Toward a new level of modeling of environmental effects on galaxies

Manuel Duarte

▶ **To cite this version:**

Manuel Duarte. Toward a new level of modeling of environmental effects on galaxies. Astrophysics [astro-ph]. Observatoire de Paris, 2014. English. NNT : . tel-02095295

HAL Id: tel-02095295
https://hal.science/tel-02095295

Submitted on 10 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THESIS

...ysique de Paris

Toward a new level of modeling of environmental effects in galaxies

École doctorale
Astronomie et Astrophysique d'Île-de-France

OBSERVATOIRE DE PARIS
9180
BIBLIOTHEQUE

PhD thesis
Astronomie et Astrophysique

DUARTE MANUEL

Director:
MAMON GARY

Jury:

VALLS-GABAUD DAVID	President
DE CARVALHO REINALDO	Reader
MARINONI CHRISTIAN	Reader
BIVIANO ANDREA	Examinator
DURRET FLORENCE	Examinator

Last modified: November 5, 2014

THESIS

Abstract

Galaxies lie in a large panel of environments from isolated galaxies, to pairs, groups or clusters. The environment is expected to have an impact on galaxy properties such as morphology, stellar formation, metallicity... Some studies already tried to quantify the importance of the global environment (linked to the dark matter halo mass) and the local environment (galaxy position in the group). These studies have shown that the environment plays a minor role except for low mass galaxies. But the quantification of the environment is difficult since detected groups in redshift space (the only one accessible by the observer) are very elongated, making it difficult to extract spherical groups in real space. If these quantification errors are too important, environment effects will not be measured correctly.

Moreover, other physical processes are at work inside groups whose relative roles are not well understood. For example, major or minor mergers (rich or poor in gas, between satellite galaxies, or after the decay of the orbit of a satellite onto the central galaxy by dynamical friction), rapid flybys harassing galaxies, stripping of the interstellar gas by ram pressure or of the gaseous reservoir by tidal forces. Although semi-analytical codes of galaxy formation from initial conditions of a ΛCDM Universe fit well a large set of observed relations, there are still some discrepancies that might be possibly explained by a lack of correct physical recipes of environmental effects in these models.

Our goal with this thesis is to have a detailed comprehension of the role of environment on galaxy properties, and finally determine the major physical processes in the modulation of these properties with both local and global environment. For this, an optimal extraction of galaxy groups from the projected phase space is necessary.

We performed a study and re-implementation of some existing group finder to estimate their strengths and weaknesses in the detection of galaxy groups.

A galaxy mock catalogue in redshift space, designed to mimic the primary spectroscopic sample of the SDSS survey was created to apply several galaxy group algorithms. An advantage is the already known membership that we can compare to galaxy groups extracted from redshift space. Semi-analytical codes of galaxy formation give us such galaxy catalogs we transformed to be coherent with the vision of an observer.

With these mock catalogues, we tested the very popular Friends-of-Friends grouping algorithm. We determined the optimal linking lengths against the set of tests and optimal criterion we developed to judge the efficiency of an algorithm. It appears that this choice of linking lengths depends on the scientific goal to do with the group catalogue.

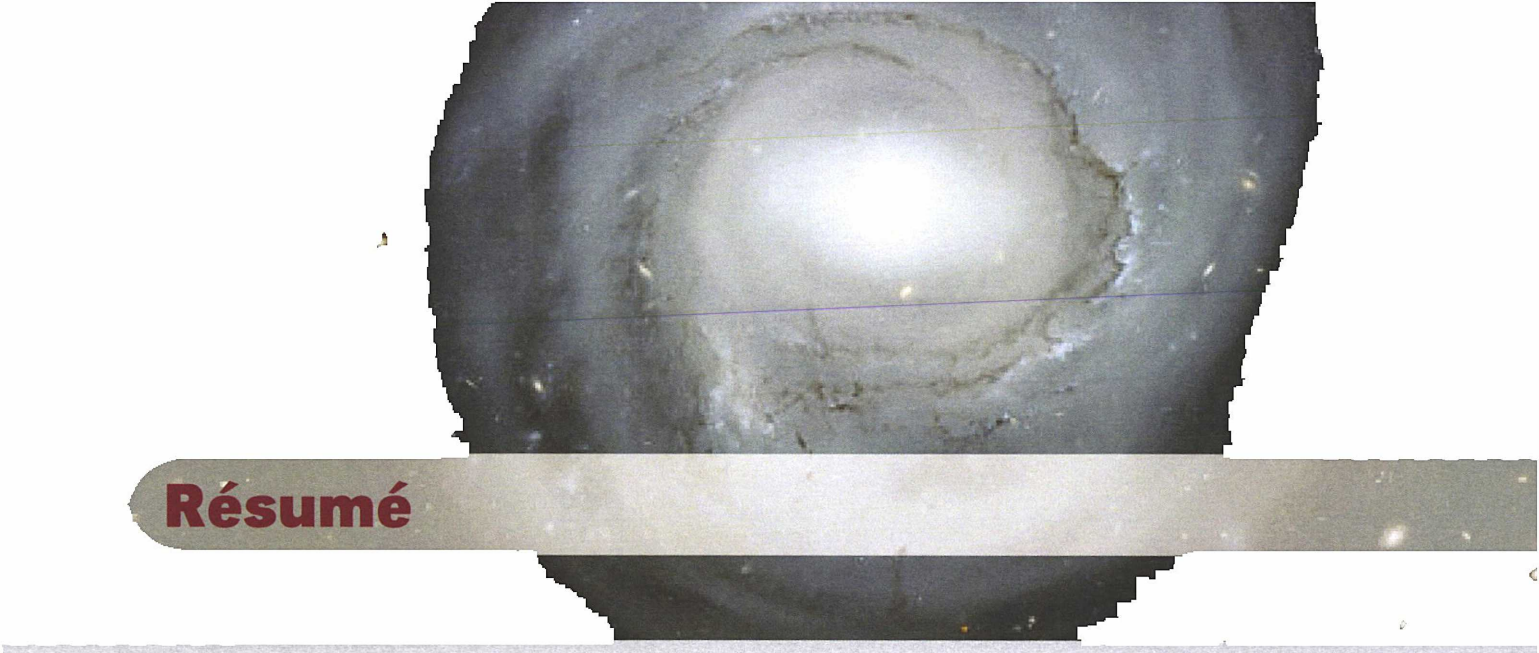
A large part of the thesis consisted on the realization of a new grouping algorithm called MAGGIE (Models and Algorithm for Galaxy Groups, Interlopers and Environment), Bayesian and probabilistic. MAGGIE uses our priors acquired with analysis of cosmological simulations for large scale structure and of observations obtained from large galaxy surveys, to better constrain the selection of galaxy groups from redshift space. Comparison of MAGGIE with the FoF algorithm shows that MAGGIE is superior in avoiding the fragmentation of real space groups, the membership selection (completeness, reliability) and in the group properties (group mass, luminosity). The better performance of MAGGIE comes from its probabilistic nature, the use of astrophysical and cosmological priors, and the use of halo abundance matching technique linking central galaxy distributions (stellar mass or luminosity) to physical properties of dark matter halos.

The future application of MAGGIE on galaxy surveys such as the Sloan Digital Sky Survey or the deeper Galaxy and Mass Assembly, taking care of their own observational problems, should improve our understanding of the modulation of galaxy properties with their global and local environments and physical processes operating inside galaxy groups.

■

THESIS

THE END



Résumé

Les galaxies reposent dans un large éventail d'environnements allant des galaxies isolées, aux paires, aux groupes ou amas. Il est donc légitime de penser que cet environnement peut influer sur les différentes propriétés des galaxies comme la morphologie, la formation stellaire, la métallicité, etc. Des études ont déjà tenté de quantifier les rôles de l'environnement global (lié à la masse du halo de matière noire du groupe) et de l'environnement local (la position de la galaxie dans le groupe). Elles ont montré que l'environnement joue un rôle mineur dans leurs propriétés excepté pour les galaxies de faible masse. Mais la quantification de l'environnement est difficile car les groupes détectés dans l'espace des redshifts (seul accessible à l'observateur) sont très allongés et ne facilitent donc pas la recherche de l'appartenance d'une galaxie à un groupe donné. Si ces erreurs de quantification sont trop importantes, les effets de l'environnement seront alors mal mesurés.

De plus, d'autres processus physiques sont à l'œuvre dans les groupes dont l'importance n'est pas tout à fait comprise. Par exemple les fusions majeures ou mineures des galaxies (riches ou pauvres en gaz, entre galaxies non centrales, ou entre une centrale et une non centrale par "déclin" de son orbite après friction dynamique), les survols rapides qui arrachent du gaz aux galaxies, le dépouillement du gaz interstellaire par la pression du gaz intra-groupe ou intra-amas, ou de celui du réservoir de gaz qui forme les disques des galaxies par des effets de marées. Bien que les modèles semi-analytiques de formation des galaxies à partir de conditions initiales d'un Univers ΛCDM représentent assez bien les observations faîtes sur les galaxies, il y a toujours des écarts qui peuvent être sûrement liés à un manque de prise en compte des effets d'environnement dans ces modèles.

On vise donc avec cette thèse à avoir une compréhension détaillée du rôle de l'environnement sur les propriétés des galaxies et finalement connaître le ou les processus physiques qui ont une importance prépondérante dans la modulation de ces propriétés avec l'environnement local et global. Pour cela, il est nécessaire de réaliser une extraction optimale des groupes de galaxies depuis l'espace des phases projeté.

Une étude et ré-implémentation de certains algorithmes de regroupement de galaxies déjà existants a été réalisée pour déterminer leur efficacité et leurs faiblesses dans la détection des groupes de galaxies.

Un catalogue de galaxies test (mock catalogue) a été réalisé pour appliquer nos divers algorithmes de regroupement sur un échantillon de galaxies certes fictif, mais avec des propriétés physiques semblables (fonction de luminosité, profil de densité des galaxies dans les groupes, biais liés au décalage vers le rouge comme indicateur de distance,...). L'avantage est que l'appartenance d'une galaxie à un groupe donné est connue à l'avance et que l'on peut donc comparer les sélections faîtes par les algorithmes à cette "réalité". Les sorties de codes semi-analytiques de formation de galaxies fournissent de tels catalogues que nous avons transformés pour convenir au point de vue d'un observateur.

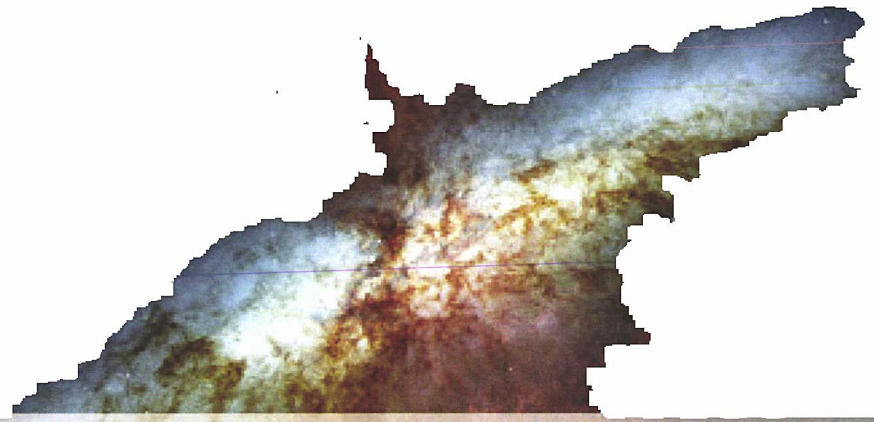
THESIS

Avec des mocks catalogues à notre disposition, nous avons pu tester et comparer divers algorithmes de regroupement à un même échantillon de galaxies et avoir une idée de leurs performances de manière quantitative et non seulement qualitative. Nous nous sommes intéressés au plus populaire algorithme de regroupement qu'est la méthode de la percolation ou algorithme amis d'amis (Friends-of-Friends, FoF ci-après). Nous avons déterminé le jeu de paramètres de liens optimums pour la sélection de groupes de galaxies avec un ensemble de tests et de critères optimaux, que nous avons développé, pour juger de l'efficacité d'un algorithme de groupes de galaxies. Il est également apparu que le choix des paramètres de liens à considérer pour un FoF dépend beaucoup de la science que l'on souhaite réaliser avec notre catalogue de groupes.

Une partie de la thèse a consisté à réaliser un tout nouvel algorithme de regroupement nommé MAGGIE (Models and Algorithm for Galaxy Groups, Interlopers and Environment), bayésien et probabiliste. MAGGIE utilise les a priori acquis à l'aide des analyses des simulations cosmologiques sur les structures à grandes échelles et les observations obtenues à partir des larges surveys sur les galaxies pour mieux contraindre la sélection des groupes de galaxies à partir de l'espace des phases projeté (biaisé par la distorsion des groupes liée au décalage vers le rouge). Les résultats de la comparaison de MAGGIE avec l'algorithme de FoF ont montré que, bien qu'équivalent dans la capacité à retrouver les galaxies membres des groupes (complétude et fiabilité), MAGGIE est bien meilleur dans l'estimation des propriétés des groupes (masses stellaires, luminosités...) grâce à la probabilité d'appartenance qui réduit l'importance des galaxies non réellement membres du groupe (interlopers). MAGGIE réduit significativement la fraction de fausses détections de groupes de galaxies, c'est-à-dire de groupes sporadiques, issus de la fragmentation par les algorithmes d'un groupe réel en plusieurs sous-groupes. L'estimation de l'environnement global est également améliorée grâce à la méthode de correspondance d'abondance (abundance matching) qui compare et lie les distributions des masses stellaires des galaxies centrales des groupes aux propriétés physiques des halos de matière noire pour une meilleure précision dans l'estimation de la masse virielle des groupes de galaxies.

Une future application de MAGGIE sur des surveys de galaxies tels que le Sloan Digital Sky Survey ou le Galaxy and Mass Assembly, en tenant compte de tous les problèmes liés aux observations de chacun d'eux, devrait nous permettre par la suite d'améliorer notre compréhension des processus physiques dans les groupes de galaxies.

THESIS



Contents

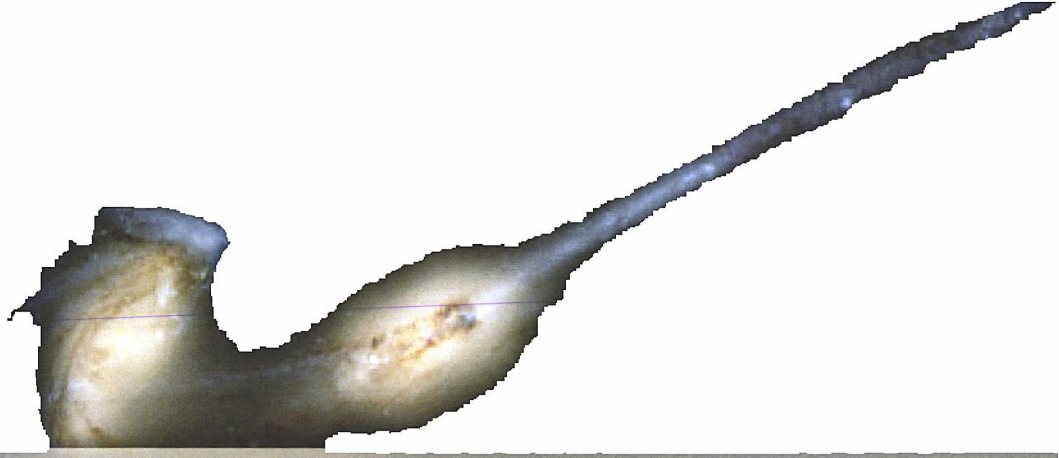
TESIS

THESIS

TESIS

THESIS

Chapter 1.



Introduction

Contents

Since the discovery of galaxies as distant objects from the Milky Way (Hubble, 1929), much work has been done to understand how they formed and what drives their observable properties (morphologies, color...) at our epoch and earlier in their evolution (Benson, 2010; Silk et al., 2013; Silk & Mamon, 2012). The combination of the structure formation of cold dark matter (CDM) particles and their history (Zentner, 2007), with the baryon physics inside dark matter halos (Kravtsov & Borgani, 2012) has been quite successful in reproducing and explaining the observations from galaxy surveys. But there are still some lacks in the galaxy formation scenario, which are headaches to solve for theorists (Weinmann et al., 2012). A frequent solution to resolve this puzzle is to introduce different recipes in galaxy formation simulations to account for the missing physics in the scenario. Such an example (and the most known problem of ΛCDM) is the overabundance of dwarf galaxies predicted by semi-analytical models (SAM) in simulations of galaxy formation. Reducing their number implies the ejection excessive baryons through several physical processes (feedback, e.g. Brooks et al. (2013), see Silk & Mamon (2012) for a review) in order to make the dwarfs not resolvable. Such typical processes are, for example, supernovae winds (Dekel & Silk, 1986; Hirschmann et al., 2013) or ram pressure stripping (Gunn & Gott, 1972). But introducing them leads to a more and more complex scenario, and doesn't allow to clearly distinguish the effect of each physical process on the galaxy evolution.

The formation and evolution of galaxies is expected to be tightly correlated to the galaxy environment. Indeed, galaxies are gregarious, living in different hosts environments from isolated galaxies, to pairs, groups, clusters and super clusters. This environment impacts on galaxy properties in different manner, at different epochs, through several physical processes. But not all them are important according to the redshift and environment of galaxies. The characterization of the major physical process at work inside environments should improve the predictions of semi-analytical models of galaxy formation and evolution (SAMs), by including more precise models and recipes in the code, directly extracted from the analysis of the observations. Moreover, this should also improve the galaxy formation scenario constructed until now, and work as a test

THESIS

for this scenario. This goal can only be achieved with an optimal selection of galaxy group and clusters.

1.1 Galaxy formation

The large scale structure of the Universe, observed in both the sky and in numerical simulations, is usually probed through galaxies and their content since the dark matter only interacts gravitationally with "ordinary" (baryonic) matter. At this time, the commonly accepted scenario for the formation of large scale structure is the hierarchical model, where small structures are created early in the history of the Universe and then merged to become more massive. This is the ΛCDM paradigm (CDM for cold dark matter): Universe is in expansion by action of dark energy (the Λ term) and structures appear through the gravitational interactions of cold dark matter, in opposition to hot dark matter where the intrinsic velocities of dark matter avoid the formation of early small structures. Then, baryons, visible and non-dominant fraction of matter, collapses inside dark matter halos, cools and forms stars.

If this process goes without nothing to stop it, the mass of galaxies should increases without limits: this is the overcooling problem (Blanchard et al., 1992; White & Rees, 1978). But baryons are not only submitted to gravitation and several processes can prevent the star formation inside such structures. At the two extremes of the halo mass function, the gas is prevented from fragmenting into stars by heating processes, avoiding the cooling of the gas to the center of the potential well of dark matter structures. This heating can be intrinsic to the gas in the halo because of the photo-ionization (Rees, 1986) or due to a pre-heating of the gas before it enters the halo (Borgani et al., 2001), acting essentially for low mass halos. Supernova explosions have also a contribution to the re-heating of the gas (Dekel & Silk, 1986; Efstathiou, 2000) for low and intermediate masses. In high mass halos, the cooling is less efficient but a large quantity of gas can still cool to form very massive galaxies. Material ejected by active galactic nuclei (AGN) is possibly an explanation for heating gas (Silk & Rees, 1998), although the mechanism through which AGN operates is not well understood.

Figure 1.1: Illustration of the ram pressure stripping experienced on a galaxy, whose interstellar gas in moved, quenching the star formation since the "fuel" of this process is dropped out.

The environment plays an important role. Galaxy over-density in groups leads to several physical processes, caused by interactions between each galaxy and/or the group. Galaxy mergers (essentially major mergers involving two galaxies of similar masses) are expected to morphologically transform galaxies to spheroidal (Bournaud et al., 2005; Mamon, 1992; Naab et al., 1999),

and to create bursts of star formation inside merging galaxies (Cox et al., 2008; Teyssier et al., 2010). On the other hand, the group environment acts too on galaxy properties. Tidal forces exerted by the group and the ram pressure stripping can remove the outer gaseous regions in orbiting galaxies leading to a quenching of the star formation (Bekki, 2014; Larson et al., 1980).

Some of these intra-group physics were already, more or less well, introduced in SAMs (Font et al., 2008; Guo et al., 2011; Lanzoni et al., 2005; Okamoto & Nagashima, 2003). But all these methods tend to over-simplify, by use of simple formulas, very complex processes depending on several parameters and the galaxy environment. A better modeling of the physics involved in galaxy group should improve the SAM and correct their difficulties in fully describing the observed Universe. This can only be achieved by optimally extracting and measuring galaxy groups from redshift space galaxy catalogs.

1.2 The importance of galaxy groups

1.2.1 Galaxy group physics

Observed galaxy groups are a direct consequence of the hierarchical growth of structure. Galaxies therein are affected by this growth since they formed in dark matter sub-halos that merged with most massive halos along the Universe expansion according to the hierarchical scenario (Lacey & Cole, 1993). So their properties must be correlated with their parent dark matter halo and reflect the history of the processes acting on it. Some evidence of such a modulation of galaxy properties with galaxy environment were already observed previously on the galaxy luminosity (Robotham et al., 2010) and stellar mass (Yang et al., 2009) functions.

Galaxies can be classified in two distinct classes: a blue class of gas rich and young stellar population and a red one, poor in gas with an old stellar population (Driver et al., 2006). This bi-modality is also visible in their morphologies where red galaxies are essentially ellipsoidal and blue galaxies are spiral. A segregation of these galaxies exists with the environment close to our epoch (low redshifts): red galaxies lie in dense environments such as clusters, while the blue population is more present in the field (outside dense environments as clusters or groups).

But some other properties lead to discrepant results. For example, the fraction of galaxies with large specific star formation rate (SSFR) doesn't show a dependence on the environment for high stellar mass galaxies according to Peng et al. (2010), but following von der Linden et al. (2010), there is clearly a trend of decline of the fraction of high SSFR for star forming galaxies towards groups center (for all galaxy masses). The results of Peng et al. (2010) are surprising since the dense environment is expected to quench the star formation in galaxies. This contradiction is possibly explained by the selection of a tracer for the environment in Peng et al. (2010) that doesn't distinguish between the two kind of environments: the *local* one related to the position of the galaxy relatively to its halo, and the *global* environment that characterizes the total mass embedded in the parent halo of the galaxy. An example is shown in Figure 1.2 where we plot the over-density as defined in Peng et al. (2010) for two different halo masses with a density profile from Navarro et al. (1996) (see Appendix B), a concentration from Macciò et al. (2008), as a function of the position relative to the halo center in units of virial radius. We chose two extremes masses (10^{12} and $10^{15}h^{-1}M_{\odot}$) to have two very distinct halos. The over-density is essentially sensitive to the local environment, but the global one has only a small effect through the concentration parameter. So, Peng et al. (2010)'s measure of environment is essentially a local measure, and can't trace global environment.

THESIS

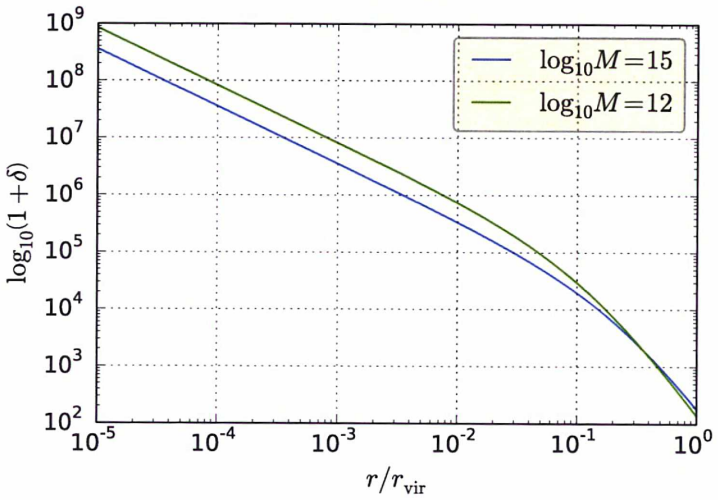


Figure 1.2: The over-density relatively to the mean density for two different halos of mass 10^{12} and $10^{15}h^{-1}M_{\odot}$ as a function of the distance to the halo center in units of virial radius $r_{\rm vir}$. A density profile from Navarro et al. (1996) is assumed with concentrations computed from Macciò et al. (2008).

Remark 1

Assuming the density profile of $p(r)$ Navarro et al. (1996), the over-density δ is:

$$\delta = \frac{\rho(r) - \rho_m}{\rho_m} \tag{1.1}$$

Using equations from Appendix B, and writing the mean density of the Universe as $\rho_m = \Omega_m \rho_c$ where Ω_m is the density fraction of matter in the Universe and ρ_c is the critical density equal to $3H_0^2/(8\pi G)$, we finally have:

$$\delta = \frac{\Delta\overline{\rho}(r/r_{\rm vir})}{3\Omega_m} - 1 \tag{1.2}$$

with Δ the value of the density in units of the critical density used to defined a halo relatively to the background, and $\overline{\rho}$ the normalized density profile as defined in Appendix B. ■

1.2.2 Galaxy groups as tests

Galaxy groups are not just limited to test and improve the models for the galaxy formation theory, but also appear in other astrophysical domains. In cosmology, they are a tool to access the cosmological parameters (Wang & Steinhardt, 1998). General relativity can be tested with them (Wojtak et al., 2011).

Unfortunately, a clean characterization of the environment from the redshift space is difficult since the redshift distortions (Jackson, 1972), called also *Fingers-of-God* (Tully & Fisher, 1978), caused by the velocity dispersion of the galaxy group can create overlapping between galaxies of foreground or background groups. But the over-density used in Peng et al. (2010) is computed from galaxy nearest neighbors in redshift space, clearly affected by interlopers because of projection effects.

1.3 Characterizing the environment

1.3.1 History

Many galaxy group catalogs were already published, usually following the first publications of data from galaxy surveys. First attempts were done with visual selections (Abell, 1958; Rose, 1976; Zwicky et al., 1961). The selection was based on a criteria for a visual over-density of galaxies.

Then the percolation or Friends-of-Friends (FoF) algorithm followed (Huchra & Geller, 1982; Nolthenius & White, 1987). One of its advantages is that it is based on a physical choice for the way to link galaxies between them in groups. A linking length is used to relate to galaxies that are closer than this distance in redshift space. The FoF algorithm requires two different linking lengths in the line-of-sight and perpendicular (plane of sky) directions to avoid the redshift distortion effect. Eke et al. (2004) and Berlind et al. (2006) published group catalogs from the application of the FoF algorithm, but taking into account, in their selection, the incompleteness induced by the galaxy surveys used.

Marinoni et al. (2002) developed a method similar to FoF but with the use of a redshift space partitioned into Voronoi cells, to have an initial seed for the over-density (Voronoi cells volume trace the galaxy density) around each galaxy. But this method suffers from the necessity to use it in small surveys in angle because of the difficulty to create a tessellation of the celestial sphere directly.

With the increasing advances in our understanding of galaxy formation processes, capacities of numerical computation and predictions of the cosmological simulations, started to appear Bayesian algorithms that used priors on galaxy groups to improve their extraction from galaxy surveys. Yang et al. (2005, 2007) developed an iterative method to select galaxy groups based on a density contrast criterion, which uses assumptions based on cosmological simulation results for the density profile of groups.

Galaxy surveys have limitations that are difficult to overcome in galaxy group algorithms. In the case of photometric redshifts surveys, probabilistic Friends-of-Friends were developed to attempt avoiding the large (and sometimes catastrophic) uncertainties in redshift measures (Liu et al., 2008). Then, probability was used to improve the membership of galaxies inside their groups, as in Domínguez Romero et al. (2012), allowing a soft affectation of galaxies to groups.

Finally, group finding algorithms continue their insertion of galaxy formation results, combining it with the advantage of geometrical methods. Muñoz-Cuartas & Müller (2012) used a FoF applied on dark matter halos associated to galaxies, with the initial assumption that all galaxies are their own halo, and so the central galaxy mass is a tracer of the density field (the most massive central galaxies are associated to the most massive halos).

1.3.2 And now...?

Current and future generations of galaxy surveys allow us to probe galaxy groups in different aspects, each of them with their improvements and limits. The Sloan Digital Sky Survey (SDSS), with around one million of spectroscoped galaxies, gives us a good overview of the density field for a large range of redshifts. But this abundance of precise redshifts as the counterpart that not all galaxies have spectroscopic redshifts, and around 5–10% of galaxies, because of the fiber collision problem (Blanton et al., 2003), need to fall back to photometric redshifts, more inaccurate. The Galaxy And Mass Assembly, at its final stage, will contain around 300 000 galaxies with a spectroscopic redshift (Hopkins et al., 2013), less than the SDSS. But the completeness of the sample will be higher than the SDSS with $\simeq 99\%$ of the sample spectroscoped and it is two magnitudes deeper. The counterpart is a less precise measurement of galaxies recession velocities (Hopkins

THESIS

et al., 2013; Robotham et al., 2011). Moreover, the adjoining angular coverage is lower because of the fragmentation of the survey regions. In consequence, galaxy group algorithms must be sufficiently flexible to be applied to and give the same results in many, different and (surprisingly) creative future galaxy survey projects. Their common limitations and advantages must be taken into account when developing it.

So we need to go beyond the usual standard and static definition of groups and work with the inevitable polluted environment of extracted galaxy groups to have a precise understanding of the major physical processes at work inside galaxy groups. We start by an overview of some common grouping algorithms, their innovations and limitations in Chapter 2. Since such algorithms must be tested in order to access their capacities in recovering the clustering from redshift space, we detailed the construction of a galaxy mock catalogue, difficulties inherent to its creation and biases introduced voluntary or not in Chapter 3. We were also interested in the most popular algorithm that is the Friends-of-Friends or percolation algorithm and performed a detailed test on its performances in Chapter 4. We present and test MAGGIE, a probabilistic Bayesian galaxy group algorithm that reduces the effects of interlopers in the galaxy group properties observed in Chapter 5. In Chapter 6, we describe our analysis of the Sloan Digital Sky Survey in the goal of a future application of MAGGIE on its database.

THESIS

Chapter 2.

Grouping algorithms

Contents

As previously discussed, a good characterization of the galaxy environment implies a good selection of galaxy groups. But galaxy observations are made in redshift space, where the velocity dispersion of galaxies inside clusters stretch the line-of-sight distribution of galaxies. Galaxies in a structure are not seen in a local region of the space, but the structure is extended in larger range of redshifts from the projected phase space, which is the only one accessible by an observer. In consequence, the extraction of a galaxy group from the redshift space is complex since a galaxy in the field can be associated to a group if it lies in a similar position on the sky but up to 10–20 virial radii in front or behind. Such a galaxy is called an *interloper* (inside the group by selection, but not pertaining to it in reality).

A large number of group catalogues were constructed with a large panel of methods for handling redshift distortions. A summary of some of such galaxy group algorithms follows, with a description of their strengths and weaknesses.

2.1 Some algorithms

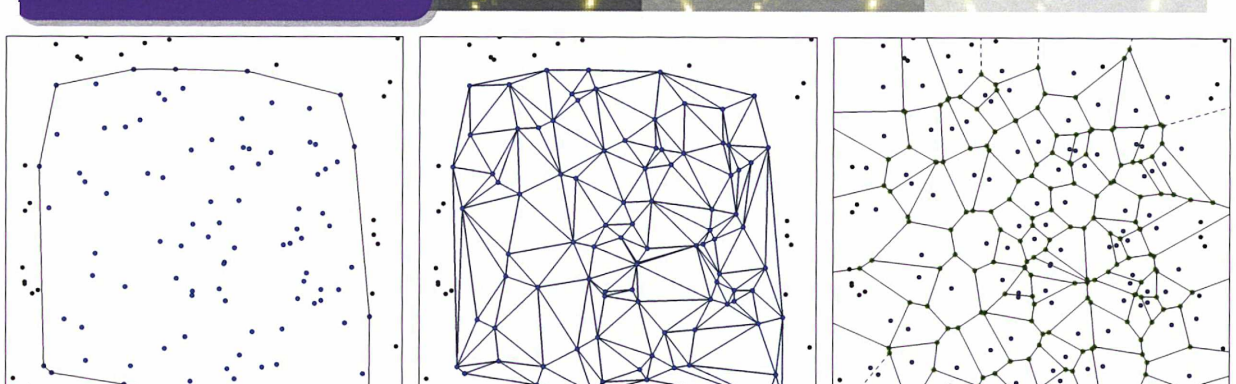
2.1.1 Marinoni et al. (2002)

2.1.1.1 Description

Marinoni et al. (2002) have introduced a group finder based on Delaunay-Voronoi tesselation. The idea is to use over-densities of galaxies in the three dimensional space (reconstructed simply from the redshift space), and use them as potential centers for groups. Over-densities are estimated by use of a Voronoi partition of space. The set of Voronoi cells forms a complete partition of space, and the volume of a cell is inversely proportional to the galaxy density around the galaxy in the cell. Then galaxies are sorted by decreasing densities in order to use them as potential galaxy groups.

The procedure for selecting galaxy groups is divided in three principal steps, with an additional phase of initialization. The latter consists on the creation of the Voronoi-Delaunay tessellation of

TESIS



(a) Convex hull (b) Delaunay (c) Voronoi

Figure 2.1: Illustration of the tessellation of space in a sub-sample of randomly positioned points. In *black* the real point distribution (reflecting the real galaxy distribution) and in *blue* the sub-sample used for the tessellation (reflecting the volume limited galaxy survey). (a) The convex hull is the set of points forming the hull of the sample. (b) The Delaunay mesh is represented by the lines interconnecting points. Each triangle of the mesh has its circumscribing circle without a point inside it by definition. (c) The Voronoi partition is the dual of the Delaunay mesh. Each node is the result of the crossing median of the Delaunay mesh. Working on a sub-sample of galaxies shows that the Delaunay mesh is not well constrained at the edges, and the Voronoi cells are affected too. A consequence is that their volumes are biased and do not correspond to the real galaxy density around them when one gets too close to borders.

the galaxy sample in three dimensional space. The Voronoi partition is the dual of the Delaunay mesh and be can deduced from it. An illustration of each set of points is given in Figure 2.1.

The first step is to search for potential groups by using the Voronoi partition. Voronoi cells have the property that their volume is inversely proportional to the local density around each point. In case of galaxies, this allows to access to local density around them. The detected high densities in the three dimensional space are used as potential group centers. Galaxies are sorted by increasing volume of their Voronoi cell, i.e. decreasing density. *First-order* galaxies, first linked to these potential groups, are searched in a 1 Mpc region, using the Delaunay triangulation to access the neighborhood of the group. If all first-order galaxies are already assigned to another group, the two structures are merged.

The second step takes into account the redshift distortions, neglected in the first step. For this, a cylindrical region is created with a base radius perpendicular to the line-of-sight, and a height of around 20 Mpc. All galaxies inside this region, not already linked as first-order galaxies, are second-order galaxies. The size of the cylinder is chosen to take into account the redshift elongation introduced in a typical group.

The third step uses the information created from the two previous steps, which are only a selection of potential groups. From the richness of those potential groups, a relation between the richness and the cylinder lengths is deduced since the number of galaxies inside a group and its virial mass are correlated. This implies that the group sample isn't affected by some incompleteness, such as luminosity incompleteness. From a constructed complete sub-sample, the relation between the richness and the characteristics sizes of the cylinder are deduced and modeled. Then, the second step is reapplied, the cylindrical region inferred from the previous relations with help of the richness of the group.

2.1.1.2 Advantages and weaknesses

The group extraction doesn't rely on physical assumptions, but uses a geometrical approach, based directly on the available galaxy sample. Moreover, there are no free parameters, since the cylindrical region is then adjusted, based on a relation between the virial radius and the group richness. This relation is adjusted in a complete sub-sample of galaxies to avoid incompleteness corrections, the algorithm should be robust under different galaxy surveys.

But the Delaunay-Voronoi tessellation has some drawbacks. The computation of the Delaunay mesh is very difficult in non-Euclidean spaces, as the redshift space, from the point of view of an observer. Moreover, the computation of the volume of the Voronoi cell is complex too, especially with non-Euclidean spaces. As a consequence, the computation must be done assuming that the redshift space is perpendicular and fixed in space (in other words, the line-of-sight direction at different location on the celestial sphere is the same). Neglecting the celestial distortions limits the application of the algorithm to a small portion of the sky of a few degrees of side.

In addition, border effects can't be neglected with the Voronoi partition of space. Since Voronoi cells form a complete partition, cells at the edges of the galaxy sample have an infinite volume size. Also, the volume of cells close to borders is biased because the distribution of galaxies is unknown beyond the sample, and the Delaunay mesh can't be fully constrained to reflect the real density of galaxies at edges. In other words, the volume of Voronoi cells near edges doesn't really reflect the local density around galaxies, since the galaxy distribution is unknown beyond the limit of the sample.

Finally, the tessellation is computed for a flux-limited sample of galaxies, but the density around galaxies is used to search high mass halos first. Since the luminous incompleteness is decreasing the observed number of galaxies with increasing redshift, the effect is that nearby groups are searched first. With the redshift distortions, the consequences of such bias in the selection aren't trivial to understand on the resulting group catalogue.

In conclusion, the Voronoi-Delaunay method of Marinoni et al. (2002) can't be really applied to recent galaxy surveys covering a large area of the sky.

2.1.2 Yang et al. (2007)

Yang et al. (2005, 2007) devised a Bayesian grouping algorithm based upon cosmological simulations. In particular, they assume that the galaxy density profile inside groups follows the Navarro et al. (1996) model that reproduces well the density profiles of halos in cosmological dark matter only simulations. A density contrast parameter is defined as the ratio between the projected density of galaxies inside a halo and the density of field galaxies (which are the interlopers). The higher is this ratio, the more likely the galaxy belongs to the group. The density of galaxies in the halo is simply the integration of the distribution function along the line-of-sight, and for interlopers, it is the integration of the mean density of the Universe along the line-of-sight over the Hubble distance. This leads them to the following definition for the density contrast:

$$P_M(R, \Delta z) = \frac{H_0 \Sigma(R)}{c \overline{\rho}} p(\Delta z) \tag{2.1}$$

where H_0 is the Hubble constant, c the speed of light, $\Sigma(R)$ the projected surface density of galaxies at the projected radius R, $\overline{\rho}$ the mean density of the Universe and $p(\Delta z)$ is the velocity distribution of galaxies in terms of redshift differences Δz with the group redshift. This definition is problematic: the density of interlopers is assumed to be constant and the same for all halos. But as described in Mamon et al. (2010), the density of interlopers is related to the position in the halo, and their line-of-sight velocity distribution isn't flat.

Using this density contrast criterion implies to have potential groups on which to apply it. For this, initially, a FoF algorithm is done on the galaxy sample but with very small linking lengths. These potential groups are whose membership must be updated using the density contrast, as described below.

For each group, the virial mass is estimated from a relation between the group luminosity and its mass. Initially, this is a constant ratio, then adjusted on the group sample itself. From it, the density contrast can be computed for each galaxy on each group. A galaxy is assigned to a group if $P_M > B$ where B is a threshold. If this condition is satisfied for multiple groups, it is assigned to the group with the highest P_M.

Then group centers and luminosities are recomputed with the new membership, iterating over the previous step until a convergence in the membership is observed.

Once the convergence is reached, the relation between the virial mass and the luminosities of groups is recomputed by abundance matching (see Chapter 5) between the distribution of group luminosities obtained from the sample and the expected distribution of virial masses assuming a halo mass function. Then the previous iterative process is done again, and this goes until a convergence is reached for the relation.

The algorithm have some drawbacks that should be technically and physically corrected to be good enough in the group extraction. Indeed, some incoherences are present in the implementation of the grouping algorithm. For example, the given formula for the computation of the virial radius is done for halos being over-densities of $\Delta = 180$ of the mean density of the Universe, while the computation of the abundance matching is done with the halo mass function of Warren et al. (2006) for the FoF mass of halos from the cosmological simulation used. The difference between the FoF mass and the virial mass is significant and should be taken into account in the grouping process.

2.1.3 Domínguez Romero et al. (2012)

Domínguez Romero et al. (2012) adapted the algorithm of Yang et al. (2007), based on a better Bayesian approach, noting that the grouping method of Yang et al. (2007) is simply a learning algorithm called K-means. Instead of hard assignments of galaxies to groups in the iterative process, this method assigns "responsibilities", equivalents of a probability of belonging to a group, weighted over all groups in the sample, using the density contrast definition above as some probability to be in the group.

First, potential groups are estimated assuming that the most luminous galaxies are linked to most massive systems. Then, galaxies are assigned to a group as satellite members if they have a density contrast superior to a chosen low threshold to allow a maximum of galaxies to belong to the group, without introducing too many interlopers since their responsibilities will be low and won't affect group properties. As in Yang et al. (2007), an iteration over the membership and the relation used to compute virial properties is done until convergence. Finally, galaxies are assigned to the group for which the responsibility is the highest.

The drawbacks of this method are essentially inherited from Yang et al. (2007), since it is an improvement of the Yang et al. algorithm.

2.1.4 Muñoz-Cuartas & Müller (2012)

Muñoz-Cuartas & Müller (2012) developed a method similar to the FoF algorithm, but applied directly on groups and not on galaxies. From an initial set of groups, a maximal circular radius is computed from the virial radius in the transverse direction to the line-of-sight. A maximal length of search for the redshift dimension is estimated from the circular velocity of the group. Those groups are sorted by decreasing masses. For each one, other groups (and their galaxies) are

merged into the current group if they belong to the ellipsoid defined by the two lengths defined above, centered on the group.

Then, group properties are computed from the membership obtained previously. The new virial masses are evaluated with an abundance matching between the group stellar masses and the halo mass function. The iteration is stopped once the number of groups doesn't change.

This method doesn't have any free parameter and doesn't rely on too many assumptions and models. Only the abundance matching can be responsible for a bias, since the virial mass is crucial in the merging of halos. As mentioned by Yang et al. (2007), the one-to-one assumption of the abundance matching creates an intrinsic dispersion in the mass estimation that is relatively low, and thus should not affect the galaxy grouping.

2.2 Discussion

We can extract common principles of the different algorithms described above. There are two approaches for the galaxy grouping: a geometrical one based only on the positional informations of galaxies and a Bayesian one using priors on group properties and galaxies therein. What emerged is that most of these algorithms make a harmonious combination of these approaches. A typical geometrical algorithm is the Friends-of-Friends (see also Chapter 4), linking galaxies between them if they are closer than a linking length. This method has the default of creating bridges between two different galaxy groups if two of their members are closer than the linking length. Adding priors to such a scheme, the membership can be improved by breaking the problematic bridges. This is a good summary for the method of Marinoni et al. (2002) or Muñoz-Cuartas & Müller (2012). Moreover, the bias in distance introduced by redshift can be reduced with the same prescriptions as done in Liu et al. (2008).

The Bayesian approach used in the described galaxy group algorithms also has its pitfalls: if the chosen models are bad, the clustering will be affected too. For example, Yang et al. (2007) assume a flat line-of-sight velocity dispersion and a flat distribution of distant interlopers in the computation of density contrast, which clearly depends on the projected radius pointed by the observer. Another problem is the way to test such algorithms. The different algorithms were tested on different galaxy mock catalogues, not constructed in the same way. In consequence, the comparison is very difficult because not operated in the same conditions. Finally, the definition of an optimal extraction, and the statistics used to assess their performances differ among the different galaxy group algorithm developers.

To go beyond such limitations that will never be completely avoided, a probabilistic approach of galaxy groups seems to be a good compromise (see Chapter 5). Moreover, the tests for algorithms must be well defined, with a common galaxy mock catalogue to perform the comparisons and a good definition of the statistics to use. How well are recovered galaxy groups? How well are they polluted by interlopers? How many selected groups are spurious? How well are the virial properties of the parent halo recovered? How well are the scaling relations recovered? These are the questions to answer in order to characterize the quality of a grouping algorithm.

THESIS

THESIS

Chapter 3.

Generating mock catalogues

Contents

3.1 Introduction

A mock catalogue is a useful tool to test galaxy group algorithms. These mocks can reproduce many properties of galaxies, e.g. clustering, luminosity function, etc, and add observational effects such as incompleteness and measurement errors. There are different methods to obtain such a mock catalogue. All of them involve cosmological simulations of dark matter halos. According to the model of galaxy formation, we can use the halo occupation distribution (HOD) to populate dark matter haloes with galaxies and putting some luminosity functions (for example) as constraints. We can, alternatively, follow galaxies in semi-analytical models (SAM) in cosmological simulations outputs in order to have statistical properties of galaxies that agree with observational results. With such realistic galaxies, we can use those simulation boxes to place an observer into it, and create a mock survey. But to have a realistic mock catalogue, it's necessary to take care of many things which will be described in the next section.

3.2 Populating dark matter halos

The first step in the construction of a galaxy mock catalogue is to populate galaxies inside their dark matter halos, according to the model of the galaxy formation and the constraints imposed by the observations. The real large scale structure of the Universe is not directly observable and we must be confident in the different cosmological simulation code outputs available to get the distribution of dark matter halos. For example, there is the Millennium-II run (Boylan-Kolchin et al., 2009) of 2160^3 dark matter particles, with a simulation box size of 100 h^{-1}Mpc allowing a relatively high resolution for the particles mass (6.9 10^6 $h^{-1}M_\odot$) and a precise determination of

TESIS

the halo mass function and the history of each individual halo. Several cosmological simulation codes exist to follow the dark matter particle distribution along the evolution of the Universe (Springel, 2005; Springel et al., 2001; Teyssier, 2002).

From the outputs of this cosmological simulations, informations on dark matter halos are extracted using halo finder codes (Knollmann & Knebe, 2009; Planelles & Quilis, 2010; Tweed et al., 2009). Several methods exist to identify such structures: the Friend-of-Friends approach (Davis et al., 1985) that tends to link between them different halos by bridges of linked dark matter particles, and doesn't identify sub-halos; the spherical over-density measuring the density field and searching around density peaks iteratively until the desired density threshold is reached (Press & Schechter, 1974a); SUBFIND of Springel et al. (2001) is similar to FoF with peaks searched within the extracted FoF halos... With all available informations on halos, they can be populated by galaxies with prescriptions of the galaxy formation model.

3.2.1 Halo occupation distribution

In the Halo Occupation Distribution method (Berlind & Weinberg, 2002; Martínez & Saar, 2002; Zehavi et al., 2011), the galaxy richness in groups is deduced on a probability distribution function depending on the halo mass, and their physical properties such as luminosity from conditional luminosity functions depending on the dynamical mass too (Yang et al., 2003). A relation between the galaxy and matter distribution is imposed by three constraints: the probability distribution $P(N|M)$ that a halo of mass M contains N galaxies, the spatial relation between the galaxy and dark matter, and the same for the velocity distribution. The galaxy distribution is assumed to be spherically symmetric, and follows that of the dark matter particles in the halos of ΛCDM cosmological simulations (e.g., NFW), the velocities are drawn from Maxwellian distributions (see Beraldo et al., 2014 for the limitations of this assumption), with radial and tangential velocity dispersions derived from the Jeans equation of local dynamical equilibrium, assuming some form for the radial variation of the velocity anisotropy.

3.2.2 Semi-analytical models

In Semi-Analytical Models (SAMs, e.g., Kauffmann et al., 1999; Roukema et al., 1997), galaxy properties (in particular stellar mass and r-band luminosity) are painted on the halos and subhalos of cosmological N body simulations across cosmic time, following well-defined physical recipes for star formation and galaxy feedback. This procedure produces galaxies that follow relatively well the observed luminosity, stellar mass functions and scaling relations.

We have chosen this second approach, because the recent SAM by Guo et al. (2011), run on the Millennium-II simulation (Boylan-Kolchin et al., 2009) fits well the z=0 observations (as shown by Guo et al.). The Millennium-II simulation involved 2160^3 particles in a box of comoving size 137 Mpc, running with cosmological parameters $\Omega_{\rm m} = 0.25$, $\Omega_\Lambda = 0.75$, $h = 0.73$, and $\sigma_8 = 0.9$. The particle mass was thus $9.5 \times 10^6 M_\odot$.

We extracted the SAM output of Guo et al. (2011) from the Guo2010a database on the German Astrophysical Virtual Observatory website. [1] The real-space groups were extracted by Guo et al. using the FoF technique applied to the particle data, with over 10^5 particles for groups of mass $> 10^{12}\mathrm{M}_\odot$. The database includes the mass within the sphere of radius r_{200}, where the mean mass density is $\Delta = 200$ times the critical density of the Universe, centered on the particle in Millennium-II simulation, within the largest sub-halo, with the most negative gravitational potential (Boylan-Kolchin et al., 2009). We slightly modified the membership of the true groups by considering only the galaxies within r_{200}.[2]

[1] http://gavo.mpa-garching.mpg.de/Millennium/Help, see Lemson & the Virgo Consortium (2006)
[2] We kept the galaxies outside the sphere of radius r_{200} as possible interlopers.

3.3 Mock structure

In all this section, we will assume that we have already in our possession a dark matter simulation box which has been populated with galaxies with one of the methods described below (SAM, HOD...). At this step, physical properties of those galaxies aren't interesting.

3.3.1 Placing boxes

The first step to make a mock catalogue is to get galaxies positions like in a survey, to get an (α, δ) frame to simulate the sky coverage of survey.

The mock catalogue must have the same volume as the galaxy survey we want to mimic. For example for the SDSS survey, we can measure redshift to a value of 0.3 and more. But the problem is that the majority of the simulation boxes have a size of around $L_{\rm box} = 100 - 300\ h^{-1}$ Mpc, letting us with a maximal redshift in our mock survey of around $H_0 L_{\rm box}/c \approx 0.025$ in the case of a box of 100 h^{-1} Mpc in size. Bigger simulations exist, and allow us to access higher redshifts, but this increasing size reduces the resolution of the simulation in particle mass and therefore we cannot have low mass halos in the simulations.

The solution is to take a "little" simulation box and to replicate it and to make some "Tetris" cube until we reach the maximal redshift we want. An example of the resulting "mock cube" is shown on Figure 3.1.

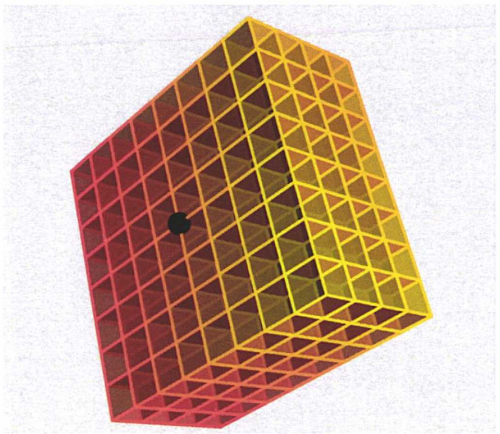


Figure 3.1: The structure of the mock catalog once we have replicated the simulation box chosen to populate dark matter halos. Each cube represents a simulation box whose galaxies were randomly rotated and translated in positions. Placing an observer at a given position (the black dot), we can access different geometries for the survey and go to higher redshift ranges than those possible with an unique simulation box.

Now, if we take an observer at some position into this big box, we can have different sky coverage for the observer. The simplest is to place the observer at a corner, which gives a solid angle of $\pi/2$ steradians. At the centre, we have a full sky coverage but we reduce the redshift extension by two. For the SDSS, as in Figure 3.1, the area of the survey is large (see Appendix 6) and we need to cover half of the sky to get the same volume.

If we want to care about redshift evolution of galaxies for the observer, we need to use other snapshots at different redshifts, simply joining cubes in comoving coordinates. Indeed, the cosmological redshift of the galaxy is deduced from the relation between the redshift and its comoving distance, equals to the comoving transverse distance (or proper motion distance) in the case of a flat Universe $\Omega_k = 0$ (Hogg, 1999). Moreover, the comoving separation R_c between two points with angular separation θ on the sky, at comoving distance D_c from the observer, are simply related by a geometrical relation $R_c = \theta D_c$. This separation θ deduced from comoving

coordinates should be the same as those of the observer working with physical coordinates. The observer wants to know the physical separation R_p between the two galaxies, so $R_p = \theta d_{\text{ang}}$ giving $R_p\,(1+z) = R_p/a\,(z) = R_c = \theta D_c$ (where $a\,(z)$ is the scale factor with $a\,(0) = 1$).

Placing boxes as described previously creates a perspective effect from the point of view of an observer (Blaizot et al., 2005), and the consequences aren't predictable in a statistical sense. To avoid this, we apply some coordinate transformations on galaxies in the initial cube like inversions, rotations and periodic translations. Rotations are multiples of $\pi/2$ around the three principal coordinates axes, because if other rotations are allowed, this create over-densities in some regions of the final mock which aren't physical. Translations are performed on the three principal axes and when galaxies are out of the initial cube, periodic conditions are applied. All of those transformations are randomly generated for each cube in the final mock catalogue.

3.3.2 Physics

3.3.2.1 Celestial coordinates

The first step to simulate this is to transform Cartesian coordinates (X, Y, Z) in the 3D space to celestial coordinates ((α, δ) frame). In our case, the origin of coordinates is the observer. Getting these coordinates is the same as computing spherical coordinates.

$$\alpha = \text{arctan2}\,(Y, X) \mod 2\pi$$

$$\delta = \text{sgn}\,(Z)\arccos\left(\frac{\sqrt{X^2+Y^2}}{\sqrt{X^2+Y^2+Z^2}}\right) \tag{3.1}$$

where sgn is the sign function[3]

3.3.2.2 Redshifts

If we keep the distance as calculated previously, the observer can still have precise determination of the distance of a galaxy. In reality, we observe it in redshift space so the redshift as distance indicator is biased by peculiar velocities. Our initial galaxy catalog allows us to get the velocity of a galaxy, so we compute the line of sight (los) velocity of this galaxy relatively to the observer.

$$v_{\text{los}} = \frac{\mathbf{OG}.\mathbf{v}_{\text{pec}}}{||\mathbf{OG}||} \tag{3.5}$$

[3]For the sign function:

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 1 \\ -1 & \text{if } x < 1 \\ 0 & \text{else} \end{cases} \tag{3.2}$$

and the arctan2 function is:

$$\text{arctan2}\,(y, x) = \begin{cases} \phi \times \text{sgn}(y) & x > 0 \\ \dfrac{\pi}{2} \times \text{sgn}(y) & x = 0 \\ (\pi - \phi) \times \text{sgn}(y) & x < 0 \end{cases} \tag{3.3}$$

with $\tan\phi = \left|\dfrac{y}{x}\right|$ particular cases:

$$\text{arctan2}\,(0, x) = \begin{cases} 0 & x > 0 \\ \text{not defined} & x = 0 \\ \pi & x < 0 \end{cases} \tag{3.4}$$

where O is the observer and G the galaxy, $\mathbf{v}_{\rm pec}$ its peculiar velocity. This velocity has a sign. The redshift is just the expression of a shift in wavelength. The observed wavelength λ is linked to the original (emitted) wavelength λ_0 by:

$$\lambda = (1+z)\lambda_0 \tag{3.6}$$

The shift caused by Universe expansion is $\lambda_{\rm cos} = (1+z_{\rm cos})\lambda_0$ where the subscript cos refer to the cosmological expansion. The shift caused by the peculiar velocity is $\lambda = (1+z_{\rm pec})\lambda_{\rm cos}$. So the observed wavelength is $\lambda = (1+z_{\rm pec})(1+z_{\rm cos})\lambda_0$. The resulting observed redshift is:

$$(1+z) = (1+z_{\rm pec})(1+z_{\rm cos}) \tag{3.7}$$

The peculiar redshift is the relativistic Doppler effect:

$$(1+z_{\rm pec}) = \sqrt{\frac{1+\beta}{1-\beta}} \tag{3.8}$$

with $\beta = v_{\rm los}/c$. The cosmological redshift is approximated by $z_{\rm cos} = H_0 D/c$ where D is the physical distance of the galaxy to the observer and H_0 the Hubble constant, in the case of a simple computation of the redshift. We can also make a more precise computation by searching the solution of $D = d_{\rm pm}(z_{\rm cos})$ with $d_{\rm pm}(z)$ the proper motion distance at the redshift z.

Applying this method to mock catalogue, we can have galaxies whose "distance" is biased by peculiar velocities in redshift space. With such a treatment, the velocity dispersion of galaxies in groups leads to the apparition of "fingers of God", an elongation of galaxy groups along the line-of-sight, as in redshift space observations.

We don't have to add the wavelength shift due to the translation of the observer relatively to the Cosmological Microwave Background (CMB). Velocities in the simulation are relative to an "absolute" frame, but our Galaxy has a movement in relation to the CMB creating an additional shift in wavelength depending on the observed region of the celestial sphere, and we should include it in our mock catalogue. But frequently, redshifts accessible in galaxy surveys are already corrected for the CMB relative motion, or can be easily pre-corrected to avoid this component. In consequence, we don't integrate it in our mock catalogue.

3.3.2.3 Survey mask

With our frame in redshift space relative to the observer, we can apply different masks on angular coordinates according to the survey we want to mimic. An example of such a mask is in Chapter 6, where we describe how to decide if a galaxy is inside the mask or not, in the case of the SDSS.

3.3.2.4 K-corrections

In reality, an observer studies galaxies in a given bandwidth in wavelength and can't use the bolometric flux of the object. With the expanding Universe, all the spectral energy distribution (SED) of galaxy is shifted. All wavelengths are shifted by the same value for a given redshift. So, knowing the luminosity L of a galaxy in a given band in reality (using the true SED), computing its apparent magnitude for an observer aren't as easy as correcting for the distance modulus. The observer in a given band sees a different part of the rest frame SED. The flux observed in the same band as the rest frame flux is maybe higher or lower. A correction for this effect is needed in the real galaxy survey to estimate the distance of an object and must be taken into account in our mock catalogue.

As explained before, this correction depends on the SED of galaxies and the band used in the survey. The common way of correcting, it when we have a multi-band photometry, is to fit the

THESIS

observed SED in those bands with theoretical templates of SEDs. Such templates can be obtained with existing programs as PEGASE (Le Borgne et al., 2004), giving us galaxy SEDs. But those programs are a little time consuming, a problem for mock when we want to run several of them. A quick alternative solution is provided by Chilingarian et al. (2010), where the K-correction is fitted on templates for SEDs as given by PEGASE in terms of a 2D polynomial of the redshift of the galaxy and its colour. The corresponding K-correction is precise for redshifts until 0.3 in different survey bands (including *ugriz* for the SDSS). This work reduces the computation of K-corrections to the use of simple polynomial relations and make our task easier.

By definition, the K-correction K for a galaxy of apparent magnitude m_X in a given band X and absolute magnitude M_X in the same band is:

$$m_X = M_X + 5\log_{10}\left(d_{\text{lum}}\,[\text{pc}]\right) - 5 + K \tag{3.9}$$

In our case, the K-correction depends on the redshift of the galaxy and its colour in apparent magnitude given two bands. So we can rewrite:

$$m_X = M_X + 5\log_{10}\left(d_{\text{lum}}\,[\text{pc}]\right) - 5 + K\left(z, m_X - m_{X'}\right) \tag{3.10}$$

where:

$$K\left(z, m_X - m_{X'}\right) = \sum_{i=0}^{N_i}\sum_{j=0}^{N_j} a_{ij} z^i (m_X - m_{X'})^j \tag{3.11}$$

and a_{ij} is a $N_i \times N_j$ matrix containing the coefficients of the two dimensional polynomial. These coefficients depend on the bands of the survey used for the colour computation.

The observer in the mock can just, in theory, access to apparent magnitude of the survey. But we don't know in advance these magnitudes, and as we can see in the expression of Equation 3.10, we need apparent magnitudes to compute apparent magnitudes. If we use the other bands of the survey, with a_{ij} coefficients, we can always write a set of equations for a galaxy which involves all apparent magnitudes of the survey. So we can write a set of non linear equations with polynomial of order N_j (redshift of the galaxy is supposed to be known). Numerically it's easy to solve this set of equations, and relatively fast with equations solvers or by iterations. In practice, the first is faster than the second method, even if both methods give similar results in apparent magnitudes.

3.3.2.5 Flux limit

We will see in Chapter 6 that spectroscoped galaxies are defined for galaxies whose apparent magnitude is less than 17.77 in the r band. So, in all the redshift sub-samples, we will miss galaxies that are not sufficiently bright. To take into account this effect, we remove galaxies not reaching the limit apparent magnitude of the survey. An additional selection on surface brightness is also done in the SDSS, but estimating this is difficult from virtual galaxy catalogs and the number of "lost" galaxies is sufficiently low to ignore this step in the construction of the mock catalog.

3.3.2.6 Spectroscopic and photometric redshifts

Sometimes, we don't have access to spectroscopic redshifts, but only to less precise photometric redshifts. In the SDSS, for example, this is due to tiling process. Fibers analysing the spectrum of galaxies cannot be closer from each other than 55″, so if for a target galaxy (selected to obtain a spectrum) there is an other galaxy closer than those 55", the tile containing all fibers doesn't have the possibility to measure the redshift of this galaxy. A very good algorithm to place tiles in order to limit the number of missed galaxies (i.e. the number of fiber collision) has been applied

to the galaxy sample of the SDSS (Blanton et al., 2003). But there is still some galaxies without spectroscoped redshifts, especially in dense regions such as the cases of groups and clusters. If we remove those galaxies from our sample, there will be a spectroscopic incompleteness with unknown effects on our results.

Unfortunately, there is no simple way to simulate this in the mock catalogue and we choose to ignore it. Just a small fraction of galaxies are not spectroscoped and this must not affect our results, when we will apply galaxy group algorithms on a real galaxy survey.

3.3.2.7 Observational errors

The way we organized the construction of the mock catalogue is useful for the introduction of observational errors. For example, we treat the case where we want to add errors on the absolute magnitude of galaxies in the final mock catalog. If we have a model for introducing such errors according to some physical galaxy properties in the virtual galaxy catalogue, we can just add them inside this virtual catalogue and magnitude errors will be reported on the mock galaxy catalogue. If errors depend on properties computed in the mock catalogue, we can simply add magnitude errors while constructing the mock catalogue. Any kind of errors can be added such as redshift measurement errors, astrometry, photometry, etc.

3.3.3 Galaxy samples

3.3.3.1 Definition

All previous steps lead to a final galaxy mock catalogue, with or without observational errors, flux limited at a given apparent magnitude. But working with flux limited samples requires correcting for missing galaxies when extracting galaxy groups. The only way to avoid errors introduced by the different choices of the model is to work with a doubly complete sample of galaxies: limited in luminosity and in volume, thus avoiding completeness issues.

We choose a minimal galaxy luminosity for our sample and the maximal redshift is computed with the maximal distance at which we can observe a galaxy with this minimal luminosity. We note that if K-correction is considered, this limit depends on the considered galaxy. A clean definition of the sample in this situation should be done by restricting a little more the redshift extent to not lose fainter galaxies. But the galaxy loss is low and we didn't consider such a case. In Table 3.1, we show the six galaxy samples that we constructed from our flux limited galaxy mock catalogue, with statistics on each of them.

3.3.3.2 Limitations

The mock catalogue is constructed from the adjoining of multiple simulation boxes, each of them having periodic boundaries. A consequence is that some galaxy groups are split by a simulation box size, and from the point of view of the observer, members are at two different locations on the celestial sphere. Inclusion of such groups in statistics leads to biased results of the performance of grouping algorithms, and a flag is used to distinguish and remove such groups.

Moreover, the limited volume extension of the survey truncates some groups. In redshift space, these limits are of two kinds: the angular mask cutting groups all along the line-of-sight (i.e. survey edges and possibly holes), and the redshift cut with a more important effect due to the elongation in redshift coordinates by the intrinsic velocity dispersion of the system. Since all the information on the group is not accessible by the observer, estimation of group properties is less precise. To avoid the degradation of the performances of grouping algorithms, we flag selected groups (after the application of a grouping algorithm) if they are close to edges of the

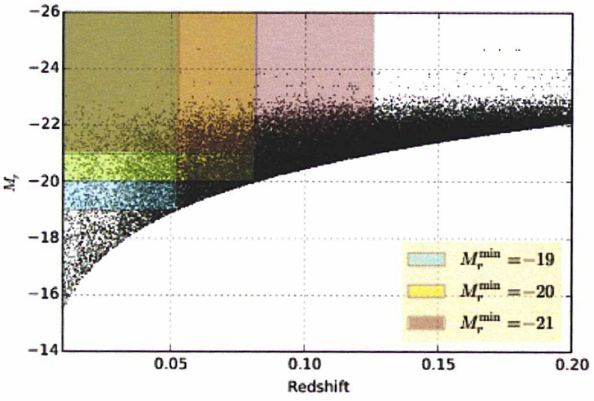


Figure 3.2: An illustration of the doubly complete galaxy samples used. Black dots represent galaxies in the mock catalogue. Colored rectangles show samples for a threshold absolute magnitude in the r band M_r of -19, -20, -21. Their sizes reflects the corresponding limits of the minimal luminosity. As we can see, in these regions, there is no need to correct for missing galaxies. All galaxies above the given threshold magnitude are visible.

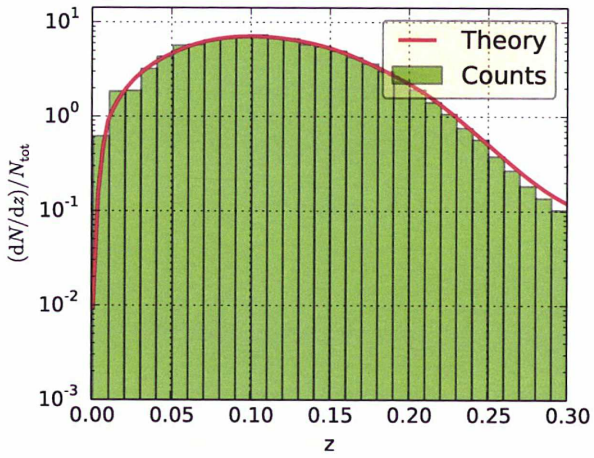


Figure 3.3: Comparison of counts in redshift of galaxies in the mock catalog with the theoretical expectation deduced from the number density of galaxies. In *red* the theory and in *green* counts directly done on the mock catalog. The small discrepancies observed at high redshift are caused by an imprecision in the computation of the integrated luminosity function, since we use an interpolation of the luminosity function, and only few bright galaxies at high redshift are visible making the integration difficult.

angular mask or edges in depth (see Chapter 4 for a detailed definition) and remove them from the statistics in tests.

3.4 Validity

Finally, we test the construction of our mock catalogue with a simple comparison with the expectation from the theoretical redshift counts. Indeed, the redshift count is:

$$\frac{\mathrm{d}N}{\mathrm{d}z} = \frac{\mathrm{d}N}{\mathrm{d}V} \times \frac{\mathrm{d}V}{\mathrm{d}z} \tag{3.12}$$

where V is the comoving volume. The first term of the *rhs* of Equation 3.12 is just the integrated luminosity function Φ:

$$\frac{\mathrm{d}N}{\mathrm{d}V} = \int_{L_{\lim}(z)}^{\infty} \frac{\mathrm{d}^2 N}{\mathrm{d}V \mathrm{d}L} \mathrm{d}L = \int_{L_{\lim}(z)}^{\infty} \Phi\left(L\right) \mathrm{d}L \tag{3.13}$$

The luminosity function is directly deduced from the virtual galaxy catalog of Guo et al. (2011).

The second term of the *rhs* of Equation 3.12 is the variation of the comoving volume with redshift (see Hogg (1999)):

$$\frac{\mathrm{d}V}{\mathrm{d}z} = D_{\mathrm{H}} \frac{d_{\mathrm{pm}}\left(z\right)^2}{E\left(z\right)} \mathrm{d}\Omega \tag{3.14}$$

where $\mathrm{d}\Omega$ is the elemental solid angle, d_{pm} the proper motion distance (or comoving distance, see Appendix G), D_H the Hubble distance c/H_0 and $E\left(z\right)$ the evolution of the Hubble constant with the redshift. The result of the theoretical prediction and the comparison with the obtained mock catalog is shown in Figure 3.3.

Table 3.1: Doubly complete mock galaxy subsamples

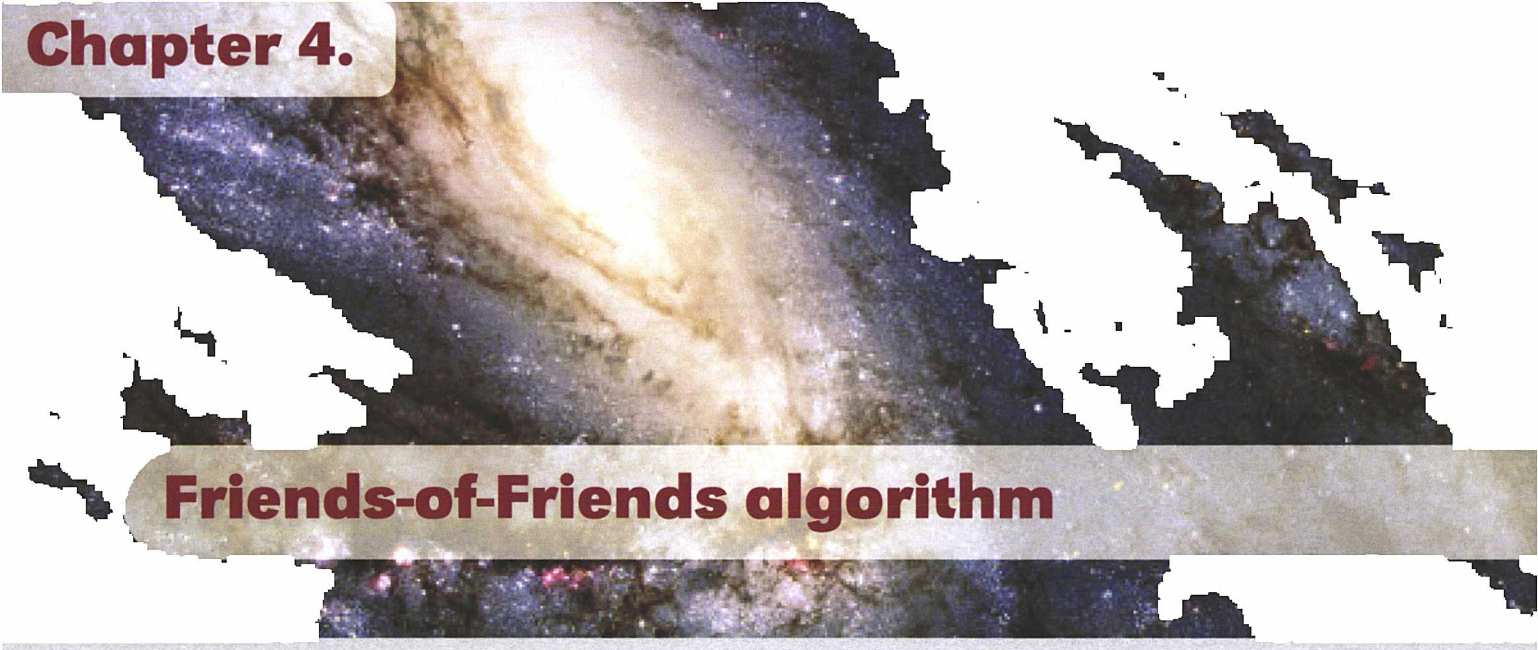
ID	$M_r^{\max}$	$L_r^{\min}/L*$	$z_{\max}$	Number	n (Mpc^{-3})	$n^{-1/3}$ (Mpc)	Fraction split
1	–18.5	0.09	0.042	47158	0.0125	4.32	5.3%
2	–19.0	0.14	0.053	72510	0.0099	4.66	6.1%
3	–19.5	0.22	0.066	112629	0.0078	5.05	6.6%
4	–20.0	0.36	0.082	166899	0.0058	5.56	7.4%
5	–20.5	0.56	0.102	213546	0.0040	6.29	8.6%
6	–21.0	0.90	0.126	245821	0.0025	7.40	9.9%

Notes: Columns are: sample, maximum r-band absolute magnitude, minimum luminosity in units of $L*$ (adopting $M* = -20.44 + 5\log h$ in the SDSS r band from Blanton et al., 2003), maximum redshift, sample size, mean density n, proxy for the mean separation to the closest neighbor, $n^{-1/3}$, and the percentage of true groups that are flagged because they are split during the simulation box transformations. The minimum redshift of each subsample is $z = 0.01$.

UTHESIS

THESIS

Chapter 4.



Friends-of-Friends algorithm

Contents

From Duarte & Mamon (2014a).

4.1 Introduction

Although several grouping algorithms has been developed recently, using our knowledge on the galaxy formation and evolution processes (as described in the Chapter 2), the Friends-of-Friends algorithm (hereafter FoF) has been the most popular grouping algorithm over time. Many catalogs of galaxy groups have been constructed from redshift space catalogs,[1] using FoF algorithms (Huchra & Geller, 1982; Merchán & Zandivarez, 2002; Nolthenius & White, 1987; Ramella, Geller, & Huchra, 1989; Trasarti-Battistoni, 1998; Berlind et al., 2006; Eke et al., 2004; Robotham et al., 2011; Tago et al., 2010; Tempel et al., 2014).

[1] Turner & Gott (1976) applied a grouping algorithm in projected space that turned out to be a Friends-of-Friends algorithm.

TESIS

4.1. INTRODUCTION

Starting with Nolthenius & White (1987), nearly all FoF group analyses on redshift space catalogs were accompanied with tests on mock galaxy catalogs derived from N-body simulations. However, not all FoF developers have applied the same tests to calibrate their linking lengths. Nolthenius & White (1987) were the first to compute the accuracy of group masses, as well as radii and velocity dispersions, crossing times and mass-to-light ratios. Ramella et al. (1989) were the first to test the recovered group multiplicity function. Frederic (1995) was the first to measure the galaxy reliability of extracted groups (comparing the FoFs of Huchra & Geller, 1982 and Nolthenius & White, 1987), as later done by Merchán & Zandivarez (2002), who also measured group completeness (against mergers of true groups) and reliability (against fragmentation of true groups). Eke et al. (2004) also tested the true group completeness and fragmentation, as well as the accuracy on group sizes and velocity dispersions. They also considered a quality criterion that amounts to a combination of galaxy completeness and reliability. Finally, Berlind et al. (2006) performed similar tests as Eke et al., with another test combining galaxy completeness and reliability. Berlind et al. noted that one cannot simultaneously optimize the accuracies on group sizes, velocity dispersions and [multiplicity function OR combined galaxy completeness/reliability].

Unfortunately, none of these studies is fully convincing: many did not perform the full suite of important tests, which we believe are true group fragmentation (group reliability) and merging (group completeness), galaxy completeness and reliability studied separately, and mass accuracy. Many have not measured the qualities of their LLs in terms of group parameters such as estimated mass and richness. Few studies have *optimized* the LLs: Eke et al. (2004) separately optimized $b_\perp$ and $b_\parallel$. Berlind et al. (2006) jointly optimized $b_\perp$ and $b_\parallel$ on a grid, for groups of 10 or more galaxies, while Robotham et al. (2011) jointly fit the LLs and their variation with density contrast and galaxy luminosity for groups of 5 or more galaxies to optimize for the product of four fairly complex measures of group and galaxy completeness and reliability. However, there is no strong agreement between the optimized LLs of Eke et al., Berlind et al., and Robotham et al. (see Table 4.1).

Moreover, we believe that in this era of large redshift surveys of $> 10^5$ galaxies, it makes little sense to extract groups from flux-limited galaxy samples, for which most current implementations of the FoF algorithm scale the maximum separations proportionally to the mean separation between neighboring field galaxies, $n^{-1/3}$. Indeed, since the minimum luminosity in flux-limited samples increases with redshift, the mean number density of galaxies decreases with redshift, and thus the mean separation between neighboring galaxies increases with redshift. Therefore, the standard implementation of the FoF algorithm leads to groups that become increasingly sparse and with increasingly higher velocity dispersion with redshift (while their multiplicity function is preserved). Alternatively, since the mean neighbor galaxy separation increases with redshift in flux-limited samples, using a fixed physical linking length leads to lower reliability at low redshift and lower completeness at higher redshifts. Moreover, grouping algorithms on flux-limited samples must evaluate the luminosity incompleteness as a function of redshift, which is difficult and imprecise (e.g., Marinoni et al., 2002; Yang et al., 2007). It is therefore much safer to consider subsamples that are complete in both distance and galaxy luminosity (as done for FoF grouping by Berlind et al., 2006, Tago et al., 2010 and Tempel et al., 2014). Admittedly, one recovers at best of order of one-quarter of the galaxies of the flux-limited sample, but one then avoids extracting a heterogeneous sample of groups (see Tempel et al., 2014) whose sizes and velocity dispersions stretch with redshift (when scaling the physical linking lengths with $n^{-1/3}$) or whose completeness and reliability vary with redshift (when adopting fixed physical linking lengths).

In the present work, we shall provide the first optimization of group LLs for doubly complete subsamples of galaxies, for six measures of the quality of the FoF grouping algorithm: minimal fragmentation and merging of true groups, maximum completeness and reliability of the galaxies of the extracted groups, and minimum bias and inefficiency in the recovered group masses. These

Table 4.1: Friends-of-Friends linking lengths and physical parameters

Authors	sample	$b_\perp$	$b_\parallel$	$b_\parallel/b_\perp$	$\delta n/n$	κ
Huchra & Geller 82	CfA	0.23	1.34	6.3	20	5.7
Ramella et al. 89	CfA2	0.14	1.9	13	80	5.8
Trasarti-Battistoni 98	PPS2	0.13	1.7	13	108	4.9
Merchan & Zand'z 02	2dFGRS	0.14	1.4	10	80	4.4
Eke et al. 04	2dFGRS	0.13	1.43	11	178	3.9
Berlind et al. 06	SDSS	0.14	0.75	5.4	86	2.3
Tago et al. 10	SDSS	0.075	0.75	10	565	1.7
Robotham et al. 11	GAMA	0.060	1.08	18	1100	2.2
Tempel et al. 14 ($M_r < -19$)	SDSS	0.11	1.1	10	178	3.0
Tempel et al. 14 ($M_r < -21$)	SDSS	0.066	0.67	10	830	1.4

Notes: The (normalized) linking lengths of Huchra & Geller (1982), Ramella et al. (1989), and Trasarti-Battistoni (1998) are derived (using Equation 4.1 and Equation 4.2) from their physical linking lengths at the fiducial distance and from the mean density at that distance, as derived by integrating the respective luminosity functions given by these authors. The linking lengths of Merchán & Zandivarez (2002) are estimated directly from the overdensity $\delta n/n$ given by these authors (using Equation 4.3), those of Tago et al. (2010) are found from the densities deduced from the numbers of galaxies counted by these authors (again with Equation 4.1 and Equation 4.2). Eke et al. (2004) provide $b_\perp$ and $b_\parallel/b_\perp$, while Berlind et al. (2006) and Tempel et al. (2014) provide $b_\perp$ and $b_\parallel$. When not provided by the authors, the overdensity $\delta n/n$ is obtained through Equation 4.3, and should be multiplied by 1.5 for a more accurate estimation (see text). Finally, the number of group velocity dispersions along the LOS, κ is obtained with Equation 4.7 assuming $\Omega_\mathrm{m} = 0.3$.

tests are performed on a wide grid of over 250 pairs of LLs. We have applied them to several doubly-complete subsamples of galaxies cut from a mock flux-limited, SDSS-like, sample of galaxies, and we analyze our results in terms of both true and estimated masses of the groups, as well as of their estimated richness.

4.2 Description

4.2.1 Predicted linking lengths and galaxy reliability

Because of the redshift distortions, the physical linking lengths are chosen to be of order of 10 times longer for the line-of-sight (LOS) separations than for the plane-of-sky (POS) ones. Moreover, for flux-limited galaxy catalogs, the physical linking lengths are scaled with the mean three-dimensional separation between neighboring galaxies, $s \simeq n^{-1/3}$, where n is the mean number density of galaxies in the Universe at a given redshift (Huchra & Geller, 1982). In other words, the FoF algorithm involves two dimensionless linking lengths (hereafter LLs):

$$b_\perp = \frac{\mathrm{Max}(S_\perp)}{s} , \tag{4.1}$$

$$b_\parallel = \frac{\mathrm{Max}(S_\parallel)}{s} , \tag{4.2}$$

where $S_\perp$ and $S_\parallel$ are the POS and LOS nearest neighbor separations, respectively.

One can relate the choice of $b_\perp$ to the minimum galaxy overdensity (in number) of the groups with

$$\frac{\delta n}{n} = \frac{3}{4\pi b_\perp^3} - 1 , \tag{4.3}$$

THESIS

(from Huchra & Geller, 1982). Hence, if galaxies are unbiased tracers of mass, i.e. $\delta n/n = \Delta/\Omega_{\rm m}$, where $\Omega_{\rm m}$ is the cosmological density parameter, then Equation 4.3 easily leads to

$$b_\perp = \left(\frac{3/(4\pi)}{\Delta/\Omega_{\rm m} + 1} \right)^{1/3} . \tag{4.4}$$

According to Equation 4.4, if one desires to have virialized groups of overdensity (relative to critical) $\Delta = 200$, one requires $b_\perp \simeq 0.07$ (for $0.24 < \Omega_{\rm m} < 0.35$). On the other hand, given $\Omega_{\rm m} = 0.279$ or 0.317, respectively obtained with the 9th-year release of the Wilkinson Microwave Anisotropy Probe (Bennett et al., 2013) and the Planck mission (Planck Collaboration et al., 2013), one deduces $\delta n/n = 352$ and 326 from Bryan & Norman's (1998) approximation for Δ at the virial radius , leading to $b_\perp \simeq 0.09$ in both cases, according to Equation 4.3.

One can also estimate the ratio of LOS to transverse LLs, as the ratio of LOS to POS group sizes caused by redshift distortions: if the LOS velocities span $\pm\kappa$ group velocity dispersions, the inferred LOS spread of distances in redshift space will be $\pm\eta\kappa\, v_{200}/H_0 = \pm\eta\kappa\sqrt{\Delta/2}\, r_{200}$ (see Mamon et al., 2010), where $\eta = \sigma_v/v_{\rm v} \simeq 0.65$ for an NFW model with realistic concentration and velocity anisotropy (Mamon et al., 2013), and where we used Equation 4.3. Therefore,

$$\frac{b_\parallel}{b_\perp} = \eta\,\kappa \sqrt{\frac{\Delta}{2}} \tag{4.5}$$

$$= \eta\,\kappa \sqrt{\frac{\Omega_{\rm m}}{2} \left(\frac{\delta n}{n} \right)} . \tag{4.6}$$

Combining Equation 4.4 and Equation 4.5, one easily deduces

$$\kappa = \sqrt{\frac{8\pi}{3}}\, \eta^{-1}\, \Omega_{\rm m}^{-1/2}\, \sqrt{b_\perp}\, b_\parallel . \tag{4.7}$$

For example, according to Equation 4.5, probing galaxies along the LOS to $\pm 1.65\,\sigma_v$ (encompassing 95% of the galaxies for Maxwellian LOS velocity distributions), for $\Delta = 200$, leads to $b_\parallel/b_\perp = 11$, hence with $b_\perp = 0.07$, one finds $b_\parallel = 0.7$ (the values are rounded off).

These theoretical LLs assume that groups are spherical and that all but one galaxy is in the center. In fact, galaxies are distributed in a more continuous fashion (especially in rich groups and clusters). One can more accurately estimate the value of the transverse LL by writing

$$\begin{aligned} b_\perp &= \frac{\mathrm{Max}(S_\perp)}{n^{-1/3}} , \\ &= \frac{\mathrm{Max}(S_\perp)}{r_{\rm vir}} \frac{r_{\rm vir}}{n_{\rm vir}^{-1/3}} \left(1 + \frac{\delta n}{n} \right)^{-1/3} , \\ &= \left(\frac{3/(4\pi)}{\Delta/\Omega_{\rm m} + 1} \right)^{1/3} \frac{\mathrm{Max}(S_\perp)}{r_{\rm vir}}\, N_{\rm vir}^{1/3} , \end{aligned} \tag{4.8}$$

where one recognizes the previous estimate of $b_\perp$ (Equation 4.4) in the first term of the right-hand side of Equation 4.8.

We estimated the value of the second term of the right-hand side of Equation 4.8 by running Monte-Carlo simulations of cylindrical groups of unit virial radius with surface density profiles obeying the (projected) NFW model of scale radius of 0.2 (i.e. concentration 5). With 10 000 realizations each for $N = 2, 4, 8, 16, 32$ and 64 galaxies within the maximum projected radius allowed for the galaxies in the simulated groups, $R_{\rm max} = r_{200} = 1$, we found that the 95th percentile for the maximum – for all galaxies of the group – distance to the nearest neighbor is

$\mathrm{Max}(S_\perp) \simeq 1.48\,N^{-0.25}$ in units of the virial radius. Inserting this value of $\mathrm{Max}(S_\perp)/r_{\rm vir}$ into Equation 4.8, with $\Delta = 200$ and $\Omega_{\rm m} = 0.25$, we predict that to obtain a completeness of 0.95, we require

$$b_\perp \simeq 0.09\,N^{0.08}\ , \tag{4.9}$$

where we took into account that, for our adopted NFW model, the ratio of the number of galaxies within the virial sphere to that within the virial cylinder is $N_{\rm vir}/N \simeq 0.80$. Equation 4.9 predicts $b_\perp = 0.10$ for $N = 4$ and $b_\perp = 0.12$ for $N = 40$, i.e. $b_\parallel = 1.1$ for $N = 4$ and $b_\parallel = 1.3$ for $N = 40$, given $b_\parallel/b_\perp = 11$ found above. In other words, Equation 4.3 underestimates $\delta n/n$ by a factor $\mathrm{Max}(S_\perp)/r_{\rm vir}\,N_{\rm vir}^{1/3} \simeq 1.4\,N^{0.08}$, i.e. by 1.5 for $N = 4$ and 1.8 for $N = 40$. The slight increase of $b_\perp$ with richness suggests that fixing $b_\perp$ will lead to the fragmentation of rich groups.

Adopting instead the virial $\delta n/n = \Delta/\Omega_{\rm m} = 326$ (Planck, see above) would lead to $b_\perp = 0.14$ for $N = 4$ and $b_\perp = 0.17$ for $N = 40$. Since, at constant Δ, $b_\perp \propto \Omega_{\rm m}^{1/3}$ (Equation 4.4), moving from $\Omega_{\rm m} = 0.25$ to $\Omega_{\rm m} = 0.3$ (a compromise between WMAP and Planck), keeping $\Delta = 200$, yields $b_\perp = 0.11$ ($N = 4$) or 0.13 ($N = 40$). According to Equation 4.5, $b_\parallel/b_\perp$ does not vary with $\Omega_{\rm m}$ at fixed Δ, hence we now obtain $b_\parallel = 1.3$.

Had we taken a maximum projected radius that is much smaller than r_{200}, we would obtain a much smaller value for $b_\perp$. Indeed, our Monte-Carlo simulations indicate that with $R_{\rm max}$ and scale radius both equal to $0.2\,r_{200}$, we find $\mathrm{Max}(S_\perp) \simeq 1.85\,N^{-0.33}$ in units of $R_{\rm max}$, hence $\mathrm{Max}(S_\perp)/r_{200} \simeq 0.37\,N^{-0.33}$. Inserting this ratio into Equation 4.8, we now obtain $b_\perp = 0.023$, independent of N. Thus, to first order, $b_\perp$ scales with $R_{\rm max}/r_{200}$. Turning the argument around, a low $b_\perp$ leads to selecting galaxies in groups with projected radii limited to a small fraction of the virial radius.

We can also predict the reliability of the galaxy membership in groups, as follows. The expected number of interlopers from the extracted group out to a LOS distance of $\pm b_\parallel n^{-1/3}$ is

$$N_{\rm int} \approx 2\,\frac{N}{200}\,\frac{b_\parallel}{b_\perp}\ , \tag{4.10}$$

where we simply stretched the group by a factor of $b_\parallel/b_\perp$ along the LOS, and where N is the number of galaxies in the real space group. For $b_\parallel/b_\perp = 11$, Equation 4.10 yields $N_{\rm int} = 0.44$ for $N = 4$ and $N_{\rm int} = 4$ for $N = 40$. Thus, the fraction of interlopers should roughly be independent of the richness hence mass of the real space group. For $b_\perp \simeq 0.1$, corresponding to groups with overdensity 200 relative to critical sampled at 95% completeness, and sampling the LOS with 95% completeness (leading to $b_\parallel/b_\perp = 11$), one then expects $N_{\rm int}/N = 0.11$. One then infers a galaxy reliability of $R = (N/N_{\rm int})/[1 + (N/N_{\rm int})] = 90\%$.

Equation 4.10 assumes that the Universe is made of spherical groups that are truncated at their virial radii. In fact, galaxy clustering brings galaxies close to groups, in a fashion that the radial number density profile pursues a gradual decrease beyond the virial radius. For NFW models of concentration of 5, the projected number of galaxies within the virial radius is $1/0.80 = 1.25$ times the number within the virial sphere. Hence the numbers of interlopers to the virial sphere should satisfy $N_{\rm int}/N = 0.25$. Then, one expects a reliability of $R = (N/N_{\rm int})/[1 + (N/N_{\rm int})] = 80\%$.

4.2.2 Previous implementations

Table 4.1 lists the dimensionless LLs for the different group FoF analyses. The values of $\delta n/n$ and κ of different FoF analyses, inferred from their LLs according to Equation 4.3 and Equation 4.6, are listed in Table 4.1. One sees that 5 of the 7 previous studies advocate $b_\perp = 0.13$ or 0.14, and two (Eke et al., 2004 and Tempel et al., 2014 for $M_r < -19$) have pairs of LLs close to our predicted values of $(b_\perp, b_\parallel) \approx (0.11, 1.3)$. The two greatest outliers are Huchra & Geller (1982), whose transverse linking length appears too large and Robotham et al. (2011), both of whose

THESIS

LLs appear too small. We will check these conclusions in Section 4.4 and Section 4.5 using our analysis of mock galaxy and group catalogs.

4.2.3 Practical implementation of the FoF algorithm

There are two issues that need to be optimally handled when writing an FoF algorithm: rapidly extracting the separations in redshift space and properly estimating the mean density.

We followed the Huchra & Geller (1982) algorithm, used in most FoF implementations. Huchra & Geller write that two galaxies with redshifts z_i and z_j and an angular separation in θ_{ij} are linked using criteria that amount to

$$\left(\frac{c}{H_0}\right)(z_i + z_j)\sin\left(\frac{\theta_{ij}}{2}\right) \leq b_\perp\, n^{-1/3}\,, \tag{4.11}$$

$$\left(\frac{c}{H_0}\right)|z_i - z_j| \leq b_\parallel\, n^{-1/3}\,. \tag{4.12}$$

We generalized[2] Equation 4.11 and Equation 4.12 to[3]

$$\frac{d_{\rm comov}(z_1) + d_{\rm comov}(z_2)}{2}\,\theta \leq b_\perp\, n^{-1/3}\,, \tag{4.13}$$

$$|d_{\rm comov}(z_1) - d_{\rm comov}(z_2)| \leq b_\parallel\, n^{-1/3}\,. \tag{4.14}$$

Thus, Huchra & Geller (1982) and Berlind et al. (2006) neglected cosmological effects. For our deepest mock SDSS catalog, at $z = z_{\rm max} = 0.125$ (Catalog 6, see Table 3.1 below), $d_{\rm comov}/(cz/H_0) = 0.97$. So, the formula $d = cz/H_0$ leads to slightly too large distances, hence to slightly too strict choices of angles and differences in redshifts.

One could argue that, since groups are virialized, one ought to use the cosmological *angular distance*, $d_{\rm ang}(z) = d_{\rm comov}(z)/(1+z)$ for the distances with which one computes the physical transverse separation in terms of the angular separation. But one should then also compress the line-of-sight distances accordingly, and we are not aware of any work doing such a compression. Hence, we chose to stick with Equation 4.13 and Equation 4.14.

Since we are working with samples that are complete in luminosity, and since they are shallow enough that evolutionary effects are small, observers can estimate the mean number density of galaxies directly from the data.

Finally, for each galaxy, we computed the maximal angular distance to define the region in which potential neighbors could be found for the given transverse linking length. With the celestial sphere grid that we have constructed (see Appendix E), we searched for galaxies obeying the criterion of Equation 4.13, and then searched for galaxies meeting Equation 4.14. The linked galaxies were then placed in a tree structure according to the Union-Find method (Tarjan & van Leeuwen, 1984). Once all galaxies were analyzed, we compressed the trees constructed with linked galaxies by replacing, in each group, the links of links with links to a single galaxy, giving us the identity of the group to which galaxies belong to. This implementation allows for a fast computation of galaxy groups for large samples of galaxies.

4.3 Analysis

We tested the FoF algorithm by running it on our mock redshift-space, doubly complete subsamples of galaxies, for a set of 16×16 geometrically-spaced pairs of LLs. By directly comparing the

[2]The *comoving distance*, $d_{\rm comov}(z) = c\int dz/H(z)$, in Equation 4.13 should really be the *proper motion distance* $d_{\rm pm}(z) = d_{\rm lum}(z)/(1+z) = (1+z)\,d_{\rm ang}(z)$, but for flat cosmologies, $d_{\rm pm}(z) = d_{\rm comov}(z)$.

[3]Equation 4.13 is similar to the relation used by Zandivarez et al. (2014), with the exception of a minor difference in projected sizes given angle.

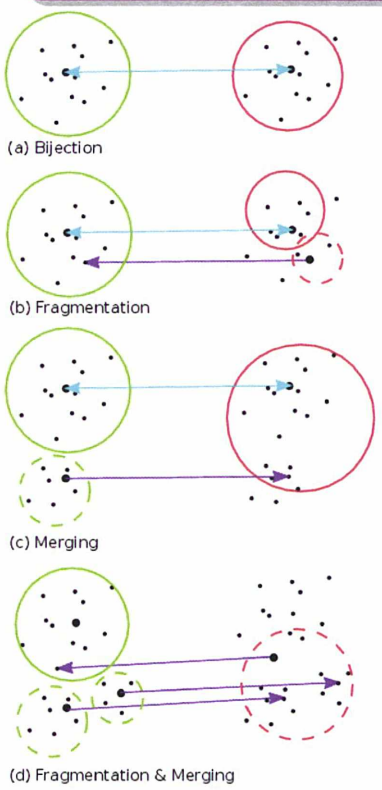


Figure 4.1: Schematic links between true groups (TGs, *green circles*) and FoF-extracted groups, (EGs, *red circles*), each with their respective most massive galaxy (*black dots*). The *solid circles* represent primary true and FoF groups, while the *dashed circles* respectively correspond to secondary true groups and FoF fragments. The *cyan double arrows* each indicate the one-to-one correspondence between the most massive galaxy in the true and extracted groups. The *purple rightwards-pointing arrows* correspond to the most massive galaxy of a true group ending up as a galaxy that is not the most massive of its extracted group. The *purple leftwards-pointing arrows* represent the cases where the most massive galaxy of an extracted group is not the most massive of its parent true group.

properties of our *extracted groups* (EGs) in redshift space with their "parent" *true groups* (TGs) in real space, we could assess the performance of the FoF in recovering the real space information from the projected phase space observations. Note that TGs can have as little as one single member galaxy. Also, galaxies in redshift space with no linked galaxies can be considered as EGs with one single galaxy.

4.3.1 Linking real space and projected redshift space

There are several ways to link the EGs and TGs. We followed Yang et al. (2007), by linking the EG to the TG that contains the EG's most massive galaxy (MMG), and conversely linking the TG to the EG that contains the TG's MMG. With this definition for linking, we could easily associate FoF groups to real groups.

4.3.2 Global tests

Our definition of the link between EGs and TGs allowed us to search for cases where there is no one-to-one correspondence between the groups in real and redshift space: a TG can suffer from *fragmentation* into several EGs, while an EG can be built from the *merging* of several TGs.

Figure 4.1 illustrates different cases (following an analogous figure in Knobel et al., 2009). The

UFES TESIS

top panel shows a one-to-one correspondence between the true and extracted groups.

We defined a fragmented TG as one that contains the MMGs of several EGs. Multiple situations can cause fragmentation of TGs. In some cases, the FoF algorithm fails to recover entire TGs, selecting instead its primary and secondary substructures (see panel Figure 4.1b). In other cases, an EG is mostly composed of galaxies from one TG, but the MMG of another TG is 'accidentally' linked to the first TG. In consequence, the EG could be linked to a TG providing only a single member galaxy to the EG, in comparison with more members arising from another TG. When fragmentation occurred, we distinguished the *primary EG*, as that whose MMG corresponds to the MMG of the parent TG, from the other EGs, which we called *fragments*.

The dual of fragmentation is merging. In this situation, an EG contains the MMGs of several TGs. Proceeding similarly as for the case of fragmentation, we denoted *primary TG* of a given EG the TG whose MMG corresponds to the MMG of that EG, denoting the other TGs as *secondary*. An example of merging is shown in Figure 4.1c. Note that a true group can be fragmented and its primary extracted group can be the result of a merger of the true group with another one, as illustrated in Figure 4.1d.

4.3.3 Local tests

Our local tests check the membership of the EGs. We defined *completeness* as the fraction of galaxies in the TG (i.e. within the sphere of radius r_{200}) that were members of the primary EG. Given this definition, it did not make sense to consider the completeness for secondary fragments, hence we limited our tests to the primary EGs.

We defined *reliability* as the fraction of galaxies in the EG that were members of the parent TG (i.e., within the sphere of radius r_{200}). Here, we also limited our tests to the primary EGs.

Mathematically speaking, these definitions of galaxy completeness, C, and reliability, R, can respectively be written as

$$C = \frac{\mathrm{TG} \cap \mathrm{EG}}{\mathrm{TG}} ,$$
$$R = \frac{\mathrm{TG} \cap \mathrm{EG}}{\mathrm{EG}} .$$

Looking at Figure 4.1, the completeness is the fraction of galaxies in the TG (left, green circles) recovered in the EG (right, red circles), while the reliability is the fraction of galaxies in the EG that belong to the TG.

These four quantities allow one to define the capacity of the FoF grouping algorithm (or any other grouping algorithm) to recover groups in real space from galaxy catalogs in redshift space.

Note that EGs that are fragments can have high reliability, while fragmentation causes primary EGs to have reduced completeness. When EGs are mergers of TGs, the secondary TGs lead to a decrease in the reliability, but can have high completeness.

4.3.4 Mass accuracy

There are many properties of groups that one wishes to recover with optimal accuracy (see Sect. 4.1). We focused here on one single property that appeared to us as the most relevant: the group total mass. We measured the masses of our EGs using the virial theorem formula of Heisler, Tremaine, & Bahcall (1985)

$$M_{\mathrm{EG}} = \frac{3\pi}{G} \langle R \rangle_{\mathrm{h}} \, \sigma_v^2 = \frac{3\pi \, N}{2\,G} \frac{\sum v_i^2}{\sum_{i<j} 1/R_{ij}} , \tag{4.15}$$

where $\langle R\rangle_{\rm h} = \langle 1/R_{ij}\rangle^{-1}$ is the harmonic mean projected separation, while σ_v is the unbiased measure of the standard deviation of the group velocities.

More precisely, we computed the accuracy of the log masses, respectively defining the *bias* and *inefficiency* as the median and equivalent standard deviation (half 16–84 interpercentile) of $\log(M_{\rm EG}/M_{\rm TG})$, where $M_{\rm TG}$ is the mass of the TG within the sphere of radius r_{200} (see Section 4.3.3).

4.3.5 Quality

It is not simple to extract a unique pair of optimal LLs from the four tests (fragmentation, merging, completeness, and reliability). To reduce the number of tests, we combined fragmentation and merging into a single *global quality* and combined completeness and reliability into a single *local quality*.

We could define our qualities by multiplying F (fragmentation) by M (merging) and similarly, C by R. However, one could alternatively multiply $1-F$ by $1-M$, etc. Instead, we chose quality estimates that minimize the distance to the perfect case. The advantage of using distance rather than multiplying probabilities is that the former gives less weight to situations where one of the two parameters is perfect and not the other. For example, consider the case $F = M = p$. With the multiplication method, we would find that $Q = p^2$ is also reached with $F = \epsilon \ll 1$, yielding $M_{\rm mult} = p^2/\epsilon$, which can be quite large (hence plenty of merging). On the other hand, with the distance method, we would find that $Q = p\sqrt{2}$ is also reached with $F = \epsilon \ll 1$ for $M_{\rm dist} \simeq p\sqrt{2}$, which is much more restrictive. In a perfect algorithm, fragmentation and merging don't occur, hence $F = M = 0$ they are null. We therefore chose to minimize the *global quality*, defined as

$$Q_{\rm global} = \sqrt{F^2 + M^2} \tag{4.16}$$

Moreover, in a perfect grouping algorithm, the EGs are fully complete and reliable, i.e. $\langle C\rangle = \langle R\rangle = 1$, where the means are over all the groups of a mass bin. We, hereafter, drop the brackets, so that C and R should now be understood as means over groups within mass bins. We then define the *local quality* as

$$Q_{\rm local} = \sqrt{(1-C)^2 + (1-R)^2}\,. \tag{4.17}$$

Both global and local qualities tend to zero for a perfect galaxy group algorithm. So the optimal LLs will be those that minimize $Q_{\rm global}$, $Q_{\rm local}$, mass bias and mass inefficiency. The maximum possible value of both qualities is $\sqrt{2}$.

4.3.6 Scope of the tests

We limit our tests to TGs containing at least 3 galaxies and that are not split by the transformations of the simulation box (see Chapter 3). Moreover, we only consider EGs with at least 3 galaxies and that do not lie near the survey edges (the virial radius, 2.3 Mpc, of a true group of log mass 15.2 in solar units, placed at $z = z_{\rm min} = 0.01$, i.e. at an angle of more than 3.°27) or redshift limits ($1.8\,v_{200} \approx 2.7\,\sigma_v$, of the same mass group, corresponding to $3073\,{\rm km\,s^{-1}}$). Typically 60% (sample 2) to 25% (sample 6) of the groups are flagged. Finally, the tests of galaxy completeness and reliability, as well as mass bias and inefficiency are restricted to primary EGs of TGs (not fragments).

4.4 Results

We have applied the FoF algorithm on near and distant doubly complete subsamples (numbers 2 and 6 in Table 3.1), repeating the tests for a grid of 16×16 geometrically-spaced pairs of LLs.

THESIS

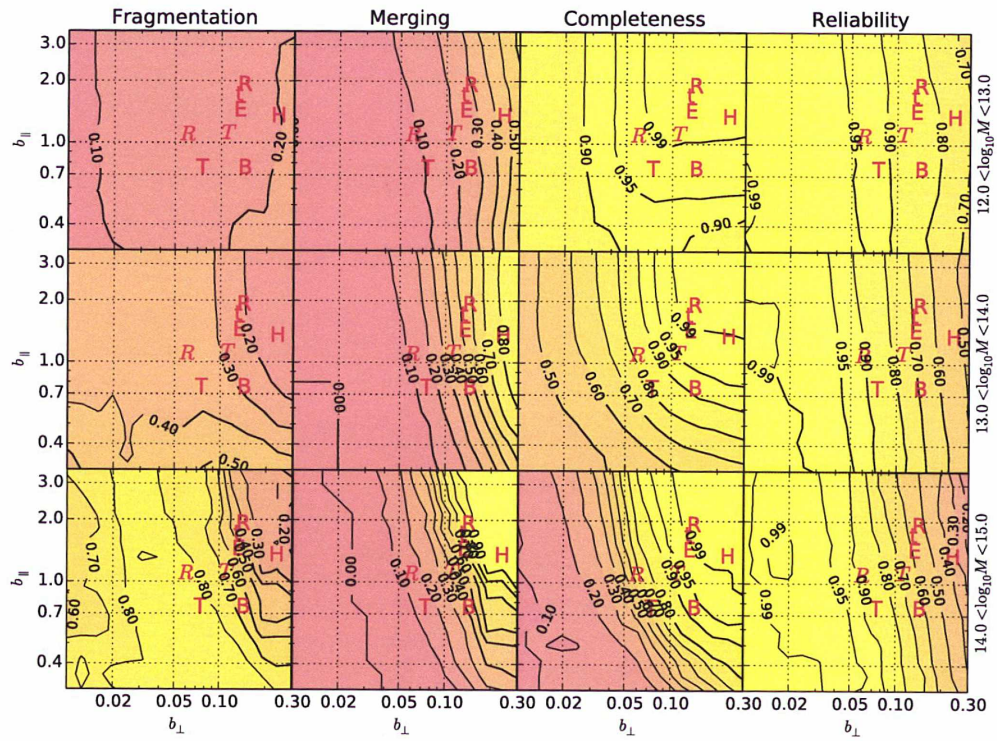


Figure 4.2: Contours of group fragmentation (*first column*) and merging (*second column*), as well as mean galaxy completeness (*third column*) and reliability (*fourth column*) computed for a 16×16 grid of linking lengths for the nearby doubly complete galaxy subsample 2 in Table 3.1. Results are shown for three bins of true group masses, for unflagged groups of at least 3 members (for both the extracted and parent groups), and further restricted to primary groups in the completeness and reliability panels. Pairs of linking lengths corresponding to previous are also shown as *red letters* (H: Huchra & Geller 1982; R: Ramella et al. 1989; t: Trasarti-Battistoni 1998; E: Eke et al. 2004; B: Berlind et al. 2006; T: Tago et al. 2010; *R*: Robotham et al. 2011; *T*: Tempel et al. 2014).

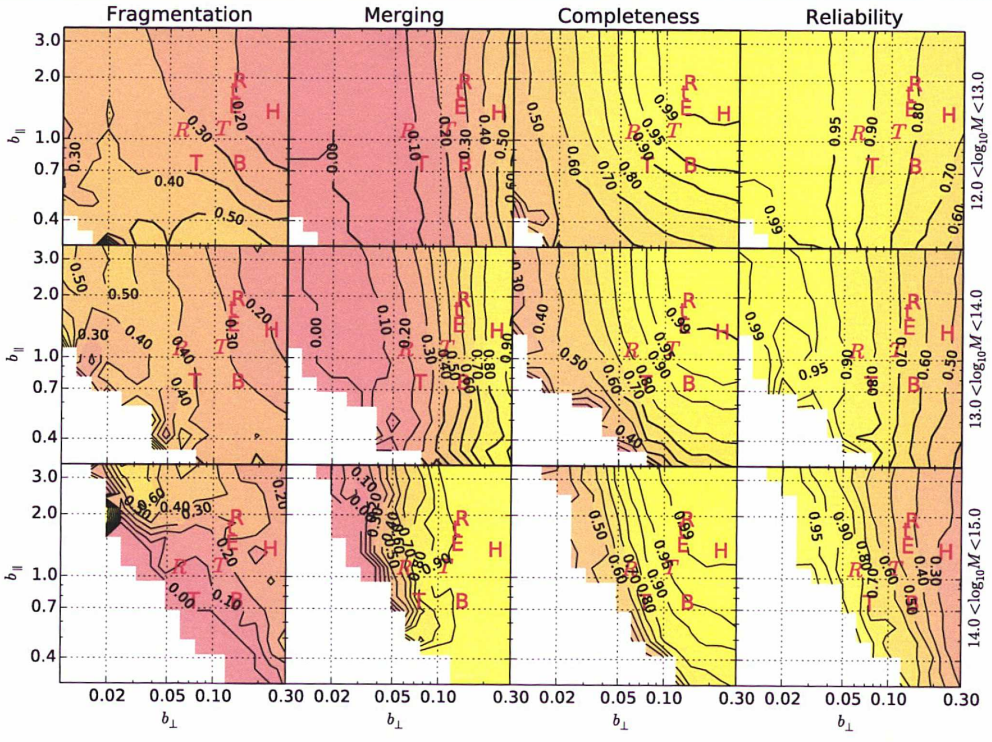


Figure 4.3: Same as Figure 4.2, but where the different rows correspond to different bins of extracted group masses estimated from the virial theorem. The white zones show cases where the linking lengths led to no unflagged groups extracted.

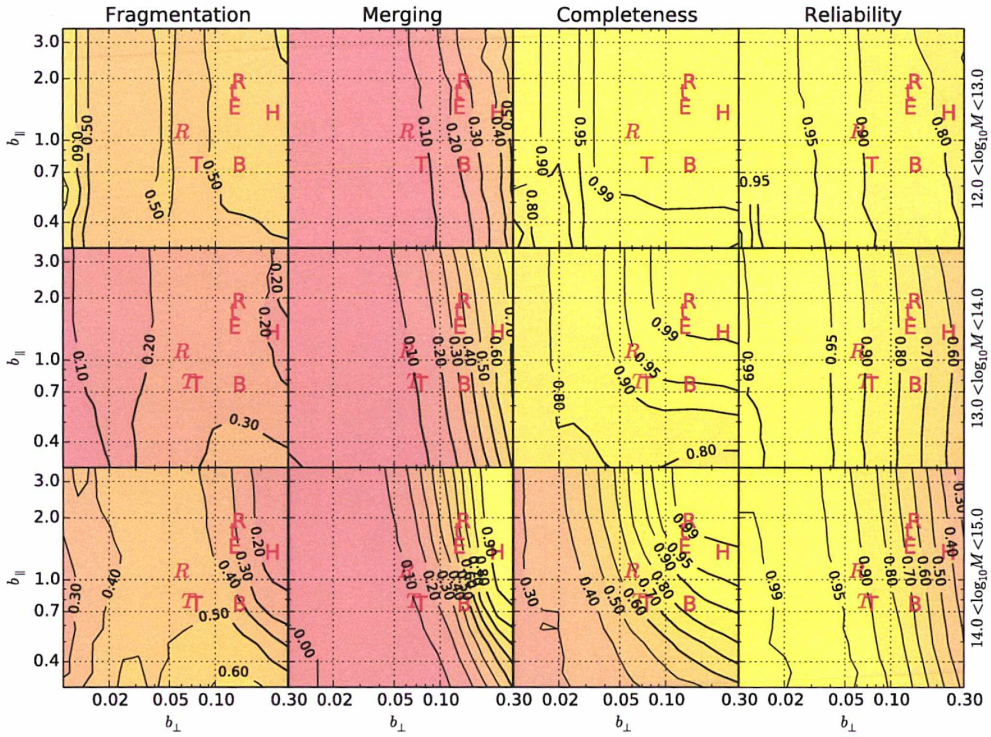


Figure 4.4: Same as Figure 4.2, but for the distant doubly complete galaxy subsample 6 in Table 3.1.

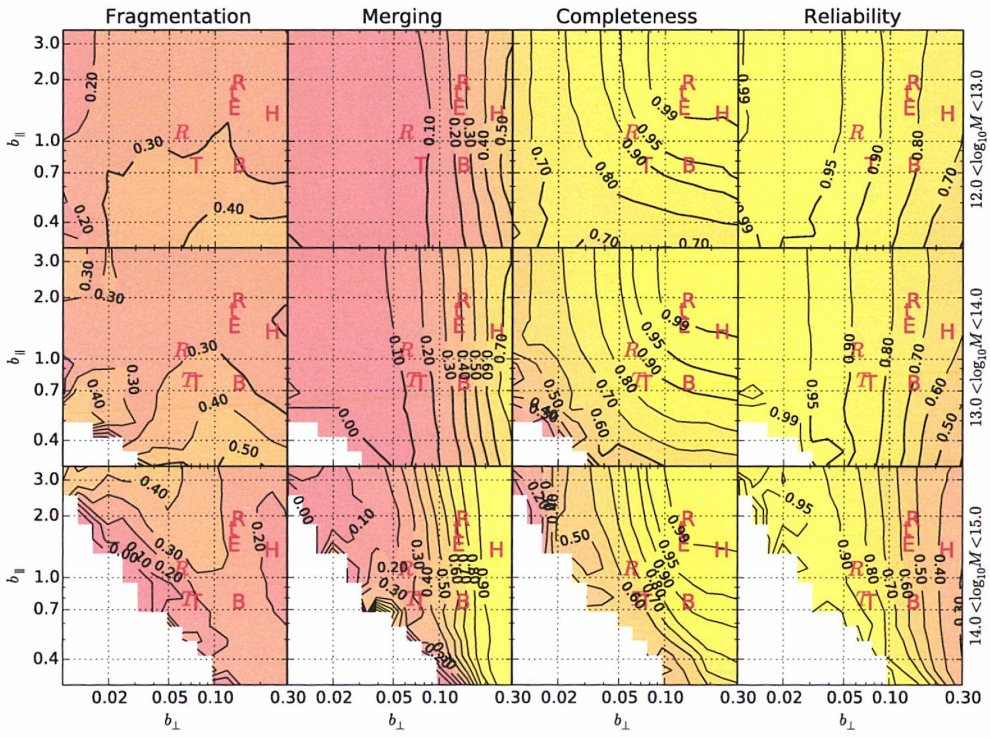


Figure 4.5: Same as Figure 4.4, but where the different rows correspond to different bins of estimated masses.

TESIS

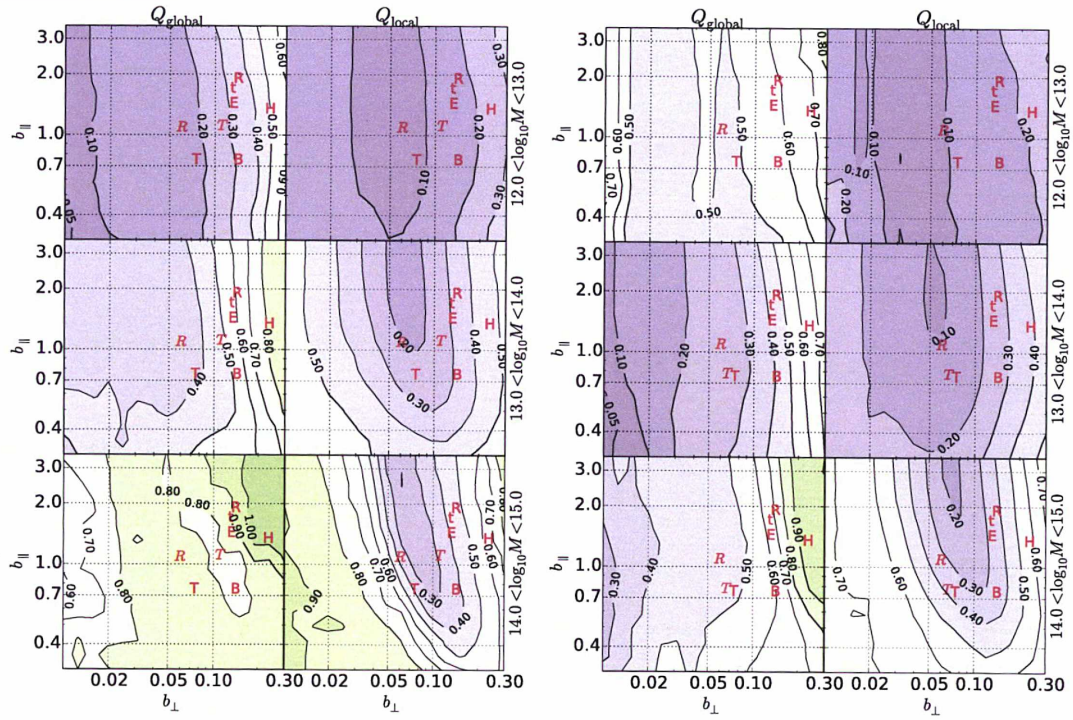


Figure 4.6: Global and local quality factors in a 16×16 grid of linking lengths for subsamples 2 (*left*) and 6 (*right*), in three bins of true masses Results are shown for unflagged groups (restricted to primary groups for $Q_{\rm local}$) of at least 3 members (in both the true and extracted group). The symbols are as in Figure 4.2

The results of our tests are shown in Figure 4.2 and Figure 4.8. The LLs of the different grouping studies listed in Table 4.1 are shown, except for Merchán & Zandivarez (2002), whose LLs nearly overlap with those of Eke et al. (2004).

4.4.1 Group fragmentation and merging

Figure 4.2 indicates that, for the nearby doubly complete subsample (number 2), fragmentation only affects the massive TGs (up to ≈80% of them for popular LLs), while Figure 4.3 shows that, for popular LLs, the fragmentation is lower (10–30%) at high EG mass, hence fragment masses tend to be small (typically 20–40% fragmentation at small and intermediate estimated masses).

On the other hand, the distant doubly complete subsample behaves in almost the opposite manner: fragmentation is most important at the lowest TG masses (roughly 50% fragmentation, Figure 4.4) and is independent of estimated EG masses (at roughly 20–30%, Figure 4.5).

In any event, fragmentation tends to decrease with greater linking lengths, as expected, although it decreases somewhat faster with increasing $b_\perp$ than with increasing $b_\parallel$.

Since merging is the dual of the fragmentation, one expects the level of merging to vary in the opposite way as fragmentation. Indeed, Figure 4.3 and Figure 4.5 indicate that merging becomes more important at higher estimated masses, respectively reaching up to 90% and 65% for high estimated masses with popular choices of LLs in subsamples numbers 2 and 6. However, Figure 4.2 and Figure 4.4 shows that the merging fraction increases only slowly with TG increasing mass, with typically 15–40% (increasing fast with $b_\perp$) of the TGs being merged with other ones. Finally, merging decreases with smaller LLs, especially with smaller $b_\perp$.

Figure 4.6 and Figure 4.7 show the $Q_{\rm global}$ quality indicator that combines fragmentation and merging into a single parameter. These figures show that decreasing $b_\perp$ leads to a better tradeoff

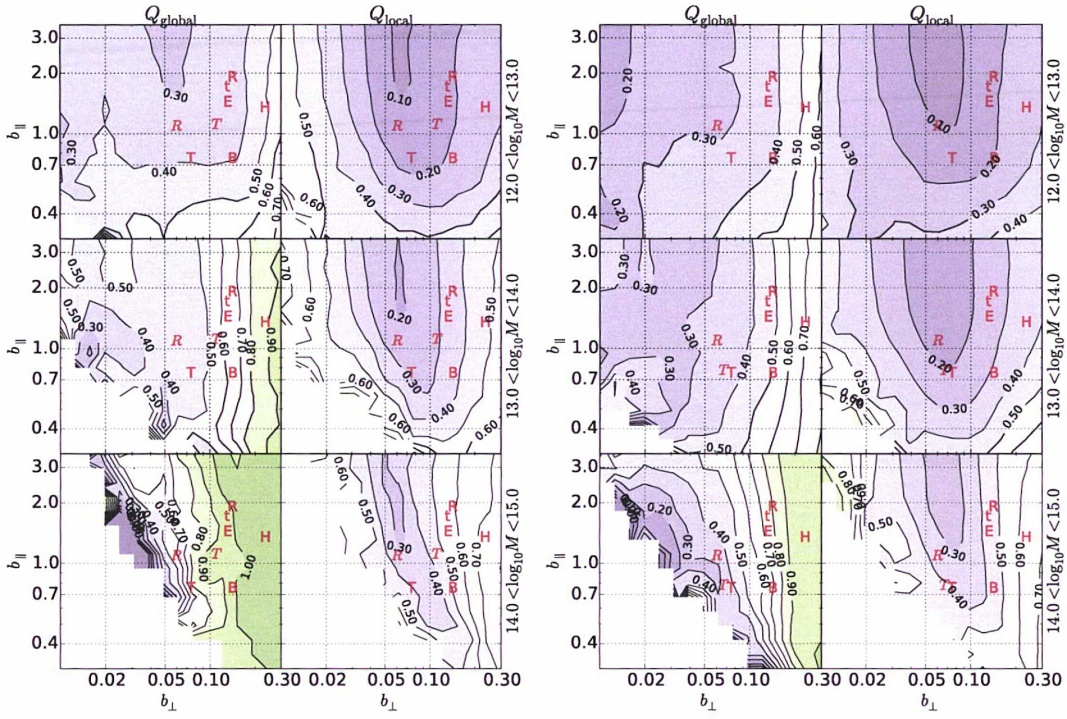


Figure 4.7: Same as Figure 4.6 but in bins of estimated masses. The white zones show cases where the linking lengths led to no unflagged groups extracted.

between fragmentation and merging, i.e. that the decrease of merging with decreasing $b_\perp$ has a stronger effect than the increase of fragmentation with decreasing $b_\perp$: the optimal $Q_{\rm global}$ is often reached for $b_\perp < 0.02$.

4.4.2 Galaxy completeness and reliability

Figure 4.2 and Figure 4.4 indicate that completeness is very high ($> 99\%$) at low TG masses, and decreases to lower values ($60 - 99\%$) at high TG mass. A weaker trend occurs when EG mass is substituted for TG mass (see Figure 4.3 and Figure 4.5). Since high mass TGs are less complete, their estimated masses should be smaller, and the EGs with high masses will be the lucky complete ones, which explains the weaker trend of completeness with EG mass. Note that we are only considering primary groups of at least 3 members. The transverse and LOS linking lengths have roughly the same impact on galaxy completeness.

The reliability of the group membership decreases with increasing EG mass (Figure 4.3 and Figure 4.5): regardless of the subsample, the reliability is 80–90% for low mass EGs, but only 50–85% for high mass EGs. The value of $b_\parallel$ has virtually no effect on galaxy reliability. We will discuss this lack of convergence of the reliability with $b_\parallel$ in Section 4.5.

Galaxy reliability also decreases with the masses of the TGs, but the trend is weaker (Figure 4.2 and Figure 4.4): as the reliability decreases from 85–95% to 60–90%, roughly independent of the subsample.

The right panels of Figure 4.6 and Figure 4.7 show that, again, the transverse LL appears to be more decisive than the LOS one when combining galaxy completeness and reliability into a single local quality factor.

THESIS

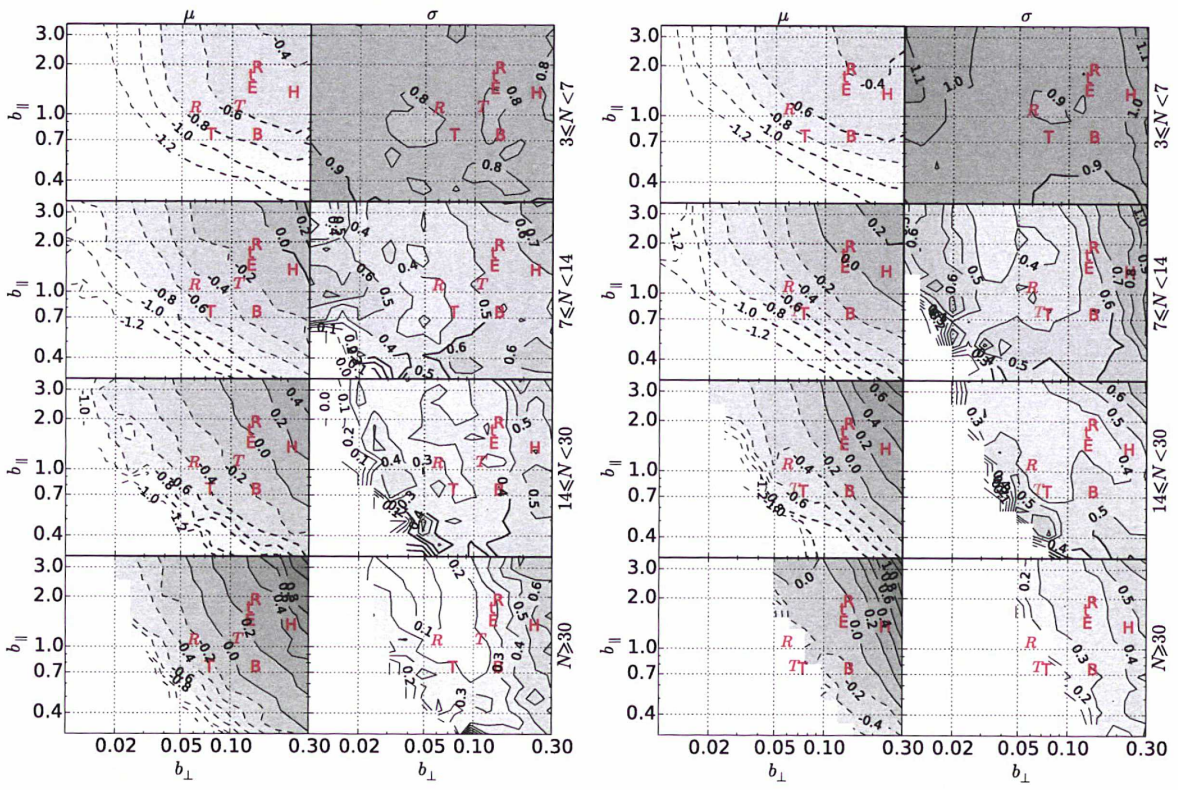


Figure 4.8: Bias (μ) and inefficiency (σ) of the group masses estimated by the virial theorem (Equation 4.15) on our 16×16 grid of linking lengths, in four bins of extracted group richness (we do not consider extracted groups for which the parent true group has ≤ 3 members). The bias and inefficiency are respectively computed as the median and half 16–84 interpercentile of $\log_{10}(M_{\rm EG}/M_{\rm TG})$. Results are shown for primary, unflagged groups. The left and right panels are respectively for galaxy subsamples 2 and 6. The symbols are as in Figure 4.2. The white zones indicate linking lengths with no unflagged groups extracted.

4.4.3 Mass accuracy

The left columns of the two panels of Figure 4.8 show that the primary EG masses recovered by the FoF algorithm are systematically biased low: for the popular choices of LLs, the bias (μ) is as strong as -0.6 ± 0.2 dex at low multiplicity ($N_{\rm EG} \leq 6$), decreasing to 0.0 ± 0.3 dex at high multiplicity ($N_{\rm EG} \geq 30$).

The right columns of the two panels of Figure 4.8 indicate that, even if the biases could be corrected for, the masses cannot be recovered to better than 0.8–0.9 dex at low multiplicity, improving to 0.2 dex at high multiplicity. The inefficiency (σ) is minimal for $b_\perp \approx 0.05$ (within a factor 2) and $b_\parallel \approx 1.0$ (low richness) or $b_\parallel \gtrsim 1.0$ (intermediate and high richness). For transverse LLs within 40% of $b_\perp = 0.1$, the inefficiency is not very insensitive to $b_\parallel$.

The situation becomes even worse when fragments are included in the statistics. In this work, we have separated the accuracy of the group masses with the occurrence of group fragmentation. But observers cannot tell if a group is a fragment or a primary EG.

4.5 Conclusions and Discussion

Before testing the FoF algorithm using a mock galaxy catalog in redshift space, we first argued on physical grounds (Section 4.2.1) that the normalized transverse linking length, ought to be $b_\perp \approx 0.10$ (slightly increasing with richness) to extract 95% of the galaxies within the virial

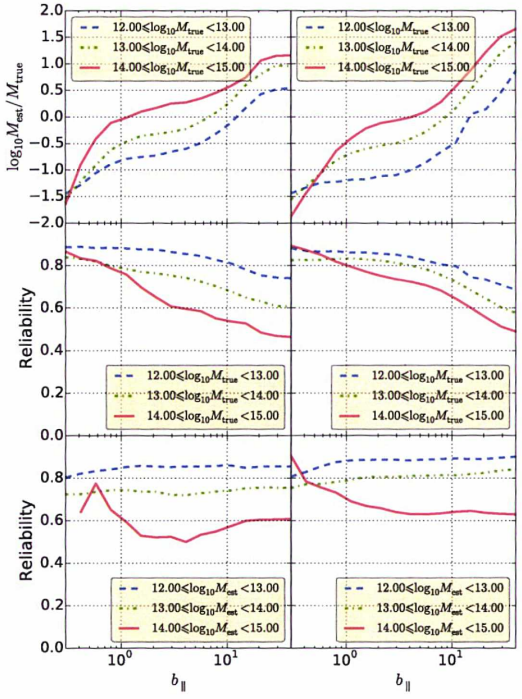


Figure 4.9: Variation of the mass bias and reliability as a function of $b_\parallel$ for $b_\perp = 0.1$, for subsamples 2 (*left*) and 6 (*right*).

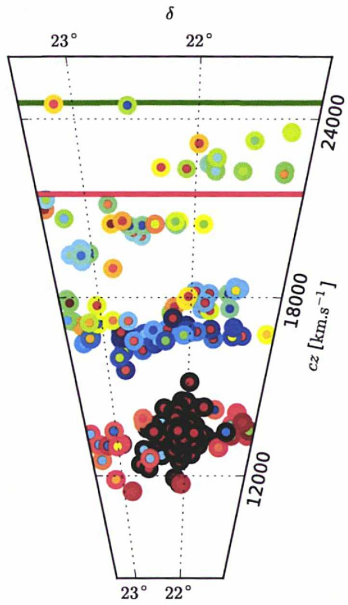


Figure 4.10: An example of group and halo for $b_\parallel = 20.8$ and $b_\perp = 0.1$ for subsample 4. The width of the cone is exaggerated by a factor of roughly 5 for illustrative purposes. *Outer* and *inner circle colors* respectively refer to the TGs and EGs. The *horizontal green* and *red lines* respectively indicate the maximum redshift, $z_{\max}$ and the redshift where galaxies are flagged for being close to $z_{\max}$. Some galaxies of the red EG, whose TG is the black one, are flagged for being close to $z_{\max}$, hence the group would not be considered in our tests.

radius of NFW true groups. We also argued that, restricting the galaxies along the line-of-sight to $\pm 1.65\,\sigma_v$ (95% of the galaxies) for groups defined to be 200 times denser than the critical density of the Universe, requires $b_\parallel / b_\perp \approx 11$, hence $b_\parallel \simeq 1.1$. These LLs are estimated from our mocks that are based upon the Millennium-II simulation that had adopted $\Omega_{\rm m} = 0.25$. Converting to $\Omega_{\rm m} = 0.3$ yields $b_\perp = 0.11$ and $b_\parallel = 1.3$. Finally, estimating the contamination by interlopers, we predict between 80% (NFW model extended outwards) to 90% (NFW model truncated to sphere plus random interlopers) galaxy reliability.

We then built a mock redshift space galaxy catalog with the properties of the flux-limited SDSS primary spectroscopic sample, from which we extracted 2 subsamples that are doubly complete in distance and luminosity (Chapter 3). We then extracted groups from both of these subsamples, running the standard FoF algorithm for 16×16 pairs of linking lengths. In each case, we measured the fraction of true groups that were fragmented in the FoF extraction process, the fraction of extracted groups that were built by the merging of several true groups, as well as the bias and inefficiency with which the group masses were extracted. Moreover, we computed the completeness and reliability of the galaxy membership relative to the spheres of radius r_{200} in which the true groups are defined.

We analyzed group fragmentation, merging, galaxy completeness and reliability, mass bias and inefficiency for two doubly complete subsamples and in bins of true and estimated mass or estimated richness (for the mass accuracy).

We found that massive true groups are more prone to fragmentation, as expected, but that, for popular choices of linking lengths, the probability of fragmentation is greatest (30%) at low

estimated mass, i.e. the fragments are of low mass. The process of fragmentation of rich (massive) groups is similar to images of large galaxies being preferentially fragmented by automatic image extraction pipelines (e.g., De Propris et al., 2007).

Group merging is low at low estimated mass, but increases drastically to reach 40–90% (for popular linking lengths) at high estimated mass. Galaxy completeness is high, typically $> 80\%$. Galaxy reliability is typically 75 to 90% depending on group mass.

Our analytical prediction of 95% completeness for $b_\perp \simeq 0.10$ is only met for groups of high true masses (Figure 4.2 and Figure 4.4). Groups of low mass will have more concentrated galaxy populations, which will lead to smaller values of $\mathrm{Max}(S_\perp)/r_{200}$, hence smaller values of $b_\perp$. Also, our analytical prediction of 80–90% reliability for groups with $b_\perp = 0.10, b_\parallel = 1.1$ is accurate for groups of all masses of the distant subsample (Figure 4.4). However, for the nearby subsample (2), our predicted reliabilities are only accurate for groups of low true masses, but optimistic for higher mass groups, for which $R \simeq 70 - 75\%$.

Group merging and galaxy reliability depend little on $b_\parallel$, especially at high transverse linking length, $b_\perp > 0.1$, where the galaxies are extracted to projected radii beyond r_{200}, hence the contamination by interlopers is mainly in the transverse direction. The lack of optimal $b_\parallel$ for galaxy reliability may seem surprising at first. We checked our analysis by measuring the reliability for $b_\perp = 0.1$, for a very wide range of $b_\parallel$ extending from 0.3 to 40. The top panels of Figure 4.9 indicate that the reliability does end up decreasing fairly fast beyond some large value of $b_\parallel \simeq 6$, i.e. beyond the limits of Figure 4.2 and Figure 4.4. The second row of panels of Figure 4.9 show a different behavior in bins of estimated mass. This is the consequence of the estimated mass increasing very fast with $b_\parallel$, as shown in the bottom panels of Figure 4.9. The increase, with increasing $b_\parallel$, of the mass bias is roughly parallel to the corresponding decrease of the reliability (in bins of TG mass). At low $b_\parallel$, the reliability decreases fairly rapidly and the mass bias increases rapidly (towards zero), then both settle into an almost constant plateau in the range $1.4 \lesssim b_\parallel \lesssim 8$, then both worsen rapidly up to $b_\parallel \simeq 25$, beyond which both saturate, because the longitudinal link is so large that one reaches the minimum and maximum redshifts of the subsample, where most groups are flagged. Massive groups that are built from TG merging can be fairly reliable if the secondary TGs have negligible mass relative to the primary one. This explains why R remains fairly high when M is high. The plateau around $b_\parallel \approx 3$ appears to represent the range of optimal longitudinal LLs.

An illustration is given in Figure 4.10, where a given EG has reached the limits of the catalog with a very large value of $b_\parallel$. Figure 4.10 also shows that interloping TGs are highly clustered. This may explain why increasing $b_\parallel$ has only a small effect on galaxy reliability: there is a void behind the main TG (black outer circles).

While fragmentation, measured in bins of true group mass, decreases with increasing $b_\parallel$, as expected (Figure 4.2 and Figure 4.4), we find that in bins of estimated mass, the fraction of groups that are (secondary) fragments increases with $b_\parallel$ (Figure 4.3 and Figure 4.5). We believe that this is caused by interlopers increasing the group estimated mass (Figure 4.9).

The masses, estimated with the virial theorem (Equation 4.15) are a strong function of the multiplicity of the extracted group. The estimated masses are systematically biased low, especially for low extracted group multiplicities (typically by a factor 4!). Similar trends have been found for FoF groups (Robotham et al., 2011) and for other, mostly dynamical, group mass estimators (Old et al., 2014). The estimated group masses are inaccurate, even after correcting for the biases: the typically errors are 0.8–0.9 dex at low multiplicity, decreasing to 0.3 dex at high multiplicity.

The optimal completeness and reliability of the galaxy membership lead to fairly extreme linking lengths, i.e. $b_\perp < 0.1$ and $b_\parallel > 2$. However, the use of such a small transverse linking length amounts to extracting the inner regions of groups, thus missing their outer envelopes. Indeed, one notices that fragmentation worsens at increasingly lower values of $b_\perp$. Therefore, our

attempt to define a local quality by combining galaxy completeness and reliability is of little use if one wishes to recover galaxies out to close to the virial radii of groups.

In fact, the optimal linking lengths depend on the scientific goal:

- statistical studies of environmental effects require high reliability (say $R > 0.9$), accurate masses and, to a lesser extent, minimal fragmentation.
- cosmographical studies of group mass functions require accurate masses, minimal group merging and fragmentation.
- studies for followups at non-optical wavelengths (e.g. X-rays), benefit from high completeness.

For statistical studies of environmental effects, it seems best to adopt $b_\perp \simeq 0.06$, $b_\parallel \approx 1.0$, for which the reliability is roughly as high as it gets for the choice of $b_\perp$: over 90% at low $M_{\rm EG}$ and over 80% at intermediate and high $M_{\rm EG}$. Then, the completeness is higher than 70% at high estimated mass and much higher at low $M_{\rm EG}$. The mass inefficiency is minimal, but with this choice of LLs, there will be virtually no EGs with more than 30 galaxies in the distant more luminous subsample (Figure 4.8).

This choice of LLs is close to that of Robotham et al. (2011), which may seem obvious since both studies used some form of optimization of the LLs. However, the details of the optimization criteria are somewhat different: Robotham et al. multiplied four criteria: basically the group completeness and reliability, which bears some resemblance to our group fragmentation and merging, but theirs is based on TG-EG pairs that have more than half their galaxies in common, as well as two measures of a combination of galaxy completeness and reliability, averaged over TGs and EGs respectively. Our analysis differs in that we directly constrained group fragmentation and merging, as well as galaxy completeness and reliability for primary fragments, and finally mass accuracy.

For cosmographical and other studies involving accurate group mass functions, it appears best to adopt $b_\perp \simeq 0.05$, $b_\parallel \simeq 2$, as lower $b_\parallel$ increases fragmentation (Figure 4.3 and Figure 4.5), while higher $b_\parallel$ causes too high group fragmentation at high EG masses. This value of $b_\parallel \simeq 2$ is in agreement with the intersection of the regions of $(b_\perp, b_\parallel)$ space that optimize both the multiplicity function and velocity dispersions obtained by Berlind et al. (2006).

Finally, for non-optical followups, for which galaxy completeness is perhaps the sole important parameter, one should privilege large linking lengths, e.g. $b_\perp \simeq 0.2$, $b_\parallel \simeq 2-4$. However, one can also adopt $b_\perp = 0.1$, $b_\parallel \simeq 2-4$, for which the completeness is greater than 95% at all masses and for both subsamples.

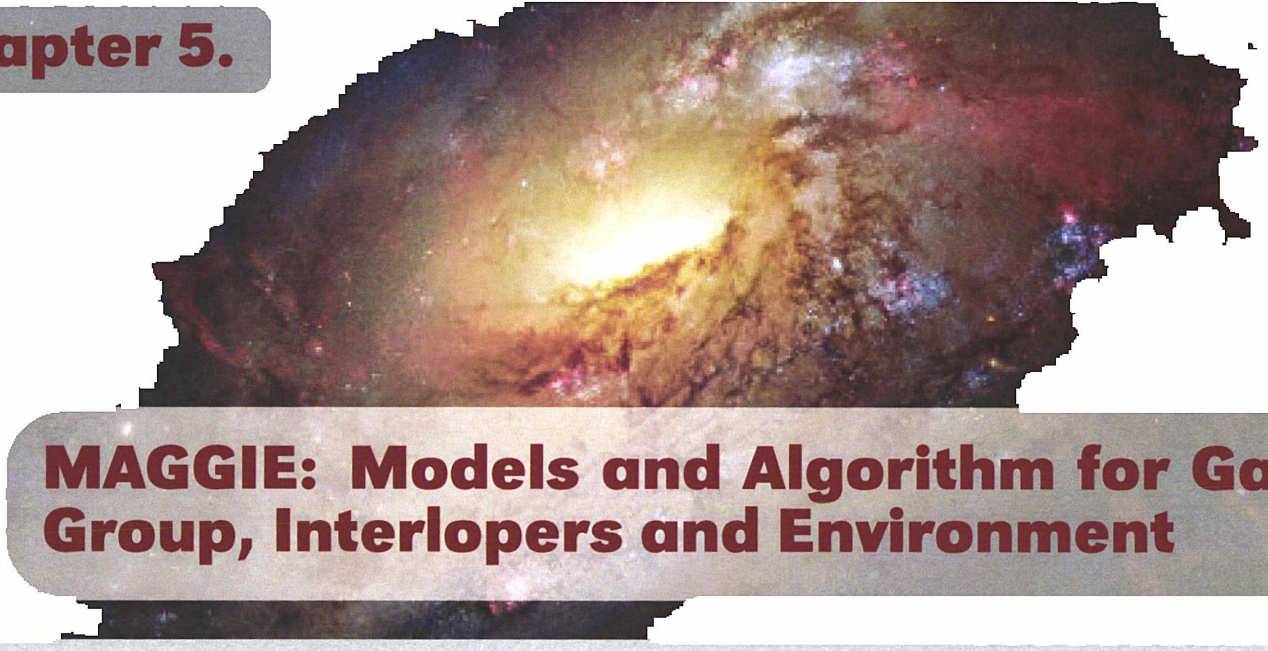
Converting from $\Omega_{\rm m} = 0.25$ (Millennium-II Simulation) to $\Omega_{\rm m} = 0.3$ (WMAP-Planck compromise), $b_\perp$ must be increased by 6% (Equation 4.4) to $b_\perp \simeq 0.07$ for the choices optimizing environmental or cosmographical studies. Since $b_\parallel/b_\perp$ is independent of $\Omega_{\rm m}$ at given Δ, $b_\parallel$ must also be increased by 6%, i.e. to $b_\parallel \approx 1.1$ for environmental studies.

We finally note that while high estimated mass group fragmentation and merging depends on the particular doubly complete subsample, galaxy completeness and reliability as well as mass accuracy depend little on the subsample. Berlind et al. (2006) had similarly concluded that the doubly complete subsample influenced little their tests of the group multiplicity function and the accuracy of projected radii and velocity dispersions.

FoF grouping techniques can be used as a first guess for other more refined grouping methods (Yang et al., 2005, 2007). In a future paper (Duarte & Mamon, 2014b), we will present another grouping algorithm, which is not an FoF, but is instead a probabilistic grouping algorithm that is built upon our current knowledge of groups and clusters (partly from X-rays and independent of FoF analyses of optical galaxy samples) and from cosmological N body simulations.

THESIS

Chapter 5.



MAGGIE: Models and Algorithm for Galaxy Group, Interlopers and Environment

Contents

5.1 Introduction

We showed in the previous chapter that the very popular grouping algorithm should have its two linking lengths optimized, and that this depends on the scientific goal of the group catalog obtained. With these limitations, it is clear why Bayesian methods appeared. Indeed, with our knowledge of the galaxy formation and evolution processes, it is possible to constrain better the galaxy grouping. With the FoF algorithm, galaxies are selected in a pure geometrical way, and their formation history doesn't matter in this selection, since only the over-density is relevant. With Bayesian algorithms, it is possible to combine geometrical and physical approaches. The history of galaxies is available by their observable properties such as luminosity, stellar mass, morphology and is used to assign galaxies to a group, in complement of the geometrical information from the density.

We already described Bayesian algorithms in Chapter 2, for example Yang et al. (2007) or Domínguez Romero et al. (2012), where similar spatial methods to the FoF are adopted, with

THESIS

priors on the density profile of galaxies inside halos to constrain the assignment. But because of observational uncertainties, model divergences, various incompletenesses..., the extraction of groups from observational data will always be affected by these problems, and the galaxy environment polluted by interlopers, creating biases in group characteristics. This leads to the blurring of the modulation of galaxy properties with their environment and of our understanding of intra-groups physical processes.

Recently, with the improvement of computer performances in terms of memory and CPU power, it becomes possible to include many priors in the computation, and to use the most computer-intensive applications of statistics. Since interlopers will still be problematic, the new powerful computer era allows for probabilistic membership of galaxies in groups. Systematic errors in galaxy surveys can be reduced or integrated in the grouping by probabilities. For example, Liu et al. (2008) used a probabilistic FoF in a galaxy survey with photometric redshifts to avoid the uncertainties inherent to this method. Domínguez Romero et al. (2012) also used "responsibilities" to improve the assignment of galaxies to groups and reduce the effect of interlopers on the observable properties of groups. In Rykoff et al. (2014), galaxies have their probabilities based on the group richness estimations.

It seems that using probabilities to describe the membership inside galaxy groups will be inevitable, because of the systematic errors and biases presents in the actual and future galaxy surveys. In particular, the modulation of the galaxy properties with their environment that we want to extract from galaxy group catalogues should be less biased by interlopers if we use probabilities as weights. Indeed, interlopers, even if they are still present in the group membership, will have a low probability to pertain to the group, and their contribution to galaxy group properties reduced.

Here is the starting point of our galaxy group algorithm called MAGGIE: Models and Algorithm for Galaxy Group, Interloper and Environment. We combine our understanding from the galaxy formation, using various models, to compute a probability for galaxies to belong to a peculiar group, and use it in the algorithm for the group extraction. Then interloper effects should be reduced in the characterization of the environment.

In the following sections, we will describe the algorithm and its implementation, the application to the SDSS and show its limitations.

5.2 Algorithm

5.2.1 Description

MAGGIE doesn't assign a galaxy to a unique group, but it assigns a probability for this galaxy to be in a given group (rather than being an interloper). With this principle, a galaxy is possibly assigned to more than one group. The goal of MAGGIE is to obtain the properties of galaxy groups in statistical and probabilistic senses. This allows users of catalogues generated by MAGGIE to compute some properties of groups, weighting galaxies in accordance to their probability.

MAGGIE is organized in an iterative way in order to be self-consistent with the data being analysed, as for learning algorithms. For this reason, we will describe the implementation of the algorithm in different steps. In what follows, we assume that we have a galaxy sample with positions (right ascension RA, declination DEC), redshifts, stellar masses or luminosities, apparent magnitudes in a given band and absolute magnitudes. It's the minimum set of data necessary.

1. We begin with seed groups before launching the iterative process. For this, we assume that the most massive galaxies (in stellar mass) are potential group centers. In an other implementation, we use the luminosity of the central (the reason is explained in Section 5.5.4). But, some intra-group physical process can lead to a false detection of the brightest galaxy

as the central one (Ebeling et al., 2013). From the galaxy sample, we sort by decreasing stellar mass (or luminosity) all galaxies and we start with the most massive (most luminous) as centre of a potential group.

2. For all our potential groups, we need to get our potential members. We are just interested in the virial sphere (of radius r_{200}) of groups. Since the unique information on groups at this step is the central galaxy, we use its stellar mass (or luminosity). At first iteration, we use the relation between halo mass and central stellar mass from Behroozi et al. (2010) (and a simple ratio relation for luminosity). We also tried other models to see the influence of this choice (see Section 5.5.1). For subsequent iterations, we use the same relation, but learned from our previous iterations. We can estimate the virial radius of the group assuming that the halo mass corresponds to the virial mass. We refer to this method as MAGGIE-m. Alternatively, we can use central galaxy luminosities to estimate the virial radius, and we call this method MAGGIE-L. Then, we select all galaxies in a cone generated by an angular separation corresponding to the virial radius physical size at the group's redshift (the redshift of the central galaxy).

3. We compute probabilities that galaxies are members of a given group. The probability is computed assuming a density profile of galaxies and dark matter in groups, and a velocity distribution of the galaxies. Considering that galaxies form in dark matter halos, we assume that galaxies in groups must follow a NFW distribution (Navarro et al., 1996), which fit well the dark matter particles distribution in ΛCDM simulations, and assume that the galaxy number density profile is proportional to the mass density profile. The detailed computation of the probability is provided in Section 5.3.

4. We compute the weighted (by probability) multiplicity, stellar mass and luminosity of groups. For this we use a probability threshold $p_{\rm mem}$ to decide if a galaxy is associated to a group, i.e. if we take the galaxy for the estimation of the group stellar mass and luminosity. This parameter will be optimized by tests. The way of computing these properties for a group is the following: we sum, using the probability weights, luminosities and stellar masses of galaxies that have an absolute magnitude less than the limiting magnitude defined by the sample, in order to be complete.

5. Using the stellar mass of the central galaxy, we can estimate the halo mass of the group. We use the abundance matching technique which assumes that there is a one-to-one relation between the central stellar mass of the group and its halo mass. It allows to compare the cumulative distribution functions (CDFs) of the two quantities. Indeed, with this assumption, the number of groups above a given central stellar mass (or luminosity) is the same as the number of groups above the corresponding halo mass. If we consider a certain halo mass function, we can predict the halo mass of a group with a given central stellar mass (or luminosity) by comparing the CDF of the data with that predicted by the halo mass function.

6. With the halo mass found for group by this abundance matching, we go back to step 2 and recompute groups with the halo mass-central stellar mass relation previously obtained. This process goes until there is a convergence in the number of groups.

If we follow this schema, there will be as many groups as galaxies. To avoid the inherent fragmentation introduced by this method, we used another threshold probability to reduce the number of groups. We allow a galaxy to be a central galaxy only if its probability to belong to another group (already determined in the loop for potential galaxy groups) is smaller than the threshold $p_{\rm cen}$. For this comparison, we consider the maximum probability among all groups in

THESIS

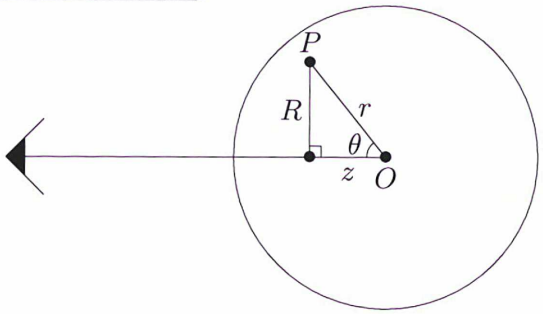


Figure 5.1: Schema illustrating a galaxy group observed from the point of view of an observer. The line-of-sight is represented by the line, corresponding to the principal axis of the cylindrical coordinates used. R is the projected radius, r the distance of a given point P to the origin of the group in O, and z the line-of-sight distance, or the height of the point in the cylindrical coordinates. The view is face on the plane defined by the position vector of the point P and the cylindrical axis.

which the galaxy maybe a member. In this way, we exclude while iterating over potential groups a large number of central galaxies, and avoid the fragmentation.

5.3 Membership probability

The membership probability is one of the most important aspect of MAGGIE. Since the observer studies galaxy groups only in projected phase space (hereafter *pps*), for defining the probability, we consider the location in the *pps* of the group with its projected radius R and the line-of-sight velocity v_z. The probability of membership to a given group, is the number of cases where we are inside the halo relative to the total number of cases. The *pps* density g is this definition of "number of cases". We can write our probability p to be in the halo as:

$$p\left(R, v_z\right) = \frac{g_h\left(R, v_z\right)}{g_h\left(R, v_z\right) + g_i\left(R, v_z\right)} \tag{5.1}$$

where g_h is the *pps* density inside the group and g_i is the foreground/background *pps* density, i.e. the interloper density.

In Section 5.3.1, we describe how to compute the probability with a general density profile and then in Section 5.3.2, we provide some analytical forms for several models.

5.3.1 General case

To compute the projected density of galaxies in the group we have to assume some models for their phase space distribution. So we use the distribution function f of the system, expressing the number of galaxies whose phase space coordinates are lying in the range $[\mathbf{r}, \mathbf{r} + \mathrm{d}\mathbf{r}]$ and $[\mathbf{v}, \mathbf{v} + \mathrm{d}\mathbf{v}]$.

$$f\left(\mathbf{r}, \mathbf{v}\right) \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{v} = \rho\left(r\right) \mathrm{d}x\mathrm{d}y\mathrm{d}z \; h_{3\mathrm{D}}\left(\mathbf{v}\right) \mathrm{d}v_x\mathrm{d}v_y\mathrm{d}v_z = \mathrm{d}^6 N \tag{5.2}$$

where $h_{3\mathrm{D}}\left(\mathbf{v}\right)$ is the 3D velocity distribution of galaxies in the group.

If we consider the line of sight as the axis of cylindrical coordinates, the density profile of the system will be simplified. For the velocity distribution, we transform Cartesian coordinates to spherical coordinates. Since both systems are related only by rotations, the Jacobian of the variable substitution for velocities is unity. Hence:

$$f\left(\mathbf{r}, \mathbf{v}\right) \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{v} = \rho\left(r\right) R\mathrm{d}R\mathrm{d}\phi\mathrm{d}z \; h_{3\mathrm{D}}\left(\mathbf{v}\right) \mathrm{d}v_r\mathrm{d}v_\theta\mathrm{d}v_\phi \tag{5.3}$$

THESIS

By definition, the projected phase space density is just the number N of galaxies with their *pps* coordinates in the ring defined by the range $R + \mathrm{d}R$ and $v_z + \mathrm{d}v_z$.

$$g_h\left(R, v_z\right) 2\pi R \mathrm{d}R\, \mathrm{d}v_z = \mathrm{d}^2 N \tag{5.4}$$

We can see that $r^2 = z^2 + R^2$, so $\mathrm{d}z = r/\sqrt{r^2 - R^2}\mathrm{d}r$. Now, to have the projected density on the sphere, we just need to integrate over the line of sight and angles:

$$g_h\left(R, v_z\right) = \int_0^{2\pi} \int_{z=-z_{\max(r)}}^{z_{\max(r)}} f\left(\mathbf{r}, \mathbf{v}\right) \mathbf{dr dv} = 2 \int_{r=R}^{r_{\mathrm{vir}}} 2\pi \frac{r\rho\left(r\right)}{\sqrt{r^2 - R^2}} R \mathrm{d}R \mathrm{d}r\, h_{3\mathrm{D}}\left(\mathbf{v}\right) \mathrm{d}v_r \mathrm{d}v_\theta \mathrm{d}v_\phi \tag{5.5}$$

We need to integrate on the velocities too in order to get the line-of-sight component, and retrieve the pps. For velocities, we make a transformation of coordinates: we pass to spherical coordinates to the coordinates defined where v_1 and v_ϕ are perpendicular to the line of sight defined by the z axes. The rotation matrix between both coordinates system is:

$$\begin{pmatrix} v_r \\ v_\theta \\ v_\phi \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} v_z \\ v_1 \\ v_\phi \end{pmatrix} \tag{5.6}$$

so the Jacobian of the transformation is unity:

$$\begin{aligned} g_h\left(R, v_z\right) &= \int_0^{2\pi} \int_{z=-z_{\max(r)}}^{z_{\max(r)}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f\left(\mathbf{r}, \mathbf{v}\right) \mathbf{dr dv} = 2 \int_{r=R}^{r_{\mathrm{vir}}} 2\pi \frac{r\rho\left(r\right)}{\sqrt{r^2 - R^2}} R \mathrm{d}R \mathrm{d}r h\left(v_z\right) \mathrm{d}v_z \\ &= \Sigma\left(R\right) \left\langle h\left(v_z | R, r\right)\right\rangle_{\mathrm{LOS}} \end{aligned} \tag{5.7}$$

where $h\left(v_z\right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h_{3\mathrm{D}}\left(\mathbf{v}\right) \mathrm{d}v_1 \mathrm{d}v_\phi$

For simplifying the equations later, we use a normalization as in Appendix B. With this normalization, we can write:

$$g_h\left(R, v_z\right) = \frac{M_{\mathrm{vir}}}{r_{\mathrm{vir}}^2} \frac{1}{2\pi} \int_{R/r_{\mathrm{vir}}}^{1} \frac{x\widehat{\rho}\left(x\right)}{\sqrt{x^2 - \left(R/r_{\mathrm{vir}}\right)^2}} \mathrm{d}x h\left(v_z\right) \tag{5.8}$$

The density of interlopers is extracted from Mamon et al. (2010), with:

$$g_i\left(R, v_z\right) = \frac{1}{2} g_i\left(R, |v_z|\right) = \frac{1}{2} \left(A \exp\left[-\frac{1}{2} \left(\frac{v_z}{\sigma_i} \right)^2 \right] + B \right) \frac{M_{\mathrm{vir}}}{r_{\mathrm{vir}}^2 v_{\mathrm{vir}}} = \widehat{g}_i\left(R, v_z\right) \frac{M_{\mathrm{vir}}}{r_{\mathrm{vir}}^2 v_{\mathrm{vir}}} \tag{5.9}$$

5.3.2 Analytical forms

In the following, we will refer to different equations of Appendix B.

If we assume that the groups are in dynamical equilibrium, the velocity distribution of galaxies should follow a Maxwellian (Gaussian) distribution. This assumption can be discussed (Beraldo et al., 2014).

In this case, the velocity distribution can be written:

$$h_{3\mathrm{D}}\left(\mathbf{v}\right) = \frac{1}{\left(2\pi\right)^{3/2} \sigma_\theta^2 \sigma_r} \exp\left(-\frac{1}{2} \left(\frac{v_r^2}{\sigma_r^2} + \frac{v_\theta^2 + v_\phi^2}{\sigma_\theta^2} \right) \right) \tag{5.10}$$

assuming that we split the three components of the velocity into three independent velocity distributions. We can transform:

$$\left(\frac{v_r^2}{\sigma_r^2} + \frac{v_\theta^2 + v_\phi^2}{\sigma_\theta^2} \right) = \mathfrak{a} v_z^2 + \mathfrak{b} v_1^2 + \mathfrak{c} v_\phi^2 + 2 v_z v_1 \mathfrak{d} \tag{5.11}$$

for the coordinate system defined in Section 5.3.1 with:

$$\begin{aligned}
\mathfrak{a} &= \left(\frac{\cos^2\theta}{\sigma_r^2} + \frac{\sin^2\theta}{\sigma_\theta^2}\right) \\
\mathfrak{b} &= \left(\frac{\cos^2\theta}{\sigma_\theta^2} + \frac{\sin^2\theta}{\sigma_r^2}\right) \\
\mathfrak{c} &= \frac{1}{\sigma_\theta^2} \\
\mathfrak{d} &= \left(\frac{1}{\sigma_r^2} - \frac{1}{\sigma_\theta^2}\right)
\end{aligned} \tag{5.12}$$

Putting Equation 5.11 in canonical form and integrating Equation 5.5 over v_ϕ and v_1 we get (Mamon et al., 2013):

$$h\left(v_z\right) = \frac{1}{\sqrt{2\pi}\sigma_z} \exp\left(-\frac{1}{2}\left(\frac{v_z}{\sigma_z}\right)^2\right) \tag{5.13}$$

since:

$$\sigma_z^2 = \sigma_r^2\left(1 - \beta\left(\frac{R}{r}\right)^2\right) \tag{5.14}$$

and β is the anisotropy profile $\beta = 1 - \sigma_\theta^2/\sigma_r^2$.

Finally, the projected density of galaxies in a halo is:

$$g_h\left(R, v_z\right) = \frac{M_{\rm vir}}{2\pi r_{\rm vir}^2 v_{\rm vir}} \int_{R/r_{\rm vir}}^{1} \frac{x\widehat{\rho}\left(x\right)}{\sqrt{x^2 - \left(R/r_{\rm vir}\right)^2}} \frac{1}{\sqrt{2\pi}\widehat{\sigma}_z} \exp\left(-\frac{1}{2}\left(\frac{\widehat{v}_z}{\widehat{\sigma}_z}\right)^2\right) \mathrm{d}x \tag{5.15}$$

where we work with velocities in units of the virial velocity v_v, i.e. $\widehat{v_z} = v_z/v_{\rm vir}$, and we use the dimensionless expression of the line-of-sight velocity dispersion (see Appendix B for details). The ratio between the interloper and halo *pps* density is:

$$\frac{g_i}{g_h}\left(R, v_z\right) = \frac{(2\pi)^{3/2}\widehat{g}_i(x_R, |\widehat{v}_z|)}{\int_0^{\mathrm{acosh}\left(\frac{c}{x_R}\right)} \frac{\left(x_R\cosh u\right)\widehat{\rho}\left(x_R\cosh u\right)}{\tilde{\sigma}_z\left(x_R, x_R\cosh u\right)} \times \exp\left(-\frac{1}{2}\frac{\widehat{v}_z^2}{\tilde{\sigma}_z^2(x_R, x_R\cosh u)}\right) \mathrm{d}u} \tag{5.16}$$

where we used the transformations $x = x_R\cosh u$ and $x_R = R/r_{\rm vir}$, to obtain a better convergence for the numerical integration. Simplifications come from the expression of the virial velocity $v_{\rm vir}^2 = \dfrac{GM_{\rm vir}}{r_{\rm vir}}$ and the dimensionless expression of the radial velocity dispersion deduced from the Jeans equation (see Appendix B).

5.3.3 Comparisons with simulations

To test our computation of the probability, we compared our theoretical expression with the dark matter particles from the Borgani et al. (2004) simulation used in Mamon et al. (2010) to deduce the *pps* density of interlopers. For this, a selection of high mass dark matter halos was performed on the cosmological simulation. Then, particles coordinates were translated to make the center of the halo the origin of the simulation box. A fictitious observer is placed on the side of the box, and all observed coordinates in phase space are computed from the observer point of view. The coordinates of all particles in the cone defined by the observer and the radius of the halo are computed in units of the virial radius and velocity. This allows to easily define a particle as an

interloper or not, with their three dimensional radial coordinate r. If in units of the virial radius, $r \leqslant 1$ means that the particle is belonging to the halo, else it's an interloper. We can stack all the particles from all the cones of each halo to create an unique halo, with numerous particles, used as a test case for our models and to estimate the interloper *pps* density.

In Figure 5.2, we show the contours of the halo membership probability (in gray for the simulation, in black for our model from the Equation 5.16) in the *pps*. For this model, we used a Gaussian velocity distribution of particles in the halo, with the anisotropy from Mamon & Łokas (2005) that fit the anisotropy profile of dark matter particles, and we assumed a NFW density profile. Moreover, we assume that the characteristic radius of the anisotropy of Mamon & Łokas (2005) is equal to the a radius at which the slope of the density profile is -2. The theory fits relatively well the data from the cosmological simulation, except that the simulation is more sharply truncated at high velocity than is our model.

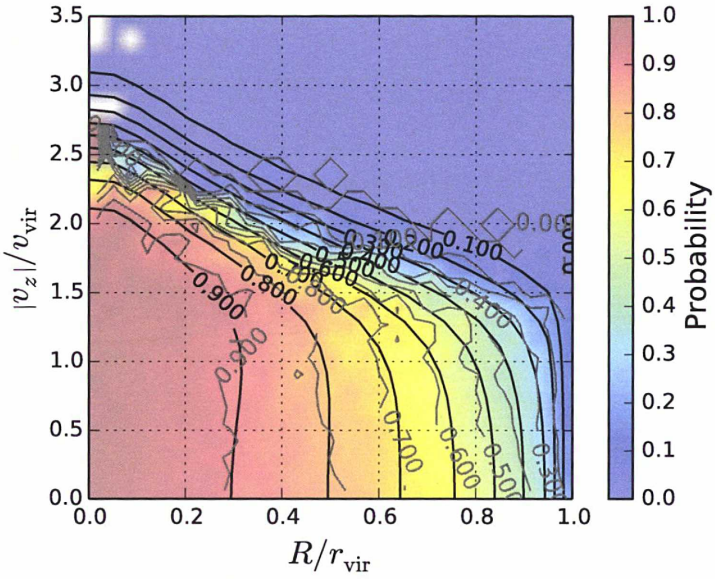


Figure 5.2: Contours of halo membership probability from the simulation of Borgani et al. (2001) as stacked by Mamon et al. (2010) and our model. In *gray* the contours obtained with particles from the cosmological simulation, and in *black* the theoretical expectation from Equation 5.16. The color scale reflects the probability from the simulation. The theoretical probability agrees with the cosmological simulation except for high velocities along the line-of-sight.

Here, we assume that a NFW density profile, since the dark matter particles of the Borgani et al. (2004) simulation follow this distribution (see Mamon et al. (2010)). But we are interested in groups of galaxies and they must follow the same distribution to apply our model. We used the galaxies from the $z = 0$ output of the Guo et al. (2011) semi-analytical code and checked that they follow a NFW density profile too. But as expected, there is a bias between the galaxies and dark matter particles. Indeed, if we fit the concentration of the NFW profile in Guo et al. (2011) and compare it to the model of Macciò et al. (2008) obtained from dark matter particles, we can see that the two functions are different. A consequence is that the link between halo mass and concentration must be adjusted for galaxies in our model. The difference between the two concentration-mass relations is shown in Figure 5.3. But we note that the modulation of the concentration with the halo mass is dependent of the cut-off in luminosity applied to the galaxy sample, making the use of a specific density profile for galaxies inadequate. We checked the influence of this choice on MAGGIE by comparing the performance with the concentration from Guo et al. (2011) and from Macciò et al. (2008). No noticeable impact is observed in the

Remark 2

We may think that the observed discrepancies in Figure 5.2 are the consequence of a bad choice for the ratio of anisotropy radius to scale radius b/a (see Appendix B) or for the concentration, but changing this value doesn't reduce them. The contours for the simulation seem to show a cut-off in the line-of-sight velocity dispersion for high velocities, as if the distribution is truncated above a given velocity. A functional form with such a property is the generalization of the Gaussian called the q-Gaussian or Tsallis distribution. Assuming such a velocity distribution, the computation of the probability involves several integrals, which is CPU time consuming. Instead, we can fit a q-Gaussian on the line-of-sight velocity distribution from the simulation and incorporate it in the probability computation. But unfortunately, this doesn't solve the problem. It seems that the number of particles with high velocities is too low to correctly define the probability to be in the virial sphere of the halo, and to compare it to theoretical expectations. The velocity distribution model isn't involved. ■

completeness, reliability and fragmentation, except on stellar masses and luminosities of groups but without being significant.

5.4 Results on mock catalogues

5.4.1 Description

For tests, we proceed as described in Duarte & Mamon (2014a); Yang et al. (2007) and Chapter 4. To link a selected group by the algorithm in redshift space to the true halo in real space, we use the most massive galaxy of the group. The true halo of a group is the true halo to which the most massive galaxy in the selected group (referred as the central galaxy) belongs to. With this link, we compute the completeness and reliability of groups relatively to this halo in real space. We define statistics used to quantify the performance of MAGGIE. The completeness C is the fraction of galaxies in the real space group (limited to the virial sphere) recovered in the selected group. The reliability R is the fraction of galaxies in the selected group present in the real space group (limited again to the virial sphere). A primary group is defined as a selected group whose central galaxy matches the central galaxy of the real space associated group, remaining groups are fragments. A complete and detailed description of the statistics can be found in Duarte & Mamon (2014a) and Chapter 4.

The reliability in the case of MAGGIE is more complex since we use probabilities for galaxies in groups. To take advantage of our probabilities, let us give a new definition for the reliability. In Duarte & Mamon (2014a), we wrote:

$$R = \frac{\mathrm{TG} \cap \mathrm{EG}}{\mathrm{EG}} = \frac{\sum_{i \in \mathrm{TG} \cap \mathrm{EG}}}{\sum_{i \in \mathrm{EG}}} \tag{5.17}$$

But many galaxies belong to our group with this definition and so we weight galaxies in the previous sum by their probabilities in order to have a coherent definition of the reliability with our probabilistic determination of groups. Our new definition in the case of a probabilistic galaxy group algorithm as MAGGIE is:

$$R = \frac{\mathrm{TG} \cap \mathrm{EG}}{\mathrm{EG}} = \frac{\sum_{i \in \mathrm{TG} \cap \mathrm{EG}} p_i}{\sum_{i \in \mathrm{EG}} p_i} \tag{5.18}$$

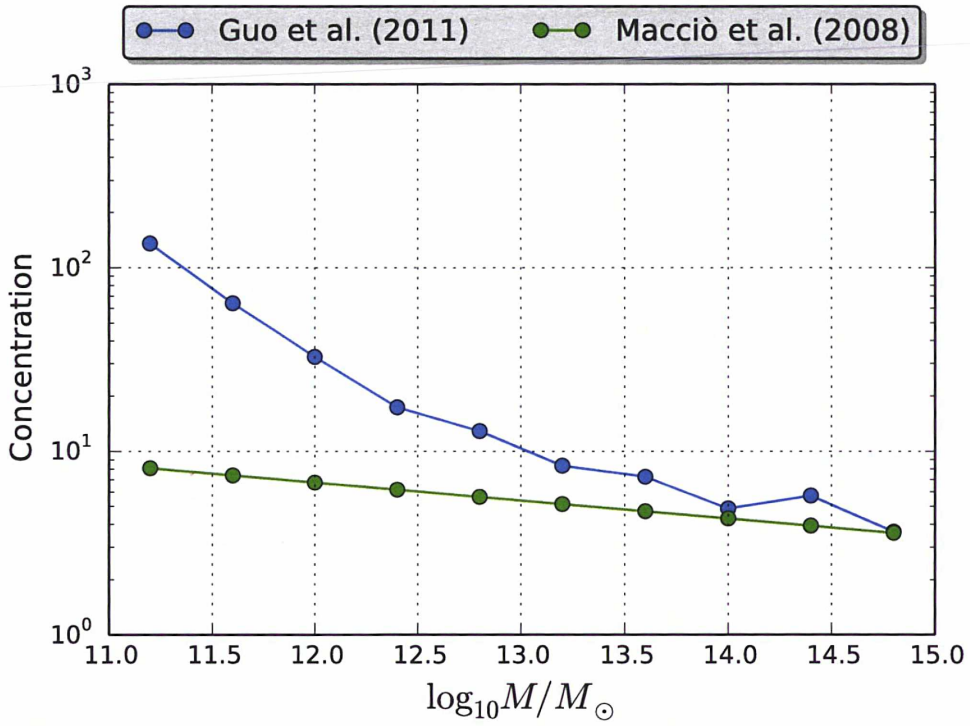


Figure 5.3: halo concentration as a function of halo mass for Macciò et al. (2008) in *green* and concentration of the galaxy population (NFW fit) from Guo et al. (2011) with $M_r \leqslant -15$ in *blue*. In low mass halos, the concentration of galaxies is much higher than that of dark matter particles.

For the completeness, since the probability doesn't introduce a bias in the selection relatively to the real group, we keep the computation as described in Duarte & Mamon (2014a), without weighting by probabilities.

Without probabilities, galaxies in groups form a complete partition of the survey since groups can be seen as disjoints sets of redshift space. But with MAGGIE and probabilities, a galaxy can be in multiple groups so the sets of groups are overlapping, and the *dual* analysis in Duarte & Mamon (2014a) for the merging of real space groups cannot be correctly done. This is because the central galaxy of a real space group can potentially belong to several extracted groups in redshift space.

5.4.2 Optimization

MAGGIE depends on two probability thresholds: the first we call central probability ($p_{\rm cen}$) constraining the fragmentation of galaxy groups by allowing galaxies to be the central galaxy of a potential galaxy group, the second one the membership probability ($p_{\rm mem}$), defining a threshold to consider or not a galaxy in a group for the computation of its properties.

In fact, a galaxy is considered as possible central galaxy while looping through ordered galaxies only if the galaxy has all its probabilities in its other groups lower than the central probability threshold. For membership, galaxies are "assigned" to a group (i.e. they are assumed to have a probability to be in this group) only if their probabilities are above the threshold membership probability.

Checking the dependence of MAGGIE to these parameters is done in the same way as we performed in Chapter 4. We computed the mean completeness, reliability, fragmentation and merging, as well as the quality factors we previously defined, for a range of threshold probabilities

$(p_{\text{central}}, p_{\text{membership}}) \in \left[10^{-15}, 0.4\right]^2$. Fortunately, results are not dependent on these thresholds. A small variation is observed only for very high values of these probabilities (above 0.1). Increasing these probabilities leads to relatively worst statistics, while keeping them small is better, but without significant variations.

We selected $p_{\text{cen}} = p_{\text{mem}} = 0.001$. Since this value is relatively small, we should notice that it is equivalent to defining the membership in the virial cone constructed with the virial radius of the group. The selection of background galaxies, far away in velocities, is avoided by p_{mem}, since with this value, when galaxies are beyond 4–5 v_{vir} from the group, they are not considered. The same happens for p_{cen} where galaxies associated to a group through their probabilities cannot be potentially the center of a new group.

When working with non-probabilistic algorithms, the set of galaxy groups is a complete partition of the space formed by galaxies. In other words, groups are non-overlapping and fragmentation is avoided naturally if the assignment is done properly. But with probabilities and our method, removing these threshold parameters, there are as many groups as galaxies. Setting the two thresholds to zero makes MAGGIE behave like non-probabilistic algorithms. Note that if we set $p_{\text{mem}} = 0$, each galaxy group will be formed of its entire virial cone, which is not desirable. The introduction of threshold probabilities is a way to make a "compatibility" between MAGGIE with its soft assignment and non-probabilistic grouping algorithms and their hard assignments.

5.4.3 Results

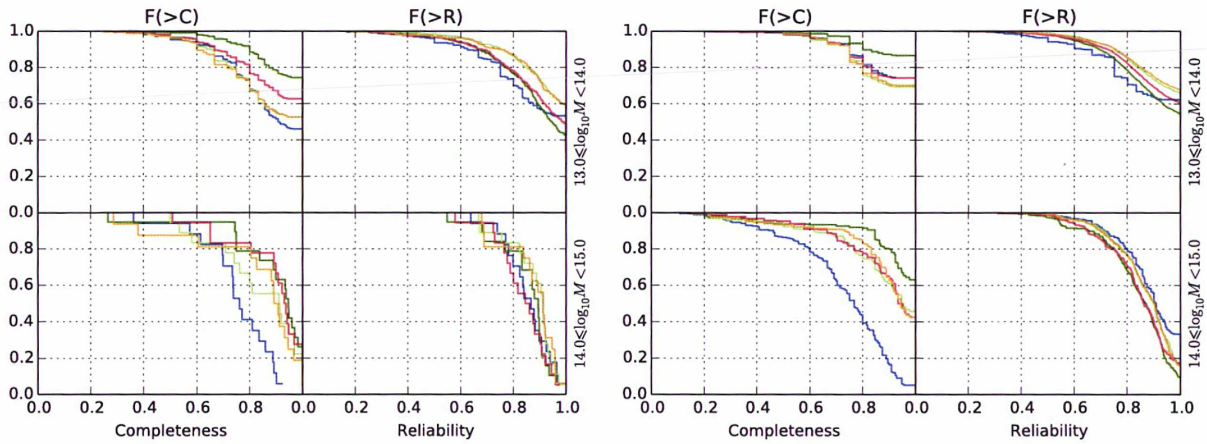
The following tests result from the application of MAGGIE on the perfect (no observational errors, no K-corrections) mock catalogue whose construction is described in Chapter 3. In this case, we assume the halo mass function extracted from the Millennium-II outputs, with an NFW galaxy number density profile identical to that for dark matter particles in their halos. The influence of these assumptions will be developed in Section 5.5, in particular taking observational errors in our mocks into account.

We compare MAGGIE with the popular FoF grouping algorithm (see Chapter 4). The set of linking lengths used for the FoF is the one defined in Duarte & Mamon (2014a) for an optimal FoF, close to the parameters used by Robotham et al. (2011), with values of $(b_\perp, b_\parallel) = (0.07, 1.1)$. This will let us see if the our probabilistic Bayesian approach improves the galaxy grouping compared to a simple geometrical approach such as FoF.

5.4.3.1 Fragmentation

Estimating the fraction of groups in the selection that are the result of the fragmentation of a real group is important since an observer using a group catalog can't distinguish the primary group from the other. In Figure 5.5, we show the fraction of fragmented groups (defined as in Chapter 4) as a function of the estimated group mass. This allows to see the expected fraction of fragmented groups by an observer using a group catalog with only information on the estimated halo mass.

MAGGIE shows much less fragmentation than FoF, for all estimated group masses, such $\log_{10} M/M_\odot \geqslant 12$ (nearby sample) or 12.3 (distant sample). This is due to the combination of the abundance matching that gives good estimates of group virial masses and the ordered search of groups from galaxy stellar masses. But fragmentation increases with the decreasing estimated mass, since with groups of few members, it is easier to make a mistake in the selection of the central galaxy of the group.



(a) Catalogue 2 (b) Catalogue 5

Figure 5.4: The cumulative distribution functions of the completeness $F(> C)$ and reliability $F(> R)$ for bins of true halo mass for two sub-samples of the mock catalogue (nearby in left and distant in right). The colored histograms are for FoF (*blue*), MAGGIE-m with no observational errors (*dark green*), MAGGIE-L with no observational errors (*light green*), MAGGIE-m with 0.2 dex errors on stellar mass (*red*) and MAGGIE-L with 0.04 dex errors on luminosites (*gold*). Details for latter cases are in Section 5.5.4. Results are shown only for primary groups (i.e. groups that are not fragments of real space halos) with the same filter as in Chapter 4.

5.4.3.2 Completeness and reliability

Figure 5.4 shows the cumulative distribution functions of the completeness C and reliability R as defined in Section 5.4.1 for MAGGIE and the FoF algorithm. Results are shown only for two doubly complete sub-samples (see Chapter 3), different with Chapter 4, where we use now catalogues 2 and 5. Only two bins in true group virial halo mass are used. MAGGIE (in dark green) shows a better behaviour in completeness for all masses in both catalogues than FoF (in blue), while the reliability is equivalent to the optimal FoF, except for high masses for the more distant catalogue, where FoF is more reliable (median reliability is 0.90 for FoF and 0.85 for MAGGIE).

5.4.3.3 Virial masses

In the Figure 5.6, we compare the estimation of the virial mass by application of the virial theorem for FoF algorithm and by abundance matching for MAGGIE. As we already discussed, the virial theorem isn't very suitable in recovering the virial masses of groups when they have a small mass (between 10^{12} and $10^{13} M_{\odot}$). We can't really see this in the bias of the estimation between the two algorithms. On the other hand, the scatter in recovered masses of groups in the distant sub-sample is lower with MAGGIE (0.25 dex) than with FoF (0.35 dex), except for $M \geqslant 10^{15} M_{\odot}$, where MAGGIE suffers from uncertainties in the abundance matching caused by small number statistics at the high end. In the nearby sub-sample, both MAGGIE and FoF produce scatter in virial mass of ≈ 0.3 dex.

5.4.3.4 Group luminosities and stellar masses

We test the deduced stellar mass and luminosity of selected groups for each algorithm. For non-probabilistic FoF, they are just the sum of the galaxy contributions. But for MAGGIE, we use the computed probabilities inside the group to weight the stellar mass and luminosity of each galaxy. If X is the property of the group, x_i the property of the galaxy i in the group with the

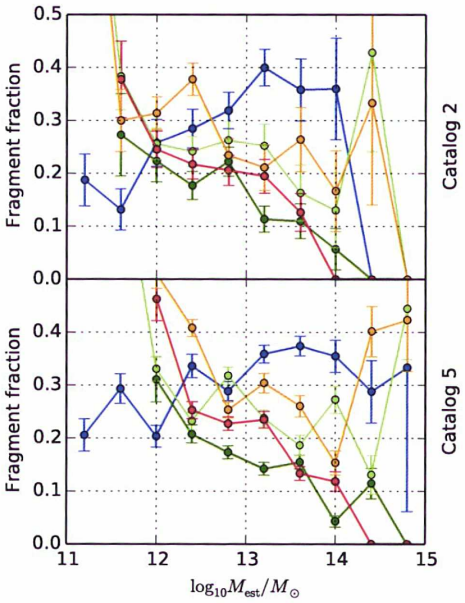
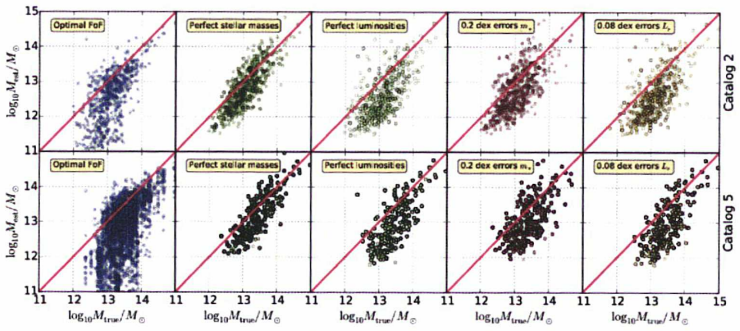
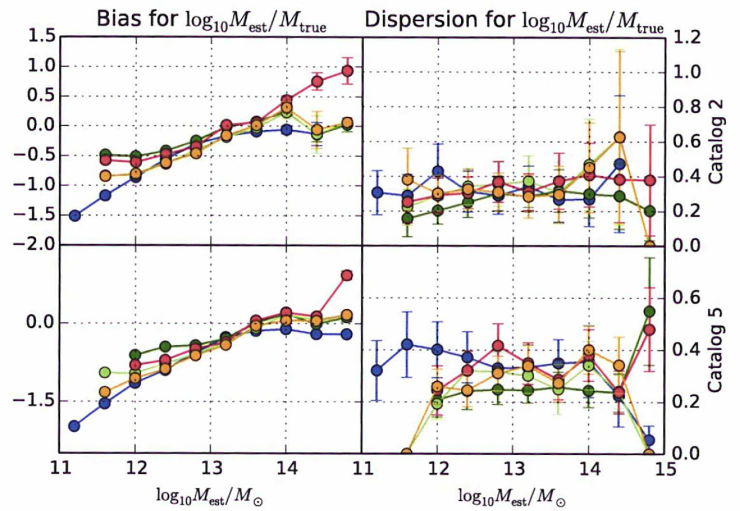


Figure 5.5: The fraction of estimated groups that are fragments in bins of estimated mass of extracted groups, for catalogues 2 and 5. Colors are the same as in Figure 5.4.



(a) Comparison of virial masses



(b) Bias and dispersion for virial masses

Figure 5.6: Comparison of the virial mass estimated by the galaxy group algorithms and the true masses obtained from the Millennium-II simulation, for catalogues 2 and 5. The top panel shows the comparison and the bottom panel bias and the dispersion of the logarithmic difference of masses. Colors are the same as in Figure 5.4.

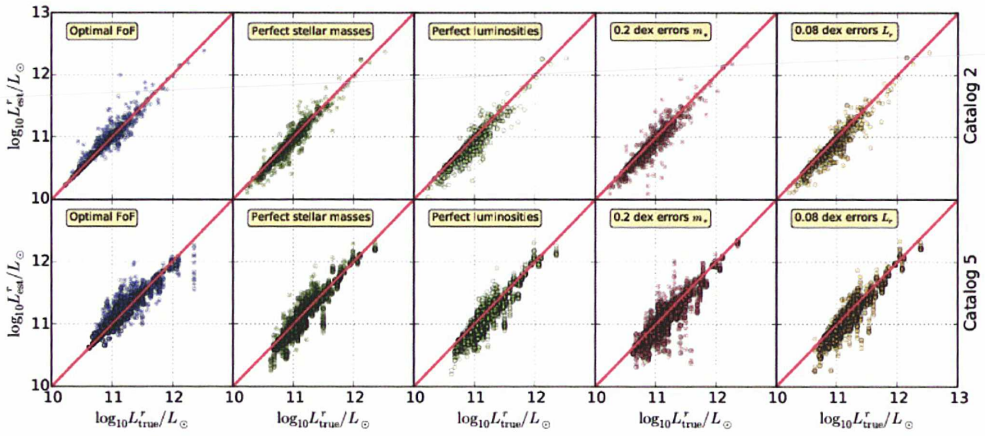
probability p_i:

$$X = \sum_i p_i x_i \tag{5.19}$$

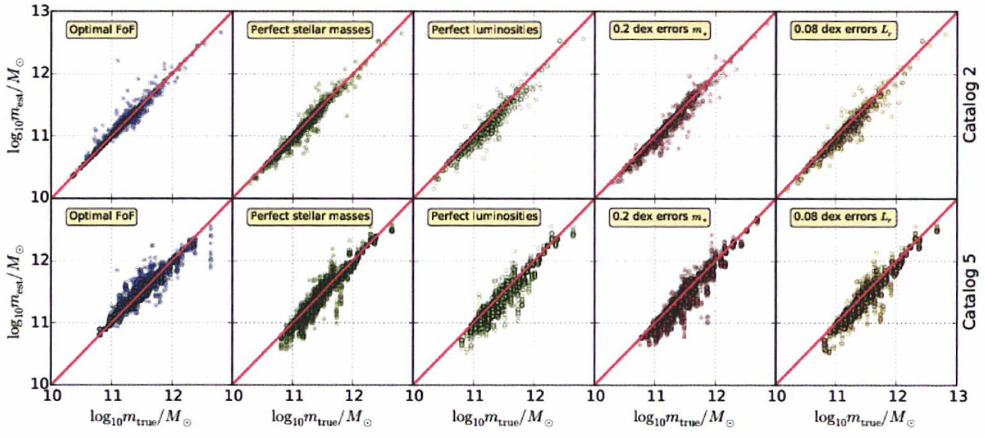
In Figure 5.7 and Figure 5.8, we compare the true luminosity of groups (computed assuming a perfect selection of groups in the sub-sample) with the luminosity computed with the galaxy membership of each algorithm. In the bottom panel, we show the bias and the dispersion of the difference between the true and estimated luminosities. These figures show the same for stellar masses. The optimal FoF algorithm has a lower bias than MAGGIE, while the scatter is lower for MAGGIE at high masses ($13.3 \leqslant \log_{10} M/M_\odot \leqslant 14.7$) thanks to the probability weighting that reduces the effect of interlopers.

5.5 Discussions

A simple comparison of MAGGIE with the most popular and geometrical grouping algorithm shows that MAGGIE is well adapted in recovering galaxy groups from redshift space catalogues.



(a) Comparison for group luminosities

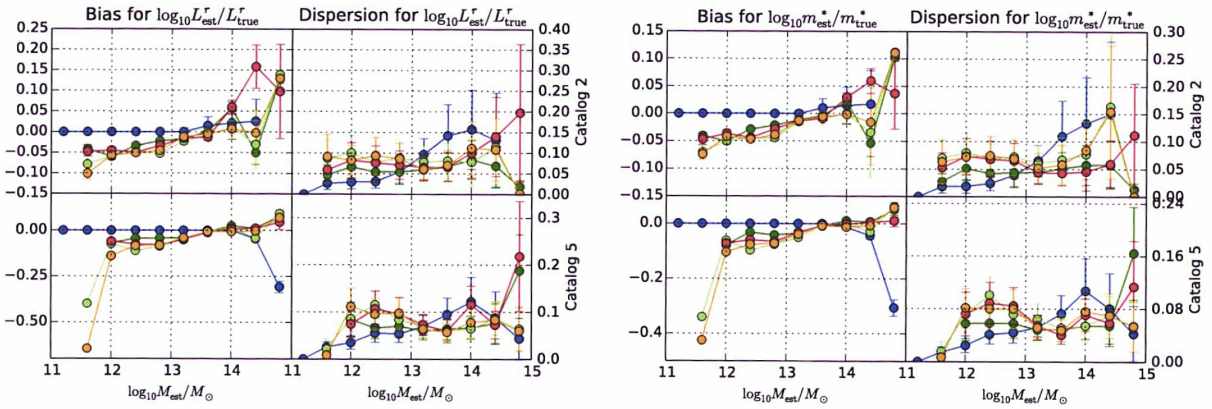


(b) Comparison for group stellar masses

Figure 5.7: Comparison of group luminosities in r band and stellar masses with the real space for primary groups in catalogues 2 and 5.. Colors are the same as in Figure 5.4.

Extracted global properties of groups are less biased and catastrophic cases avoided by using probabilities as weights to smooth the estimation. The membership inside these groups is better too since the completeness shows that MAGGIE selects a large part of galaxies from the real group, without polluting it by interlopers (as shown by the reliability). Moreover, the importance of interlopers is reduced still by using probabilities. The abundance matching technique is also a very good way to contribute to this galaxy group extraction, since the virial mass estimation relies only on group or galaxy properties, which are observables certainly biased and uncertain, but with less importance than biased geometrical informations (velocity dispersion, richness...). On the contrary, a geometrical based group finder such as FoF performs well when the number of galaxies is important because interlopers act as a small noise in the group membership, even with their relatively important presence at high halo masses for the FoF algorithm. Hence, velocity dispersion and harmonic radius are more efficient with high richness and the virial theorem thus becomes more precise.

This comparison is done in the case where the data on galaxies is perfect, in the sense that there are no observational errors and we perfectly know the various scaling relations used in our models. But the behaviour of MAGGIE is unknown in the real situation of an observer, with a limited knowledge in these models. In the following sections, we study the robustness of the



(a) Bias and dispersion for luminosities

(b) Bias and dispersion for stellar masses

Figure 5.8: Bias and dispersion for group luminosities and stellar masses for catalogues 2 and 5. Colors are the same as in Figure 5.4.

performance of MAGGIE under pertubations, i.e. in cases where we modify our initial estimate of virial radius, aw well as the halo mass function, and when we take into account observational errors in the galaxy luminosities and stellar masses.

5.5.1 Prior halo mass — central stellar mass relation

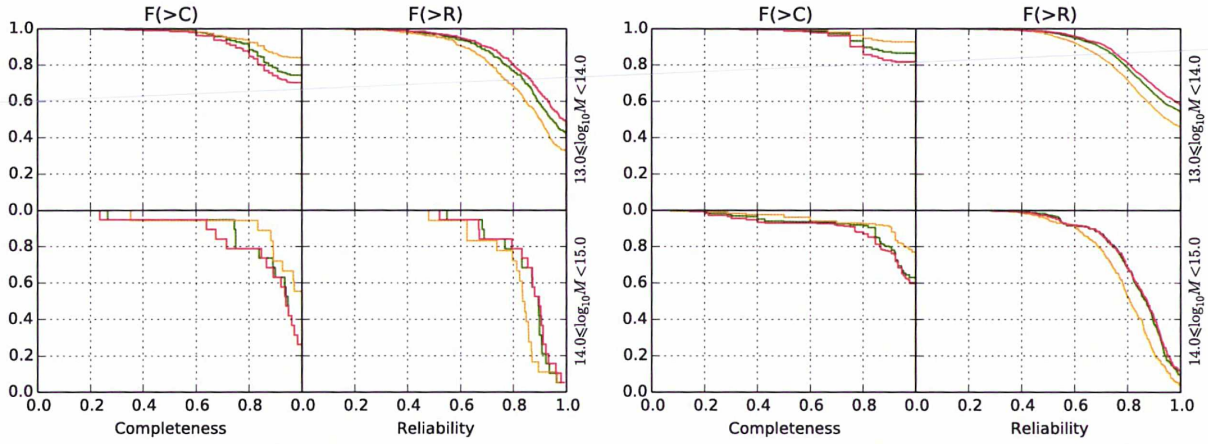
We tested the choice for the initial relation between the halo mass and the central stellar mass of groups to see its effects on MAGGIE. We used the relation from Behroozi et al. (2010) against a simple ratio relation with different values for the ratio. Extracted groups are insensitive to this choice, if we keep this choice with physical values. The iterative process corrects a bad assumption in our initial guess.

5.5.2 Influence of the halo mass function model

The estimation of the virial mass (radius) is a crucial step of MAGGIE (and other Bayesian methods). A biased estimate of group masses will affect observed trends of galaxy properties with the global environment.

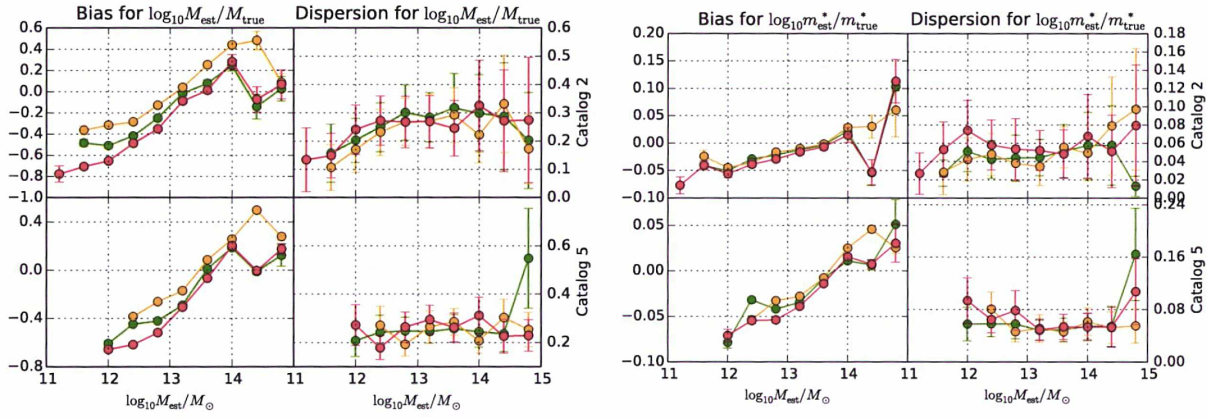
Our mass computation needs to be precise in the largest mass range possible, and independent of the pollution of groups by interlopers. The abundance matching technique seems to be a good way to estimate the virial mass of galaxy group halos. In principle, it seems more biased than using the luminosities or stellar masses of groups, but since the central galaxy in a selected group is well recovered, this is a quantity less affected by interlopers and so the halo mass estimation will be good enough. But since there is a saturation of the relation between the halo mass and the central stellar mass at high halo mass, we expect that the estimation will be poorer for high masses than other methods.

Most halo mass functions described in the literature fit the FoF mass of the halos instead of the spherical over-density mass, which is related to the virial mass of the halo. Since we used the galaxy catalogue from Guo et al. (2011), whose semi-analytical code was applied onto the Millennium-II run, we fit the virial halo mass function directly on its output. We show it in Figure C.2 where we plot the FoF mass function (in red) and the virial mass function (in black) for halos in the Millennium-II simulations at redshift zero. Virial masses are lower than FoF masses so we don't use existing models of halo mass functions displayed too on the figure. The way of computing such halo mass functions is described in Appendix C.

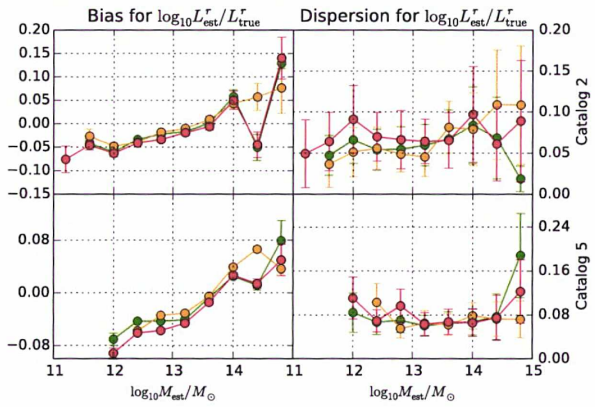


(a) Catalogue 2 (b) Catalogue 5

Figure 5.9: The cumulative distribution functions of the completeness and reliability for comparison of the perfect case of MAGGIE-m in *green* with the halo mass function model of Warren et al. (2006) in *orange* and Courtin et al. (2011) in *red*. The filter applied on groups is the same as in Chapter 4. The three different halo mass functions lead to similar group memberships.



(a) Group halo masses (b) Group stellar masses



(c) Group luminosities

Figure 5.10: Group properties compared to the perfect case of MAGGIE-m (no observational errors) using the halo mass function measured in the Millennium-II output, for both Warren et al. (2006) and Courtin et al. (2011) halo mass functions. Colors and filter are the same as in Figure 5.9. As seen in Figure 5.9, differences are not really significant.

The robustness of MAGGIE against the choice of the halo mass function is important because this choice will affect the completeness and reliability of our selected groups and their properties too in a non obvious way. Indeed, the halo mass function is a prior in MAGGIE and doesn't reflect necessary the reality. We apply an equivalent of the perturbation method to test the stability of MAGGIE under a bad choice of model. We used two halo mass functions very different of the halo mass function measured on the Millennium-II simulation: Warren et al. (2006) and Courtin et al. (2011). Those models are fits of the FoF halo mass function from different cosmological simulations. This is not the same as the virial mass but ideal for a perturbation test. The result of the application of MAGGIE with these models is shown in Figure 5.9 and Figure 5.10.

Comparisons are performed against the perfect case of MAGGIE-m (no observational errors) in green with the halo mass function directly fitted on the Millennium-II, perfect stellar masses and luminosities for galaxies. In orange, halo mass function of Warren et al. (2006) and in red, that one of Courtin et al. (2011). The influence of the halo mass function is very small on the completeness and reliability for all catalogues and for group properties. The fragmentation is not shown but behaves like in other plots, not affected by the choice of halo mass function.

5.5.3 Influence of cosmological parameters

The distances used by MAGGIE depends on the choice of cosmological parameters. For example, when computing the projected radius of a galaxy at the redshift of the group (i.e. its plane-of-sky distance to group center), we implicitly need to compute the luminosity distance which is cosmology dependent. We assume in our case a flat Universe and in this case, it is computed using just elliptic integrals (Eisenstein, 1997; Liu et al., 2011) Moreover, the different analytical halo mass functions tested in Sect 5.5.2 all assumed the same cosmological parameters as in our mock (based on those from the Millennium-II simulation). The observer may choose a slightly different set of cosmological parameters. We therefore now run MAGGIE on our mock, assuming slightly incorrect cosmological parameters, to test how sensitive is its performance on the correct choice of cosmological parameters.

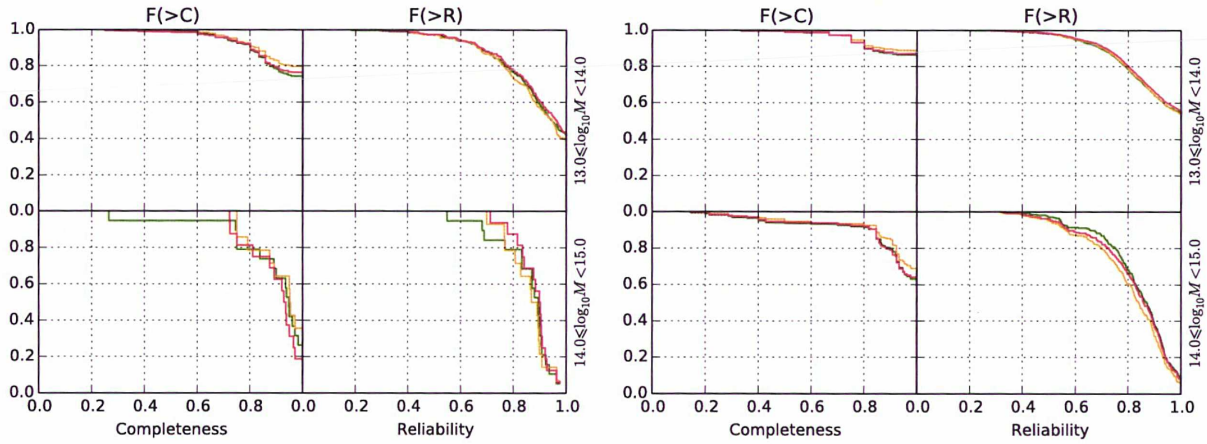
We ran MAGGIE-m with the "true" cosmology (from Millennium-II simulation) and two "false" cosmologies (Planck and WMAP9) to compare results. As expected, the importance of the cosmology is low, of the order of statistical errors, as seen in Figure 5.11 and Figure 5.12.

5.5.4 Influence of observational errors

Our way of sorting galaxies by mass uses our prior on the galaxy formation scenario. Indeed, the stellar mass of the central galaxy of a dark matter halo is correlated to its virial mass. But the relation is saturated at high halo masses. So the intrinsic precision is affected by this choice. Moreover, estimates of stellar masses from observations are not very precise and can significantly differ according to the chosen model for computing them. We show the differences between several spectral models present in the SDSS data base, with the bias and dispersion for each distribution, in Chapter 6.

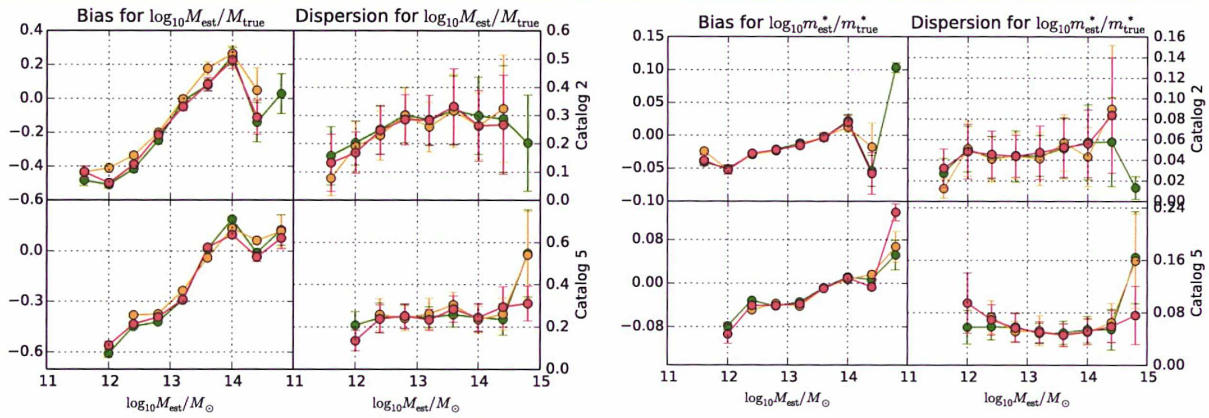
Typically, the errors in the estimation of stellar mass is roughly 0.2 dex. We introduce such errors in the stellar masses of the mock catalogue to estimate the effect of the bad estimation. We generated Gaussian errors without bias and dispersion of 0.2 dex. The application of MAGGIE-m on these galaxy mock catalogues is shown on Figure 5.4, Figure 5.7, Figure 5.8, Figure 5.5 and Figure 5.6 in red.

The effect of a 0.2 dex error on stellar masses (red) is to slightly increase group fragmentation, which remains well below the level found for the FoF algorithm. The completeness of MAGGIE-m is reduced to a level between that of the perfect MAGGIE-m and the optimal FoF. The effect of stellar mass errors on the reliability is less important since probabilities reduce the importance

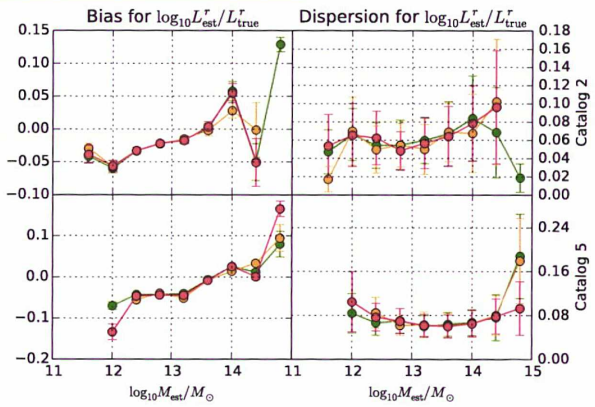


(a) Catalogue 2 (b) Catalogue 5

Figure 5.11: The cumulative distribution functions of the completeness and reliability for comparison of the perfect case of MAGGIE-m (no observational errors) (*green*) to different cosmologies *orange* for Planck Collaboration et al. (2013) and *red* for Bennett et al. (2013). Errors in the cosmological parameters don't have a significant impact on the group extraction.



(a) Group halo masses (b) Group stellar masses



(c) Group luminosities

Figure 5.12: Group properties compared to the perfect case (no observational errors) of MAGGIE-m for both Planck Collaboration et al. (2013) and Bennett et al. (2013) cosmologies. Colors are the same as in Figure 5.11. Again, the differences are not significant.

of interlopers introduced by the decrease in completeness. In fact, the reliability is, surprisingly, slightly higher once the stellar mass errors are incorporated. Finally, galaxy group properties (luminosity, stellar and halo mass) are not very affected by these errors in stellar masses, except for high halo masses, where they are biased high and more scattered. In this case, if the most massive galaxy has not its mass well estimated, the estimation of the halo mass is bad and probabilities can't "correct" this effect properly. This is visible in Figure 5.6 where, for the nearby sub-sample, the bias in the estimation of the halo mass is very important for high group masses and the dispersion is increased in all group masses.

One may wonder whether it would be better to use another tracer for the halo mass such as the central luminosity, which is less affected by observational errors. Despite its very high quality, the SDSS survey is not immune to errors on galaxy luminosity or stellar mass. Writing the r-band absolute magnitude of a galaxy as:

$$M_r = r - \mu(z) - k_r(z) - A_r \tag{5.20}$$

where μ is the distance modulus, while r, k_r, and A_r are the apparent magnitude, k-correction and extinction, all in the r band. The photometric errors are expected to be less than 0.05 mag, i.e. less than 0.02 dex on luminosity. The error caused by the uncertain distance can be written:

$$\epsilon(\log L_r) = \frac{1}{\ln 10}\left[\left(\frac{\epsilon(v)}{cz}\right)^2 + \left(\frac{\sigma(v_\mathrm{p})}{cz}\right)^2\right]^{1/2} \lesssim 0.056\,\mathrm{dex} \tag{5.21}$$

for $\epsilon(v) \simeq 30\,\mathrm{km\,s^{-1}}$, $\sigma(v_\mathrm{p}) \simeq 200\,\mathrm{km\,s^{-1}}$, and $z > 0.01$ (where the assumption of zero difference in peculiar velocity between the galaxy and the observer dominates the error). Finally, according to Figure 2 of Chilingarian et al. (2010), the intrinsic scatter in the k-correction is of order 0.015 mag, i.e. 0.006 dex. Admittedly, the k-correction of Chilingarian et al. suffers from some catastrophic errors, but since 99.9% of the galaxies with $z < 0.12$ have k-corrections between -0.15 and 0.25, it suffices to impose these limits to k_r. Finally, since SDSS spans high galactic latitudes the uncertainty o the Galactic extinction is of order 0.075 mag (the median error of SDSS galaxies), i.e. 0.03 dex. The uncertainty on internal extinction is more difficult to measure, but can be estimated to be 0.1 mag, i.e. 0.04 dex. Combining these 6 errors (photometry, redshift, assumption of no peculiar velocity, k-correction, Galactic extinction and internal extinction) in quadrature, we deduce that the error on luminosity is of order of 0.08 dex.

The inclusion of the luminosity instead of stellar mass in the group extraction process of MAGGIE is quite simple. The abundance matching between the virial mass and the central stellar mass is replaced by an abundance matching between the virial mass and the central luminosity. Intrinsically, using luminosities instead of stellar masses in the inference of group virial masses is expected to be less precise because the relation between the luminosity of the central galaxy and the halo mass is more saturated for high mass groups. But the loss at high mass should be offset by the 2.5 times greater precision.

Comparing the perfect case of MAGGIE using galaxy luminosity (light green) to the perfect case of stellar masses (dark green) shows, as expected, that the completeness is worse for luminosities (the reliability is a little better since the completeness has decreased). But group properties are not really affected still by the use of probabilities to avoid bad membership. The fragmentation is worse too since groups aren't entirely recovered (missing galaxies are considered as belonging to fragment groups). For group mass estimations, the bias induced by using luminosities is comparable to the perfect case of stellar masses. It is only in the dispersion that we observe the counterparts, specifically for high group masses, due to the uncertainties introduced by the saturation in the relation between the central luminosity and the virial mass in this range of halo masses. But it is somewhat better than using stellar masses with errors.

Adding errors following a Gaussian distribution without bias and a dispersion of 0.08 dex on luminosities, we compare it (light orange) to the perfect case of the luminosity. As expected, introduced errors do not have a real impact, because the behaviour of light orange and light green curves are roughly identical.

The negative point of using luminosities instead of stellar masses is that it seems to increase the fragmentation of true groups. But this fragmentation remains lower than that found for the FoF group. We should discuss a little how we make a match between the real space and the observed space. To say which galaxy is the central of a group in the real space, we can't directly use the one given by Guo et al. (2011) since in our mock catalogue, there is a magnitude limit removing a large number of galaxies not sufficiently luminous. The central is not necessarily the most luminous of the group and the flux limit can possibly hide us the central while the group is visible with help of some of its galaxies. To define the central galaxy in real space, we use the most massive in stellar mass of the group taking into account only galaxies within the complete sample used. Without such a treatment, we could possibly increase the fragmentation artificially by a lack of central galaxy in the sample. With MAGGIE-L, the central galaxy in extracted groups has a strong chance to be the most luminous (not necessarily the most massive in stellar mass) and the match with real space groups will frequently say that the central galaxy of the extracted group is not the same as the true group, resulting in a frequent fragmentation in our tests. This is what we observe in the results of MAGGIE-L.

A possible conclusion is that observational errors are very important when working with Bayesian galaxy group algorithms based on physical priors, contrary to geometrical based algorithms, where such uncertainties do not influence their performances.

5.5.5 Conclusion

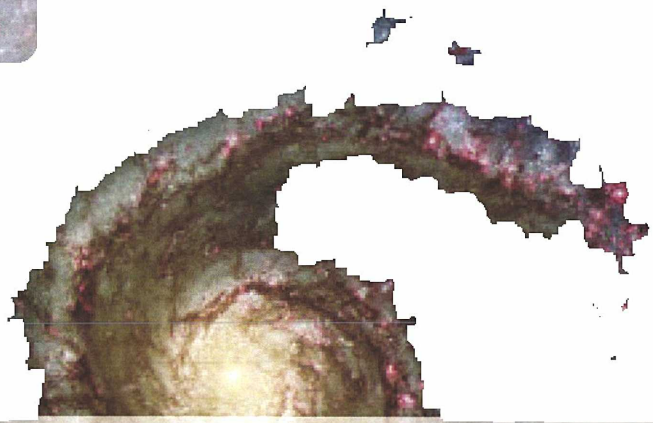
MAGGIE performs quite well in comparison to the optimal FoF. The use of probabilities to recover group properties is very useful to reduce the effect of inevitable interlopers present in the group membership. But when applied on data with uncertainties, its performances are reduced compared to tests with a perfect knowledge of the various needed observables. Although there are some counterparts for MAGGIE on realistic data, we note that globally it performs better than the FoF algorithm applied on perfect data. The extracted membership is better than the FoF and the virial mass estimation by abundance matching compared to the simple virial theorem used generally with FoF.

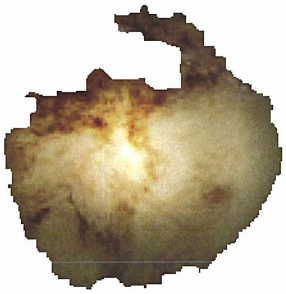
This makes MAGGIE a suitable grouping algorithm to be applied on large galaxy surveys such as the Sloan Digital Sky Survey (SDSS) and the Galaxy And Mass Assembly (GAMA).

THESIS

THESIS

Chapter 6.





SDSS-DR10 analysis

Contents

6.1 Introduction

An application of MAGGIE on a real galaxy survey implies an analysis of the galaxy sample. We must understand the various incompletenesses it suffers in order to be able to correct them. Here we describe the analysis we performed on the Sloan Digital Sky Survey, with the various problems we encountered.

6.2 Analysis

6.2.1 Definitions

In SDSS, stripes are bands of observations along great circles of the survey. Each of them is composed of six parallel scanlines (of 13 arcmin wide) with gaps of approximately the same width between them. Two stripes make a single stripe of 2.5°. Each scanline include all the data (in *ugriz*), and is divided in fields (that can overlap). So when accessing an observation at a given position in the sky, we access a specific field. A given observation is completely defined by its run number, the number of the camcol of the scanline and by the field number.

The pipeline of the SDSS is applied for the objects extraction. They are detected as pixel over-densities relative to the background. With this method, multiple real and different objects can be seen as a single object. They are linked by their pixels as galaxies using Friends-of-Friends algorithm. A deblending algorithm is then applied to resolve child objects from their parents

THESIS

(defined as the first detection). Then a resolve algorithm is applied to extract the best object when multiple fields are overlapping.

There are numerous object flags that are useful to select well observed galaxies. In the `PhotoObjAll` table, there is a `clean` for a predefined selection of the most common good flags, which facilitates the selection of galaxies.

There can be many problems with the photometry, with cases of bright galaxies with sky levels not well estimated and missing faint galaxies for example. Most of these known problems are corrected in the recent releases (DR9 and DR10).

Old releases worked with a spectrograph of 640 fibers, with collisions at $55''$, while the new BOSS survey works with a 1000-fiber spectrograph but with a greater collision size of $64''$. The coverage of the old releases should be used for the new BOSS, so its better to use latest releases. Moreover, the pipeline used for the spectrum had changed and improved along releases.

Following definitions given in the SDSS website, we can define two coordinate systems in the survey.

Great Circle: This coordinates system is define with two angles (μ, ν). Coordinates are relatives to one stripe so they can be used when working with galaxies inside a stripe region.

Survey Coordinates: It's an other system similar to celestial coordinates but "centred" on the contiguous block of galaxies of the survey. Coordinates are written (λ, η). The range of these coordinates is: $-\frac{\pi}{2} < \eta < \frac{\pi}{2}$ and $-\pi < \lambda < \pi$.

We will work only with survey coordinates as they allow us to easily define a mask for the SDSS. The celestial coordinates and survey coordinates are the same system of coordinates, except that one is a particular rotation of the other. The relations between the two systems are:

6.2.1.1 Survey coordinates to celestial coordinates

$$\begin{aligned} \delta &= \arcsin\left(\cos\lambda \sin\left(\eta + \delta_0\right)\right) \\ \alpha &= \operatorname{atan2}\left(\sin\lambda, \cos\lambda\cos\left(\eta + \delta_0\right)\right) + \alpha_0 \end{aligned} \tag{6.1}$$

with $(\alpha_0, \delta_0)_{(\alpha,\delta)} = (185°, 32.5°)_{(\alpha,\delta)} = (0, 0)_{(\lambda,\eta)}$.

6.2.1.2 Celestial coordinates to survey coordinates

The inverse transformation is:

$$\begin{aligned} \eta &= \operatorname{atan2}\left(\sin\delta, \cos\delta\cos\left(\alpha - \alpha_0\right)\right) - \delta_0 \\ \lambda &= \arcsin\left(\cos\delta \sin\left(\alpha - \alpha_0\right)\right) \end{aligned} \tag{6.2}$$

with $(\alpha_0, \delta_0)_{(\alpha,\delta)} = (185°, 32.5°)_{(\alpha,\delta)} = (0, 0)_{(\lambda,\eta)}$. Periodic conditions must be applied to angles found by the latter equation:

$$\begin{cases} \eta \to \eta + 180° \ \lambda \to 180° - \lambda & \text{if } \eta < -90° \text{ or } \eta > 90° \\ \eta \to \eta - 360° & \text{if } \eta > 180° \\ \lambda \to \lambda - 360° & \text{if } \lambda > 180° \end{cases} \tag{6.3}$$

6.2.1.3 Stripe number

Stripes have a constant width of 2.5° along the η coordinate. So, stripe number n of a galaxy with η coordinate is:

$$n = \text{floor}\left(\frac{\eta + 58.75°}{2.5°}\right) \tag{6.4}$$

6.2.2 Galaxy selection

Many tables in the SDSS save galaxies and other objects properties extracted from images of the survey. These tables are the results of different selections in objects extracted in images. When crossing objects between images of the survey that overlap, there are some differences in positions for the same object. So there are possibilities that an object is observed twice or more. In many of those tables, there is no object duplicated.

In the SDSS database, the `Galaxy` view is a selection from the `PhotoPrimary` for objects flagged as *galaxy*, with `type`=3. The `Galaxy` view contains the photometric parameters (no redshifts or spectroscopic parameters) measured for resolved primary objects. But we have other useful informations to link with tables that give us photometric and spectroscopic redshifts. There is the `specobjid` entry to link with spectroscopic redshifts in the table `SpecObj` which doesn't contain duplicates (it's a clean table of `SpecObjAll` with clean redshifts). If `specobjid=0`, the galaxy doesn't have a spectroscopic redshift (the galaxy wasn't spectroscoped). The `objid` allows to link to the `Photoz` table which contains all photometric redshifts for galaxies in the `Galaxy` table. Estimation is based on a robust fit on spectroscopically observed objects with similar colors and inclination angle. There is also the `PhotozRF` where estimates are based on the Random Forest technique. Galaxies in the SDSS are limited to $m_r < 17.77$ and a given surface brightness. So we need to apply the same flux limitations when selecting galaxies on the `Galaxy` table. A possible `SQL` query for selecting galaxies in this table and link them with redshift tables is for spectroscoped galaxies:

```
SELECT G.ra, G.dec, G.petroMag_u, G.petroMag_g, G.petroMag_r,
G.petroMag_i, G.petroMag_z, G.specobjid, G.objid, Z.z, Z.Zerr
FROM Galaxy AS G
JOIN SpecObj AS Z ON Z.specobjid=G.specobjid
WHERE G.specobjid!=0
AND G.petroMag_r-G.extinction_r<17.77
```

and for galaxies which couldn't be spectroscoped:

```
SELECT G.ra, G.dec, G.petroMag_u, G.petroMag_g, G.petroMag_r,
G.petroMag_i, G.petroMag_z, G.specobjid, G.objid, Z.z, Z.Zerr
FROM Galaxy AS G, Photoz AS Z
WHERE G.specobjid=0
AND G.objid=Z.objid
AND G.petroMag_r-G.extinction_r<17.77
```

Stripe limits are given in the table `StripeDefs` but they represent the limits that were planned at the beginning of the survey, not the actually observed limits.

THESIS

Some planned regions aren't still observed, so we need to define other limits in λ coordinates for incomplete stripes. We find, by hand, the new limits of stripes which contains spectroscoped galaxies. Now, the survey mask is like in Figure 6.1. We will consider just galaxies in this mask in order to find groups in the SDSS.

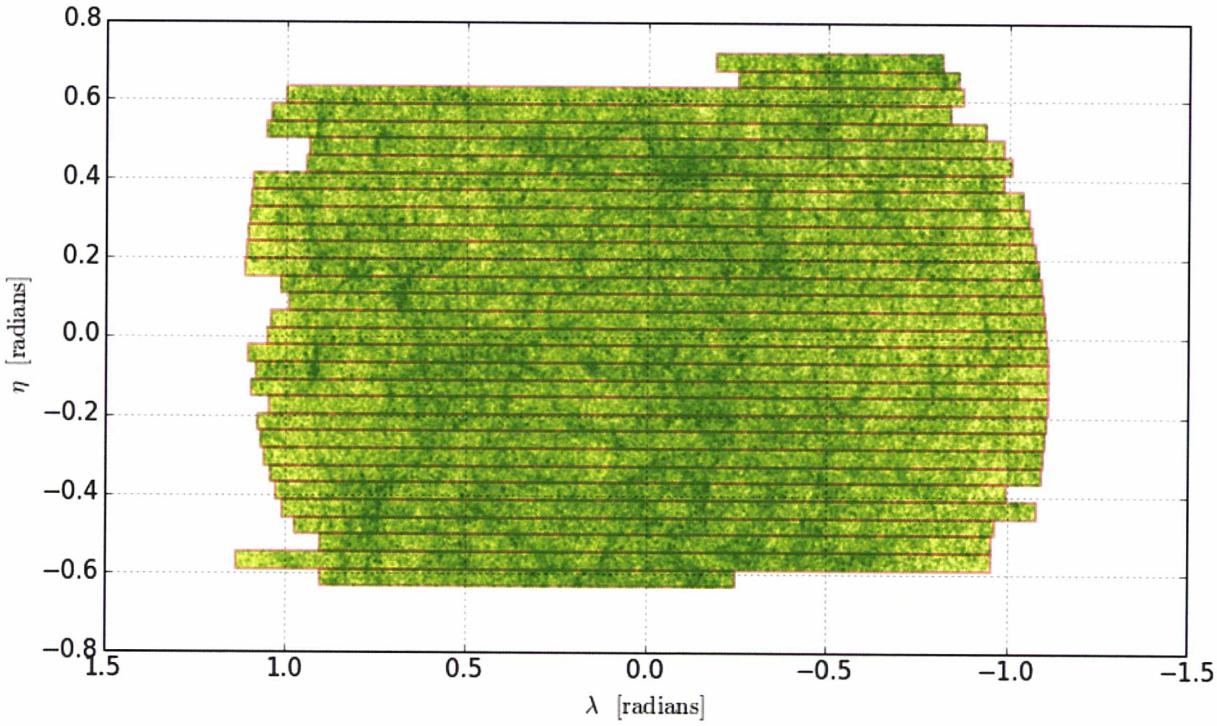


Figure 6.1: Galaxies in the SDSS DR10 with stripes limits defined by hand. The red lines limits of the stripes make the SDSS mask used to identify edges.

6.2.2.1 Flags in the SDSS

Galaxy photometry can have some troubles in the SDSS. In the general case, those objects are flagged with `clean` property which indicates by 1 that the photometry is OK and by 0 when there is a problem. Details of the problems are in the bit flag. But for groups, we need to select all galaxies, even if they are not clean, or our groups will suffer incompleteness in their membership and their physical properties such as luminosity, stellar mass... will be biased.

However, we have to take into account the error on the redshift estimation using `zErr`. For photometric redshifts, if `zErr` is too high, we can use `nnAvgZ`, which is the average redshift of galaxies in the neighbourhood of the considered galaxy. It can be better if the photometric redshift is strongly different from its value.

`SpecObjAll` contains duplicates and bad data. But `SpecObj` contains just clean spectra. The field `zWarning` can be used to decide if we keep a redshift or not.

6.2.3 Fibre collision estimation

We need a sample of galaxies for which we can easily characterize borders and where all galaxies are present given the flux limit of the survey. But there is the problem of missing galaxies due to fibre collisions. But our algorithm is tested on a "perfect" mock catalogue. In order to know the

behaviour of the algorithm with these problematic galaxies, we need to implement the effect of fibre collisions in our mock catalogue.

In the SDSS, galaxy spectra are obtained on fibers using a plate of 1.5° diameter. But on the plate, the number of fibres is limited. Moreover, each portion of the sky can't be spectroscoped multiple times, because the SDSS had to cover a predefined portion of the sky in a fixed number of years. Although spectroscopic runs may overlap, there are galaxies that can't be spectroscoped. Indeed, while fibres collect spectra in a 3″ diameter field, their coatings prevent two fibres of lying close than 55″ from one another. When galaxies are closer than this distance, one (or more) of those galaxies aren't spectroscoped. We can see this fibre collision effect in Figure 6.2, where we have taken the nearest neighbour of a galaxy on the celestial sphere, and determined the differences in angular positions and redshift between the two galaxies. As expected, the number of galaxies that are closer than 55″ is much less than what would be extrapolated from greater separations. There are still some galaxies because the overlapping of runs allows to observe galaxy spectra below this limit.

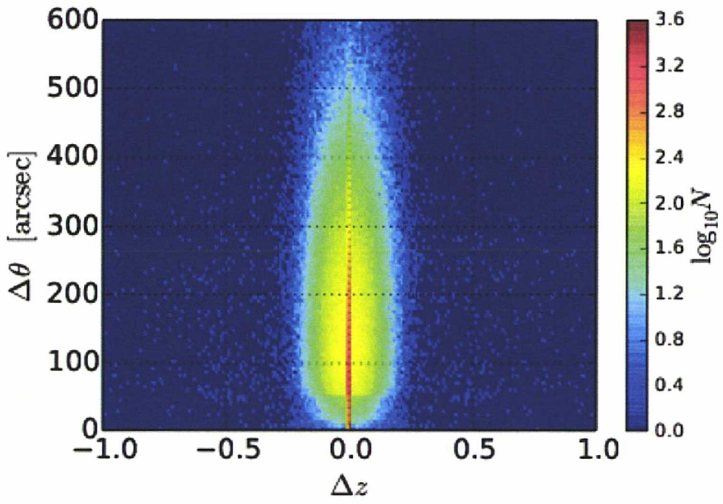


Figure 6.2: Distribution of spectroscoped galaxies in the SDSS DR8 in angular size and redshift differences with the nearest neighbour galaxy.

Nevertheless, the dense regions with more than one galaxy per 55″ diameter circle are partially incomplete in the SDSS spectroscopic sample.

We tried to implement this selection effect in our mock catalogue. For this, we computed the local density in the field, taking all galaxies (spectroscoped or not) in the neighbourhood of 1.5° around each galaxy, and at the same time, we determine the fraction of galaxies that do not have a spectroscopic redshift, to see the relation between spectroscopic completeness and photometric galaxy number density. We expect to deduce a relation between the density field and the fraction of fibre collisions. In the mock catalogue, we compute the same density field and we apply the spectroscopic completeness relation estimated in the SDSS sample to the mock. We have to remove galaxies that are close to survey edges, because otherwise, there are missing galaxies and the spectroscopic completeness will be affected. Edge galaxies are those lying closer than 1.5 deg from the survey edges, which we measure in practice by generating XXX random points within a circle of 1.5 deg radius around each galaxy.

We didn't see the trend we expected with the density field, so we thought that it can be due to the large area in which we compute the fraction of spectroscoped galaxies and we ran the same with a radius of 0.3°, but without success too.

TESIS

Remark 3

We can generate samples of points at an angular distance d to a point at position (α_0, δ_0) using formulas of the spherical triangle. If we define a triangle by the pole, the point (α_0, δ_0) and the point whose we want coordinates (α, δ), we can write the following relations using the spherical triangle and its dual:

$$\begin{aligned} \sin\delta &= \sin\delta_0 \cos d + \cos\delta_0 \sin d \cot\gamma \\ \sin\delta_0 \cos\gamma &= \cos\delta_0 \cot d - \sin\gamma \cot(\alpha - \alpha_0) \end{aligned} \tag{6.5}$$

where γ is like a polar angle, which have all the values between 0 and 2π. We can rewrite:

$$\begin{aligned} \delta &= \arcsin(\sin\delta_0 \cos d + \cos\delta_0 \sin d \cos\gamma) \\ \alpha - \alpha_0 &= \arctan\left(\frac{\sin\gamma}{\cos\delta_0 \cot d - \sin\delta_0 \cos\gamma}\right) \end{aligned} \tag{6.6}$$

There are problems at poles. For a γ_0 limit, angles can't be recovered with above formulas. Indeed, the problem appears when $\tan\Delta\alpha \to \infty$. So:

$$\cos\delta_0 \cot d - \cos\gamma_0 \sin\delta_0 = 0 \tag{6.7}$$

implying:

$$\cos\gamma_0 = \frac{1}{\tan d \tan\delta_0} \tag{6.8}$$

So to handle these limit cases, we summarize the correction for the differences in right ascensions by:

$$\Delta\alpha \to \Delta\alpha + \pi \quad \text{if} \quad \text{sign}(\delta_0)\cos\gamma \geqslant \text{sign}(\delta_0)\cos\gamma_0 \tag{6.9}$$

Another way to draw circles on the sphere is to consider the point for which we want to know celestial coordinates around a given angular distance as the pole of a new coordinate system. In this system, points at a given distance of our central point are just points with $\pi/2 - \delta$ and α running between 0 and 2π. We now can determine cartesian coordinates of those points in this system and apply a rotation to go from the "real" system to the system where the central point is the pole. This can be easily done if we know the axis of rotation and the angle using quaternions, which is numerically more efficient than Euler angles. ■

Moreover, including photometric redshifts in the mock catalogue and in MAGGIE is very complex. For example, we measured the bias and dispersion of the distribution of differences between spectroscoped and photometric redshifts in the SDSS. Figure 6.3 shows that while the dispersion remains roughly constant, the bias increases with the spectroscoped redshift. So some effects are not still under control when computing photometric redshifts, and we should avoid their use in galaxy group algorithms when possible. In the case of surveys where spectroscopic redshifts are not available, the photometric redshifts should be as clean as possible.

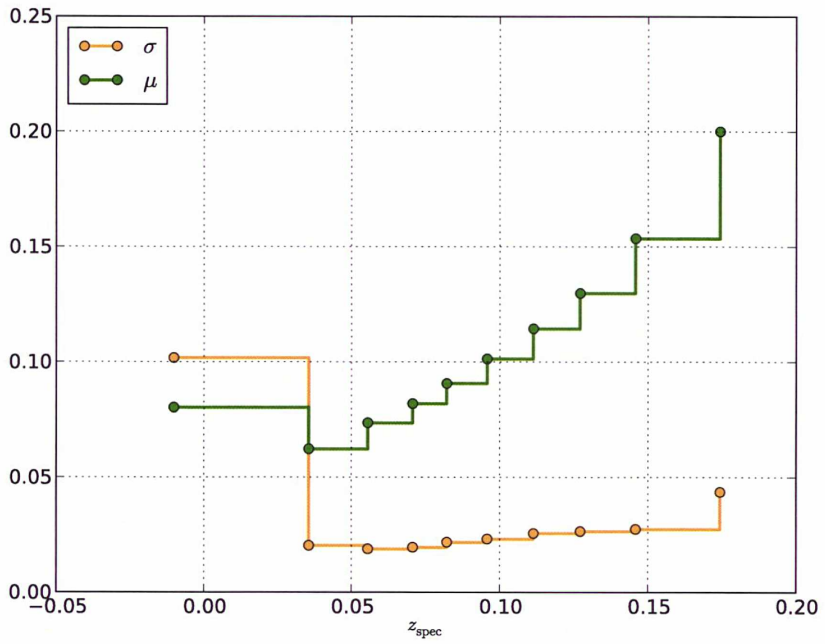


Figure 6.3: Bias (μ) and scatter (σ) of $z_{\text{phot}} - z_{\text{spec}}$.

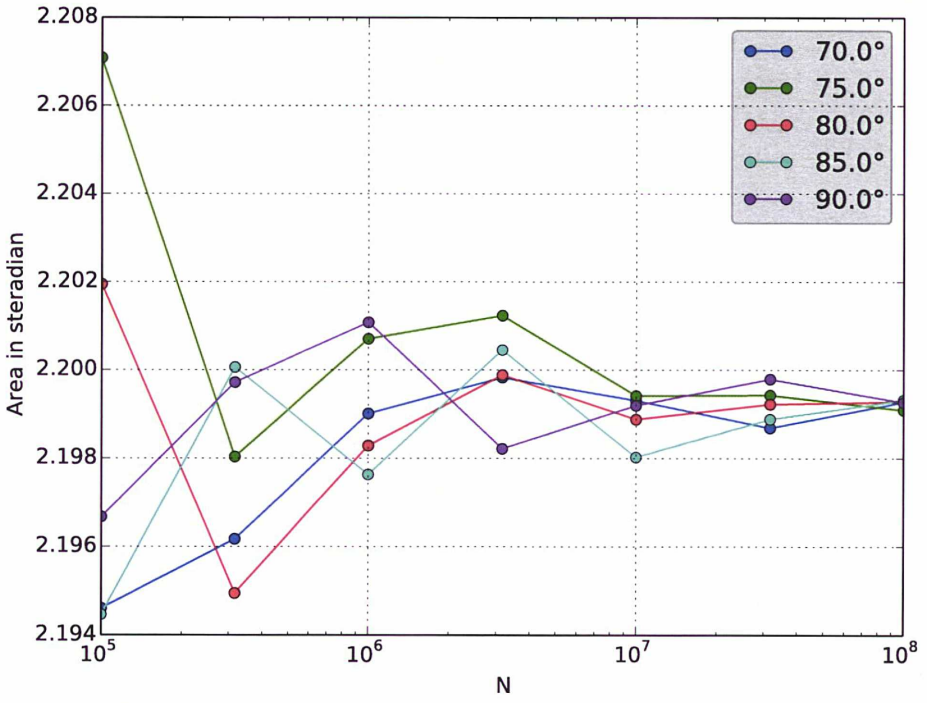


Figure 6.4: Determination of the area of the SDSS for our selection with a Monte Carlo process. Results converge on a value of 2.1993 ± 0.0001 steradians (i.e. roughly $7220 \pm \text{deg}^2$).

THESIS

6.3 Coverage of the SDSS

For many computations in this thesis, we need to determine the solid angle covered by our galaxy sample. In the SDSS, the mask we constructed allows us to do it easily by a Monte Carlo process.

First, we generate a number N of points around a point of coordinates (α_0, δ_0) with a maximal angular separation $\theta_{\max}$ which is larger than the maximal angular separation in our sample. The fraction of points falling inside the mask gives us the fraction of the generated area corresponding to the mask. This area is just $\mathcal{S} = \int_0^{\theta_{\max}} \int_0^{2\pi} \sin\theta \mathrm{d}\theta \mathrm{d}\phi = 2\pi\,(1 - \cos\theta_{\max})$. We made this calculation for different cone angles $\theta_{\max}$ and for different number of points to see if we have a convergence in the value of the area. Figure 6.4 shows that our geometry has a solid angle of $7220 \pm 1\mathrm{deg}^2$ but this required five simulations with 10^8 points.

Remark 4

Generating points uniformly on the celestial sphere around a point of coordinates (α_0, δ_0) to an angular distance d can be done by assuming that this point is the upper pole of an other spherical system. In this situation, points follow $0 \leqslant \theta \leqslant d$ and $0 \leqslant \phi \leqslant 2\pi$, assuming spherical coordinates and not celestial one. The azimuthal ϕ coordinates are generated as $2\pi U_1$ where U_1 is a random variable following an uniform distribution between 0 and 1. The latitude θ coordinates, follow $\left(p(\theta) = \frac{1}{2}\sin\theta\right)$ and are generated by $\theta = \arccos(2U_2 - 1)$, where U_2 is a variable following an uniform distribution with values between 0 and 1.

Then, the points are rotated by quaternions to (α_0, δ_0). The rotation axis is just the cross product between the pole vector and the vector defined by (α_0, δ_0), and the rotation angle is $\frac{\pi}{2} - \delta_0$. ■

6.4 Galaxy stellar masses

In SDSS, contrary to coordinates, magnitudes or redshifts, stellar masses are not measured by the SDSS pipelines. Instead, several teams have applied stellar population models to the spectra and corrected their stellar masses from the area subtended by the spectroscopic fiber to the entire galaxy, using the apparent magnitudes within the fiber (fiberMag) and that extrapolated to the entire galaxy (petroMag or modelMag). Indeed, contrary to coordinates, magnitudes or redshifts, the stellar mass is not a direct observable. Its estimation is based on the application of various stellar population models on the galaxy spectrum observed by the SDSS. Several models exist, but they do not provide the same estimation for a given galaxy. In Figure 6.5, we compare eight models to have an order of the inaccuracy of the stellar mass: FSPSGranWideDust, FSPSGranWideNoDust, FSPSGranEarlyDust and FSPSGranEarlyNoDust from Conroy et al. (2009), PassivePort and StarFormingPort from Maraston et al. (2009), PCAWiscM11 and PCAWiscBC03 from Chen et al. (2012) and MPA-JHU from Brinchmann et al. (2004); Kauffmann et al. (2003); Tremonti et al. (2004).

The principal discrepancies between the models come essentially from the various stellar population synthesis (SPS) models involved in the fit of the galaxy spectrum, necessary for the stellar mass estimation. But each model has also some internal variations. For example, Conroy et al. (2009) assume an early star formation in galaxies for its FSPSGranEarlyNoDust (without dust

extinction correction) and FSPSGranEarlyDust (with dust extinction correction), while FSPS-GranWideDust and FSPSGranWideNoDust assume an extended star formation history. As we can see, differences are relatively important: models using different SPS have large dispersion in their estimation, while when using the same SPS, stellar masses are coherent. Some models are also biased between each other, but bias can be corrected and not considered in our analysis. Generally, models agree to better than 0.3 dex, i.e. errors on individual masses are of $0.3/\sqrt{2} = 0.2$ dex. In particular, the MPA-JHU masses agree with all others to typically better than 0.2 dex in σ.

6.5 Final galaxy sample

All previous sections are showing something important in the SDSS data: observational errors can be important, and the automatic processing of this data sometimes leads to false detections, artefacts..., making analysis and corrections complex.

Fortunately, recently in their FoF analysis of galaxy groups in the SDSS-DR10, Tempel et al. (2014) had to deal too with such problems and the contamination they introduce. Major problems are stars classified as galaxies, nearby large galaxies fragmented into several galaxies or poor photometry of some galaxies due to bright stars or bad sky level estimation in the neighbourhood. They performed an impressive filtering on the sample by visually checking 30000 galaxies that were potentially problematic galaxies. Tempel et al. (2014) thus checked the following:

- 10000 apparently brightest galaxies (in r band). For galaxies brighter than $m_r < 13.5$, about 10% of the objects were spurious. For galaxies $13.5 < M_r < 14.5$, about 1% were spurious entries; this fraction decreases with luminosity;
- 5000 intrinsically brightest galaxies in the sample ($< 1\%$ were spurious);
- 3000 intrinsically faintest galaxies in the sample (to ensure the correctness of the faint-end of the luminosity function);
- all the sources with the spectroscopic class QSO;
- all the objects with `bestobjid` missing or not GALAXY. For these objects, they used `fluxobjid` if the matched photometric object was classified as a galaxy;
- all the objects for which the difference between r band point spread function (PSF) magnitude and model magnitude was smaller than 0.25 (thus further excluding some of the stellar sources in the catalogue);
- all the galaxies with the difference between r band Petrosian and model magnitudes greater than 0.4;
- all the galaxy pairs that were closer than $5'$ (in order to remove double/multiple entries);
- the entries where the colour indices $g - r$, $r - i$, and $g - i$ had extreme values.

Finally, Tempel et al. (2014) removed around 600 galaxies, while 1400 other galaxies were flagged as having a bad photometry. We decided to use their galaxy sample, since it covers exactly the same area we use and their conscientious clean up of the SDSS-DR10 is difficult to surpass.

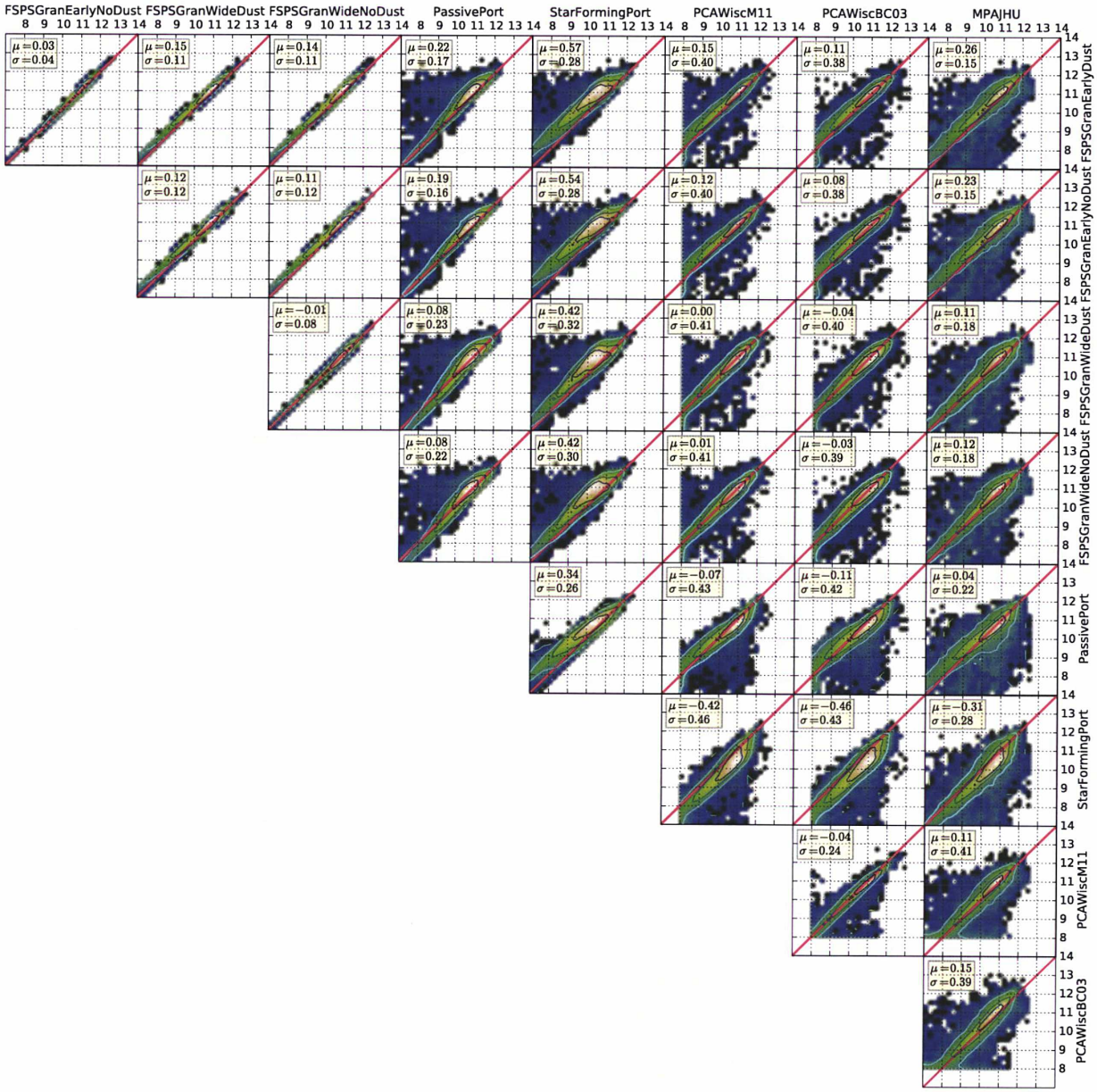


Figure 6.5: Comparison between stellar mass models applied onto galaxies from SDSS. Contours show the offsets in units of the scatter σ. Ordinates and abscissas are the logarithmic stellar masses of galaxies in the solar units. The upper left box shows the bias (μ) and dispersion (σ) of the logarithmic difference between both models. The models are FSPSGranWideDust, FSPSGranWideNoDust, FSPSGranEarlyDust and FSPSGranEarlyNoDust (Conroy et al., 2009), PassivePort and StarFormingPort (Maraston et al., 2009), PCAWiscM11 and PCAWiscBC03 (Chen et al., 2012) and MPA-JHU (Brinchmann et al., 2004; Kauffmann et al., 2003; Tremonti et al., 2004).

6.5.1 Stellar masses

Stellar masses are a major component of our algorithm, but Tempel et al. (2014) did not work with them, letting us the choice of the stellar masses to use. In Figure 6.5, we show that there are large differences between available models in the SDSS database. Estimations of some models are different from the other and should not be used. A way to deal with this problem is, for each galaxy in the sample, to use the median of the stellar mass for all models. But sometimes, we don't have access to the stellar mass of a galaxy and removing it from the sample will create supplementary incompleteness. In such situations, we provide by default the photometrically-based stellar mass estimation of Bell et al. (2003). Those fitting formulas allow to get the stellar mass of a galaxy directly from its color and luminosity. Several colors are available to make the computation. We show in Figure 6.6 the stellar mass distribution for several colors used on the formula of Bell et al. (2003). The $r-z$ color creates fewer outliers in stellar masses than other bands. A possible explanation is that the magnitude bands involved in the computation are less sensitive to dust extinction and thus provide a more accurate estimation of stellar mass. So, we adopt the stellar mass from $r-z$ color for those galaxies without spectral mass estimates in the SDSS database.

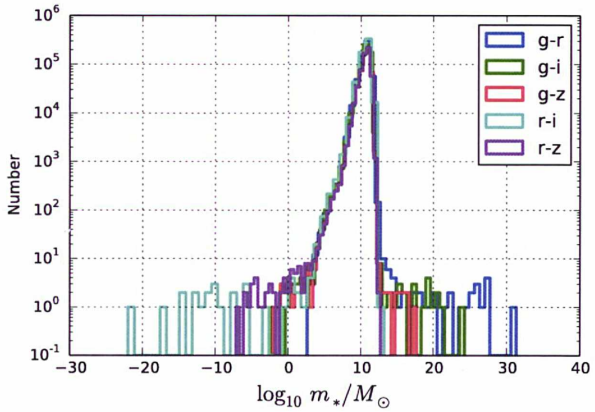


Figure 6.6: The distribution of stellar masses for galaxies on the SDSS with the median of models described in Figure 6.5 and the default value for galaxies without stellar mass estimations from Bell et al. (2003) for different magnitude colors. The number of non-physical values for stellar masses is reduced by using the $r-z$ color, less affected by dust extinction and hence more accurate.

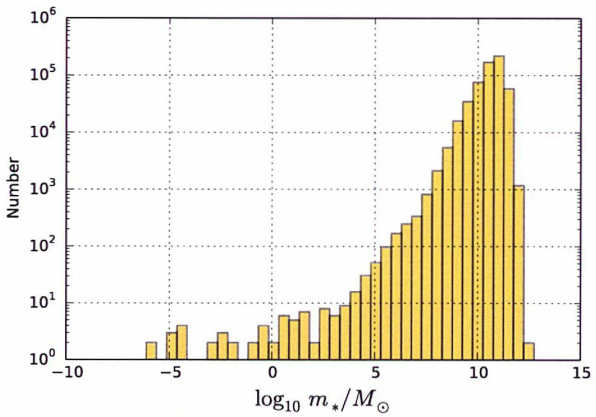


Figure 6.7: The distribution of stellar masses in solar units for our SDSS galaxy sample once chosen the $r-z$ color magnitude as default estimation. There are no high mass galaxies, but some stellar masses have unphysically low stellar masses. We keep them to avoid introducing supplementary incompleteness in the galaxy sample.

The resulting stellar mass distribution from our galaxy sample is shown in Figure 6.7. No galaxies have too high stellar masses, but a lot of them are very low and seems to not be physical. But we can't remove them without introducing an incompleteness hence we keep them in the sample.

6.5.2 Star formation rate

Measures of the star formation rate suffer the same problems as stellar masses: the different models (from the same teams) do not necessarily agree with one another. The comparison of the SFR measures (shown as the specific star formation rate, SSFR = SFR divided by stellar mass) is shown on Figure 6.8. Some models disappeared: PCAWiscM11 and PCAWiscBC03 do not provide SFR estimates for galaxies, while PassivePort and StarFormingPort produce null SFR

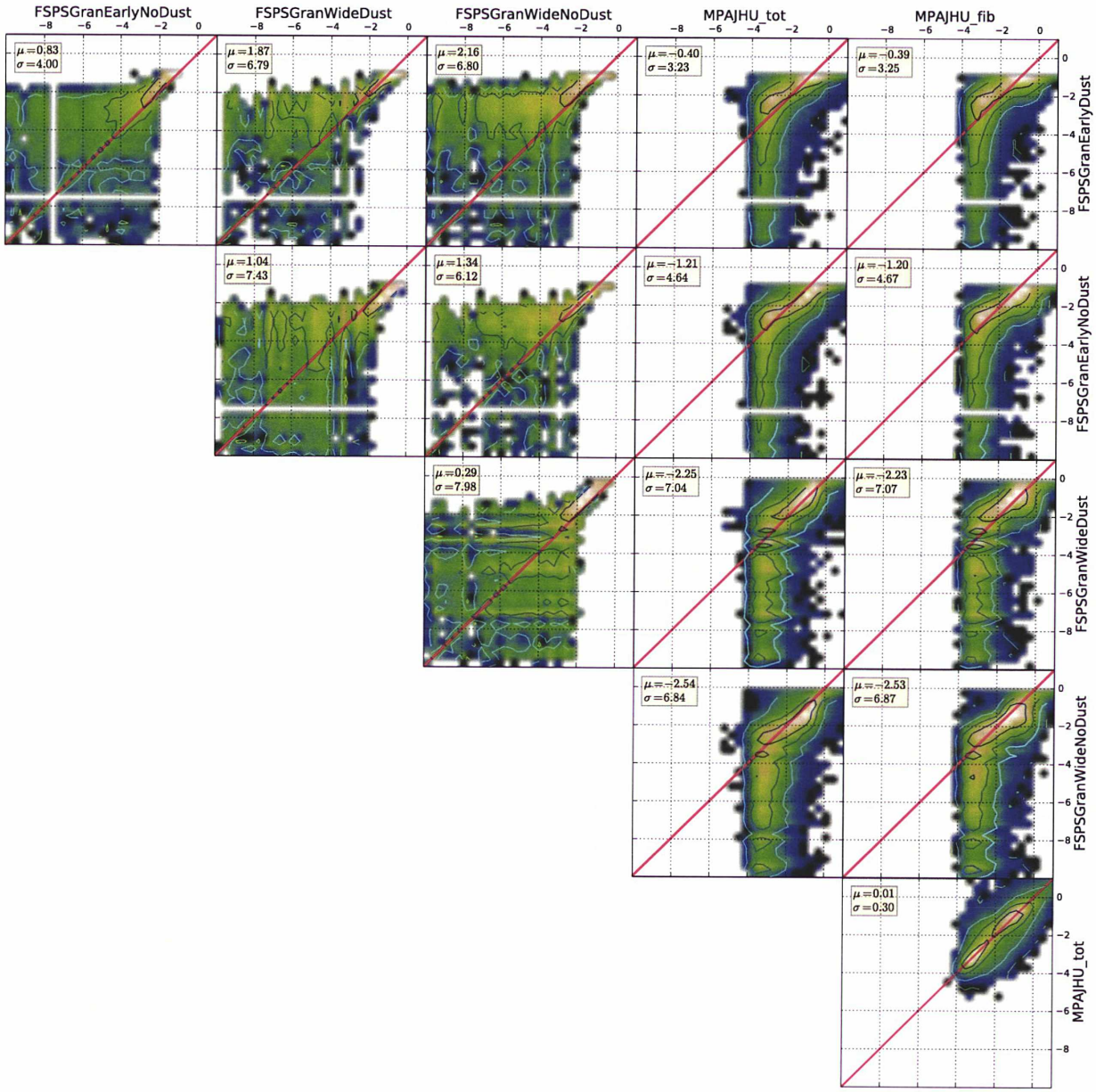


Figure 6.8: Comparison of several specific star formation rate (SSFR) measures from different models. FSPSGranWideDust, FSPSGranWideNoDust, FSPSGranEarlyDust and FSPSGranEarlyNoDust from Conroy et al. (2009) and MPA-JHU Brinchmann et al. (2004); Kauffmann et al. (2003); Tremonti et al. (2004). Two variants exist for MPA-JHU according to if the estimation is based on the region of the fiber (suffixed *fib*) or is also extrapolated (suffixed *tot*). Axes are $\log_{10}$ SSFR in units of Gyr^{-1}. We show also the bias and dispersion of the log-difference of models. FSPS models are not very consistent between one another and we do not use them. MPA-JHU are relatively coherent but we should prefer the *total* estimation since with the fiber estimation not all the stellar population of the galaxy is probed.

values for too many galaxies. MPA-JHU has several estimates of the SFR: one based only on informations acquired by the fiber pointing to the galaxy to get its spectrum (suffixed by *fib*) and the other where an extrapolation of the informations is done outside the aperture (suffixed by *tot*).

Figure 6.8 shows that the SSFR values from the different FSPS models are not consistent with one another, hence we did not use them in our analysis. The bias and scatter between both models of MPA-JHU are small, making them good measure of the SSFR of galaxies. Our preference goes to the *total* model, since the extrapolation used by the authors (at roughly constant SSFR) must be better than none.

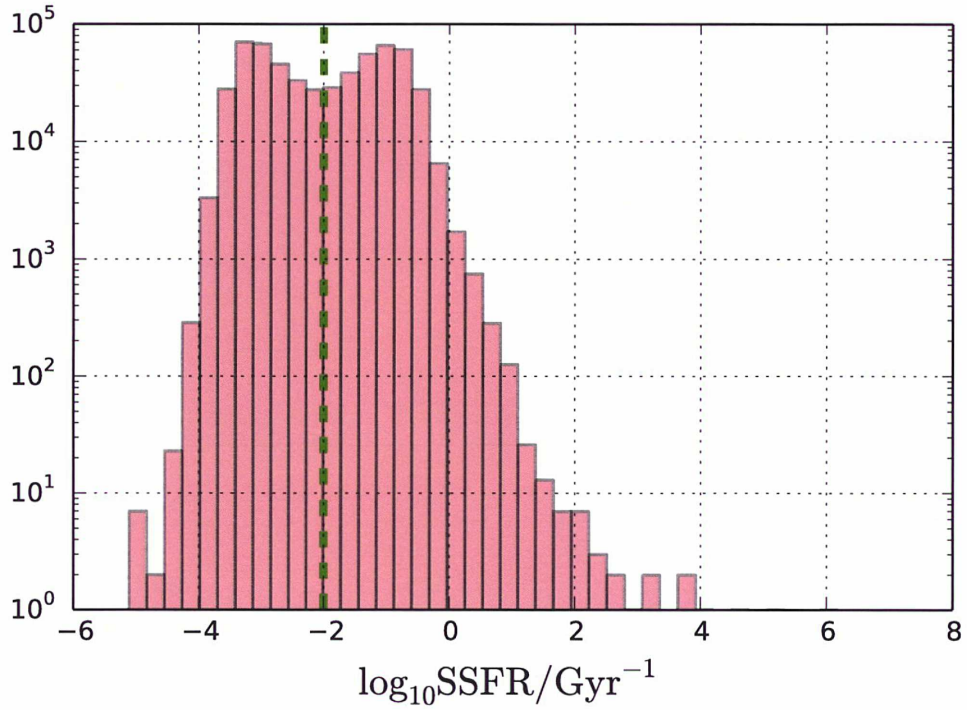
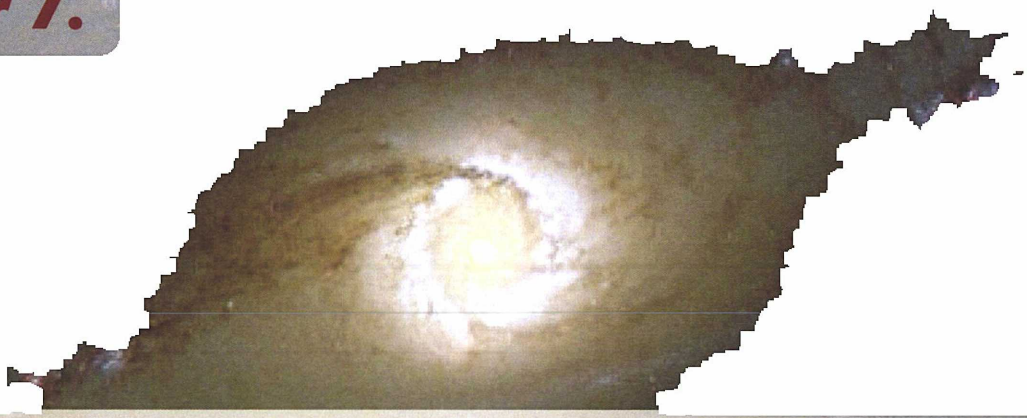


Figure 6.9: The distribution of SSFR from the MPA-JHU model. We can see a bi-modality in the distribution, splitting galaxies into star forming and passive galaxies (the *green dashed* line shows the separation).

The resulting distribution of the SSFR in Figure 6.9 shows that galaxies can be classified into two categories: a star forming population where an important fraction of the galaxy stellar mass is produced in a time scale of 1 Gyr and a passive one, where ongoing or recent star formation represents a small fraction of its stellar mass. The green line in Figure 6.9 shows this limit. It will be useful when searching if there is a modulation with the environment of the fraction of young galaxies.

THESIS

Chapter 7.



Conclusions and perspectives

7.1 Conclusions

The optimal extraction of galaxy groups from redshift space is not an easy task. The observer has to deal with observational errors, projection effects and bias to perform such an optimal grouping. We have argued that all previously created galaxy group algorithms are imperfect in the sense that with their assumptions, there are some lacks in extracted groups, explaining the apparition of two different kind of algorithms: Bayesian and geometrical.

We constructed a galaxy mock catalogue to test several grouping algorithms. We tested the Friends-of-Friends algorithm to understand what is the optimal set of linking lengths. We conclude that the choice of optimal linking lengths depends on the science one wishes to do.

We created MAGGIE, a Bayesian galaxy group finder, using probabilities to constrain the membership in groups. The virial radii are estimated from either the stellar mass (MAGGIE-m) or the luminosity (MAGGIE-L) of the central galaxy. We show by tests on our mock catalogues that both implementations of MAGGIE perform better on perfect data with no observational errors than the optimal FoF algorithm. We also show that Bayesian algorithms as MAGGIE are more sensitive to the quality of observational data than geometrical ones as FoF. Nevertheless, both MAGGIE-m (with 0.02 dex errors in stellar masses) and MAGGIE-L (with 0.08 dex errors in observed luminosities) perform better than the optimal FoF, except at very high group masses, where the abundance matching technique used in MAGGIE becomes inaccurate.

The application of MAGGIE on real galaxy surveys implies a full understanding of the possible incompletenesses of these surveys. The analysis of the SDSS-DR10 indicates that correcting for luminous and spectroscopic incompletenesses is very important but also very difficult, since the extraction of galaxy groups implies no missing galaxies. Incomplete membership can affect the grouping, but also the informations obtained from their analysis. Indeed, environmental effects we wish to observe in galaxy groups (essentially through the SSFR) can be very sensitive to the way incompleteness is handled. Therefore, MAGGIE is a very powerful tool for galaxy group analysis, but we have to apply it carefully on the analysed data or the interpretation of results can be biased.

7.2 Perspectives

We plan to run MAGGIE on the SDSS-DR10 and publish optimized galaxy groups in different doubly complete subsamples in redshift and luminosity, and to re-assess the modulation of sSFR, etc... with local and global environments. If a modulation of galaxy properties is observed, we will model it and apply it in semi-analytical codes to see if it reduces discrepancies between observations and outputs of such codes. It will be a new measure of quenching of star formation

THESIS

with global and local environments. We also wish to run MAGGIE on the deeper GAMA redshift survey to be able to extract the evolution in time of environmental effects on galaxies. We already have some preliminary results on this modulation applying MAGGIE on the SDSS. Figure 7.1 and Figure 7.2 show the median SSFR and the fraction of non-passive galaxies with local and global environment for the SDSS, and the same for group catalogues of Tempel et al. (2014) in Figure 7.3 and Figure 7.4. We use two complete catalogues to show this modulation: catalogue 3 to have sufficient statistics in number of galaxies and catalogue 5 to see the behaviour at larger redshifts.

The modulation of the SSFR and fraction of young galaxies is very dependent of the catalogue of groups used. Since these results between the two algorithms, a deeper analysis must be done to understand from where the discrepancies come from.

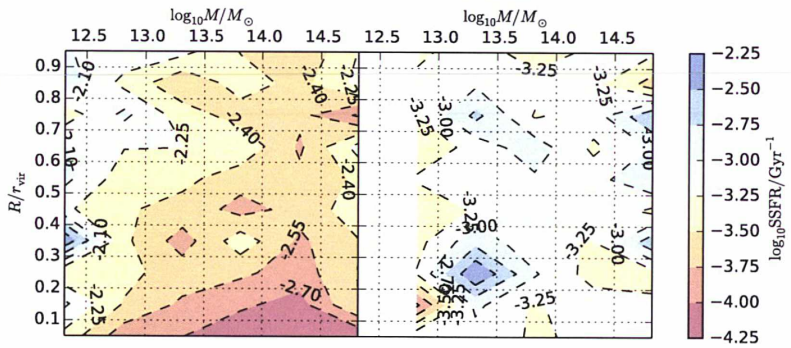
Still for MAGGIE, we plan to improve the galaxy grouping by the use of the red-blue segregation of galaxies. We can use some priors for the modulation of the fraction of blue galaxies in groups to adapt the probability computation according to the class of the galaxy. Then, we can iteratively reduce the impact of our initial model for the blue fraction by using the informations obtained by MAGGIE to de-project the red-blue segregation observed in groups. For next iterations, we can re-use our new real space model in the probability computation, and do it until the convergence of memberships.

In parallel, we plan to launch a collaborative project with other grouping algorithm developers. We will propose to each developer (and myself) to apply their algorithms to a set of mock catalogs constructed in the same way to avoid cosmic variance on the results, for blind tests. Then, we will run the same tests on each algorithm in order to have a clear understanding of the strengths and weaknesses of each of them. It will be the first time that galaxy grouping algorithms will be compared in the same conditions.

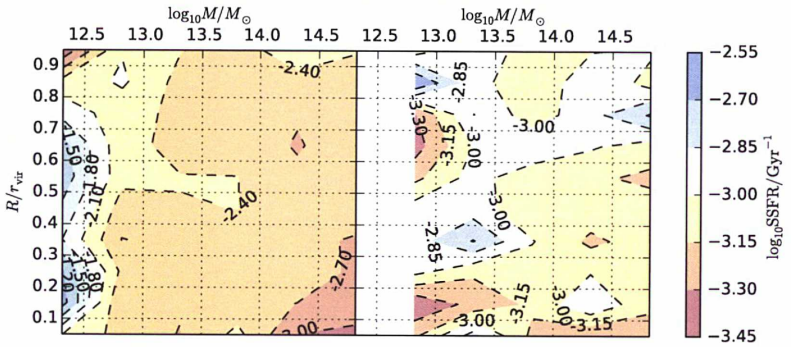
Given that imperfect grouping algorithms wash out the observable environmental effects, it is also interesting to know if there is a limit to recover the real space modulation of galaxy properties (such as specific star formation rate) with environment when trying to extract it from projected redshift space. This can be easily done by imposing ourselves a modulation in the outputs of galaxy formation codes, and then construct galaxy mock catalogues in redshift space. We can then see if the imposed modulation is recovered in the observations, and if galaxy group algorithms introduce biases in some cases. This will allow one to determine the maximum level of environmental dependence of star formation quenching that is consistent with the observations.

In continuation with the thesis work, we can try to theoretically explain the observed dependency of galaxy properties with their environments. Using hydrodynamical simulations of galaxies in groups, we wish to understand and model intra-cluster physical processes (ram pressure stripping, tidal stripping...). This will imply running academic simulations independently for each physical process, then model as a function of the different input parameters (local and global environment essentially). And by trying to switch them on-off in semi-analytical codes of galaxy formation, determine their relative importance on galaxy properties (sSFR, bulge to disk ratios...).

THESIS

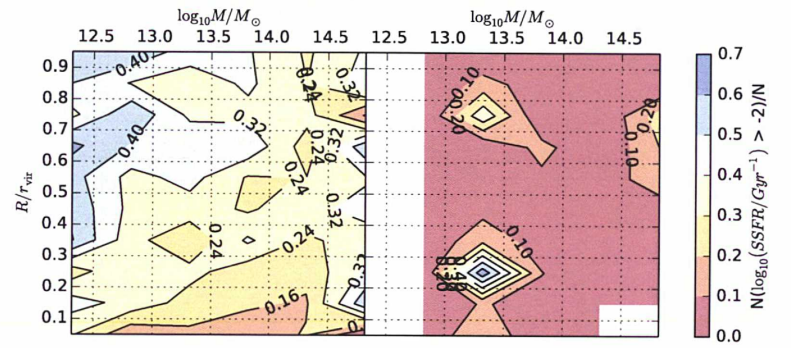


(a) Catalogue 3

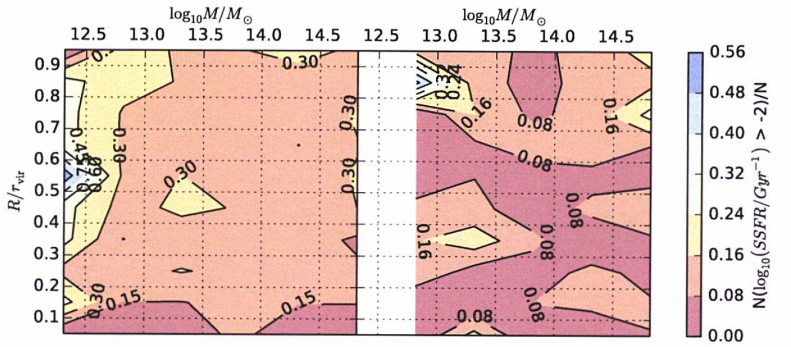


(b) Catalogue 5

Figure 7.1: Mean SSFR for galaxies with $10 \leqslant \log_{10} m_* < 11$ (left panel) and $11 \leqslant \log_{10} m_* < 12$ as a function of the projected radius in units of virial radius (local environment) and of the virial mass in solar units (global environment), for galaxy groups found with MAGGIE.

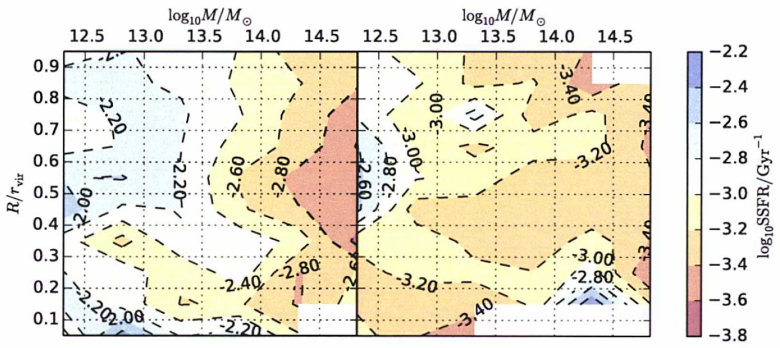


(a) Catalogue 3

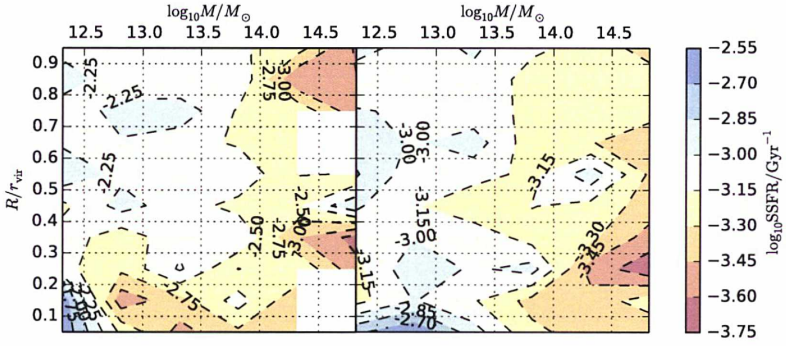


(b) Catalogue 5

Figure 7.2: Fraction of galaxies classified as star forming galaxies according the criterion of Section 6.5.2 for the same range in stellar masses as in Figure 7.1, with galaxy group from MAGGIE.

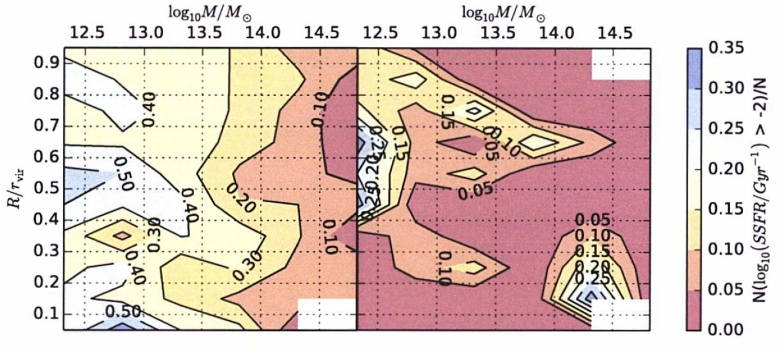


(a) Catalogue 3

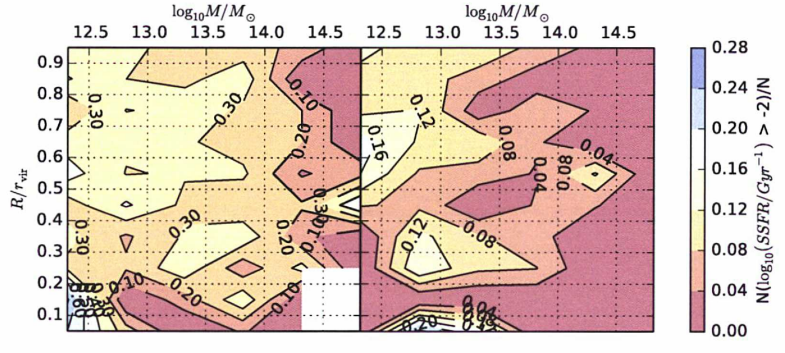


(b) Catalogue 5

Figure 7.3: Same as Figure 7.1 but for galaxy groups from Tempel et al. (2014).



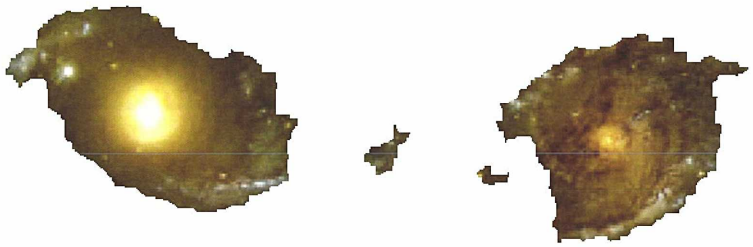
(a) Catalogue 3



(b) Catalogue 5

Figure 7.4: Same as Figure 7.2 but for galaxy groups from Tempel et al. (2014).

Appendix A.



MAGGIE's adventures

While developing MAGGIE, we tried several methods with the goal of improving the extraction of galaxy groups. Many of those methods weren't viable because of bad performances when applied to the galaxy mock catalogue we constructed, or because their implementation were too complex by technical imperatives or some physical knowledges not easily accessible. We describe some of these methods in the following sections. We describe also what we think are possible future improvements of MAGGIE.

A.1 Flux-limited algorithm

A.1.1 Problem

For MAGGIE, we tried to create a flux-limited version of the algorithm to apply it to a large range of luminosities and redshift.

The problem is that we must correct for missing low-luminosity galaxies. One way is to take into account the luminosity function of galaxies in the sample, and with that assumption, one can correct for the fraction of missing galaxies expected at a given redshift. Assuming that the luminosity function is $\phi(L)$, this fraction can be written:

$$f(L_{\mathrm{lim}}(z)) = \frac{\int_{L_{\mathrm{lim}}(z)}^{\infty} L\phi(L)\,\mathrm{d}L}{\int_{L_{\mathrm{thresh}}}^{\infty} L\phi(L)\,\mathrm{d}L} \tag{A.1}$$

where $L_{\mathrm{lim}}(z)$ is the minimal luminosity that a galaxy should have to be observed at the redshift z, given the observed magnitude limit m_{lim} of the survey:

$$L_{\mathrm{lim}}(z) = \left(\frac{d_{\mathrm{lum}}(z)}{10\mathrm{pc}}\right)^2 10^{0.4(M_\odot - m_{\mathrm{lim}})} \tag{A.2}$$

with $M_\odot$ the absolute magnitude of the Sun in the same band as the limit magnitude m_{lim}, $d_{\mathrm{lum}}(z)$ the luminosity distance at the redshift z and L_{thresh} is the minimal luminosity of the sample.

We expect the galaxy environment to modulate galaxy properties such as their luminosity. Correcting for missing galaxies in all groups in the same way is in consequence not ideal. We can modulate the luminosity function with the group total mass by transforming the luminosity function into a conditional luminosity function. Since our estimations of the virial mass M of the group is good, we can use it as modulation parameter and so:

$$\phi(L) \rightarrow \phi(L|M) \tag{A.3}$$

THESIS

In this way, the previous missing fraction correction should be accurate enough for the correction. But for this to work, we must be able to determine properly this modulation with the halo mass.

A.1.2 Modulation of the luminosity function with the global environment

In galaxy groups, we separate galaxies into two classes: centrals and satellites. Centrals are expected to be the most massive galaxies in groups, and probably the most luminous. A consequence is that if we can't see the central galaxy, we can't see other galaxies in the group and the correction is not needed because we don't know how to correct for incompleteness. So for the correction, we simply need to constrain the distribution of luminosities in satellite galaxies. In practice, we have to choose a functional for this conditional luminosity function (CLF) which can be easily fitted and integrated to determine the correction factor in our group luminosities.

There are two kinds of luminosity functions widely used. The Schechter function can be written as:

$$\phi\left(L\right) = \phi^{*}\left(\frac{L}{L_{*}}\right)^{\alpha} \exp\left(-\frac{L}{L_{*}}\right) \tag{A.4}$$

where α characterizes the slope in log-space of the function, L_{*} the luminosity of turn-off and ϕ^{*} is the normalization of the function.

In studies of the galaxy sample from the SDSS survey as in Blanton et al. (2005), the LF has been well fitted by a double Schechter functional form which can be written:

$$\phi\left(L\right) = \left[\phi 1^{*}\left(\frac{L}{L_{*}}\right)^{\alpha_1} + \phi_2^{*}\left(\frac{L}{L_{*}}\right)^{\alpha_2}\right] \exp\left(-\frac{L}{L_{*}}\right) \tag{A.5}$$

This model allows for two galaxy populations in luminosity with different faint end slopes α_1 and α_2 but same high end luminosity cutoff L_{*}.

Now we assume that the CLF have the same form of Equation A.5. The dependence on the group mass M is done with the parameters of the double Schechter (DS). For example $\alpha_1 \rightarrow \alpha_1\left(M|\theta\right)$, where the functional form of this dependence is not given explicitly here, and θ is a set of parameters relative to the function used to describe the dependence with group mass. The number of parameters in θ can vary greatly, depending on the function used.

The form of this dependence cannot be determined in advance when we want to fit the CLF on the data. For example in the SDSS, we have to know in advance the properties of the groups in order to choose a dependence for the parameters of the DS with virial mass. For testing the viability of this method, we have to select a functional that describes correctly the modulation of the parameters with the group mass, and samples of galaxies that can give us these informations are present in outputs of semi-analytical models (SAM). To validate this method of correction for incompleteness, we test it on galaxy mock catalogues.

A.1.3 Parameter estimation

When working with distribution functions, it is common and better to use the maximum likelihood estimation. We define $p_i\left(L_i|\boldsymbol{\theta}\right)$ as the probability to get the luminosity L_i given the parameters $\boldsymbol{\theta}$, so it is a probability density function. To determine it, we calculate the number of galaxies in the sample which are between L_i and $L_i + \mathrm{d}L_i$, compared to the total number of points in the set:

$$p_i\left(L_i|\boldsymbol{\theta}\right)\mathrm{d}L_i\mathrm{d}V = \frac{\mathrm{d}^2 N_i}{N_{\mathrm{tot}}} \tag{A.9}$$

Remark 5

We consider a set of independent data $\{X\}$ drawn from distribution following the probability distribution function p, dependent of parameters θ. If we assume that observations are independent and identically distributed, the probability to obtain the given set of observations given the parameters θ is just the joint probability function of the observations. We define it as the likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}|X) = \prod_i p_i(X_i|\boldsymbol{\theta}) \tag{A.6}$$

To obtain the most probable parameters allowing the probability function p to correctly fit the data, we need to find the given set of parameters θ maximizing the likelihood function.

If we consider Bayesian statistics, the likelihood is defined as $p(X|\boldsymbol{\theta})$ and the Bayes's theorem gives that we need to maximize for the given set of data:

$$p(\boldsymbol{\theta}|X) = \frac{\prod_i p_i(X_i|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(X)} \tag{A.7}$$

where $p(\boldsymbol{\theta})$ is the prior distribution of the parameters and $p(X)$ is the probability to obtain the set of data. But we can see that *our* likelihood is in reality the posterior distribution which is proportional to the likelihood in the definition of Bayesian statistics, multiplied by a prior. If we take a constant for the prior (probability equal for each value of the parameter), since the probability to obtain the data is constant, using directly the likelihood defined in Equation A.6, the obtained parameters after the maximization are the same.

Numerically, it's more convenient to use the logarithm of the likelihood in order to prevent numerical problems when calculating the likelihood and the product in Equation A.6 becomes a sum. Often, numerical methods for optimization minimize of function instead of maximizing it so we put a minus sign in front of it:

$$-\log \mathcal{L}(\boldsymbol{\theta}|X) = -\sum_i \log\left(p_i(X_i|\boldsymbol{\theta})\right) \tag{A.8}$$

By definition of the CLF, which is the number of galaxies in the sample between L and $L+\mathrm{d}L$ at a given halo mass M, we can write:

$$\mathrm{d}^2 N = \phi(L|M)\,\mathrm{d}L\mathrm{d}V \tag{A.10}$$

So we can write:

$$p_i(L_i|\boldsymbol{\theta})\,\mathrm{d}L_i\mathrm{d}V = \frac{\phi(L_i|M)}{N_{\mathrm{tot}}}\mathrm{d}L_i\mathrm{d}V \tag{A.11}$$

and the total number of galaxies in this kind of halo (with mass M) is just:

$$N_{\mathrm{tot}} = \int_{\mathcal{V}}\int_{L_{\mathrm{thres}}}^{\infty} \phi(L|M)\,\mathrm{d}L\mathrm{d}V \tag{A.12}$$

where $\mathcal{V}$ is the volume of the galaxy sample, and L_{thres} is the minimal luminosity used for the sample. If a physical superior limit of luminosity exists, it should replace the infinity in the integration to not allow a probability to have luminosities superior to this limit.

In the case of the simple Schechter, the total number is:

$$N_{\text{tot}} = \Gamma\left(1+\alpha, \frac{L_{\text{thres}}}{L_*}\right)\phi^*\mathcal{V} \tag{A.13}$$

and for the double Schechter:

$$N_{\text{tot}} = \left[\Gamma\left(1+\alpha, \frac{L_{\text{thres}}}{L_*}\right) + \frac{\phi_2^*}{\phi_1^*}\Gamma\left(1+\alpha, \frac{L_{\text{thres}}}{L_*}\right)\right]\mathcal{V} \tag{A.14}$$

where $\Gamma(a,x) = \int_x^\infty \exp(-t)\, t^{a-1}\mathrm{d}t$ is the incomplete gamma function (see Appendix F for its computation with negative values of a). Then the computation of the density function p is easy in each case and we can do the minimization of the likelihood to estimate the best fit parameters $\hat{\boldsymbol{\theta}}$.

Remark 6

There are many ways of doing such a minimization. When the probability density isn't too complex, $\hat{\boldsymbol{\theta}}$ can be determined analytically. But in this case, with the DS, the incomplete gamma function prevents us to do it in this way. So we are constrained to use numerical methods in order to minimize the likelihood. Many algorithms exist to do this job like Powell's method, Newton-Raphson's method, etc..., but they share the same problem: when they find a minimum, we don't know if it is the global minimum or if it is a local minimum. The result depends on the initial starting point of the algorithm in the parameter space. Some other methods use Monte-Carlo methods to do a better exploration of this parameter space, allowing some "jumps" to other regions in order to see if there isn't a better minimum. An example of such an algorithm is the simulated annealing method which implement the cooling of a material, where the function to minimize becomes the energy of the system, and a fictive temperature T is introduced to allow some temperature jumps. But it is not always sure that we get the global minimum. Moreover, we can't easily determine errors on the estimation of the parameters, except using bootstraps or jackknife techniques which need many estimation of the parameters varying the sample which may be expensive in calculation time.

Another way is too estimate the posterior distribution of the parameter $\boldsymbol{\theta}$ by using the Markov Chains Monte Carlo method (MCMC). From it we can estimate the errors of choosing $\hat{\boldsymbol{\theta}}$ since we can estimate the distribution of the parameters.

We tested a large number of such methods for the minimization and it seems to be the Nelder-Mead (or simplex) algorithm that gives the better estimation of the best fit parameters $\hat{\boldsymbol{\theta}}$. ■

A.1.4 Tests on mock catalogues

There are two steps to determine the dependence of the luminosity function on the group mass. First, we have to determine what is the best functional form to fit this dependency which can be done on a complete sample of galaxies. Secondly, we can see if we can recover this parametrization and modulation with a flux-limited sample of this galaxies to know if the method works well when applied in a real survey.

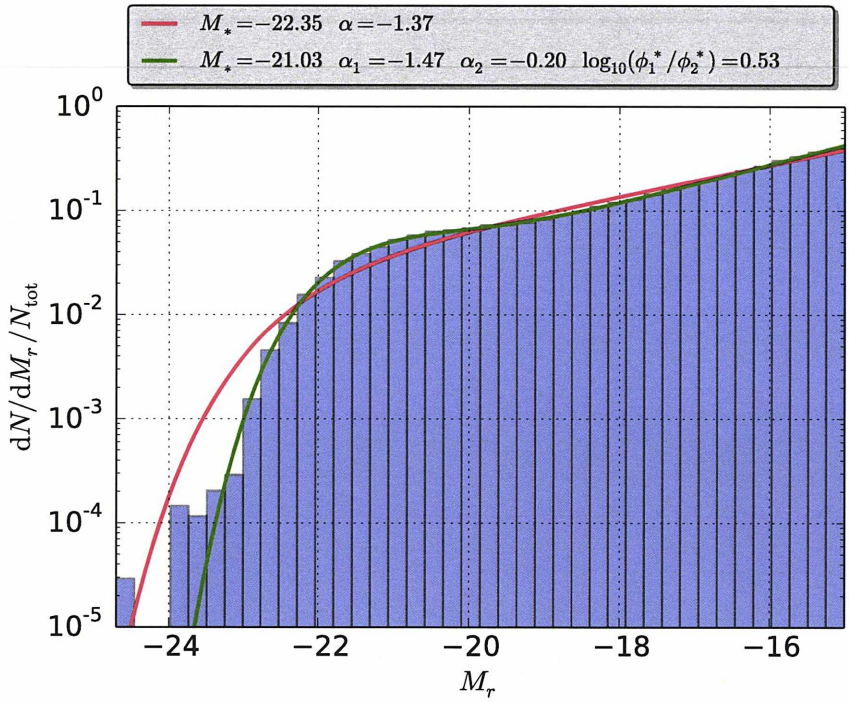


Figure A.1: Fits of the galaxy luminosity distribution of the GUO2010A catalogue in the r band. We fitted the simple Schechter function in green and the double Schechter function in red over the data in blue. Values in the legend correspond to the best fit parameters for each model, as described in the text.

A.1.4.1 Complete sample

We use a sample of galaxies, complete in luminosity, taken from the outputs of the SAM of Guo et al. (2011) applied on dark matter halos from the Millennium II run. We limit our sample of galaxies from this catalogue to galaxies with a luminosity such that the absolute magnitude in the r band is $M_r < -15$. For each galaxy, we have the virial mass of the halo (group) containing this galaxy (this is a cheat in comparison with running a group finder, but serves for illustrative purposes).

First, we determine what is the best model for the luminosity function. We tried to adjust a simple Schechter and a double Schechter. Results are shown on Figure A.1. The double Schechter fits better the data than the simple Schechter because we can constrain the two populations of galaxies. We see that there is a faint population with a high faint end slope and a brighter population with a lower slope. Differences with data for bright galaxies is due to the fact that the number of galaxies with $M_r < -24$ is very low, in some bins there is just one galaxy. As expected, both Schechter and double Schechter functionals are adequate models for the luminosity function.

We want to see the modulation of the parameters with the halo mass. We take galaxies in bins of logarithmic halo mass, and we compute the parameters that fit well the data in each, as previously. This modulation is represented in Figure A.2.

A.1.4.2 Flux limited sample

With a flux-limited sample, we just need to rewrite the normalization to take into account the total number of galaxies observed for a given redshift z. This can be proven by rewriting the probability density in terms of the cumulative distribution. The probability that a galaxy have a

Table A.1: Simple Schechter fit on the real space and on the redshift space mock catalogue.

	M_*	α
Real space	-22.34	-1.37
Redshift space	-22.40	-1.31

Table A.2: Double Schechter fit on the real space and on the redshift space mock catalogue.

	M_*	α_1	α_2	$\log_{10}(\phi_2^*/\phi_1^*)$
Real space	-21.02	-1.47	-0.19	0.53
Redshift space	-21.09	-1.43	-0.05	0.57

magnitude $\mathcal{M}$ superior (fainter) than M is given by:

$$P(\mathcal{M} > M|z) = \frac{\int_{-\infty}^{M} \phi(M') f(M') \, \mathrm{d}M'}{\int_{-\infty}^{\infty} \phi(M') f(M') \, \mathrm{d}M'} \tag{A.15}$$

where f is the completeness function:

$$f(M) = \begin{cases} 1, & M^{\text{bright}} \leq M \leq M^{\text{faint}} \\ 0, & \text{else} \end{cases} \tag{A.16}$$

Calculating the probability density is straightforward:

$$P(\mathcal{M} > M|z) = \int_{-\infty}^{M} p(M'|z) \, \mathrm{d}M' \tag{A.17}$$

and so:

$$p(M|z) = \frac{\partial P(\mathcal{M} > M|z)}{\partial M} \tag{A.18}$$

Finally:

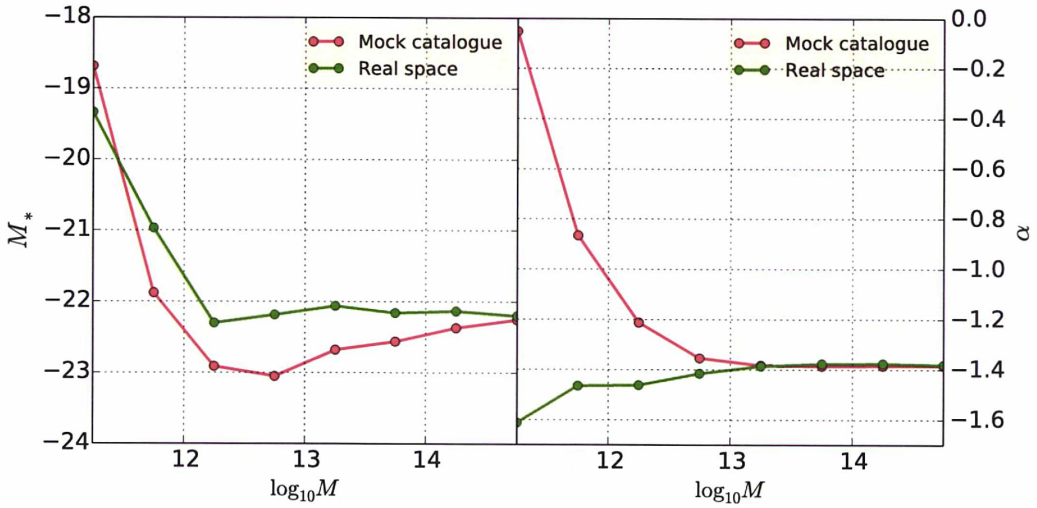
$$p(M_i|z_i) = \frac{\phi(M_i)}{\int_{M_{\text{bright}}(z_i)}^{M_{\text{faint}}(z_i)} \phi(M') \, \mathrm{d}M'} \tag{A.19}$$

and this defines the new likelihood in the case of a flux limited sample.

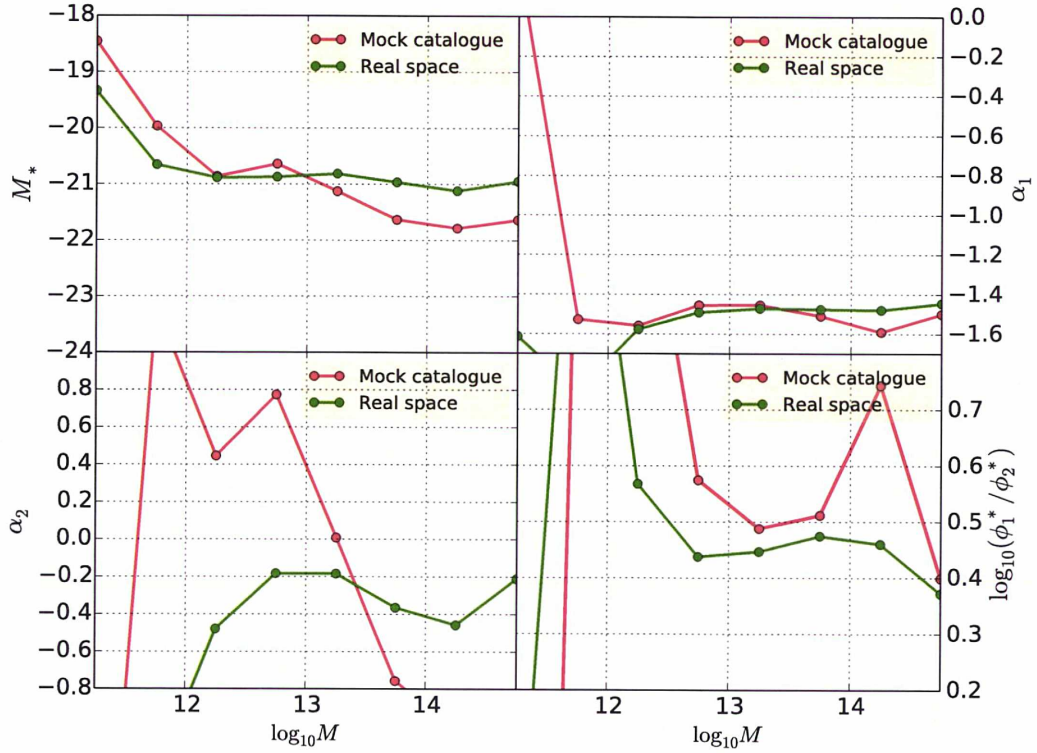
The result of the application of the MLE method on our mock redshift space catalogue (see Chapter 3) is shown on Figure A.2.

The parameters are more or less well recovered in flux-limited space. The bright population is poorly recovered since its faint end slope is very badly estimated and so is the ratio between the two populations too. In the simple Schechter fit, the discrepancies with the real space appear for low mass halos. Such groups are formed of faint galaxies disappearing at high redshift because of the magnitude limit. Thus, there are fewer galaxies for the statistics of low mass groups and the fit is poor. Moreover, the random filtering of boxes in the mock creation increases the cosmic variance of the data for low mass groups. The ratio in number between the two populations is roughly of 5%, and the statistics are always poor for each bin of halo mass.

To verify this assumption, and not to incriminate a bad implementation of the MLE for flux-limited samples, we applied the method on perfect samples of galaxies. We generated an Universe with a given luminosity function and applied the flux limit to galaxies in this region. We



(a) For a Schechter distribution



(b) For a double Schechter distribution

Figure A.2: Modulation of the parameters of both Schechter and double Schechter luminosity distributions with the halo mass obtained from the redshift space mock catalog (in red) and from the real space mock data (in green) from Guo et al. (2011).

TESIS

Remark 7

Generating galaxies following a given luminosity function is done by the inverse transform sampling method. Let suppose that F is a cumulative distribution function. This function is monotonic. Let U be an random variable following a uniform distribution over $[0, 1]$. If we define $Y = F^{-1}(U)$, this random variable follows the distribution of F. By definition, the cumulative distribution function of Y is $p(Y \leq x) = p(F^{-1}(U) \leq x)$. Since the function is monotonic, $p(F^{-1}(U) \leq x) = p(U \leq F(x))$. The last expression is the cumulative distribution function for uniform distribution applied to the variable $F(x)$, which is directly equal to $F(x)$.

The cumulative distribution function for the simple Schechter is:

$$F(L) = \frac{\Gamma\left(\alpha+1, \frac{L_{\min}}{L_*}\right) - \Gamma\left(\alpha+1, \frac{L}{L_*}\right)}{\Gamma\left(\alpha+1, \frac{L_{\min}}{L_*}\right) - \Gamma\left(\alpha+1, \frac{L_{\max}}{L_*}\right)} \tag{A.20}$$

and for the double Schechter:

$$F(L) = \frac{\gamma_{\alpha_1}[L] + \frac{\phi_2^*}{\phi_1^*}\gamma_{\alpha_2}[L]}{\gamma_{\alpha_1}[L_{\max}] + \frac{\phi_2^*}{\phi_1^*}\gamma_{\alpha_2}[L_{\max}]} \tag{A.21}$$

with:

$$\gamma_\alpha[X] = \Gamma\left(\alpha+1, \frac{L_{\min}}{L_*}\right) - \Gamma\left(\alpha+1, \frac{X}{L_*}\right) \tag{A.22}$$

Clearly, we cannot invert such cumulative distribution functions analytically. By interpolating them in the range of luminosities to generate, we can do a numerical inversion and obtain the precious random variables following the Schechter distributions. This is fast and precise enough.

The double Schechter can also be generated by two populations of simple Schechter functions. If N_i is the number of galaxies following the distribution with parameters α_i and ϕ_i^*, the ration between the two population is:

$$\frac{N_2}{N_1} = \frac{\phi_2^*}{\phi_1^*} \times \frac{\gamma_{\alpha_2}[L_{\max}]}{\gamma_{\alpha_1}[L_{\max}]} \tag{A.23}$$

with $N_{\text{tot}} = N_1 + N_2$. But its easier to take the cumulative distribution function of the double Schechter, otherwise we need to shuffle the resulting two single Schechter populations. ■

are able to recover the simple Schechter parameters used to generate the distribution in the flux limited sample. Using a double Schechter distribution for the generation of galaxy luminosities, its parameters are more difficult to recover, essentially for the bright population whose number is low relative to the faint one.

The results are also dependent of the initial guess chosen for the minimization. This is a problem if we want to iteratively correct for missing galaxies in MAGGIE, since we need it to

be robust against this choice. Indeed, the group population is varying in the iterative process and the modulation of galaxy properties with groups will evolve, as will our assumptions on the luminosity function parameters.

Since it can be very difficult to correct our groups in a flux-limited sample, we will restrict our analysis to doubly complete samples, where corrections for luminosity incompleteness are not required.

A.2 Red and blue galaxies

Galaxies form a bimodal distribution, mainly separated into red and blue ones (representing low and high SSFR). From previous studies, their distributions inside galaxy groups are not the same. Incorporating this segregation into MAGGIE should improve the group selection and our measures of the environment.

Incorporating red versus blue galaxies can be done inside the membership probability. We compute a different probability if the galaxy is red or blue, by adjusting the models according to the galaxy color. Such models are updated in the iterative process, in order to get a relative independence of our results to the adopted models.

Taking again the computation of the probability, if we know the fraction of blue or red galaxies at a given radius to the group center, we can multiply the density profile by this fraction, giving us the projected phase space density of blue or red galaxies in halos. We have:

$$g^i_{\rm halo}\left(R, v_z\right) = \int_R^{r_v} f_i\left(r\right) \nu\left(r\right) \frac{r}{\sqrt{r^2 - R^2}} h\left(v_z | R, r\right) {\rm d}r \tag{A.24}$$

where $h\left(v_z|R,r\right)$ is the line of sight velocity distribution and $f_i\left(r\right)$ is the fraction of i galaxies, with $i \in \{\text{red, blue}\}$. The fraction is a model whose parameters must be fitted to the data for each iteration with the set of extracted groups. Since its a distribution function, it implies the use of MLE with numerical computation of the integral, and a double integral for the normalization of the density since:

$$g^i_{\rm halo}\left(R, v_z\right) = \frac{{\rm d}^2 N_i}{2\pi R {\rm d}R {\rm d}v_z} \tag{A.25}$$

Supposing this computation can be easily done, there are still two problems.

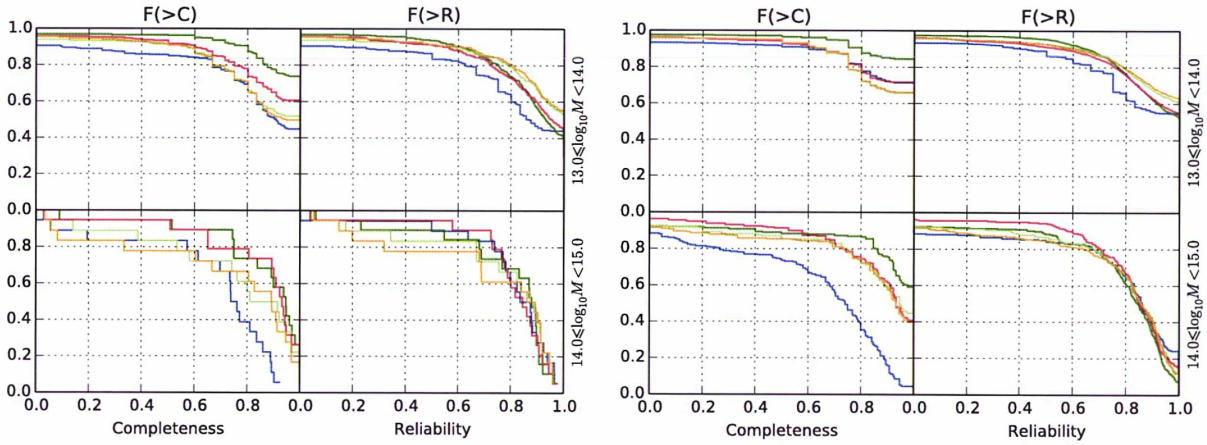
- The NFW profile is dependent of the concentration. Several studies shows that the concentration of blue and red galaxies inside clusters are different. For example, Guo et al. (2012) find that around red central SDSS galaxies, the red satellites have a concentration $c = 3.2 \pm 0.4$ while the blue satellites have $c = 1.7 \pm 0.2$, which is significantly lower.

- The probability needs to be normalized with the projected phase space density of interlopers. But we have no idea of their distribution when red and blue galaxies are separated. This needs to be extracted from mock catalogues, leading to densities not universal and dependent of the semi-analytical code used, the chosen cut-off in magnitude for the complete sample used. In a first approximation, the fraction of interlopers that are, say blue, should be independent of projected radius R and line-of-sight velocity v_z.

A.3 Abundance matching

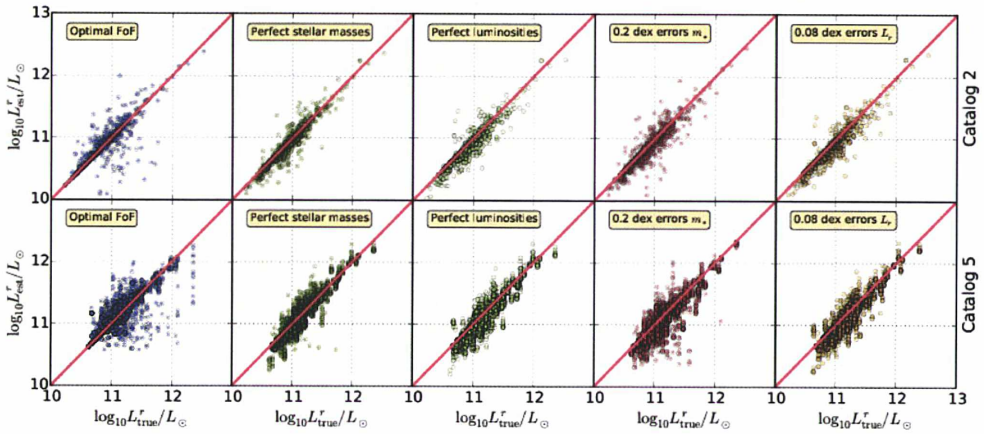
In MAGGIE, the virial mass estimation by abundance matching is performed between the central stellar mass or luminosity and the halo mass function. But the relation between the stellar mass and the halo mass is saturated for high masses, making the relation relatively flat, and so the

THESIS

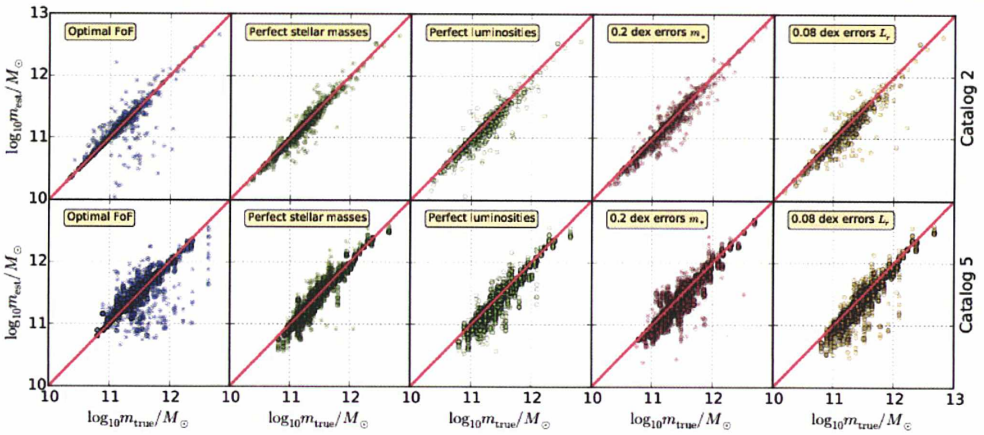


(a) Catalogue 2 (b) Catalogue 5

Figure A.3: Same as Figure 5.4, but with primary groups defined as the most massive in halo mass of groups linked to a real group.



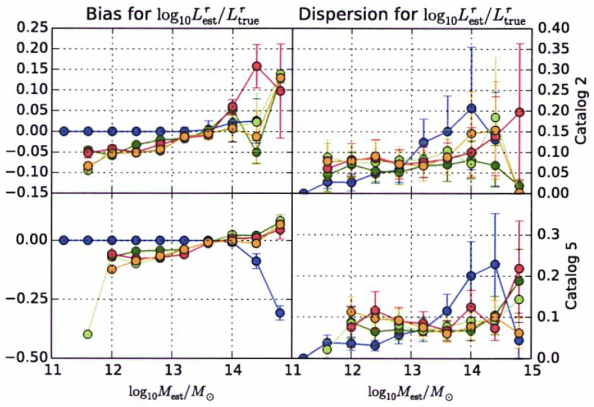
(a) Comparison for group luminosities



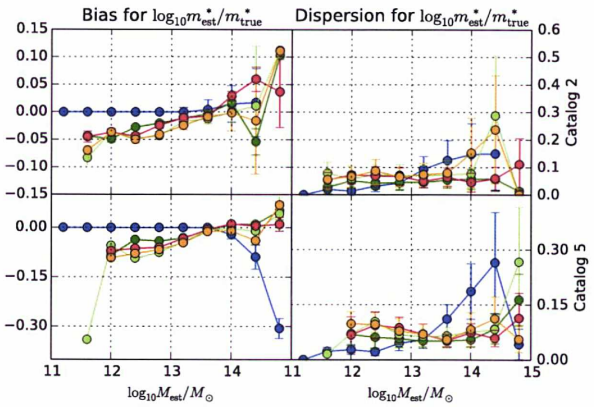
(b) Comparison for group stellar masses

Figure A.4: Same as Figure 5.7, but with primary groups defined as the most massive in halo mass of groups linked to a real group.

THESIS



(a) Bias and dispersion for luminosities



(b) Bias and dispersion for stellar masses

Figure A.5: Same as Figure 5.8, but with primary groups defined as the most massive in halo mass of groups linked to a real group.

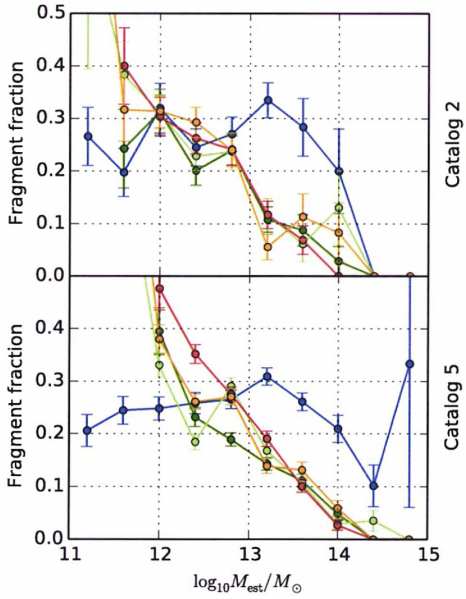
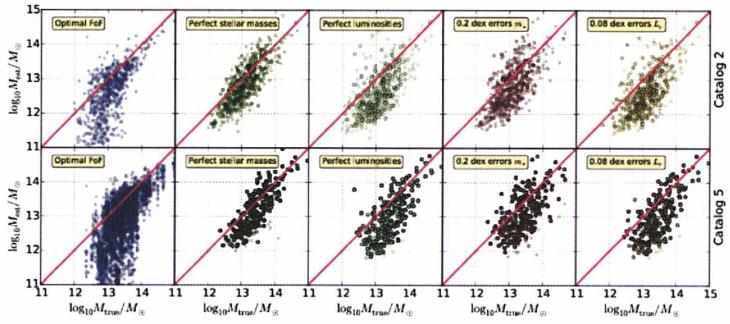
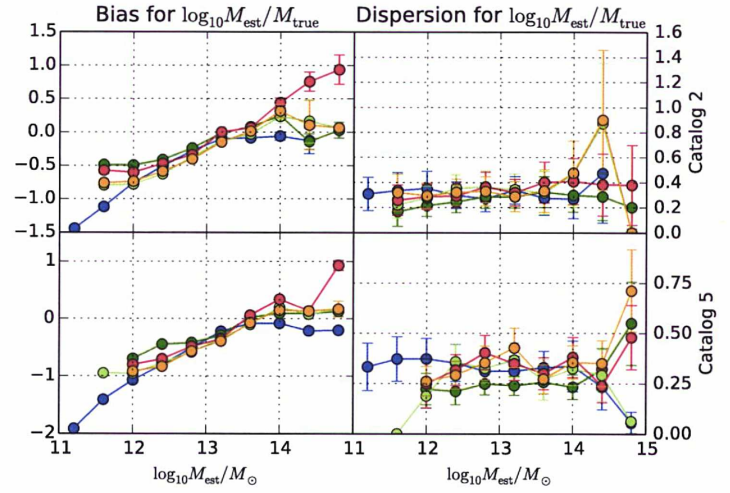


Figure A.6: Same as Figure 5.5, but with primary groups defined as the most massive in halo mass of groups linked to a real group.



(a) Comparison of virial masses



(b) Bias and dispersion for virial masses

Figure A.7: Same as Figure 5.6, but with primary groups defined as the most massive in halo mass of groups linked to a real group.

Appendix B.

Density profiles

B.1 Introduction

In this chapter, we provide details on the computation of the density profiles and their derived quantities. We define here the different normalizations used along the thesis for some popular density profiles.

B.1.1 Definitions

The number of galaxies in a sphere of radius r with a density profile in number $\nu(r)$ is the case of a spherical symmetry:

$$N(r) = \int_0^r 4\pi r'^2 \nu(r') \mathrm{d}r' \tag{B.1}$$

To start, we define some dimensionless functions to facilitate the computations.

$$\begin{aligned} N(r) &= N(a)\, \widetilde{N}(r/a) \\ \nu(r) &= \frac{N(a)}{4\pi a^3} \widetilde{\nu}(r/a) \end{aligned} \tag{B.2}$$

with a the radius at which the logarithmic slope of the density profile is equal to -2. We also define the same relations for a virial normalization.

$$\begin{aligned} N(r) &= N_v\, \widehat{N}(r/r_{\rm vir}) \\ \nu(r) &= \frac{N_v}{4\pi r_{\rm vir}^3} \widehat{\nu}(r/r_{\rm vir}) \end{aligned} \tag{B.3}$$

We also define the concentration c as the ratio between the virial radius $r_{\rm vir}$ and the radius a, i.e. $c = r_{\rm vir}/a$. We define $r_{\rm vir} = r_{200}$ for simplicity. We should note that the 'slope' normalization and the virial normalization can be linked together simply by setting $c = 1$ on the definition of the virial normalization, in order words the normalization radius that is $r_{\rm vir}$ becomes the slope radius a.

B.2 Density profiles

B.2.1 Navarro et al. (1996)

The NFW density profile is:

$$\nu(r) = \frac{\nu_0}{r(r+a)^2} \tag{B.4}$$

THESIS

with ν_0 a constant density.

We can write by integrating previous relations with $\int_0^1 x^2\widetilde{\nu}(x)\,\mathrm{d}x = \int_0^1 x^2\widehat{\nu}(x)\,\mathrm{d}x$ and searching for the constant ν_0:

$$\widehat{\nu}(x) = \frac{1}{\ln 2 - 1/2} \quad \frac{1}{x(1+x)^2} \tag{B.5}$$

$$\widetilde{N}(x) = \frac{1}{\ln 2 - 1/2} \quad \left(\ln(1+x) - \frac{x}{x+1}\right) \tag{B.6}$$

$$\widehat{\nu}(x) = \frac{1}{\ln(1+c) - c/(1+c)} \quad \frac{1}{x(1/c+x)^2} \tag{B.7}$$

$$\widehat{N}(x) = \frac{1}{\ln(1+c) - c/(1+c)} \quad \left(\ln(1+xc) - \frac{xc}{xc+1}\right) \tag{B.8}$$

B.2.2 Einasto

For an Einasto density profile:

$$\nu(r) = \nu_0 \exp\left[-\left(\frac{r}{b}\right)^{1/m}\right] \tag{B.9}$$

Writing the definition of the a radius with this density profile, we have:

$$\left(\frac{1}{b}\right)^{1/m} = 2m\left(\frac{1}{a}\right)^{1/m} \tag{B.10}$$

leading to the following normalizations:

$$\widetilde{\nu}(x) = \frac{(2m)^{3m}}{m\gamma(3m, 2m)} \exp\left(-2mx^{1/m}\right) \tag{B.11}$$

$$\widetilde{N}(x) = \frac{\gamma\left(3m, 2mx^{1/m}\right)}{\gamma(3m, 2m)} \tag{B.12}$$

$$\widehat{\nu}(x) = \frac{(2m)^{3m}}{m\gamma\left(3m, 2mc^{1/m}\right)} \exp\left(-2m(xc)^{1/m}\right) \tag{B.13}$$

$$\widehat{N}(x) = \frac{\gamma\left(3m, 2m(xc)^{1/m}\right)}{\gamma\left(3m, 2mc^{1/m}\right)} \tag{B.14}$$

B.2.3 Generalized NFW

If any previous density profiles isn't sufficient to describe the distribution of dark matter particles or galaxies inside the halos, a solution is possibly to fit a generalized NFW profile, whose the density is:

$$\nu(r) = \frac{\nu_0}{r^\alpha (r+a)^{\beta-\alpha}} \tag{B.15}$$

In this case:

$$\widetilde{N}(x) = \frac{\mathcal{B}_{-x}(3-\alpha, 1+\alpha-\beta)}{\mathcal{B}_{-1}(3-\alpha, 1+\alpha-\beta)} \tag{B.16}$$

therefore:

$$\widetilde{\nu}(x) = \frac{1}{(-1)^{\alpha+1}\mathcal{B}_{-1}(3-\alpha, 1+\alpha-\beta)} \quad \frac{1}{x^{\alpha}(1+x)^{\beta-\alpha}} \tag{B.17}$$

For the virial normalization:

$$\widehat{N}(x) = \frac{\mathcal{B}_{-xc}(3-\alpha, 1+\alpha-\beta)}{\mathcal{B}_{-c}(3-\alpha, 1+\alpha-\beta)} \tag{B.18}$$

$$\widehat{\nu}(x) = \frac{1}{(-1)^{\alpha+1}\mathcal{B}_{-c}(3-\alpha, 1+\alpha-\beta)} \quad \frac{1}{(xc)^{\alpha}(1+xc)^{\beta-\alpha}} \tag{B.19}$$

where $\mathcal{B}$ is the function defined as:

$$\mathcal{B}(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 t^{a-1}(1+t)^{b-1}\mathrm{d}t \tag{B.20}$$

and its incomplete version is:

$$\mathcal{B}_z(a,b) = \int_0^z t^{a-1}(1+t)^{b-1}\mathrm{d}t \tag{B.21}$$

B.3 Radial velocity dispersion

Galaxies in groups (and their associated dark matter halos) are assumed to be a system of particles only submitted to the gravitation. Neglecting mergers and other physical processes inside galaxy groups, the number of galaxies doesn't evolve in phase space, and the distribution function is constant along the evolution of the system. In this case, we can use the collisionless Boltzmann equation to extract dynamical properties of galaxy groups.

The Jeans equation is the first velocity momentum of the Boltzmann equation. In a spherical symmetry, assuming stationarity, the Jeans equation is:

$$\frac{\partial\left[\nu(r)\sigma_r^2(r)\right]}{\partial r} + \frac{2\beta(r)}{r}\left[\nu(r)\sigma_r^2(r)\right] = -\nu(r)\frac{GM(r)}{r^2} \tag{B.22}$$

where $\beta(r)$ is the radial profile of velocity anisotropy $\beta = 1 - \sigma_\theta^2/\sigma_r^2$.

We can compute the radial velocity dispersion using Equation B.22 for a spherical system at equilibrium. The solution to this equation is given by:

$$\nu(r)\sigma_r^2(r) = \int_r^\infty K_r(r,s)\nu(s)\frac{GM(s)}{s^2}\mathrm{d}s \tag{B.23}$$

with $K_r(r,s)$ the kernel of the integral defined as:

$$K_r(r,s) = \exp\left[2\int_r^s \beta(r)\frac{\mathrm{d}t}{t}\right] \tag{B.24}$$

There are two ways of normalizing the radial velocity dispersion according to the normalization used for the density and mass profiles. We show it for the virial normalization for illustration:

$$\widehat{\sigma}_r^2(x) = \frac{1}{\widehat{\nu}(x)}\int_x^\infty K_r(x,s)\widehat{\nu}(s)\frac{\widehat{M}(s)}{s^2}\mathrm{d}s \tag{B.25}$$

with:

$$\sigma_r^2(r) = \frac{GM_{\rm vir}}{r_{\rm vir}} \widehat{\sigma}_r^2(r/r_{\rm vir}) \tag{B.26}$$

We are interested only in the NFW profile in the thesis, since it is accurate enough to adjust the model. If we want an analytical form for $\sigma_r(r)$, we need to choose a model for the anisotropy profile $\beta(r)$. We provide here some expressions of the radial velocity dispersion, assuming the NFW density profile, for a useful anisotropy model.

B.3.1 Mamon & Łokas (2005)

This model is of the form:

$$\beta(r) = \frac{1}{2}\frac{r}{r+\mathfrak{b}} \tag{B.27}$$

where $\mathfrak{b}$ is a characteristic radius of the model. Introducing this expression in Equation B.25, we obtain:

$$\begin{aligned}\widehat{\sigma}_r^2(x) = &\frac{c/[6\,y(y+b)]}{\ln(c+1)-c/(c+1)} \\ &\times \Big\{6(3b-2)(y+1)^2y^2{\rm Li}_2(-y) - 3by^2(y+1)^2\ln y + 3(3b-2)y^2(y+1)^2\ln^2(y+1) \\ &+ 3(y+1)\left[b\left(y^3-5y^2-3y+1\right)+2y(2y+1)\right]\ln(y+1) \\ &+\pi^2(3b-2)y^4 + \left(6\pi^2 b - 21b - 4\pi^2 + 12\right)y^3 + \left[3\left(\pi^2-9\right)b - 2\pi^2 + 15\right]y^2 - 3by\Big\}\end{aligned} \tag{B.28}$$

with $b = c\mathfrak{b}/r_{\rm vir}$, $x = r/r_{\rm vir}$ and $y = cx$.

B.4 Line of sight velocity variance

We will compute in this section the line of sight velocity dispersion of galaxies in a general spherical density profile, and then compute it specifically for an NFW profile. This is useful to make cuts at $\pm\kappa\sigma_{\rm LOS}(R)$ in the *pps*.

By definition, the variance is the mean of the squared quantity. We use a general density profile which is invariant under rotations $\nu(r)$. In our case, we make this mean on the line of sight, so:

$$\sigma_{\rm LOS}^2(R) = \frac{\int_{-\infty}^{\infty} v_{\rm LOS}^2 \nu(r)\,{\rm d}z}{\int_{-\infty}^{\infty}\nu(r)\,{\rm d}z} \tag{B.29}$$

But in the group, $r^2 = R^2 + z^2$ so:

$$\sigma_{\rm LOS}^2(R) = \frac{2\int_R^{r_{\rm max}} v_{\rm LOS}^2 \dfrac{\nu(r)r}{\sqrt{r^2-R^2}}{\rm d}r}{2\int_R^{r_{\rm max}} \dfrac{\nu(r)r}{\sqrt{r^2-R^2}}{\rm d}r} \tag{B.30}$$

The denominator is by definition the projected density surface along the line of sight and we denote it

$$\Sigma(R) = 2\int_R^{r_{\rm max}} \frac{\nu(r)r}{\sqrt{r^2-R^2}}{\rm d}r \tag{B.31}$$

Normally the integration is for $r_{\rm max}\to\infty$ but in our case we want to restrict to a limited region in the group (to the virial sphere precisely).

In the same coordinate system as previously, the line of sight velocity can be expressed in spherical coordinates as:

$$v_{\mathrm{LOS}} = v_r \cos\theta - v_\theta \sin\theta \tag{B.32}$$

We suppose that we are at the equilibrium and so that there is no flow in the group in consequence we can neglect means of velocities. In terms of velocity variance we have now:

$$\Sigma(r)\sigma^2_{\mathrm{LOS}}(R) = 2\int_R^{r_{\max}} \left(\sigma_r^2(r)\cos^2\theta + \sigma_\theta^2\sin^2\theta\right)\frac{\nu(r)r}{\sqrt{r^2-R^2}}\mathrm{d}r \tag{B.33}$$

If we want to use the anisotropy parameter $\beta(r) = 1 - \sigma_\theta^2(r)/\sigma_r^2(r)$ in case of sphericity, we can write:

$$\Sigma(r)\sigma^2_{\mathrm{LOS}}(R) = 2\int_R^{r_{\max}} \left(1-\beta(r)\frac{R^2}{r^2}\right)\frac{\nu(r)\sigma_r^2(r)r}{\sqrt{r^2-R^2}}\mathrm{d}r \tag{B.34}$$

We can compute the radial velocity dispersion using the Jeans equation for a spherical system at equilibrium.

B.4.1 Mamon & Łokas (2005) anisotropy

With the decomposition of the integral over the domain of integration, we can write:

$$\begin{aligned}\Sigma(R){\sigma_{\mathrm{LOS}}}^2(R) =& 2\int_R^{r_v}\frac{(s+a)}{s^2}\nu(s)GM(s)\mathrm{d}s \\ &\times\left(\int_R^s\left(\frac{r}{r+a}-\frac{1}{2}\left(\frac{R}{r+a}\right)^2\right)\frac{1}{\sqrt{r^2-R^2}}\mathrm{d}r\right) \\ &+2\int_{r_v}^{\infty}\frac{(s+a)}{s^2}\nu(s)GM(s)\mathrm{d}s \\ &\times\left(\int_R^{r_v}\left(\frac{r}{r+a}-\frac{1}{2}\left(\frac{R}{r+a}\right)^2\right)\frac{1}{\sqrt{r^2-R^2}}\mathrm{d}r\right)\end{aligned} \tag{B.35}$$

where we are setting $r_{\max}$ to r_v. So now we can write:

$$\begin{aligned}{\sigma_{\mathrm{LOS}}}^2(R) =& {v_v}^2\frac{c/2}{\widetilde{M}(c)\widetilde{\Sigma}(R/a,c)} \\ &\times\left(\int_{R/a}^{c}K\left(x\frac{a}{R},\frac{a}{R}\right)\widetilde{\nu}(x)\frac{\widetilde{M}(x)}{x}\mathrm{d}x + I\left(c\frac{a}{R},\frac{a}{R}\right)J(c)\right)\end{aligned} \tag{B.36}$$

$$I(u,u_a) = \begin{cases} -u_a\mathrm{sign}(u_a-1)\dfrac{{u_a}^2-1/2}{|{u_a}^2-1|^{3/2}}C^{-1}\left(\dfrac{1+uu_a}{u+u_a}\right) & \\ \qquad +\mathrm{acosh}u + \dfrac{1/2}{u_a+u}\dfrac{\sqrt{u^2-1}}{{u_a}^2-1}, & u_a \neq 1 \\ \mathrm{acosh}u - \sqrt{\dfrac{u-1}{u+1}}\left(\dfrac{8+7u}{6(1+u)}\right), & u_a = 1 \end{cases} \tag{B.37}$$

with:

$$K(u,u_a) = \left(1+\frac{u_a}{u}\right)I(u,u_a) \tag{B.38}$$

and:

$$C^{-1}(X) = \begin{cases} \text{acosh}X & u_a > 1 \\ \text{acos}X & u_a < 1 \end{cases} \tag{B.39}$$

We also have an other integral:

$$J(y) = \int_y^\infty \frac{x+1}{x^2}\widetilde{\nu}(x)\widetilde{M}(x)\mathrm{d}x \tag{B.40}$$

In the case of an NFW profile, this can be expressed in an analytical way:

$$\begin{aligned} J(y) =& \frac{2}{3y^2(1+y)(\ln 4-1)^2}\left(y\left(-3+y\left(-9+\pi^2\left(1+y\right)\right)\right)\right. \\ &+3y^3\ln\left(1+\frac{1}{y}\right)+3\ln\left(1+y\right)\left(1-y+y^2\left(1+y\right)\ln\left(1+y\right)\right) \\ &\left.-3y^2\ln\left(y\left(1+y\right)\right)+6y^2\left(1+y\right)\mathrm{Li}_2\left(-y\right)\right) \end{aligned}$$

where the dilogarithm function is defined in our case as:

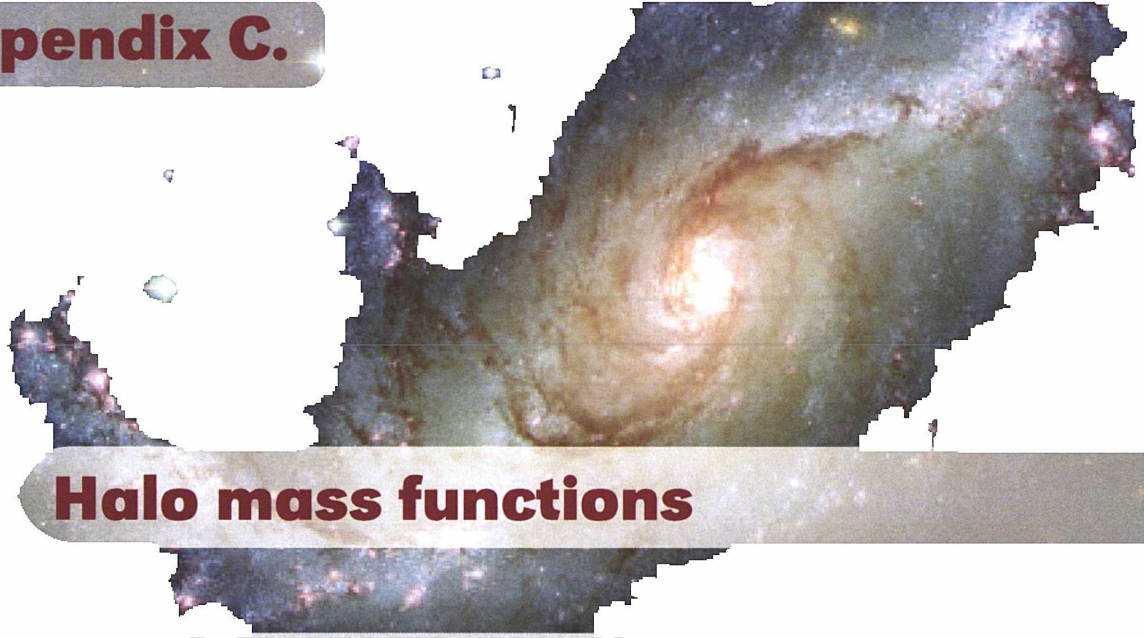
$$\mathrm{Li}_2\left(z\right) = -\int_0^1 \frac{\ln\left(1-zt\right)}{z}\mathrm{d}t \tag{B.41}$$

For the NFW model, Mamon et al. (2010) provide the expression of $\widetilde{\Sigma}$:

$$\begin{aligned} \widetilde{\Sigma}(X,c) &= \frac{1}{2\ln 2-1}\int_X^c \frac{\mathrm{d}x}{(1+x)^2\sqrt{x^2-X^2}} \\ &= \frac{1}{2\ln 2-1}\begin{cases} \dfrac{1}{(1-X^2)^{3/2}}\cosh^{-1}\left[\dfrac{c+X^2}{(c+1)X}\right]-\dfrac{1}{(c+1)}\dfrac{\sqrt{c^2-X^2}}{1-X^2} & \text{if } 0<X<1 \\ \dfrac{\sqrt{c^2-1}(c+2)}{3(c+1)^2} & \text{if } X=1<c \\ \dfrac{1}{(c+1)}\dfrac{\sqrt{c^2-X^2}}{X^2-1}-\dfrac{1}{(X^2-1)^{3/2}}\cos^{-1}\left[\dfrac{c+X^2}{(c+1)X}\right] & \text{if } 1<X<c \\ 0 & \text{if } X=0 \text{ or } X>c \end{cases} \end{aligned} \tag{B.42}$$

THESIS

Appendix C.



Halo mass functions

Contents

C.1 Theory

C.1.1 Definition

By definition, the halo mass function by unit of comoving volume is the number of halos with mass M between M and $M + \mathrm{d}M$. If N is the number of halos in comoving volume V, the halo mass function $\phi(M)$ can be written:

$$\phi(M) = \frac{\mathrm{d}^2 N}{\mathrm{d}M \mathrm{d}V} = \frac{\mathrm{d}n}{\mathrm{d}M} \tag{C.1}$$

In this case, n can be the comoving density of halos, or the cumulative distribution function of the density. In the latter case, we have:

$$n(M, z) = \int_0^M \phi(M, z)\, \mathrm{d}M \tag{C.2}$$

and so:

$$\frac{\mathrm{d}n}{\mathrm{d}M} = \frac{\mathrm{d}}{\mathrm{d}M} \int_0^M \phi(M, z)\, \mathrm{d}M = \frac{\mathrm{d}}{\mathrm{d}M} (\Phi(M, z) - \Phi(0, z)) = \phi(M, z) \tag{C.3}$$

where Φ is a primitive of ϕ.

According to Press & Schechter (1974b), the halo mass function is:

$$\phi(M, z) = \frac{\rho_m(z)}{M^2} \frac{\mathrm{d} \ln \sigma^{-1}}{\mathrm{d} \ln M} \nu \exp\left(-\nu^2/2\right) \tag{C.4}$$

where $\nu = \delta_c(z)/\sigma(M)$, δ_c a threshold parameter, ρ_m the mean density of the Universe and σ the standard deviation of density fluctuations.

THESIS

C.1.2 In practice

Jenkins et al. (2001) found a general way to fit halo mass functions from different cosmological simulations done with different cosmologies, allowing an easy comparison between the different redshifts and cosmologies. The halo mass function is related to the standard deviation of density fluctuations σ, which is a function of halo mass. A function $f(\sigma)$ has been introduced for that, which is the fraction of matter that is inside a halo of mass M by units of $\ln \sigma^{-1}$. So:

$$f(\sigma) = \frac{\mathrm{d}\rho/\rho_m(z)}{\mathrm{d}\ln\sigma^{-1}} = \frac{M}{\rho_m(z)}\frac{\mathrm{d}n}{\mathrm{d}\ln\sigma^{-1}} \tag{C.5}$$

and the halo mass function is:

$$\phi(M,z) = \frac{\mathrm{d}\ln\sigma^{-1}}{\mathrm{d}M}\frac{\rho_m(z)}{M}f(\sigma) = \frac{\rho_m(z)}{M^2}\left|\frac{\mathrm{d}\ln\sigma}{\mathrm{d}\ln M}\right| f(\sigma) \tag{C.6}$$

where the computation of σ involves the power spectrum $P(k)$ and the Fourier space representation of the real space top hat filter $\tilde{W}(k)$:

$$\sigma^2(M) = \frac{1}{2\pi^2}\int_0^\infty P(k)\,\tilde{W}^2(kR)\,k^2\mathrm{d}k \tag{C.7}$$

and its logarithmic derivative is:

$$\frac{\mathrm{d}\ln\sigma}{\mathrm{d}\ln M} = \frac{R}{12\pi^2\sigma^2}\int_0^\infty \frac{\mathrm{d}W^2(kR)}{\mathrm{d}(kR)}k^3P(k)\,\mathrm{d}k \tag{C.8}$$

By definition, if we assume that all the matter is contained in dark matter halos, summing over all the possible variance gives us the total mass, so:

$$\int_{-\infty}^{\infty} f(\sigma)\,\mathrm{d}\ln\sigma^{-1} = 1 \tag{C.9}$$

The variance of density fluctuations follows the evolution of the linear perturbations and must grow with them, so we need to multiply it by the growth rate to extend the expression to other redshifts.

C.1.3 Window function

By definition, σ is the variance of mass within a sphere of radius R containing mass M with the mean density of the Universe. For this, a top-hat filter is used in real space corresponding to the sphere of radius R. Its expression in the Fourier space is:

$$W(kR) = \frac{3\left[\sin(kR) - kR\cos(kR)\right]}{(kR)^3} \tag{C.10}$$

and we can explicitly write the derivative:

$$\frac{\mathrm{d}W^2(x)}{\mathrm{d}x} = \left[\sin x - x\cos x\right] \times \left[\sin x\left(1 - \frac{3}{x^3}\right) + 3\frac{\cos x}{x}\right] \tag{C.11}$$

THESIS

C.1.4 Power spectrum

In the context of the theory of small perturbations, over-densities, expressed as $\delta(\mathbf{x}) = (\rho(\mathbf{x}) - \overline{\rho})/\overline{\rho}$ grow linearly if they are small (i.e. $\delta \ll 1$). The power spectrum is the second moment of the probability distribution function of the density perturbation field expressed in Fourier space:

$$P(k) \propto \left\langle |\delta_{\mathbf{k}}|^2 \right\rangle \tag{C.12}$$

Inflation models predicts a power spectrum of the form:

$$P(k) \propto k^n \tag{C.13}$$

where n is the spectral index, close to 1. Different kinds of matter contribute to the power spectrum and in general diverge from this simple model. The transfer function $T(k)$ accounts for this, as a correction to the inflation model:

$$P(k) \propto k^n T^2(k) \tag{C.14}$$

The transfer function is sensitive to the model of dark matter matter and the density of baryons through Ω_b. The transfer function is difficult to compute precisely and thus we depend on the CAMB program (Lewis et al., 2000).

The normalization of the power spectrum is not predictable by the theory, hence must be set by confrontation to the observations. For this, the variance of the density field of the fluctuations within the smoothing window function of radius $R = 8h^{-1}$Mpc is used, and by comparison with the value obtained from observations with the galaxy distribution or other method, the power spectrum is fully determined.

If we want to compute the power spectrum at different epochs, we must apply the growth factor to the power spectrum, under the assumptions of linear evolution of fluctuations. In this case $\delta(\mathbf{x}, a) = D(a)\,\delta_i(\mathbf{x})$, where $D(a)$ is the growth factor and $a = 1/(1+z)$ the scale factor. For the growing mode of perturbations:

$$D(z) = \frac{5\Omega_m}{2} E(z) \int_z^\infty \frac{(1+z')}{E^3(z')} \mathrm{d}z' \tag{C.15}$$

with:

$$E(z) = \frac{H(z)}{H_0} = \sqrt{\Omega_m (1+z)^3 + \Omega_\Lambda} \tag{C.16}$$

for a flat Universe (see Carroll et al. (1992); Hogg (1999)). The growth factor to apply to the power spectrum is:

$$d(z) = \frac{D(z)}{D(z=0)} \tag{C.17}$$

The power spectrum and the variance of the standard deviation of the perturbations evolve in the following way:

$$\begin{aligned} P(k, z) &= d^2(z)\, P(k, 0) \\ \sigma(M, z) &= d(z)\, \sigma(M, 0) \end{aligned} \tag{C.18}$$

OBSERVATOIRE DE PARIS

C.2 In practice

C.2.1 Approximation

As described above, the computation of the density fluctuation variance involves computing an integral that must be done numerically, since the power spectrum doesn't have an analytical form. Hence, the halo mass function involves evaluating two integrals numerically, which is time consuming. Luckily, van den Bosch (2002) has provided a good approximation for the standard deviation of the density field:

$$\sigma\left(M\right) = \sigma_8 \frac{f\left(u\right)}{f\left(u_8\right)} \tag{C.19}$$

where:

$$f\left(u\right) = 64.087\left(1 + 1.074u^{0.3} - 1,581u^{0.4} + 0.954u^{0.5} - 0.185u^{0.6}\right)^{-10} \tag{C.20}$$

with:

$$\begin{aligned} u &= 3.80410^{-4}\Gamma\left(\frac{Mh}{\Omega_{m,0}}\right)^{1/3} \\ u_8 &= 32\Gamma \\ \Gamma &= \Omega_{m,0}h\exp\left[-\Omega_b\left(1+\sqrt{2h}/\Omega_{m,0}\right)\right] \end{aligned} \tag{C.21}$$

Now, with this approximation, we can compute the derivative of σ and:

$$\left(M\frac{\mathrm{d}\ln\sigma}{\mathrm{d}M}\right)^{-1} + \frac{1}{2} = \frac{\left(-0.000310111X^{1.7} + 0.00225895X^{1.6} - 0.00505879X^{1.5} - 0.1X^{1.2}\right)}{\left(-0.000328357X^{1.8} + 0.00310111X^{1.7} - 0.0090358X^{1.6} + 0.0101176X^{1.5}\right)} \tag{C.22}$$

with:

$$X = \Gamma\sqrt[3]{\frac{hM}{\Omega_{m,0}}} \tag{C.23}$$

In Figure C.1, we show the comparison between the variance computed with the transfer function obtained from the CAMB program and the approximation of van den Bosch (2002). The approximation diverges from the theoretical computation involving the transfer function, but when used for the computation of the halo mass function, discrepancies are not significant.

C.2.2 Halo mass function models

We put in Table C.1 some popular models for the halo mass function, and used in the thesis for various comparisons. A detailed list can be found in Murray et al. (2013).

We used these models of halo mass function to compare them to the halo mass function obtained directly from the data of Boylan-Kolchin et al. (2009) extracted on the Millennium-II database. Most of these models are defined using FoF halo masses, resulting directly from the sum of the dark matter particle masses constituting the halo. But for MAGGIE, we are just interested on the mass within radius r_{200}. As we can see in Figure C.2, both definitions are very different.

For the abundance matching technique in MAGGIE, we need to use the halo mass function in black in Figure C.2. The fit done on the data is shown in green. Discrepancies are important at high virial masses because of the low number of halos at such masses, giving a poorly constrained relation. But this fit is sufficient and gives relatively good results when using it with MAGGIE (see Chapter 5).

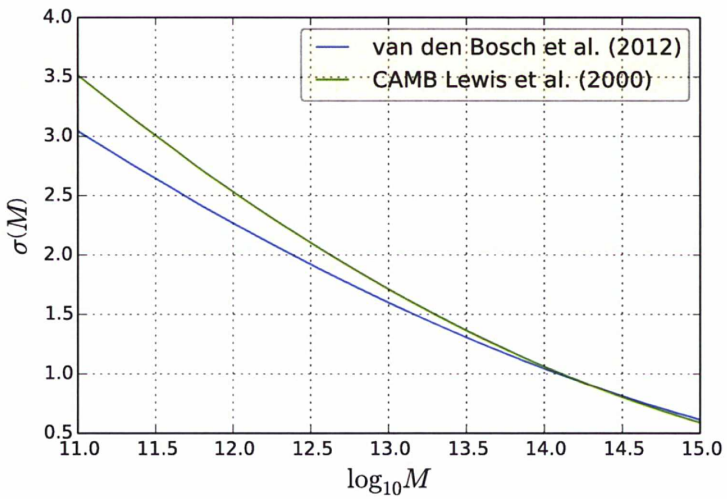


Figure C.1: Comparison between the variance of the density fluctuation computed numerically from the transfer function obtained from CAMB (Lewis et al., 2000) in *green* and the approximation from van den Bosch (2002) in *blue*. Models diverge at low masses but the computation of the halo mass function seems to not be affected by this divergences.

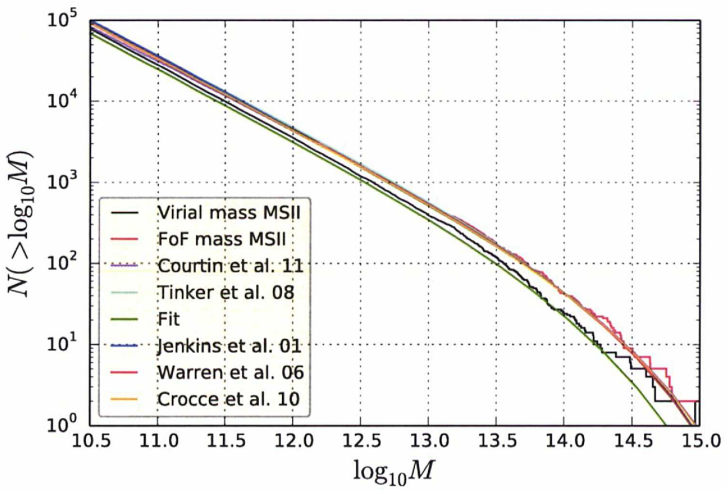


Figure C.2: Cumulative halo mass function for the output of the Millennium-II run for FoF masses (red) and virial masses (black), and some models of halo mass functions. We show the fit we have done on the virial mass function (green). Virial and FoF masses are very different and we must be careful when using it in MAGGIE, using just groups at equilibrium.

Table C.1: A table of some models for the halo mass function.

Model	$f(\sigma)$	Parameters
Warren et al. (2006)	$f(\sigma) = A\left(\sigma^{-a} + b\right)\exp\left(-\frac{c}{\sigma^2}\right)$	$A = 0.7234$, $a = 1.625$, $b = 0.2538$, $c = 1.1982$
Courtin et al. (2011)	$f(\sigma) = A\left(\frac{2a}{\pi}\right)^{\frac{1}{2}}\frac{\delta_c}{\sigma}\left(1 + \left(\frac{\sqrt{a}\delta_c}{\sigma}\right)^{-2p}\right) \times \exp\left(-\frac{\delta_c^2 a}{2\sigma^2}\right)$	$A = 0.348$, $a = 0.695$, $p = 0.1$, $\delta_c = 1.673$
Crocce et al. (2010)	$f(\sigma) = A\left(\sigma^{-a} + b\right)\exp\left(-\frac{c}{\sigma^2}\right)$	$A(z) = 0.58(1+z)^{-0.13}$, $a(z) = 1.37(1+z)^{-0.15}$, $b(z) = 0.3(1+z)^{-0.084}$, $c(z) = 1.036(1+z)^{-0.024}$
Jenkins et al. (2001)	$f(\sigma) = A\exp\left(-\left\lvert\ln\sigma^{-1} + a\right\rvert^b\right)$	$A = 0.315$, $a = 0.61$, $b = 3.8$
Tinker et al. (2008)	$f(\sigma) = A\left(\left(\frac{\sigma}{b}\right)^{-a} + 1\right)\exp\left(-\frac{c}{\sigma^2}\right)$	$A(z) = A_0(1+z)^{-0.14}$, $a(z) = a_0(1+z)^{-0.06}$, $b(z) = b_0(1+z)^{-\alpha}$, $\log_{10}\alpha = -\left[\frac{0.75}{\ln(\Delta/75)}\right]^{1.2}$, $A_0 = 0.1\log_{10}\Delta - 0.05$, $a_0 = 1.43 + (\log_{10}\Delta - 2.3)^{1.5}$, $b_0 = 1.0 + (\log_{10}\Delta - 1.6)^{-1.5}$, $c_0 = 1.2 + (\log_{10}\Delta - 2.35)^{1.6}$

Appendix D.

q-Gaussians (or Tsallis) distributions

Contents

Our first computations of the probability of membership for MAGGIE were not satisfying when compared to the probability estimated directly from the cosmological simulation. Probability contours let us think in a stronger truncation in the velocity distribution for high velocities in the simulation than that of the assumed Gaussian distribution, creating discrepancies between our expectations and data. A natural distribution with this property is the q-Gaussian or Tsallis distribution, which has an additional free parameter, measuring the departures from Gaussianity. It can lead to stronger truncation in the velocity distribution. But the resulting distribution function of the particle system becomes quite complex and the analytical computations difficult in some cases, resumed in the following sections.

D.1 q-Gaussian (or Tsallis) distributions

Tsallis (1988) developed a generalization of the entropy, non extensive, in terms of an additional parameter q leading to systems that are not at the equilibrium. According to Hansen et al. (2006), the radial and tangential velocity distributions of dark matter particles in halos in cosmological simulations do not exactly follow a Gaussian distribution, but are more accurately adjusted by a Tsallis distribution. A possible explanation is that halos are not fully at the equilibrium at our epoch.

Radial v_r and tangential v_t velocity distributions at a given point of coordinates $\mathbf{r}$ are of the following forms:

$$\frac{\mathrm{d}N}{\mathrm{d}v_r} = A_r \left(1 - B_r \left(\frac{v_r}{\sigma_r}\right)^2\right)^{\alpha_r} \tag{D.1}$$

TESIS

$$\frac{\mathrm{d}N}{\mathrm{d}v_t} = A_t v_t \left(1 - B_t \left(\frac{v_t}{\sigma_t}\right)^2\right)^{\alpha_t} \tag{D.2}$$

In the two above equations, integrating over the corresponding velocities gives us the number of particles at a given coordinates by unit of volumes, in other words the density profile $\nu(\mathbf{r})$. Integrating over the second moment, we have the velocity dispersion, another constraint to determine the normalization factor.

$$\int f(\mathbf{r},\mathbf{v})\,\mathrm{d}\mathbf{v} = \int \frac{\mathrm{d}N}{\mathrm{d}v_i}\mathrm{d}v_i = \nu(\mathbf{r}) \tag{D.3}$$

$$\int {v_i}^2 f(\mathbf{r},\mathbf{v})\,\mathrm{d}\mathbf{v} = \int {v_i}^2 \frac{\mathrm{d}N}{\mathrm{d}v_i}\mathrm{d}v_i = {\sigma_i}^2 \nu(\mathbf{r}) \tag{D.4}$$

In the case of the tangential velocity $v_t^2 = v_\theta^2 + v_\phi^2$, the dispersion is simply $\sigma_t^2 = \sigma_\theta^2 + \sigma_\phi^2$. In the following sections, we will assume isotropy and so by symmetry $\sigma_\theta = \sigma_\phi$.

For the integration over v_r, we need to take into account the two cases $B_r > 0$ and $B_r < 0$.

$$\int_{-\infty}^{\infty} \left(1 + x^2\right)^{\alpha} \mathrm{d}x = \frac{\sqrt{\pi}\Gamma\left(-\frac{1}{2} - \alpha\right)}{\Gamma(-\alpha)} \qquad \alpha < -\frac{1}{2} \tag{D.5}$$

$$\int_{-1}^{1} \left(1 - x^2\right)^{\alpha} \mathrm{d}x = \frac{\sqrt{\pi}\Gamma(1+\alpha)}{\Gamma\left(\frac{3}{2} + \alpha\right)} \qquad \alpha > -1 \tag{D.6}$$

Equation D.3 gives us normalizations by substitution to have the equations Equation D.6:

$$\frac{A_r\sqrt{\pi}\sigma_r\Gamma\left(-\frac{1}{2} - \alpha_r\right)}{\sqrt{-B_r}\Gamma(-\alpha_r)} = \nu(\mathbf{r}) \qquad \alpha_r < -\frac{1}{2} \tag{D.7}$$

$$\frac{A_r\sqrt{\pi}\sigma_r\Gamma(1+\alpha_r)}{\sqrt{B_r}\Gamma\left(\frac{3}{2} + \alpha_r\right)} = \nu(\mathbf{r}) \qquad \alpha_r > -1 \tag{D.8}$$

For second moment equations, we use the following equations:

$$\int_{-\infty}^{\infty} x^2\left(1 + x^2\right)^{\alpha} \mathrm{d}x = \frac{\sqrt{\pi}\Gamma\left(-\frac{3}{2} - \alpha\right)}{2\Gamma(-\alpha)} \qquad \alpha < -\frac{3}{2} \tag{D.9}$$

$$\int_{-1}^{1} x^2\left(1 - x^2\right)^{\alpha} \mathrm{d}x = \frac{\sqrt{\pi}\Gamma(1+\alpha)}{\Gamma\left(\frac{5}{2} + \alpha\right)} \qquad \alpha > -1 \tag{D.10}$$

giving us:

$$\frac{A_r\sqrt{\pi}{\sigma_r}^3\Gamma\left(-\frac{3}{2} - \alpha\right)}{2(-B_r)^{3/2}\Gamma(-\alpha)} = {\sigma_r}^2\nu(\mathbf{r}) \qquad \alpha_r < -\frac{3}{2} \tag{D.11}$$

$$\frac{A_r\sqrt{\pi}{\sigma_r}^3\Gamma(1+\alpha_r)}{2{B_r}^{3/2}\Gamma\left(\frac{5}{2} + \alpha_r\right)} = {\sigma_r}^2\nu(\mathbf{r}) \qquad \alpha_r > -1 \tag{D.12}$$

Finally, normalizations are:

$$B_r = \frac{1}{3 + 2\alpha_r} \tag{D.13}$$

$$A_r = \frac{\nu(\mathbf{r})\,\Gamma(-\alpha_r)}{\Gamma\left(-\frac{1}{2} - \alpha_r\right)} \frac{1}{\sqrt{-(3 + 2\alpha_r)\,\pi}\sigma_r} \qquad \alpha_r < -\frac{3}{2} \tag{D.14}$$

$$A_r = \frac{\nu(\mathbf{r})\,\Gamma\left(\frac{3}{2} + \alpha_r\right)}{\Gamma(1 + \alpha_r)} \frac{1}{\sqrt{(3 + 2\alpha_r)\,\pi}\sigma_r} \qquad \alpha_r > -1 \tag{D.15}$$

THESIS

Integrating other tangential velocities, we use following results:

$$\int_0^{\infty} x\left(1+x^2\right)^{\alpha}\mathrm{d}x = -\frac{1}{2\left(1+\alpha\right)} \qquad \alpha < -1 \tag{D.16}$$

$$\int_0^{1} x\left(1-x^2\right)^{\alpha}\mathrm{d}x = \frac{1}{2\left(1+\alpha\right)} \qquad \alpha > -1 \tag{D.17}$$

$$\int_0^{1\,\mathrm{or}\,\infty} x^3\left(1\pm x^2\right)^{\alpha}\mathrm{d}x = \frac{1}{2\left(1+\alpha\right)\left(2+\alpha\right)} \qquad \alpha > -1\,\mathrm{or}\,\alpha < -2 \tag{D.18}$$

giving:

$$B_t = \frac{1}{\left(2+\alpha_t\right)} \tag{D.19}$$

$$A_t = \frac{\nu\left(\mathbf{r}\right)\left(1+\alpha_t\right)}{\sigma_t\left(2+\alpha_t\right)} \tag{D.20}$$

D.2 Choice of a distribution function

The global velocity distribution is the combination of the radial and tangential distributions, not as easy as wanted.

D.2.1 Similar to Gaussian case

We choose a form similar ot the Gaussian case, but not identical in the sense of a strict product of independent variables. We have:

$$f\left(\mathbf{r},\mathbf{v}\right) = A\left(\alpha\right)\left[1 - B\left(\alpha\right)\left[\left(\frac{v_r}{\sigma_r}\right)^2 + \left(\frac{v_\theta}{\sigma_\theta}\right)^2 + \left(\frac{v_\phi}{\sigma_\phi}\right)^2\right]\right]^{\alpha} \tag{D.21}$$

We find normalizations in the same way as previously. We must worry in the choice of the limiting velocity for the integration in the case where $B\left(\alpha\right) > 0$. But if we choose to impose a limit on one velocity, the constraints on the two others becomes the consideration of the three velocities as a single one with the good substitution in the integration. We just consider the triplet of velocities following the constraint $\left(\frac{v_r}{\sigma_r}\right)^2 + \left(\frac{v_\theta}{\sigma_\theta}\right)^2 + \left(\frac{v_\phi}{\sigma_\phi}\right)^2 < 1$. In consequence, we find for the normalizations:

$$A\left(\alpha\right) = \frac{\nu\left(\mathbf{r}\right)}{\left|5/2+\alpha\right|^{3/2}(2\pi)^{3/2}\sigma_r{\sigma_\theta}^2}\frac{\Gamma\left(-\alpha\right)}{\Gamma\left(-3/2-\alpha\right)} \qquad \alpha < -\frac{5}{2} \tag{D.22}$$

$$A\left(\alpha\right) = \frac{\nu\left(\mathbf{r}\right)}{\left|5/2+\alpha\right|^{3/2}(2\pi)^{3/2}\sigma_r{\sigma_\theta}^2}\frac{\Gamma\left(5/2+\alpha\right)}{\Gamma\left(1+\alpha\right)} \qquad \alpha > -1 \tag{D.23}$$

$$B\left(\alpha\right) = \frac{1}{5+2\alpha} \tag{D.24}$$

Integrating over v_t, we get the radial distribution and the tangential distribution when integrating over v_r. The definitions of the q-Gaussian distributions imply $\alpha_r = 1+\alpha$ et $\alpha_t = 1/2+\alpha$. But fitting these distributions to the data of the cosmological simulation gives different values of α for the radial and tangential distribution, hence the model of the global velocity distribution is not adapted and we need to find an other expression.

THESIS

Remark 8

However, if we want to obtain the velocity distribution along the line-of-sight we need to integrate this form on the two velocities perpendicular to the line-of-sight. Computations are simple if the quadratic form is transformed into a canonical one, useful with the previous integral definitions. ■

D.2.2 Separable joint velocity distribution

We treat the case:

$$f(\mathbf{r},\mathbf{v}) = A\left(1 - B_r\left(\frac{v_r}{\sigma_r}\right)^2\right)^{\alpha_r} v_t\left(1 - B_t\left(\frac{v_t}{\sigma_t}\right)^2\right)^{\alpha_t} \tag{D.25}$$

Constraints give:

$$B_t = \frac{1}{2(2+\alpha_t)} \tag{D.26}$$

$$B_r = \frac{1}{(3+2\alpha_r)} \tag{D.27}$$

and following cases, we obtain for A in the Table D.1 the different normalizations.

Table D.1: Table of coefficient for the normalization in different cases.

	$\alpha_r < -3/2$	$\alpha_r > -1$
$\alpha_t < -2$	$\frac{(1+\alpha_t)\,\nu(\mathbf{r})\,\Gamma(-\alpha_r)}{\sqrt{-\pi(3+2\alpha_r)}\sigma_r\sigma_t^2(2+\alpha_t)\,\Gamma(-1/2-\alpha_r)}$	$\frac{(1+\alpha_t)\,\nu(\mathbf{r})\,\Gamma(3/2+\alpha_r)^{3/2}}{\sqrt{-\pi(3+2\alpha_r)}\sigma_r\sigma_t^2(2+\alpha_t)\,\Gamma(1+\alpha_r)\,\Gamma(5/2+\alpha_r)^{1/2}}$
$\alpha_t > -2$	$\frac{(1+\alpha_t)\,\nu(\mathbf{r})\,\Gamma(-\alpha_r)}{\sqrt{-\pi(3+2\alpha_r)}\sigma_r\sigma_t^2(2+\alpha_t)\,\Gamma(-1/2-\alpha_r)}$	$\frac{(1+\alpha_t)\,\nu(\mathbf{r})\,\Gamma(3/2+\alpha_r)^{3/2}}{\sqrt{-\pi(3+2\alpha_r)}\sigma_r\sigma_t^2(2+\alpha_t)\,\Gamma(1+\alpha_r)\,\Gamma(5/2+\alpha_r)^{1/2}}$

Integrating to obtain the radial and tangential distribution, we see that we have two different values of α, each one equal to the tangential and radial α. This seems to correspond with what observed in the cosmological simulation. The problem is that to obtain the line-of-sight velocity distribution, the equation doesn't have an analytical expression, and in consequence not useful in the computation of the probability of MAGGIE. So we choose to abandon this model. Moreover, in simulations, it turns out that α_r depends on α_t hence the joint distribution $f(\mathbf{r},\mathbf{v})$ is not separable.

D.3 Generating q-Gaussian distributions

D.3.1 One dimension

The distribution function is expressed as:

$$f(\mathbf{r},\mathbf{v}) = A\left(1 - B\left(\frac{(v-\mu)^2}{\sigma^2}\right)\right)^{\frac{q}{1-q}} \tag{D.28}$$

where we replaced α with its equivalent q which is $q/(1-q)$, making the modeling easier since there is no cut in values of the q parameter. The distribution is centered in μ with dispersion σ. According to Thistleton et al. (2006), random numbers following the q-Gaussian distribution are expressed as:

$$Z_1 = \sqrt{-2\left(\frac{3-q}{1+q}\right)\ln_{\frac{3q-1}{q+1}} U_1}\cos\left(2\pi U_2\right) \tag{D.29}$$

$$Z_2 = \sqrt{-2\left(\frac{3-q}{1+q}\right)\ln_{\frac{3q-1}{q+1}} U_1}\sin\left(2\pi U_2\right) \tag{D.30}$$

with:

$$\ln_q x = \frac{x^{1-q}-1}{1-q} \tag{D.31}$$

where U_1 and U_2 are two random variables following a uniform distribution, with their values between 0 and 1. To generate a one dimensional Tsallis with dispersion σ and mean μ, the following random variables are sufficient:

$$Z = \sigma Z_i + \mu \tag{D.32}$$

where $i \in \{1, 2\}$.

D.3.2 Two dimensional case

In the case of the tangential distribution:

$$f\left(\mathbf{r}, \mathbf{v}\right) = Av\left(1 - B\left(\frac{(v-\mu)^2}{\sigma^2}\right)\right)^{\frac{q}{1-q}} \tag{D.33}$$

We force $\mu = 0$ for an easier computation. The cumulative distribution function is:

$$F_X\left(x\right) = \int_0^x Av\left(1 - B\left(\frac{v}{\sigma}\right)^2\right)^{\frac{q}{1-q}} \mathrm{d}v \tag{D.34}$$

If U is an uniform random variable between 0 and 1, hence $U = F_X\left(X\right)$. Inverting the relation, we find X following the distribution. In all cases where $q < 1$ and $1 < q < 2$, we can find:

$$X = \sigma\sqrt{2\left(\frac{2-q}{1-q}\right)\left(1 - U^{1-q}\right)} \tag{D.35}$$

With the mean of the distribution:

$$X = \mu + \sigma\sqrt{2\left(\frac{2-q}{1-q}\right)\left(1 - U^{1-q}\right)} \tag{D.36}$$

D.4 Cumulative distribution functions

The cumulative distribution function can be useful in the situation where we search parameters fitting an unknown distribution by the Kolmogorov-Smirnov method for example.

THESIS

D.4.1 One dimensional case

By definition of the cumulative distribution function:

$$F_X(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x \tag{D.37}$$

with $f(x)$ the distribution function. With the dispersion σ and the bias μ:

$$F_X(x) = \frac{A_r\sigma}{\sqrt{|B_r|}}\left(\frac{\sqrt{\pi}}{2}f^1(\alpha) + \frac{\sqrt{|B_r|}}{\sigma}(x-\mu)\,{}^2\mathrm{F}_1\left(\frac{1}{2}, -\alpha, \frac{3}{2}, -B_r\left(\frac{x-\mu}{\sigma}\right)^2\right)\right) \tag{D.38}$$

where A_r and B_r are the same coefficients found previously, ${}^2\mathrm{F}_1$ the Gaussian hypergeometric and f^1 a function such:

$$f^1(\alpha) = \frac{\Gamma\left(-\frac{1}{2}-\alpha\right)}{\Gamma(-\alpha)} \quad \alpha < -\frac{1}{2} \tag{D.39}$$

$$= \frac{\Gamma(1+\alpha)}{\Gamma(-\alpha)} \quad \alpha > -1 \tag{D.40}$$

D.4.2 Two dimensional case

In this case the computation is easier:

$$F_X(x) = \frac{A_t\sigma^2}{B_t}\frac{1}{2(1+\alpha)}\left(1 - \left(1 - B_t\left(\frac{x-\mu}{\sigma}\right)^2\right)^{1+\alpha}\right) \tag{D.41}$$

with A_t and B_t identical to the coefficients determined previously.

THESIS

Appendix E.

QuadTree on celestial sphere

E.1 Introduction

The extraction of galaxy groups from redshift space involves various algorithms to search for galaxies in a given region of the sky. Methods as those used in numerical simulations for searching dark matter halos can be applied. Such techniques often use a partition of the space to make a brute force computation of the distance between particles only on a small portion of the three dimensional space. Same partitioning of the celestial sphere can be done, but the non-euclidean metric of celestial coordinates make the task a little harder.

E.2 QuadTree

The principle of the QuadTree is to make a partition of the space (celestial sphere in our case). Each created partition will be partitioned too if the number of galaxies in it is superior to a limit we define at the creation of the QuadTree. If the number of levels in the refinement is superior to a given limit, we stop the refinement.

This is clearly a tree structure, since the partitions, called nodes, are subdivided into other nodes. This allow to rapidly search for galaxies in a given region since we can easily determine which node intersect a given region.

E.2.1 Construction

The construction is straightforward with the description above. We start by defining the limits in the (α, δ) plane for the region to refine. This region is the root node. Then, the following instructions are applied recursively.

- We determine in which child node each point is falling inside. We keep an array of the identities of points in the tree to which each node point to. In this array, identities are ordered according to the node of the point. So, at the end of the tree construction, the array of the identities will be structured in the same as the tree, allowing for optimization of the memory and for future searches of points.
- If the maximal level of refinement is reached, we no longer subdivide the node.
- If the number of points in the child node is superior to the fixed limit, we subdivide the node in four other nodes.
- Go to the brother of the node.

THESIS

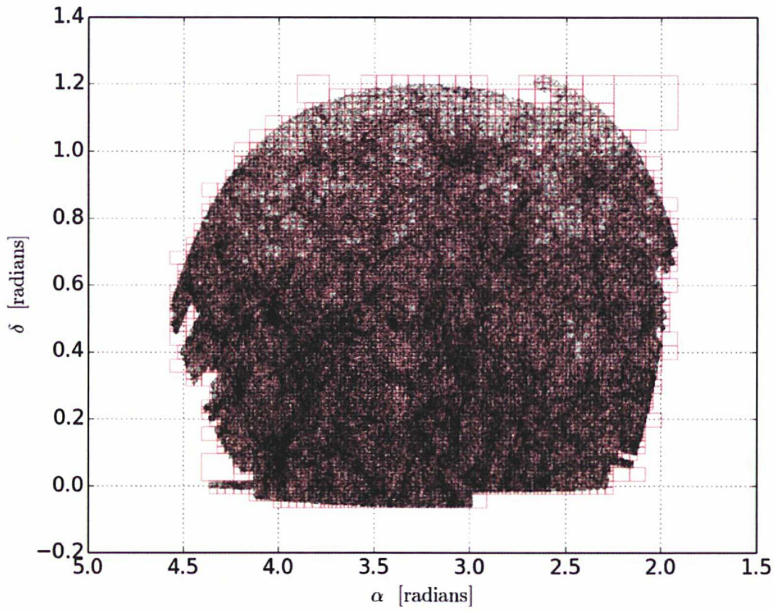


Figure E.1: A simple illustration of a QuadTree generated for the SDSS adjoining block of galaxies. The tree is more refined in the regions at higher surface density.

For optimization, we keep just nodes that are not empty, linking together brother nodes.

During the construction of the node, we also keep the information of their spatial geometry such as extremal coordinates in right ascension and declination, center position, half width in each axis to avoid useless computations when searching points on the celestial sphere.

At this stage, we make a simple partition of the space as in any other QuadTree, without caring about the special metric involved.

An illustration of a QuadTree generated for the galaxies in the adjoining block of the SDSS is shown in Figure E.1.

E.2.2 Searching in a given region

To search in a given region, we go recursively through the tree structure, finding all nodes that intersect it. An improvement can be done by computing if a node is entirely contained by the searching region. If yes, we can directly use the pointer to the identities array to include its points, without descending more in the tree.

It is easy to determine whether the region and a node intersect is easy, since they are defined as two rectangles in a two dimensional space.

The rectangular region is defined in the declination axis simply by taking the central declination coordinate and adding it the angular distance for the research region, since no distortions are present along this axis. For the right ascension, we need to know the maximal separation between the central point and the spherical circle generated by the angular distance. For this extremal case point, it is clear that the corresponding meridian is tangent to the spherical circle. So, in the spherical triangle formed by our central point, the extremal point and the pole, we have a supplementary constraint. The sinus formula applied to it gives us:

$$\Delta\alpha = \operatorname{asin}\left(\frac{\sin d}{\cos \delta_0}\right) \tag{E.1}$$

where d is the angular radius inside which we are searching for points and δ_0 is the declination of the central point around which we search. Our rectangular area is completely defined, and the intersections with the nodes of the tree are easy to compute.

The case of the periodic search is complex. If the rectangular region fall outside the periodic limits (inferior to 0 or superior to 2π in right ascension on the celestial sphere), we need to duplicate the search region and make the intersection with nodes for two regions instead of one. This is a little time consuming but is the only way to handle correctly the periodic case.

E.2.3 k nearest neighbors

The k nearest neighbors in the celestial sphere uses the implementation of the search in a given region of the sky.

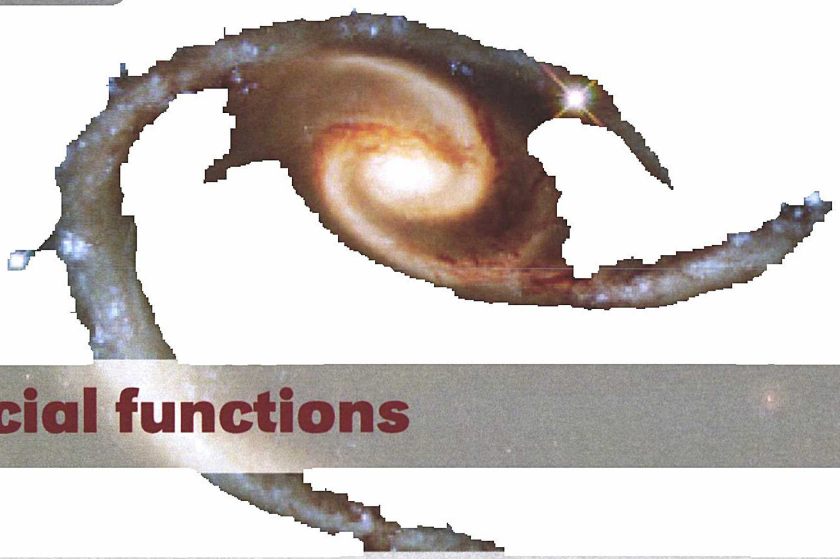
We find the leaf node to which our central point belongs to. A particular attention must be done since we keep only non empty nodes. If the point belongs to an empty one, we affect to it the parent node. In each case, we take the parent node of the found node and search points inside it. Their identities are added to a queue of size k in ascending order of distance to the central point.

We define a search region with this most distant point and fill again the queue with points of this region. If the number of points found is inferior to k, we take the parent node and redo the same computation until the queue is entirely filled with the k nearest neighbors.

THESIS

THESIS

Appendix F.



Special functions

F.1 Legendre elliptic integral function

F.1.1 Introduction

The Legendre elliptic integral function appears naturally when evaluationg distances like the luminosity distance in a flat Universe. But the most of the time, this function is not used directly because of the difficulty in its implementation. When it is already adapted, it is only for special values. In following sections, we described how to use the NSWC implementation of elliptic integrals.

F.1.2 Luminosity distance

Our goal is to easily and precisely compute the luminosity distance given a cosmology according to the formula:

$$d_L(z) = \frac{c(1+z)}{H_0} \int_0^z \frac{\mathrm{d}t}{\sqrt{\Omega_m (1+t)^3 + \Omega_\Lambda}} \tag{F.1}$$

Making the change of variable $u = 1/t$ and defining $s = \sqrt{3}(1-\Omega_m)/\Omega_m$ gives us:

$$d_L(z) = \frac{c(1+z)}{H_0\sqrt{s\Omega_m}} \left[T(s) - T\left(\frac{s}{1+z}\right) \right] \tag{F.2}$$

with:

$$T(x) = \int_0^x \frac{\mathrm{d}u}{\sqrt{u^4+u}} \tag{F.3}$$

As described in Liu et al. (2011), this integral is an elliptic integral, and such integrals can be expressed in terms of Carlson symmetric forms $R_F(x_1, x_2, x_3)$:

$$R_F(x_1, x_2, x_3) = \frac{1}{2} \int_0^\infty \frac{\mathrm{d}t}{\sqrt{(t+x_1)(t+x_2)(t+x_3)}} \tag{F.4}$$

With help of the reduction theorem, we can write:

$$T(x) = 4R_F(m, m+3+2\sqrt{3}, m+3-2\sqrt{3}) \tag{F.5}$$

where:

THESIS

$$m(x) = \frac{2\sqrt{x^2 - x + 1}}{x} + \frac{2}{x} - 1 \tag{F.6}$$

F.2 Incomplete gamma function

F.2.1 Introduction

By definition, the incomplete gamma function $\Gamma(a, x)$ is:

$$\Gamma(a, x) = \int_x^\infty e^{-t} t^{a-1} \mathrm{d}t \tag{F.7}$$

By default, many algorithms used to evaluate incomplete gamma functions do not allowed for $a < 0$. Moreover, we need to use an algorithm that do not use the "simple" gamma function $\Gamma(a)$:

$$\Gamma(a) = \int_0^\infty e^{-t} t^{a-1} \mathrm{d}t \tag{F.8}$$

because that function have singularities for negative values of a where a is an integer, as we can see in figure Figure F.1. So we need an algorithm that does not involve the gamma function for

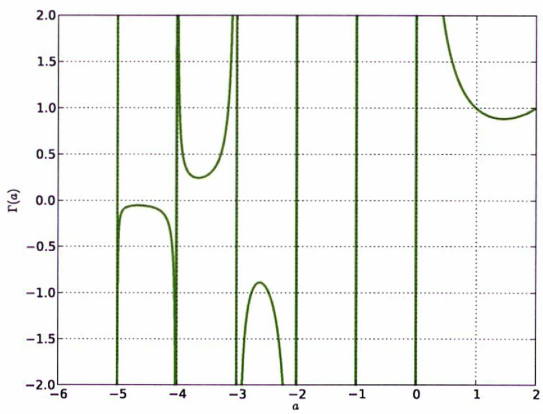


Figure F.1: The gamma function showing singularities at zero and negative integers.

negative values. Here is described such algorithm.

F.2.1.1 Theory

The best way to compute the incomplete gamma function for a negative values is to use recurrence relations. Let us define:

$$\Gamma(a + 1, x) = \int_x^\infty e^{-t} t^a \mathrm{d}t \tag{F.9}$$

Defining $u' = e^{-t}$ and $v = t^a$, we can use integration by parts:

$$\Gamma(a + 1, x) = \left[-e^{-t} t^a\right]_x^\infty + a \int_x^\infty e^{-t} t^{a-1} \mathrm{d}t \tag{F.10}$$

The term in square brackets is always zero at infinity, $\forall a$, and the second member of the right hand side of the previous equation lets appear the definition of the incomplete gamma function.

THESIS

So the recurrence relation for the incomplete gamma function is:

$$\Gamma(a+1,x) = e^{-x}x^a + a\Gamma(a,x) \tag{F.11}$$

We can see that computing the incomplete gamma function for $a \leqslant 0$ can be done with a recursive function using the function at higher values of a.

$$\Gamma(a,x) = \frac{\Gamma(a+1,x) - e^{-x}x^a}{a} \tag{F.12}$$

The previous equation shows that there is still a problem for integer values of a because if $a = -2$ for example, at a moment in the recursion, we have a value of 0 for a which create problems. If we refer to Abramowitz & Stegun (1964), the definition of the elliptical integral is:

$$E_n(z) = \int_1^\infty e^{-zt}t^{-n}\mathrm{d}t \tag{F.13}$$

for integer values of n. If we change the variable in the integral to $t' = zt$, we can rewrite the equation to have:

$$E_n(z) = z^{n-1}\Gamma(1-n,z) \tag{F.14}$$

so:

$$\Gamma(a,x) = x^a E_{1-a}(x) \tag{F.15}$$

for $a \leq 0$ and a integer. Now we have a good computation for the incomplete gamma function. But numerically, there is still a problem near integer negative values of a. If a is very close to an integer value, then at some moment in the recursion, a will be very small. So $1/a$ can be greater than the overflow value for the machine. To avoid this, we add a condition for a when it is near zero.

An other definition of the incomplete gamma function is:

$$\Gamma(a,x) = \Gamma(a) - \gamma(a,x) = \Gamma(a)\left(1 - P(a,x)\right) \tag{F.16}$$

with:

$$\gamma(a,x) = \int_0^x e^{-t}t^a\mathrm{d}t \tag{F.17}$$

In Press et al. (1992) exists a precise computation of the function $P(a,x)$. We can remark that this function is not required in the recursion if we have already access to a function that computes the incomplete gamma function for positive values of a.

We provide below the algorithm for computing the incomplete gamma function without loss of precision and without numerical problems for negative values of a.

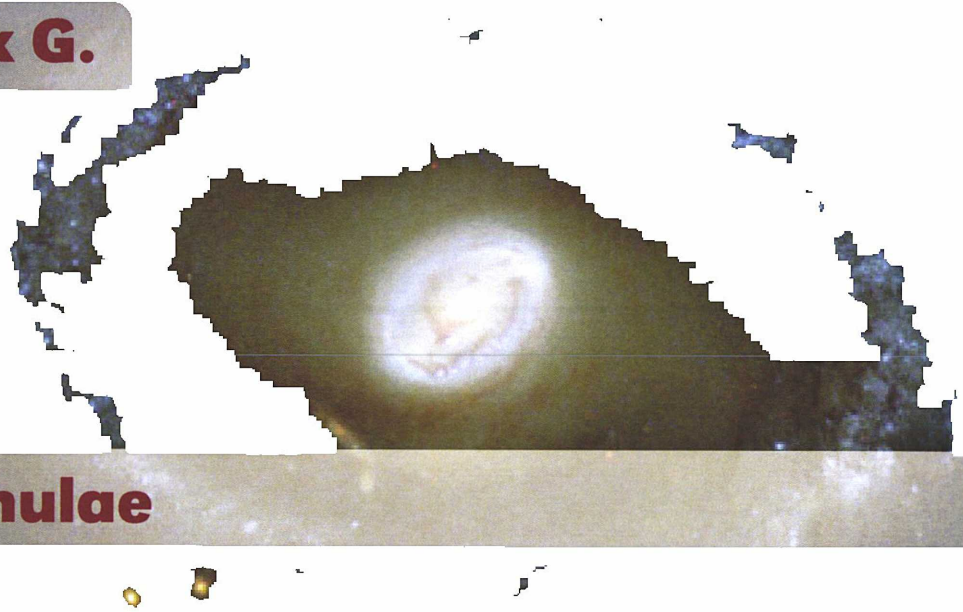
THESIS

F.2.1.2 Numerical

```
def gammainc(a, x):

    """
    To compute the incomplete gamma function
    without loss of precision or without numerical
    problems. OF is the value of the overflow for
    the machine and expint( n, x ) the function
    which computes the integral function for n and x.
    """

    import numpy as np

    if x >= 0.:
        if a <= 1.:
            if a == int(a) or OF * abs(a) < 1 :
                return x ** int(a) * expint(1 - int(a), x)
            else :
                return (gammainc(a + 1, x) -
                    np.exp(-x) * (x ** a)) / a

        else :
            return gamma(a) (1 - P(a, x))
            # or call the function which computes
            # the incomplete gamma function for
            # positive values of a
```

THESIS

Appendix G.



Formulae

G.1 Introduction

In this appendix are described the formulae used in all computations realized during my thesis. Its just a simple way to share and verify that the job is done correctly. References to those formulae are indicated too, in order to improve search when some doubts are present.

G.2 Formulas

The luminosity distance is defined as the relation between the galaxy flux S and its absolute luminosity L by:

$$d_{\mathrm{lum}} = \sqrt{\frac{L}{4\pi S}} \tag{G.1}$$

An analytical precise computation is not possible and while numerical computations exist, they are computationally slow. Some other analytical approximations of this distance were created. For example, Wickramasinghe & Ukwatta (2010) provided an approximation for flat Universe good to 0.3% is available for a range of values in Ω_Λ compatible with WMAP and Planck results:

$$d_{\mathrm{lum}}(z) = \frac{c}{3H_0}\frac{1+z}{\Omega_\Lambda^{1/6}(1-\Omega_\Lambda)^{1/3}}\left[\Psi\left(x\left(0,\Omega_\Lambda\right)\right) - \Psi\left(x\left(z,\Omega_\Lambda\right)\right)\right] \tag{G.2}$$

with:

$$\Psi(x) = 3x^{1/3}2^{2/3}\left[1 - \frac{x^2}{252} + \frac{x^4}{21060}\right] \tag{G.3}$$

$$x(\alpha) = \ln\left(\alpha + \sqrt{\alpha^2+1}\right) \tag{G.4}$$

$$\alpha(z,\Omega_\Lambda) = 1 + 2\frac{\Omega_\Lambda}{1-\Omega_\Lambda}\frac{1}{(1+z)^3} \tag{G.5}$$

The other distances are simply linked to this luminosity distance. The angular distance d_{ang} and the proper distance d_{pm} related with $d_{\mathrm{lum}}(z) = (1+z)^2 d_{\mathrm{ang}}(z) = (1+z)d_{\mathrm{pm}}(z)$.

The element of comoving volume is expressed using the Robertson-Walker metric as:

$$\mathrm{d}V = \frac{c}{H(z)}d_{\mathrm{pm}}(z)^2 \mathrm{d}\Omega \mathrm{d}z \tag{G.6}$$

THESIS

The evolution of the fraction of matter, and dark energy is the following:

$$\Omega_m(z) = \Omega_{m,0}\frac{(1+z)^3}{E(z)^2} \tag{G.7}$$

$$\Omega_\Lambda(z) = \frac{\Omega_{\Lambda,0}}{E(z)^2} \tag{G.8}$$

where z is the redshift and the subscript 0 refers to the actual value of the parameter.

The distance modulus represents the magnitude difference between the observed flux of the galaxy and what it would be if the galaxy were at a distance of 10 pc:

$$\mu(z) = 5\log_{10}\left(\frac{d_{\text{lum}}(z)}{10\,\text{pc}}\right) \tag{G.9}$$

where z is the redshift of the galaxy and d_{lum} is the luminosity distance.

The apparent magnitude m of galaxy in the perfect case where isn't K-correction, extinction, is just:

$$m = M + \mu(z) \tag{G.10}$$

where M is the absolute magnitude of this galaxy in the same band of m and $\mu(z)$ is the distance modulus at redshift z.

Magnitudes are defined at a given constant which is the same for each object so:

$$M - M_\odot = -2.5\log_{10}\left(\frac{L}{L_\odot}\right) \tag{G.11}$$

where M is absolute magnitude, L the luminosity of the object and $\odot$ refers to Sun's quantities. We can determined the luminosity by this relation which gives:

$$\frac{L}{L_\odot} = 10^{0.4(M_\odot - M)} \tag{G.12}$$

For galaxies at a given redshift z, we can see all galaxies with an absolute magnitude (using equation (G.10)):

$$M < m_{\text{lim}} - \mu(z) \tag{G.13}$$

where m_{lim} is the apparent magnitude limit for a survey.

The virial radius r_Δ is defined as the radius at which the density is Δ times the critical density of the Universe. So we have:

$$\rho(r_\Delta) = \Delta\rho_c \tag{G.14}$$

with $\rho_c = \dfrac{3H(z)^2}{8\pi G}$.

If we suppose that the density is constant in this radius, we have:

$$\Delta\frac{3H(z)^2}{8\pi G} = \frac{M_\Delta}{4\pi r_\Delta^3/3} \tag{G.15}$$

where M_Δ is the virial mass. We can now defined three quantities, the virial mass as:

$$M_\Delta = \frac{\Delta H(z)^2 r_\Delta^3}{2G} \tag{G.16}$$

the virial radius as:

$$r_\Delta = \left(\frac{2GM_\Delta}{\Delta H(z)^2}\right)^{1/3} \tag{G.17}$$

and the virial velocity as:

$$v_\Delta = \sqrt{\frac{GM_\Delta}{r_\Delta}} = \sqrt{\frac{\Delta}{2}} H(z)\, r_\Delta \tag{G.18}$$

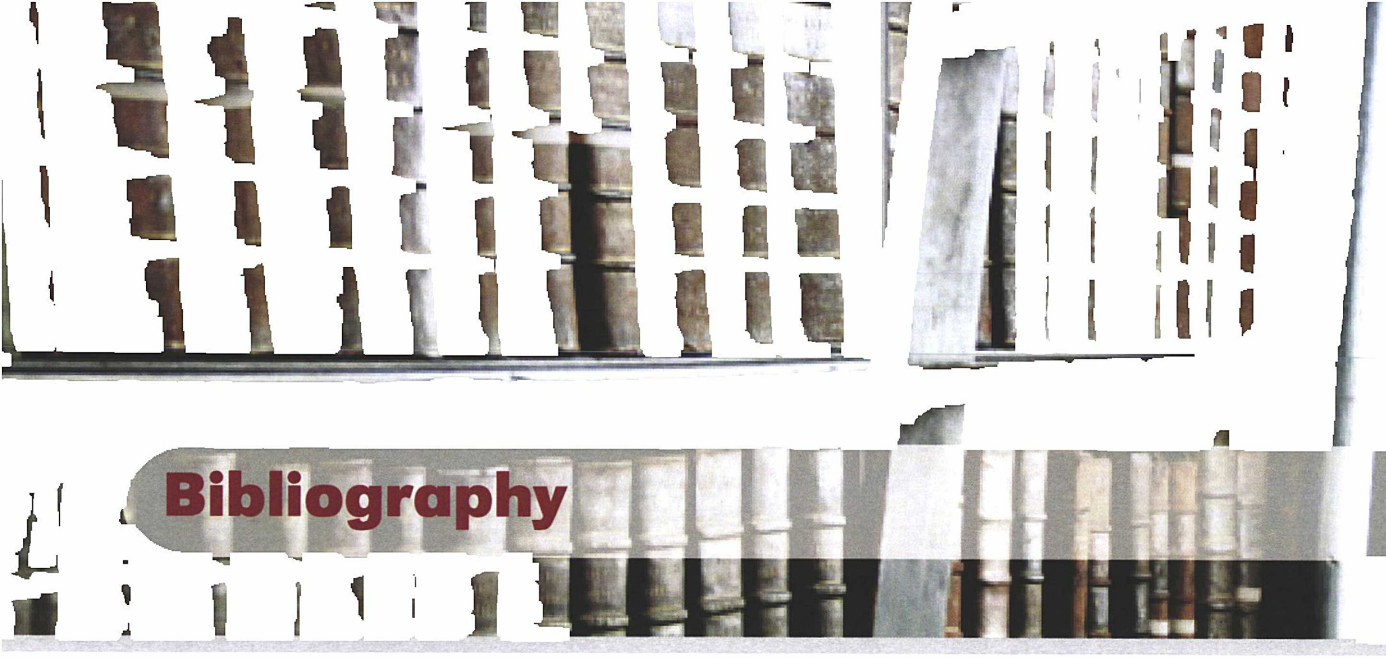
Sometimes, the density at the virial radius isn't defined in relation with the critical density but instead with mean density of the Universe. So the equation (G.14) becomes:

$$\rho(r_\Delta) = \Delta\rho_m = \Delta\Omega_m\rho_c \tag{G.19}$$

We can treat this situation in the same way as previously, but formally with $\Delta \to \Delta\Omega_m$.

THESIS

THESIS



Bibliography

Abell, G.O., 1958, The Distribution of Rich Clusters of Galaxies., ApJS, 3, 211

Abramowitz, M., Stegun, I.A., 1964, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, Dover, New York, ninth Dover printing, tenth GPO printing edn.

Behroozi, P.S., Conroy, C., Wechsler, R.H., 2010, A Comprehensive Analysis of Uncertainties Affecting the Stellar Mass-Halo Mass Relation for $0 < z < 4$, ApJ, 717, 379

Bekki, K., 2014, Galactic star formation enhanced and quenched by ram pressure in groups and clusters, MNRAS, 438, 444

Bell, E.F., McIntosh, D.H., Katz, N., Weinberg, M.D., 2003, The Optical and Near-Infrared Properties of Galaxies. I. Luminosity and Stellar Mass Functions, ApJS, 149, 289

Bennett, C.L., Larson, D., Weiland, J.L., Jarosik, N., Hinshaw, G., Odegard, N., Smith, K.M., Hill, R.S., Gold, B., Halpern, M., Komatsu, E., Nolta, M.R., Page, L., Spergel, D.N., Wollack, E., Dunkley, J., Kogut, A., Limon, M., Meyer, S.S., Tucker, G.S., Wright, E.L., 2013, Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Final Maps and Results, ApJS, 208, 20

Benson, A.J., 2010, Galaxy formation theory, Phys. Rep., 495, 33

Beraldo, L.J., Mamon, G.A., Duarte, M., Peirani, S., Boué, G., 2014, Anisotropic q-Gaussian velocity distributions in LambdaCDM halos, MNRAS, submitted, arXiv:1310.6756

Berlind, A.A., Frieman, J., Weinberg, D.H., Blanton, M.R., Warren, M.S., Abazajian, K., Scranton, R., Hogg, D.W., Scoccimarro, R., Bahcall, N.A., Brinkmann, J., Gott, III, J.R., Kleinman, S.J., Krzesinski, J., Lee, B.C., Miller, C.J., Nitta, A., 2006, Percolation Galaxy Groups and Clusters in the SDSS Redshift Survey: Identification, Catalogs, and the Multiplicity Function, ApJS, 167, 1

Berlind, A.A., Weinberg, D.H., 2002, The Halo Occupation Distribution: Toward an Empirical Determination of the Relation between Galaxies and Mass, ApJ, 575, 587

Blaizot, J., Wadadekar, Y., Guiderdoni, B., Colombi, S.T., Bertin, E., Bouchet, F.R., Devriendt, J.E.G., Hatton, S., 2005, MoMaF: the Mock Map Facility, MNRAS, 360, 159

Blanchard, A., Valls-Gabaud, D., Mamon, G.A., 1992, The origin of the galaxy luminosity function and the thermal evolution of the intergalactic medium, A&A, 264, 365

Blanton, M.R., Lin, H., Lupton, R.H., Maley, F.M., Young, N., Zehavi, I., Loveday, J., 2003, An Efficient Targeting Strategy for Multiobject Spectrograph Surveys: the Sloan Digital Sky Survey "Tiling" Algorithm', AJ, 125, 2276

Blanton, M.R., Lupton, R.H., Schlegel, D.J., Strauss, M.A., Brinkmann, J., Fukugita, M., Loveday, J., 2005, The Properties and Luminosity Function of Extremely Low Luminosity Galaxies, ApJ, 631, 208

Borgani, S., Governato, F., Wadsley, J., Menci, N., Tozzi, P., Lake, G., Quinn, T., Stadel, J., 2001, Preheating the Intracluster Medium

THESIS

in High-Resolution Simulations: The Effect on the Gas Entropy, ApJ, 559, L71

Borgani, S., Murante, G., Springel, V., Diaferio, A., Dolag, K., Moscardini, L., Tormen, G., Tornatore, L., Tozzi, P., 2004, X-ray properties of galaxy clusters and groups from a cosmological hydrodynamical simulation, MNRAS, 348, 1078

Bournaud, F., Jog, C.J., Combes, F., 2005, Galaxy mergers with various mass ratios: Properties of remnants, A&A, 437, 69

Boylan-Kolchin, M., Springel, V., White, S.D.M., Jenkins, A., Lemson, G., 2009, Resolving cosmic structure formation with the Millennium-II Simulation, MNRAS, 398, 1150

Brinchmann, J., Charlot, S., White, S.D.M., Tremonti, C., Kauffmann, G., Heckman, T., Brinkmann, J., 2004, The physical properties of star-forming galaxies in the low-redshift Universe, MNRAS, 351, 1151

Brooks, A.M., Kuhlen, M., Zolotov, A., Hooper, D., 2013, A Baryonic Solution to the Missing Satellites Problem, ApJ, 765, 22

Bryan, G.L., Norman, M.L., 1998, Statistical Properties of X-Ray Clusters: Analytic and Numerical Comparisons, ApJ, 495, 80

Carroll, S.M., Press, W.H., Turner, E.L., 1992, The cosmological constant, ARA&A, 30, 499

Chen, Y.M., Kauffmann, G., Tremonti, C.A., White, S., Heckman, T.M., Kovač, K., Bundy, K., Chisholm, J., Maraston, C., Schneider, D.P., Bolton, A.S., Weaver, B.A., Brinkmann, J., 2012, Evolution of the most massive galaxies to z= 0.6 - I. A new method for physical parameter estimation, MNRAS, 421, 314

Chilingarian, I.V., Melchior, A.L., Zolotukhin, I.Y., 2010, Analytical approximations of K-corrections in optical and near-infrared bands, MNRAS, 405, 1409

Conroy, C., Gunn, J.E., White, M., 2009, The Propagation of Uncertainties in Stellar Population Synthesis Modeling. I. The Relevance of Uncertain Aspects of Stellar Evolution and the Initial Mass Function to the Derived Physical Properties of Galaxies, ApJ, 699, 486

Courtin, J., Rasera, Y., Alimi, J.M., Corasaniti, P.S., Boucher, V., Füzfa, A., 2011, Imprints of dark energy on cosmic structure formation - II. Non-universality of the halo mass function, MNRAS, 410, 1911

Cox, T.J., Jonsson, P., Somerville, R.S., Primack, J.R., Dekel, A., 2008, The effect of galaxy mass ratio on merger-driven starbursts, MNRAS, 384, 386

Crocce, M., Fosalba, P., Castander, F.J., Gaztañaga, E., 2010, Simulating the Universe with MICE: the abundance of massive clusters, MNRAS, 403, 1353

Davis, M., Efstathiou, G., Frenk, C.S., White, S.D.M., 1985, The evolution of large-scale structure in a universe dominated by cold dark matter, ApJ, 292, 371

De Propris, R., Conselice, C.J., Liske, J., Driver, S.P., Patton, D.R., Graham, A.W., Allen, P.D., 2007, The Millennium Galaxy Catalogue: The Connection between Close Pairs and Asymmetry; Implications for the Galaxy Merger Rate, ApJ, 666, 212

Dekel, A., Silk, J., 1986, The origin of dwarf galaxies, cold dark matter, and biased galaxy formation, ApJ, 303, 39

Domínguez Romero, M.J.d.L., García Lambas, D., Muriel, H., 2012, An improved method for the identification of galaxy systems: measuring the gravitational redshift by dark matter haloes, MNRAS, 427, L6

Driver, S.P., Allen, P.D., Graham, A.W., Cameron, E., Liske, J., Ellis, S.C., Cross, N.J.G., De Propris, R., Phillipps, S., Couch, W.J., 2006, The Millennium Galaxy Catalogue: morphological classification and bimodality in the colour-concentration plane, MNRAS, 368, 414

Duarte, M., Mamon, G.A., 2014a, How well does the Friends-of-Friends algorithm recover

group properties from galaxy catalogues limited in both distance and luminosity?, MNRAS, 440, 1763

Duarte, M., Mamon, G.A., 2014b, MAGGIE: Models and Algorithms for Galaxy Groups, Interlopers and Environment, in preparation

Ebeling, H., Stephenson, L.N., Edge, A.C., 2013, Jellyfish: Evidence of extreme ram-pressure stripping in massive galaxy clusters, ArXiv e-prints

Efstathiou, G., 2000, A model of supernova feedback in galaxy formation, MNRAS, 317, 697

Eisenstein, D.J., 1997, An Analytic Expression for the Growth Function in a Flat Universe with a Cosmological Constant, ArXiv Astrophysics e-prints

Eke, V.R., Baugh, C.M., Cole, S., Frenk, C.S., Norberg, P., Peacock, J.A., Baldry, I.K., Bland-Hawthorn, J., Bridges, T., Cannon, R., Colless, M., Collins, C., Couch, W., Dalton, G., de Propris, R., Driver, S.P., Efstathiou, G., Ellis, R.S., Glazebrook, K., Jackson, C., Lahav, O., Lewis, I., Lumsden, S., Maddox, S., Madgwick, D., Peterson, B.A., Sutherland, W., Taylor, K., 2004, Galaxy groups in the 2dFGRS: the group-finding algorithm and the 2PIGG catalogue, MNRAS, 348, 866

Font, A.S., Bower, R.G., McCarthy, I.G., Benson, A.J., Frenk, C.S., Helly, J.C., Lacey, C.G., Baugh, C.M., Cole, S., 2008, The colours of satellite galaxies in groups and clusters, MNRAS, 389, 1619

Frederic, J.J., 1995, Testing the accuracy of redshift-space group-finding algorithms, ApJS, 97, 259

Gunn, J.E., Gott, III, J.R., 1972, On the Infall of Matter Into Clusters of Galaxies and Some Effects on Their Evolution, ApJ, 176, 1

Guo, Q., Cole, S., Eke, V., Frenk, C., 2012, Satellite galaxy number density profiles in the Sloan Digital Sky Survey, MNRAS, 427, 428

Guo, Q., White, S., Boylan-Kolchin, M., De Lucia, G., Kauffmann, G., Lemson, G., Li, C., Springel, V., Weinmann, S., 2011, From dwarf spheroidals to cD galaxies: simulating the galaxy population in a ΛCDM cosmology, MNRAS, p. 164-

Hansen, S.H., Moore, B., Zemp, M., Stadel, J., 2006, A universal velocity distribution of relaxed collisionless structures, jcap, 1, 14

Heisler, J., Tremaine, S., Bahcall, J.N., 1985, Estimating the masses of galaxy groups — Alternatives to the virial theorem, ApJ, 298, 8

Hirschmann, M., Naab, T., Davé, R., Oppenheimer, B.D., Ostriker, J.P., Somerville, R.S., Oser, L., Genzel, R., Tacconi, L.J., Förster-Schreiber, N.M., Burkert, A., Genel, S., 2013, The effect of metal enrichment and galactic winds on galaxy formation in cosmological zoom simulations, MNRAS, 436, 2929

Hogg, D.W., 1999, Distance measures in cosmology, ArXiv Astrophysics e-prints

Hopkins, A.M., Driver, S.P., Brough, S., Owers, M.S., Bauer, A.E., Gunawardhana, M.L.P., Cluver, M.E., Colless, M., Foster, C., Lara-López, M.A., Roseboom, I., Sharp, R., Steele, O., Thomas, D., Baldry, I.K., Brown, M.J.I., Liske, J., Norberg, P., Robotham, A.S.G., Bamford, S., Bland-Hawthorn, J., Drinkwater, M.J., Loveday, J., Meyer, M., Peacock, J.A., Tuffs, R., Agius, N., Alpaslan, M., Andrae, E., Cameron, E., Cole, S., Ching, J.H.Y., Christodoulou, L., Conselice, C., Croom, S., Cross, N.J.G., De Propris, R., Delhaize, J., Dunne, L., Eales, S., Ellis, S., Frenk, C.S., Graham, A.W., Grootes, M.W., Häußler, B., Heymans, C., Hill, D., Hoyle, B., Hudson, M., Jarvis, M., Johansson, J., Jones, D.H., van Kampen, E., Kelvin, L., Kuijken, K., López-Sánchez, Á., Maddox, S., Madore, B., Maraston, C., McNaught-Roberts, T., Nichol, R.C., Oliver, S., Parkinson, H., Penny, S., Phillipps, S., Pimbblet, K.A., Ponman, T., Popescu, C.C., Prescott, M., Proctor, R., Sadler, E.M., Sansom, A.E., Seibert, M., Staveley-Smith, L., Sutherland, W., Taylor, E., Van Waerbeke, L., Vázquez-Mata, J.A., Warren, S., Wijesinghe, D.B., Wild, V., Wilkins, S., 2013, Galaxy And

THESIS

Mass Assembly (GAMA): spectroscopic analysis, MNRAS, 430, 2047

Hubble, E.P., 1929, A spiral nebula as a stellar system, Messier 31., ApJ, 69, 103

Huchra, J.P., Geller, M.J., 1982, Groups of galaxies. I - Nearby groups, ApJ, 257, 423

Jackson, J.C., 1972, A critique of Rees's theory of primordial gravitational radiation, MNRAS, 156, 1P

Jenkins, A., Frenk, C.S., White, S.D.M., Colberg, J.M., Cole, S., Evrard, A.E., Couchman, H.M.P., Yoshida, N., 2001, The mass function of dark matter haloes, MNRAS, 321, 372

Kauffmann, G., Colberg, J.M., Diaferio, A., White, S.D.M., 1999, Clustering of galaxies in a hierarchical universe - I. Methods and results at z=0, MNRAS, 303, 188

Kauffmann, G., Heckman, T.M., White, S.D.M., Charlot, S., Tremonti, C., Brinchmann, J., Bruzual, G., Peng, E.W., Seibert, M., Bernardi, M., Blanton, M., Brinkmann, J., Castander, F., Csábai, I., Fukugita, M., Ivezic, Z., Munn, J.A., Nichol, R.C., Padmanabhan, N., Thakar, A.R., Weinberg, D.H., York, D., 2003, Stellar masses and star formation histories for 10^5 galaxies from the Sloan Digital Sky Survey, MNRAS, 341, 33

Knobel, C., Lilly, S.J., Iovino, A., Porciani, C., Kovač, K., Cucciati, O., Finoguenov, A., Kitzbichler, M.G., Carollo, C.M., Contini, T., Kneib, J.P., Le Fèvre, O., Mainieri, V., Renzini, A., Scodeggio, M., Zamorani, G., Bardelli, S., Bolzonella, M., Bongiorno, A., Caputi, K., Coppa, G., de la Torre, S., de Ravel, L., Franzetti, P., Garilli, B., Kampczyk, P., Lamareille, F., Le Borgne, J.F., Le Brun, V., Maier, C., Mignoli, M., Pello, R., Peng, Y., Perez Montero, E., Ricciardelli, E., Silverman, J.D., Tanaka, M., Tasca, L., Tresse, L., Vergani, D., Zucca, E., Abbas, U., Bottini, D., Cappi, A., Cassata, P., Cimatti, A., Fumana, M., Guzzo, L., Koekemoer, A.M., Leauthaud, A., Maccagni, D., Marinoni, C., McCracken, H.J., Memeo, P., Meneux, B., Oesch, P., Pozzetti, L., Scaramella, R., 2009, An Optical Group Catalog to z = 1 from the zCOSMOS 10 k Sample, ApJ, 697, 1842

Knollmann, S.R., Knebe, A., 2009, AHF: Amiga's Halo Finder, ApJS, 182, 608

Kravtsov, A.V., Borgani, S., 2012, Formation of Galaxy Clusters, ARA&A, 50, 353

Lacey, C., Cole, S., 1993, Merger rates in hierarchical models of galaxy formation, MNRAS, 262, 627

Lanzoni, B., Guiderdoni, B., Mamon, G.A., Devriendt, J., Hatton, S., 2005, GALICS- VI. Modelling hierarchical galaxy formation in clusters, MNRAS, 361, 369

Larson, R.B., Tinsley, B.M., Caldwell, C.N., 1980, The evolution of disk galaxies and the origin of S0 galaxies, ApJ, 237, 692

Le Borgne, D., Rocca-Volmerange, B., Prugniel, P., Lançon, A., Fioc, M., Soubiran, C., 2004, Evolutionary synthesis of galaxies at high spectral resolution with the code PEGASE-HR. Metallicity and age tracers, A&A, 425, 881

Lemson, G., the Virgo Consortium, 2006, Halo and Galaxy Formation Histories from the Millennium Simulation: Public release of a VO-oriented and SQL-queryable database for studying the evolution of galaxies in the LambdaCDM cosmogony, arXiv:astro-ph/0608019

Lewis, A., Challinor, A., Lasenby, A., 2000, Efficient Computation of Cosmic Microwave Background Anisotropies in Closed Friedmann-Robertson-Walker Models, ApJ, 538, 473

Liu, D.Z., Ma, C., Zhang, T.J., Yang, Z., 2011, Numerical strategies of computing the luminosity distance, MNRAS, 412, 2685

Liu, H.B., Hsieh, B.C., Ho, P.T.P., Lin, L., Yan, R., 2008, A New Galaxy Group Finding Algorithm: Probability Friends-of-Friends, ApJ, 681, 1046

Macciò, A.V., Dutton, A.A., van den Bosch, F.C., 2008, Concentration, spin and shape of

dark matter haloes as a function of the cosmological model: WMAP1, WMAP3 and WMAP5 results, MNRAS, 391, 1940

Mamon, G.A., 1992, Are cluster ellipticals the products of mergers?, ApJ, 401, L3

Mamon, G.A., Biviano, A., Boué, G., 2013, MAMPOSSt: Modelling Anisotropy and Mass Profiles of Observed Spherical Systems - I. Gaussian 3D velocities, MNRAS, 429, 3079

Mamon, G.A., Biviano, A., Murante, G., 2010, The universal distribution of halo interlopers in projected phase space. Bias in galaxy cluster concentration and velocity anisotropy?, A&A, 520, A30

Mamon, G.A., Łokas, E.L., 2005, Dark matter in elliptical galaxies - II. Estimating the mass within the virial radius, MNRAS, 363, 705

Maraston, C., Strömbäck, G., Thomas, D., Wake, D.A., Nichol, R.C., 2009, Modelling the colour evolution of luminous red galaxies - improvements with empirical stellar spectra, MNRAS, 394, L107

Marinoni, C., Davis, M., Newman, J.A., Coil, A.L., 2002, Three-dimensional Identification and Reconstruction of Galaxy Systems within Flux-limited Redshift Surveys, ApJ, 580, 122

Martínez, V.J., Saar, E., 2002, Statistics of the Galaxy Distribution, Chapman & Hall, CRC, chapter 7.8

Merchán, M., Zandivarez, A., 2002, Galaxy groups in the 2dF Galaxy Redshift Survey: the catalogue, MNRAS, 335, 216

Muñoz-Cuartas, J.C., Müller, V., 2012, Galaxy groups and haloes in the seventh data release of the Sloan Digital Sky Survey, MNRAS, 423, 1583

Murray, S.G., Power, C., Robotham, A.S.G., 2013, HMFcalc: An online tool for calculating dark matter halo mass functions, Astronomy and Computing, 3, 23

Naab, T., Burkert, A., Hernquist, L., 1999, On the Formation of Boxy and Disky Elliptical Galaxies, ApJ, 523, L133

Navarro, J.F., Frenk, C.S., White, S.D.M., 1996, The structure of cold dark matter halos, ApJ, 462, 563

Nolthenius, R., White, S.D.M., 1987, Groups of galaxies in the CfA survey and in cold dark matter universes, MNRAS, 225, 505

Okamoto, T., Nagashima, M., 2003, Environmental Effects on Evolution of Cluster Galaxies in a Λ-dominated Cold Dark Matter Universe, ApJ, 587, 500

Old, L., Skibba, R.A., Pearce, F.R., Croton, D., Muldrew, S.I., Muñoz-Cuartas, J.C., Gifford, D., Gray, M.E., der Linden, A.v., Mamon, G.A., Merrifield, M.R., Müller, V., Pearson, R.J., Ponman, T.J., Saro, A., Sepp, T., Sifón, C., Tempel, E., Tundo, E., Wang, Y.O., Wojtak, R., 2014, Galaxy cluster mass reconstruction project - I. Methods and first results on galaxy-based techniques, MNRAS, 441, 1513

Peng, Y.j., Lilly, S.J., Kovač, K., Bolzonella, M., Pozzetti, L., Renzini, A., Zamorani, G., Ilbert, O., Knobel, C., Iovino, A., Maier, C., Cucciati, O., Tasca, L., Carollo, C.M., Silverman, J., Kampczyk, P., de Ravel, L., Sanders, D., Scoville, N., Contini, T., Mainieri, V., Scodeggio, M., Kneib, J.P., Le Fèvre, O., Bardelli, S., Bongiorno, A., Caputi, K., Coppa, G., de la Torre, S., Franzetti, P., Garilli, B., Lamareille, F., Le Borgne, J.F., Le Brun, V., Mignoli, M., Perez Montero, E., Pello, R., Ricciardelli, E., Tanaka, M., Tresse, L., Vergani, D., Welikala, N., Zucca, E., Oesch, P., Abbas, U., Barnes, L., Bordoloi, R., Bottini, D., Cappi, A., Cassata, P., Cimatti, A., Fumana, M., Hasinger, G., Koekemoer, A., Leauthaud, A., Maccagni, D., Marinoni, C., McCracken, H., Memeo, P., Meneux, B., Nair, P., Porciani, C., Presotto, V., Scaramella, R., 2010, Mass and Environment as Drivers of Galaxy Evolution in SDSS and zCOSMOS and the Origin of the Schechter Function, ApJ, 721, 193

Planck Collaboration, Ade, P.A.R., Aghanim, N., Armitage-Caplan, C., Arnaud, M., Ashdown, M., Atrio-Barandela, F., Aumont, J., Baccigalupi, C., Banday, A.J., et al., 2013,

THESIS

Planck 2013 results. XVI. Cosmological parameters, A&A, submitted, arXiv:1303.5076

Planelles, S., Quilis, V., 2010, ASOHF: a new adaptive spherical overdensity halo finder, A&A, 519, A94

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1992, Numerical Recipes in Fortran 77: The Art of Scientific Computing, vol. 1, Cambridge University Press, 2nd edn.

Press, W.H., Schechter, P., 1974a, Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation, ApJ, 187, 425

Press, W.H., Schechter, P., 1974b, Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation, ApJ, 187, 425

Ramella, M., Geller, M.J., Huchra, J.P., 1989, Groups of galaxies in the Center for Astrophysics redshift survey, ApJ, 344, 57

Rees, M.J., 1986, Lyman absorption lines in quasar spectra - Evidence for gravitationally-confined gas in dark minihaloes, MNRAS, 218, 25P

Robotham, A., Phillipps, S., de Propris, R., 2010, The variation of the galaxy luminosity function with group properties, MNRAS, 403, 1812

Robotham, A.S.G., Norberg, P., Driver, S.P., Baldry, I.K., Bamford, S.P., Hopkins, A.M., Liske, J., Loveday, J., Merson, A., Peacock, J.A., Brough, S., Cameron, E., Conselice, C.J., Croom, S.M., Frenk, C.S., Gunawardhana, M., Hill, D.T., Jones, D.H., Kelvin, L.S., Kuijken, K., Nichol, R.C., Parkinson, H.R., Pimbblet, K.A., Phillipps, S., Popescu, C.C., Prescott, M., Sharp, R.G., Sutherland, W.J., Taylor, E.N., Thomas, D., Tuffs, R.J., van Kampen, E., Wijesinghe, D., 2011, Galaxy and Mass Assembly (GAMA): the GAMA galaxy group catalogue (G^3Cv1), MNRAS, 416, 2640

Rose, J.A., 1976, A catalogue of southern clusters of galaxies., A&AS, 23, 109

Roukema, B.F., Quinn, P.J., Peterson, B.A., Rocca-Volmerange, B., 1997, Merging history trees of dark matter haloes - A tool for exploring galaxy formation models, MNRAS, 292, 835

Rykoff, E.S., Rozo, E., Busha, M.T., Cunha, C.E., Finoguenov, A., Evrard, A., Hao, J., Koester, B.P., Leauthaud, A., Nord, B., Pierre, M., Reddick, R., Sadibekova, T., Sheldon, E.S., Wechsler, R.H., 2014, redMaPPer. I. Algorithm and SDSS DR8 Catalog, ApJ, 785, 104

Silk, J., Di Cintio, A., Dvorkin, I., 2013, Galaxy formation, ArXiv e-prints

Silk, J., Mamon, G.A., 2012, The current status of galaxy formation, Research in Astronomy and Astrophysics, 12, 917

Silk, J., Rees, M.J., 1998, Quasars and galaxy formation, A&A, 331, L1

Springel, V., 2005, The cosmological simulation code GADGET-2, MNRAS, 364, 1105

Springel, V., Yoshida, N., White, S.D.M., 2001, GADGET: a code for collisionless and gasdynamical cosmological simulations, New Ast, 6, 79

Tago, E., Saar, E., Tempel, E., Einasto, J., Einasto, M., Nurmi, P., Heinämäki, P., 2010, Groups of galaxies in the SDSS Data Release 7 . Flux- and volume-limited samples, A&A, 514, A102

Tarjan, R.E., van Leeuwen, J., 1984, Worst-case Analysis of Set Union Algorithms, J. ACM, 31, 2, 245

Tempel, E., Tamm, A., Gramann, M., Tuvikene, T., Liivamägi, L.J., Suhhonenko, I., Kipper, R., Einasto, M., Saar, E., 2014, Flux- and volume-limited groups/clusters for the SDSS galaxies: catalogues and mass estimation, A&A, 566, A1

Teyssier, R., 2002, Cosmological hydrodynamics with adaptive mesh refinement. A new high resolution code called RAMSES, A&A, 385, 337

Teyssier, R., Chapon, D., Bournaud, F., 2010, The Driving Mechanism of Starbursts in Galaxy Mergers, ApJ, 720, L149

THESIS

Thistleton, W., Marsh, J.A., Nelson, K., Tsallis, C., 2006, Generalized Box-Muller method for generating q-Gaussian random deviates, eprint arXiv:cond-mat/0605570

Tinker, J., Kravtsov, A.V., Klypin, A., Abazajian, K., Warren, M., Yepes, G., Gottlöber, S., Holz, D.E., 2008, Toward a Halo Mass Function for Precision Cosmology: The Limits of Universality, ApJ, 688, 709

Trasarti-Battistoni, R., 1998, Loose groups of galaxies in the Perseus-Pisces survey, A&AS, 130, 341

Tremonti, C.A., Heckman, T.M., Kauffmann, G., Brinchmann, J., Charlot, S., White, S.D.M., Seibert, M., Peng, E.W., Schlegel, D.J., Uomoto, A., Fukugita, M., Brinkmann, J., 2004, The Origin of the Mass-Metallicity Relation: Insights from 53,000 Star-forming Galaxies in the Sloan Digital Sky Survey, ApJ, 613, 898

Tsallis, C., 1988, Possible generalization of Boltzmann-Gibbs statistics, Journal of Statistical Physics, 52, 479

Tully, R.B., Fisher, J.R., 1978, Nearby small groups of galaxies, in: Longair, M.S., Einasto, J. (eds.), , Large Scale Structures in the Universe, vol. 79 of IAU Symposium, p. 31

Turner, E.L., Gott, III, J.R., 1976, Groups of galaxies. I. A catalog., ApJS, 32, 409

Tweed, D., Devriendt, J., Blaizot, J., Colombi, S., Slyz, A., 2009, Building merger trees from cosmological N-body simulations. Towards improving galaxy formation models using subhaloes, A&A, 506, 647

van den Bosch, F.C., 2002, The universal mass accretion history of cold dark matter haloes, MNRAS, 331, 98

von der Linden, A., Wild, V., Kauffmann, G., White, S.D.M., Weinmann, S., 2010, Star formation and AGN activity in SDSS cluster galaxies, MNRAS, 404, 1231

Wang, L., Steinhardt, P.J., 1998, Cluster Abundance Constraints for Cosmological Models with a Time-varying, Spatially Inhomogeneous Energy Component with Negative Pressure, ApJ, 508, 483

Warren, M.S., Abazajian, K., Holz, D.E., Teodoro, L., 2006, Precision Determination of the Mass Function of Dark Matter Halos, ApJ, 646, 881

Weinmann, S.M., Pasquali, A., Oppenheimer, B.D., Finlator, K., Mendel, J.T., Crain, R.A., Macciò, A.V., 2012, A fundamental problem in our understanding of low-mass galaxy evolution, MNRAS, 426, 2797

White, S.D.M., Rees, M.J., 1978, Core condensation in heavy halos - A two-stage theory for galaxy formation and clustering, MNRAS, 183, 341

Wickramasinghe, T., Ukwatta, T.N., 2010, An analytical approach for the determination of the luminosity distance in a flat universe with dark energy, MNRAS, 406, 548

Wojtak, R., Hansen, S.H., Hjorth, J., 2011, Gravitational redshift of galaxies in clusters as predicted by general relativity, Nature, 477, 567

Yang, X., Mo, H.J., van den Bosch, F.C., 2003, Constraining galaxy formation and cosmology with the conditional luminosity function of galaxies, MNRAS, 339, 1057

Yang, X., Mo, H.J., van den Bosch, F.C., 2009, Galaxy Groups in the SDSS DR4. III. The Luminosity and Stellar Mass Functions, ApJ, 695, 900

Yang, X., Mo, H.J., van den Bosch, F.C., Jing, Y.P., 2005, A halo-based galaxy group finder: calibration and application to the 2dFGRS, MNRAS, 356, 1293

Yang, X., Mo, H.J., van den Bosch, F.C., Pasquali, A., Li, C., Barden, M., 2007, Galaxy Groups in the SDSS DR4. I. The Catalog and Basic Properties, ApJ, 671, 153

Zandivarez, A., Díaz-Giménez, E., Mendes de Oliveira, C., Ascaso, B., Benítez, N., Dupke, R., Sodré, L., Irwin, J., 2014, Assessing the reliability of friends-of-friends groups on the

future Javalambre Physics of the Accelerating Universe Astrophysical Survey, A&A, 561, A71

Zehavi, I., Zheng, Z., Weinberg, D.H., Blanton, M.R., Bahcall, N.A., Berlind, A.A., Brinkmann, J., Frieman, J.A., Gunn, J.E., Lupton, R.H., Nichol, R.C., Percival, W.J., Schneider, D.P., Skibba, R.A., Strauss, M.A., Tegmark, M., York, D.G., 2011, Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity, ApJ, 736,

Zentner, A.R., 2007, The Excursion Set Theory of Halo Mass Functions, Halo Clustering, and Halo Growth, International Journal of Modern Physics D, 16, 763

Zwicky, F., Herzog, E., Wild, P., Karpowicz, M., Kowal, C.T., 1961, Catalogue of galaxies and of clusters of galaxies, Vol. I

THESE

OBSERVATOIRE DE PARIS

THESIS

Abstract

Galaxies lie in a large panel of environments from isolated galaxies, to pairs, groups or clusters. The environment is expected to have an impact on galaxy properties such as morphology, stellar formation, metallicity... Some studies already tried to quantify the importance of the global environment (linked to the dark matter halo mass) and the local environment (galaxy position in the group). These studies have shown that the environment plays a minor role except for low mass galaxies. But the quantification of the environment is difficult since detected groups in redshift space (the only one accessible by the observer) are very elongated, making it difficult to extract spherical groups in real space. If these quantification errors are too important, environment effects will not be measured correctly.

Moreover, other physical processes are at work inside groups whose relative roles are not well understood. For example, major or minor mergers (rich or poor in gas, between satellite galaxies, or after the decay of the orbit of a satellite onto the central galaxy by dynamical friction), rapid flybys harassing galaxies, stripping of the interstellar gas by ram pressure or of the gaseous reservoir by tidal forces. Although semi-analytical codes of galaxy formation from initial conditions of a ΛCDM Universe fit well a large set of observed relations, there are still some discrepancies that might be possibly explained by a lack of correct physical recipes of environmental effects in these models.

Our goal with this thesis is to have a detailed comprehension of the role of environment on galaxy properties, and finally determine the major physical processes in the modulation of these properties with both local and global environment. For this, an optimal extraction of galaxy groups from the projected phase space is necessary.

We performed a study and re-implementation of some existing group finder to estimate their strengths and weaknesses in the detection of galaxy groups.

A galaxy mock catalogue in redshift space, designed to mimic the primary spectroscopic sample of the SDSS survey was created to apply several galaxy group algorithms. An advantage is the already known membership that we can compare to galaxy groups extracted from redshift space. Semi-analytical codes of galaxy formation give us such galaxy catalogs we transformed to be coherent with the vision of an observer.

With these mock catalogues, we tested the very popular Friends-of-Friends grouping algorithm. We determined the optimal linking lengths against the set of tests and optimal criterion we developed to judge the efficiency of an algorithm. It appears that this choice of linking lengths depends on the scientific goal to do with the group catalogue.

A large part of the thesis consisted on the realization of a new grouping algorithm called MAGGIE (Models and Algorithm for Galaxy Groups, Interlopers and Environment), Bayesian and probabilistic. MAGGIE uses our priors acquired with analysis of cosmological simulations for large scale structure and of observations obtained from large galaxy surveys, to better constrain the selection of galaxy groups from redshift space. Comparison of MAGGIE with the FoF algorithm shows that MAGGIE is superior in avoiding the fragmentation of real space groups, the membership selection (completeness, reliability) and in the group properties (group mass, luminosity). The better performance of MAGGIE comes from its probabilistic nature, the use of astrophysical and cosmological priors, and the use of halo abundance matching technique linking central galaxy distributions (stellar mass or luminosity) to physical properties of dark matter halos.

The future application of MAGGIE on galaxy surveys such as the Sloan Digital Sky Survey or the deeper Galaxy and Mass Assembly, taking care of their own observational problems, should improve our understanding of the modulation of galaxy properties with their global and local environments and physical processes operating inside galaxy groups.

■