and Frédéric Kaplan, Analyse de réseaux sur les Confessions de Rousseau. *Humanités délivrées*, Lausanne, Switzerland, October 1-2, 2013.

Yannick Rochat, Frédéric Kaplan and Cyril Bornet, A social network analysis of Rousseau's autobiography "Les Confessions". *Digital Humanities*, Lincoln, Nebraska, USA, July 15-19, 2013.

Yannick Rochat, Martin Grandjean and Vincent Barbay, R, twitteR and igraph. *Just-in Time Sociology*, Lausanne, Switzerland, December 4, 2012.

Jorge Peña and Yannick Rochat, Evolutionary games on 2-mode networks. *International Sunbelt Social Network Conference*, Riva del Garda, Italy, June 29 - July 4, 2010.

Yannick Rochat, David Jilli and Henri Volken, Centrality and community detection in a mobile social network. *6th UK Social Networks Conference*, Manchester, United Kingdom, April 14-16, 2010.

Yannick Rochat, A social network analysis of scientific collaboration in swarm intelligence : a question of centrality. *ASNA*, Zürich, Switzerland, September 15-17, 2010.

Yannick Rochat, Closeness Centrality Extended to Unconnected Graphs: the Harmonic Centrality Index. *ASNA*, Zürich, Switzerland, August 26-28, 2009.

Master thesis

Yannick Rochat, *Géométrie de l'espace hyperbolique dans le modèle du demi-espace supérieur*. Sous la direction de Klaus-Dieter Semmler, September 2007.