Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Γραφοθεωρητική Προσέγγιση του Κοσμικού Δικτύου

Διπλωματικη Εργασια

του

ΔΗΜΗΤΡΙΟΥ Κ. ΚΕΛΕΣΗ

Επιβλέπων: Δημήτρης Φωτάκης
Αν. Καθηγητής Ε.Μ.Π.

Εργαστηριο Επιστημης Υπολογιστων, Λογικης και Αλγοριθμων (CoReLab)
Αθήνα, 9 Ιουλίου 2020

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Επιστήμης Υπολογιστών, Λογικής και Αλγορίθμων (CoRe-Lab)

Γραφοθεωρητική Προσέγγιση του Κοσμικού Δικτύου

Διπλωματικη Εργασια

του

ΔΗΜΗΤΡΙΟΥ Κ. ΚΕΛΕΣΗ

Επιβλέπων: Δημήτρης Φωτάκης
Αν. Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 9 Ιουλίου 2020.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Δημήτρης Φωτάκης	Αριστείδης Παγουρτζής	Γιώργος Στάμου
Αν. Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2020

(Υπογραφή)

.....

Δημητρης Κ. Κελεσης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Επιστήμης Υπολογιστών, Λογικής και Αλγορίθμων (CoRe-Lab)

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κ. Δ.Φωτάκη για την συνεργασία μας, καθώς και τα όσα μου προσέφερε, όχι μόνο στα στενά πλαίσια της διπλωματικής. Αλλά κυρίως γιατί μου έδειξε πως είναι και πως πρέπει να αντιμετωπίζει κανείς την έρευνα και τις επιτυχίες, αλλά και τις αποτυχίες σε αυτή. Τον κ. Θ.Ρασσιά για την συνεργασία μας από το πρώτο έτος και το χρόνο που αφιέρωσε απαντώντας κάθε φορά σε οποιοδήποτε ερώτημα μου ή σχολιάζοντας κάθε πρόταση ή λύση που του πρότεινα σχετικά με ασκήσεις ή άλυτα προβλήματα, κατευθύνοντας με ταυτόχρονα διαρκώς. Αποτέλεσε σημαντικό παράδειγμα για μένα από την πρώτη στιγμή της γνωριμίας μας, αλλά και καθοριστικό παράγοντα για την μετέπειτα εξέλιξη μου. Τον κ. Σ.Βασιλάκο και στην ομάδα του European University of Cyprus (κ. Ευσταθίου, κ. Παπαδοπούλου) για την συνεργασία μας στο πλαίσιο της παρούσας διπλωματικής εργασίας, καθώς οι εποικοδομητικές συζητήσεις μας, αλλά και οι συμβουλές τους φάνηκαν ιδιαίτερα χρήσιμες. Τους κ. Γ.Παλιούρα και κ. Α.Αρτίκη για την συνεργασία, την εμπιστοσύνη και την ελευθερία κινήσεων που μου έδωσαν στον Δημόκριτο, σε όλη τη διάρκεια της συνεργασίας μας.

Θα ήθελα ακόμη να ευχαριστήσω τους φίλους μου για την υπομονή και την στήριξη τους αυτά τα χρόνια (αναφέρονται με τυχαία σειρά): Γιάννης Κ., Γιώργος Γ., Έλενα Θ., Δήμητρα Α., Ανδρέας Κ., Χάρης Π., Διονύσης Σ., Δανάη Ε., Γραμματική Τ., Γιάννης Σ., Μάριος Π., Κωνσταντίνος Σ., Δημήτρης Χ., Δημήτρης Ζ., Φωτεινή Κ., Δήμητρα Κ.

Θα ήθελα επίσης να ευχαριστήσω τον παππού μου Σπύρο και τις γιαγιάδες μου Μαρία και Σταματία για την διαρκή και απλόχερη προσφορά τους καθ΄όλη τη διάρκεια της ζωής μου τόσο στον υλιστικό τομέα, αλλά κυρίως στον ηθικοπλαστικό.
Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου Κωνσταντίνο και Παναγιώτα, αλλά και την αδελφή μου Μαρία για τη διαρκή στήριξη και τον καθοριστικό ρόλο που έπαιξαν στη διαμόρφωση του ανθρώπου που είμαι σήμερα.

Δημήτριος Κ. Κελέσης
Αθήνα, 07 Ιουνίου 2020

Μια αστραπή η ζωή μας... μα προλαβαίνουμε
Ν.Καζαντζάκης

Στους γονείς μου

Περίληψη

Η ανάλυση του κοσμολογικού δικτύου είναι ένα ενεργό πεδίο έρευνας με πολλές εφαρμογές. Ο απώτατος σκοπός του είναι η κατανόηση των βαθιά ριζωμένων φυσικών νόμων και δυνάμεων που διέπουν την δημιουργία και την εξέλιξη του από την απαρχή του χρόνου. Πολλές μέθοδοι και τεχνικές έχουν προταθεί για να εξαχθούν τα ιδιαίτερα χαρακτηριστικά σχετικά με τους γαλαξίες, τα διαγαλαξιακά οικοδομήματα και τις σχέσεις μεταξύ των γαλαξιών. Σε αυτή τη διπλωματική εργασία θα προσπαθήσουμε να μελετήσουμε το κοσμικό δίκτυο από μία διαφορετική προοπτική που δεν είναι τόσο κοντά στον τρόπο προσέγγισης από τους αστροφυσικούς. Θα προσπαθήσουμε να συνδυάσουμε την θεωρία γραφημάτων και τα εργαλεία της με το κοσμικό δίκτυο, έτσι ώστε να μοντελοποιήσουμε την κατασκευή και διάρθρωση του σύμπαντος με τέτοιο τρόπο ώστε να μπορέσουμε να προσφέρουμε μία ποικιλία χαρακτηριστικών στους αστροφυσικούς για παραπάνω μελέτη. Στο ίδιο σκεπτικό, θα εφαρμόσουμε επίσης μεθόδους τοπολογικής κατηγοριοποίησης τροποποιημένες κατάλληλα ώστε να ταιριάζουν στην φύση των κοσμολογικών δεδομένων ώστε να μπορέσουμε να τα ερμηνεύσουμε σε μία πιο ενοποιημένη μορφή. Σε αυτή την εργασία προτείνουμε και εφαρμόζουμε κυρίως καινοτόμες ιδέες ώστε να προσεγγίσουμε το πρόβλημα, καθώς σχεδόν όλοι οι αλγόριθμοι, ακόμη και όσοι είχαν οριστεί από άλλους, τροποποιήθηκαν καταλλήλως στο πλαίσιο του ενδιαφέροντος μας. Κάποιες από τις ιδέες φαίνονται υποσχόμενες αν γίνουν περαιτέρω τροποποιήσεις, ενώ άλλες έχουν ήδη εξάγει ενδιαφέροντα αποτελέσματα. Σε αυτή τη διπλωματική εργασία παρουσιάζουμε την χρήση και τα αποτελέσματα από εργαλεία στην θεωρία γραφημάτων εφαρμοσμένα στο κοσμικό δίκτυο, όπως το Gravity Lattice και το Gravity Fields. Επιπλέον, παρουσιάζονται τα αποτελέσματα από τη χρήση αλγορίθμων χωρικής κατηγοριοποίησης, όπως το Gravity Lattice Filtering και ο τροποποιημένος αλγόριθμος clustering ABACUS. Κύρια συνεισφορά της παρούσας εργασίας είναι η δημιουργία μοντέλων που ενσωματώνουν γραφοθεωρητικές και αστροφυσικές γνώσεις ώστε να προσεγγίσουν το κοσμικό δίκτυο αποτελεσματικά και γρήγορα, εξοικονομώντας έτσι υπολογιστικό χρόνο.

Λέξεις Κλειδιά

Κοσμολογικό Δίκτυο, Σχαλε φρεε δίκτυα, Ιεραρχική ομαδοποίηση, Graphons, Βαρύτητα, Γαλαξίες, Πεδία, Χωρική ταξινόμηση, Πλέγμα, Filtering, GANs

Abstract

Analysis of the Cosmic Web is an active field of research with a wide variety of applications. The ultimate purpose of it is to understand the deeply rooted forces and laws that govern the creation and evolution of the universe since its beginning. Many methods have been proposed in order to extract features about the galaxies, their structures and the relation among them. On this thesis we will try to study the cosmic web from a different perspective not so close to the astrophysical one. We will attempt to combine graph theory and its tools with the cosmic web in order to model the universal formation in a way that it will offer a variety of new features to the astrophysicists. On the same manner we will also apply spatial categorization methods altered in a way to fit the cosmological data in order to interpret them in a more compact way. Throughout this thesis we mainly suggest and apply novel methods of approximating the problem as almost every algorithms used, even the predefined ones, are altered in the context of interest. Some of these approaches seem promising if further modifications are made, while others have already made interesting results. Throughout this thesis we present the usage and results of graph tools in cosmic web like Gravity Graphons and Gravity Fields. Moreover we present the results from using spatial categorization algorithms like Gravity Lattice Filtering and modified clustering algorithm ABACUS. The main contribution of this thesis is the creation of models that integrate graphical and astrophysical knowledge in order to approach the cosmic web fast end effectively, saving computational time from supercomputers.

Keywords

Cosmic Web, Scale Free Networks, Hierarchical clustering, Graphons, Gravity, Galaxies, Fields, Spatial Clustering, Lattice, Filtering, GANs

Contents

Εκτενής Περίληψη στα Ελληνικά

Σε αυτό το κεφάλαιο θα παρουσιάσουμε εκτενώς την εργασία που εκπονήσαμε στο πλαίσιο της παρούσας διπλωματικής εργασίας μέσα από μία εκτεταμένη περίληψη στα Ελληνικά.

Συνεισφορά

Η κύρια συνεισφορά της παρούσας εργασίας εντοπίζεται στην παροχή μοντέλων με σκοπό την προσέγγιση του κοσμικού δικτύου. Το πεδίο έρευνας της κοσμολογίας αν και ιδιαίτερα ενεργό δεν προσφέρει κάποια αντικειμενική συνάρτηση μέτρησης της επιτυχίας κάθε μοντέλου στην προσέγγιση των ιδιοτήτων και των κατασκευών του κοσμολογικού δικτύου. Άμεσες προεκτάσεις και επόμενα βήματα μας θα αποτελέσει η εξαγωγή και οριστικοποίηση μίας τέτοιας αντικειμενικής συνάρτησης. Στο πλαίσιο της παρούσας εργασίας παρέχουμε γραφικά μοντέλα που ενσωματώνουν αλγοριθμικές έννοιες ώστε να προσεγγίσουμε το σύμπαν. Το κύριο πλεονέκτημα μας σε σχέση με τις υπόλοιπες εργασίες στο χώρο βρίσκεται στο γεγονός ότι οι προσεγγίσεις μας είναι αποδεδειγμένα γρήγορες και τα αποτελέσματα τους αρκετά ικανοποιητικά. Χρήση των μεθόδων μας θα οδηγούσε σε εξοικονόμηση υπολογιστικού χρόνου που απαιτείται υπό τις υπάρχουσες συνθήκες για την εξαγωγή συμπερασμάτων σχετικά με τις γαλαξιακές κατασκευές και την υφή του σύμπαντος. Τέλος η εργασία μας διαφοροποιείται από εκείνη των Barabási et.el. [32] καθώς εμείς πραγματοποιούμε και υπολογιστικές διαδικασίες πάνω στα προκύπτοντα γραφήματα.

Θεωρητικό Υπόβαθρο

Στα πλαίσια αυτής της ενότητας θα παρουσιάσουμε το απαιτούμενο από τον αναγνώστη θεωρητικό υπόβαθρο ώστε να μπορέσει να παρακολουθήσει στη συνέχεια τις έννοιες καθώς και τις μεθόδους προσέγγισης του προβλήματος της ανάλυσης του κοσμικού δικτύου. Η απαιτούμενες θεωρητικές γνώσεις χωρίζονται σε δύο κατηγορίες, τις γραφοθεωρητικές και τις σχετικές με την χωρική ταξινόμηση.

Γραφοθεωρητικές Γνώσεις

Σε ότι αφορά τις γραφοθεωρητικές γνώσεις απαιτούμενη είναι η εξοικείωση με τα scale free δίκτυα καθώς και τα small world δίκτυα. Σχετικά με τα παραπάνω απαιτείται η γνώση

σχετικά με την δομή και τα ιδιαίτερα χαρακτηριστικά γνωρίσματα τους. Συγκεκριμένα τα δίκτυα αυτού του είδους χαρακτηρίζονται από μία ειδική συνάρτηση που προσεγγίζει σε πολύ ικανοποιητικό βαθμό τη κατανομή των βαθμών των κόμβων τους. Αυτή δίνεται ως:

$$\mathrm{P(k)} = k^{-\gamma}$$

όπου το γ είναι η εκθετική παράμετρος και καθορίζει την συμπεριφορά του δικτύου κατά κύριο λόγο. Επίσης κρίσιμο στοιχείο ενός δικτύου είναι η ικανότητα μας να μπορέσουμε να διακρίνουμε κοινότητες εντός αυτού. Με τον όρο κοινότητες αναφερόμαστε σε μία ομαδοποίηση των κόμβων του δικτύου ώστε κόμβοι που ανήκουν στην ίδια ομάδα να μοιάζουν περισσότερο μεταξύ τους παρά με κόμβους διαφορετικής ομάδας. Για να καταφέρουμε να κάνουμε την παραπάνω ομαδοποίηση υπάρχουν πολλές μέθοδοι αλλά στο πλαίσιο της ανάλυσης του κοσμικού δικτύου θα χρησιμοποιήσουμε κυρίως την ιεραρχική ομαδοποίηση. Επιλέγουμε το παραπάνω καθώς θεωρούμε με βάση τις γνώσεις μας και την αντίληψη μας για το φυσικό κόσμο ότι υπάρχει μίας μορφής ομοιομορφία ανεξάρτητα από το επίπεδο ανάλυσης που επιλέγουμε να δούμε τον κόσμο.

Απαραίτητη για την ακόλουθη ανάλυση των μεθόδων μας σχετικά με τα κοσμικά δίκτυα είναι η γνώση μίας μορφής ορίου γραφικής ακολουθίας η οποία καλείτε Graphon. Αυτού του είδους το όριο γραφικής ακολουθίας έχει την ιδιότητα να μπορεί να προσεγγίζει οριακά το ν-οστό στοιχείο μας ακολουθίας γραφημάτων και να έχει τα ίδια ή περίπου τα ίδια χαρακτηριστικά σχετικά με αυτό. Ορίζεται ως μία συνάρτηση στο $[0, 1]^2$ που κάνει αντιστοίχηση στο [0, 1]. Συγκεκριμένα για να πραγματοποιηθεί αυτό σε κάθε κόμβο ανατίθεται μία τιμή που έχει επιλεγεί τυχαία από μία ομοιόμορφη κατανομή και επιλέγεται μία συνάρτηση W ως συνάρτηση του Graphon. Εν τέλει για την τελική απόφαση σχετικά με την ύπαρξη ακμής μεταξύ κορυφών του γραφήματος αυτή γίνεται πιθανοτικά με την πιθανότητα ύπαρξης ακμής ανάμεσα σε δύο κορυφές με τιμές α, β να δίνεται από την τιμή W(α, β).

Κοσμολογικές Γνώσεις

Σχετικά με την δομή του σύμπαντος απαιτούνται κάποιες βασικές γνώσεις ώστε να μπορέσει ο αναγνώστης να κατανοήσει τους λόγους που μας οδήγησαν στην επιλογή κάποιων παραμέτρων ή και στην αγνόηση κάποιων άλλων. Η κοσμολογία ως επιστήμη μελετά την γέννηση, την εξέλιξη και την δομή του σύμπαντος καθώς επίσης και τις δυνάμεις και τις αιτίες που οδήγησαν σε αυτή την εξέλιξη και διέπουν την πορεία του καθολικά. Χρονολογικά το σύμπαν ξεκίνησε με την μεγάλη έκρηξη και η διαδρομή μέχρι την σημερινή του μορφή περιλάμβανε τον διαχωρισμό των τεσσάρων βασικών δυνάμεων, με τη βαρύτητα να είναι η πρώτη που αποσχίστηκε από τις υπόλοιπες τέσσερις, καθώς και τη δημιουργία των πρώτων πυρήνων, ατόμων αλλά και χημικών στοιχείων. Στη μακραίωνη πορεία του το σύμπαν χρειάστηκε πολλά χρόνια μέχρις ότου να φτάσει στο σημείο να δημιουργήσει τις πρώτες κατασκευές, αστέρια, γαλαξίες, και φυσικά ακόμη περισσότερα χρόνια μέχρι να φτάσει στο σημείο να δημιουργήσει

τα μεγάλα οικοδομήματα που πλέον είναι ορατά σε αυτό. Σε όλες τις παραπάνω διαδικασίες σημαντικό ρόλο, ίσως τον πιο σημαντικό έπαιξε αφενός η υπεροχή της ύλης έναντι της αντι-ύλης που οδήγησε στην ύπαρξη του σύμπαντος και όχι ενός άντι-σύμπαντος΄ αλλά μετά από αυτό η βαρυτική δύναμη και η αλληλεπίδραση της με την ύλη. Η ύλη εκ φύσεως οδηγούνταν σε εσωτερική κατάρρευση κάτω από τις δυνάμεις της βαρύτητας, γεγονός που ήταν ο κυρίαρχος παράγοντας που οδήγησε στην δημιουργία μεγάλων σωμάτων όπως τα αστέρια, στη συνέχεια ο ρόλος της βαρύτητας εμφανίζεται σημαντικότατος στην ομαδοποίηση των αστεριών σε γαλαξίες και των γαλαξιών σε σμήνη καθώς όλες οι παραπάνω διαδικασίες έλαβαν τόπο κυρίως λόγω της βαρύτητας και της επιρροής της πάνω στην ύλη.

Οι κοσμολογικές κατασκευές που θα επιχειρήσουμε να παρατηρήσουμε και να ταξινομήσουμε διακρίνονται κυρίως σε δύο μεγάλες κατηγορίες: τα κενά (περιοχές με ελάχιστους η καθόλου γαλαξίες που έχουν συνήθως σφαιρικό σχήμα) και τα τείχη (περιοχές που αποτελούνται από γαλαξίες) και χωρίζονται επιμέρους σε σμήνη γαλαξιών, με μεγάλη πυκνότητα μάζας και γαλαξίες σε κοντινή απόσταση και σε filaments που αποτελούν μακρόστενες περιοχές που έχουν γαλαξίες σε πυκνότητες μικρότερες από τα σμήνη και αποτελούν ουσιαστικά τις αρτηρίες που ενώνουν διαφορετικά σμήνη γαλαξιών, αλλά και τα όρια για τις κενές περιοχές. Κάθε μία από τις τρεις παραπάνω κατηγορίες έχει τα δικά της εγγενή χαρακτηριστικά που την ξεχωρίζουν από τις υπόλοιπες σχετικά με την τυπική μάζα, ακτίνα και σχήμα.

Σε ότι αφορά τα δεδομένα που θα χρησιμοποιήσουμε σε αυτή τη διπλωματική εργασία αυτά θα αντληθούν από δύο προσομοιώσεις που είναι ευρέως διαδεδομένες και γνωστές στον χώρο ως οι πιο επιτυχημένες και αποτελούν την κύρια πηγή δεδομένων για όλους τους ερευνητές που θέλουν να δοκιμάσουν τις ιδέες τους στο χώρο. Αυτές έγιναν με την χρήση του εργαλείου GADGET-2 και είναι οι: Millennium Simmulation II και Illustris Project. Στο πρώτο μέρος της εργασίας μας χρησιμοποιούμε την πρώτη ενώ από ένα σημείο και μετά επιλέξαμε να χρησιμοποιήσουμε μόνο τα δεδομένα της δεύτερης καθώς όπως αναγράφεται σε σχετικές δημοσιεύσεις προσφέρουν καλύτερη ποιότητα και πληθώρα πληροφοριών. Παράλληλα ο τρόπος προσομοίωσης στην δεύτερη περίπτωση εμπεριέχει περισσότερα από τα φυσικά χαρακτηριστικά που καθόρισαν την τελική μορφή του σύμπαντος. Για λόγους παρουσίασης στα πλαίσια της διπλωματικής χρησιμοποιήσαμε δεδομένα χαμηλής ανάλυσης από την δεύτερη προσομοίωση όπως έχει γίνει και σε άλλες εφαρμογές σύμφωνα με την βιβλιογραφία μας.

Η ομαδοποίηση των δεδομένων είναι απαραίτητη στο πλαίσιο του κοσμικού δικτύου καθώς όπως παρουσιάσαμε στο παραπάνω θεωρητικό υπόβαθρο σκοπός μας θα είναι να εντοπίσουμε αποτελεσματικά τα διάφορα είδη περιοχών στο κοσμικό δίκτυο και να τα κατηγοριοποιήσουμε. Για να το κάνουμε αυτό θα χρησιμοποιήσουμε μίας μορφής ταξινόμηση που ονομάζεται χωρική ταξινόμηση. Για να καταλήξει σε αποτελέσματα χρησιμοποιεί την θέση των δεδομένων σε κάποιο χώρο διαστάσεων και την μεταξύ τους απόσταση ορισμένη σύμφωνα με κάποια μετρική. Έτσι καταφέρνει να χωρίσει ομάδες δεδομένων τα στοιχεία των οποίων είναι αρκετά κοντά μεταξύ τους σύμφωνα με τη μετρική αυτή αλλά ταυτόχρονα και μακριά από τα στοιχεία άλλων ομάδων. Στον κλάδο αυτό έχουν αναπτυχθεί πολλοί αλγόριθμοι και εμείς θα κάνουμε χρήση κάποιων από αυτούς με κατάλληλες τροποποιήσεις ώστε να ταιριάζουν στο πλαίσιο του ενδιαφέροντος μας.

Προσέγγιση του Κοσμολογικού Δικτύου με χρήση Γραφικής Θεωρίας

Για την προσέγγιση του δικτύου με χρήση γραφικής θεωρίας χρησιμοποιήσαμε μεθόδους που παρουσιάστηκαν προηγουμένως με κατάλληλες τροποποιήσεις.

Βαρυτικά Graphons

Πραγματοποιήσαμε μία τροποποίηση της θεωρίας σχετικά με τα Graphons ώστε να ανταποκριθούμε στις απαιτήσεις του κοσμικού δικτύου. Συγκεκριμένα αναθέσαμε σε κάθε γαλαξία που αποτελεί κορυφή στο κοσμικό δίκτυο ως τιμή του την μάζα του και στη συνέχεια ως πιθανότητα ύπαρξης ακμής ανάμεσα σε δύο γαλαξίες (έστω α και β) δώσαμε την τιμή που προέκυψε από το softmax(gravity(α, β», δηλαδή η πιθανότητα ύπαρξης ακμής μεταξύ δύο γαλαξιών εξαρτάται άμεσα και σε μεγάλο βαθμό από το μέγεθος της βαρυτικής δύναμης που ασκείται ανάμεσα τους. Μία ακόμη τροποποίηση για την βελτίωση της παραπάνω μεθόδου είναι η δημιουργία γραφήματος κοντινότερων γειτόνων και η εφαρμογή της παραπάνω μεθόδου μόνο στους κοντινότερους γείτονες όπως αυτοί θα προκύψουν από το γράφημα. Η παραπάνω μέθοδος αποτελεί μία μέθοδο που είναι καινούρια και όσο γνωρίζουμε δεν έχει εφαρμοστεί ξανά στο πρόβλημα προσέγγισης του κοσμικού δικτύου. Επίσης προτείνεται μία ακόμη καινοτόμα ιδέα που θα συμπεριλάβει περισσότερους νόμους εκτός της βαρύτητας μέσω ενός σταθμισμένου αθροίσματος των δυνάμεων που προκύπτουν από την εφαρμογή των νόμων αυτών.

Γάμμα Ιεραρχική Ανάλυση

Σε αυτό το σημείο προτείνουμε μία καινοτόμα προσέγγιση για την ανάλυση του κοσμικού δικτύου σε επίπεδα. Υποθέτουμε ότι σε κάθε επίπεδο ανάλυσης του δικτύου αυτό παρουσιάζει χαρακτηριστικά scale free δικτύου και άρα κάθε επίπεδο ανάλυσης θα διέπεται από κάποιο χαρακτηριστικό γ, όπου γ είναι η εκθετική παράμετρος της κατανομής βαθμών κόμβων. Θεωρούμε επίσης ότι σε κάθε επίπεδο ανάλυσης αν και θα υπάρχουν διαφορετικές τιμές για το γ σε κάθε συνεκτική συνιστώσα αυτές θα έχουν μικρή διακύμανση από την μέση τιμή τους. Για το λόγο αυτό θεωρούμε την μέση αυτή τιμή ως αντιπρόσωπο για το επίπεδο ανάλυσης. Για το λόγο αυτό πραγματοποιούμε μία διαδικασία αφαίρεσης ακμών με σκοπό να εισχωρήσουμε σε βαθύτερα επίπεδα ανάλυσης του κοσμικού δικτύου και στη συνέχεια επιχειρούμε να ανακατασκευάσουμε το κοσμικό δίκτυο ακολουθώντας την αντίστροφη πορεία και χρησιμοποιώντας τα γ που εξαγάγουμε σε κάθε επίπεδο ανάλυσης. Για να επιτύχουμε υψηλή ταχύτητα εκτέλεσης στον παραπάνω αλγόριθμο χρησιμοποιούμε μέγιστη πιθανοφάνεια για την εκτιμηση του γ ανά επίπεδο. Επίσης επιλέγουμε ποιες ακμές θα αφαιρέσουμε με γνώμονα την ελαχιστοποίηση της διακύμανσης του γ σε κάθε επίπεδο.

Βαρυτικά Πεδία

Στην μέθοδο αυτή χρησιμοποιούμε την έννοια του πεδίου. Για να ορίσουμε την έννοια αυτή θεωρούμε ότι κάθε γαλαξίας δρα σαν πηγή που εκπέμπει βαρυτικό πεδίο και αν δύο πεδία

υπεισέρχονται το ένα μέσα στο άλλο τότε οι πηγές αυτών θα πρέπει να ενώνονται με ακμή. Για να το πραγματοποιήσουμε αυτό χρησιμοποιούμε τον αλγόριθμο:

Algorithm 1 Gravity Fields

Result: Gravity Field Network's Edges

```
M = all masses
P = all positions
HM = all half mass radii
edges =[]
a = 1e-32
for pair in all pairs do
    i, j = pair
    dist = euclidean(i, j)
    rad sum = M[i] * HM[i] + M[j] * HM[j]
    grav = gravity force(i, j)
    gravity = a * grav
    if gravity * rad sum >= dist then
        edges.append([i, j])
    end
end
return edges
```

Schwarzschild Ακτίνα

Για την παραπάνω προσέγγιση χρησιμοποιούμε ως ακτίνα επιρροής κάθε γαλαξία την ακτίνα που ορίζεται ως Schwarzschild radius και είναι στενώς συνδεδεμένη με την μάζα του κάθε γαλαξία καθώς αν ο γαλαξίας συμπιεστεί σε ακτίνα μικρότερη αυτής τότε μετατρέπεται σε μαύρη τρύπα. Χρησιμοποιώντας αυτή αν δύο γαλαξίες βρίσκονται σε απόσταση μικρότερη από κάποιο πολλαπλάσιο του αθροίσματος των ακτινών τους τότε θα ενώνονται με ακμή.

Γενετικός Αλγόριθμος

Δοκιμάσαμε με χρήση γενετικού αλγορίθμου να ανακαλύψουμε την σχέση που θα καθορίζει αν θα υπάρχει ακμή μεταξύ δύο γαλαξιών. Για να το κάνουμε αυτό ορίσαμε μία γενική σχέση που μοιάζει με τη βαρύτητα αλλά δίνει τη δυνατότητα τόσο στις μάζες όσο και στην απόσταση μεταξύ των γαλαξιών να εμφανιστούν σε μεγαλύτερες δυνάμεις. Πραγματοποιήσαμε δειγματοληψία στους γαλαξίες που είχαμε στην διάθεση μας και με κριτήριο υπεροχής του κάθε χρωμοσώματος (που ουσιαστικά όριζε μία διαφορετική συνάρτηση σχέσης μεταξύ των γαλαξιών) την δημιουργία scale free δικτύου δοκιμάσαμε να εφαρμόσουμε έναν γενετικό αλγόριθμο. Τα αποτελέσματα δεν έμοιαζαν σε μεγάλο βαθμό με τον τύπο της βαρύτητας αλλά αυτό πιθανώς οφείλεται σε μικρό δείγμα κατά τη δειγματοληψία.

Προσέγγιση του Κοσμολογικού Δικτύου με χρήση Χωρικής ταξινόμησης

Σε αυτή την ενότητα θα παρουσιάσουμε τις μεθόδους ταξινόμησης και εύρεσης των μεγάλων κατασκευών στο κοσμικό δίκτυο με χρήση των μεθόδων χωρικής ταξινόμησης.

Octree

Η παρούσα μέθοδος στηρίζεται στην διαρκή και επαναλαμβανόμενη διαίρεση των περιοχών ενός κύβου σε μικρότερους υπο-κύβους. Αυτό ήταν χρήσιμο στο πλαίσιο της τοπολογικής ταξινόμησης για την εξαγωγή των κατασκευών του σύμπαντος. Αφού δημιουργήσουμε την δενδροειδή απεικόνιση του κύβου με τους υπο-κύβους στους ενδιάμεσους και στους τελικούς κόμβους εισάγουμε τους γαλαξίες. Έτσι τελικά κάθε κύβος θα έχει μάζα ίση με την συνισταμένη των μαζών των γαλαξιών που ανήκουν σε αυτόν και ο όγκος του είναι καθορισμένος γεωμετρικά από τον τρόπο κατασκευής του. Έτσι μπορούμε άμεσα να εξάγουμε την πυκνότητα του και με βάση αυτή να καθορίσουμε τι είδους κατασκευή υπάρχει εντός του κύβου. Φυσικά στην δενδροειδή απεικόνιση των διαδοχικών επιπέδων από κύβους μπορεί να εφαρμοστεί κάποιου είδους ΄κλάδεμα΄ με γνώμονα την πυκνότητα ώστε τα τελικά φύλλα του δέντρου να έχουν πυκνότητα μεγαλύτερη από κάποιο καθορισμένο όριο. Τέλος θα ήταν πιο βολική η χρήση άλλου γεωμετρικού στερεού και ειδικότερα κυλίνδρων με μεταβλητό ύψος και ακτίνα αλλά η προσέγγιση αυτή ενέχει πολλές δυσκολίες αναφορικά με την υπολογιστική γεωμετρία που απαιτείται ώστε ο αρχικός κύβος να γεμίσει με κατάλληλου μεγέθους μικρότερους κυλίνδρους.

Βαρυτικό Πλέγμα

Σε αυτή τη προσέγγιση χρησιμοποιούμε μία πλεγμοειδή κατασκευή χωρίς ακμές στον αρχικό τρισδιάστατο κύβο ώστε αυτός να γεμίσει. Σε κάθε κορυφή του πλέγματος τοποθετούμε δοκιμαστικά φορτία μάζας 1kg και εισάγουμε διαδοχικά τους γαλαξίες. Για κάθε γαλαξία έχουμε ορίσει μία ακτίνα επιρροής του και για όσες από τις δοκιμαστικές σφαίρες βρίσκονται εντός αυτής της ακτίνας υπολογίζουμε την βαρυτική δύναμη που τους ασκεί ο γαλαξίας αυτός. Έχοντας επαναλάβει την παραπάνω διαδικασία για όλους τους γαλαξίες τους αφαιρούμε από τον κύβο καθώς θεωρούμε ότι έχουν πλέον αφήσει το βαρυτικό τους αποτύπωμα. Στη συνέχεια είτε μετακινούμε στους τρεις άξονες τις δοκιμαστικές σφαίρες κατά απόσταση ανάλογη της συνισταμένης βαρυτικής δύναμης που αυτές έχουν δεχθεί είτε για να έχουμε καλύτερα αποτελέσματα μετράμε πόσοι γαλαξίες επηρεάζουν κάθε σφαίρα και αυτό θα είναι το αναγνωριστικό της. Στη συνέχεια φιλτράρουμε τις δοκιμαστικές σφαίρες με βάση αυτό το χαρακτηριστικό. Έτσι στο τέλος μπορούμε να απομονώσουμε τις μεγάλες κατασκευές του σύμπαντος ενώ ταυτόχρονα να αποκλείσουμε να εμφανιστούν οι κενές περιοχές ή οι περιοχές με μικρή πυκνότητα γαλαξιών.

HDBSCAN

Εφαρμόσαμε τον γνωστό αλγόριθμο HDBSCAN τοπολογικής ταξινόμησης για τα δεδομένα του κοσμολογικού δικτύου. Για να προσαρμόσουμε την εφαρμογή τους καθώς και για να δώσουμε τιμές στις παραμέτρους του που να έχουν άμεση σχέση με τα φυσικά παρατηρούμενα γεγονότα στο σύμπαν αξιοποιήσαμε τις παρατηρήσεις. Έτσι ορίσαμε το ελάχιστο πλήθος σημείων για μία ομάδα σύμφωνα με τον τυπικό αριθμό γαλαξιών σε ένα σμήνος, το θόρυβο που θα αποκλείσει ο αλγόριθμος σύμφωνα με το ποσοστό των κενών περιοχών που υπάρχουν στο σύμπαν σύμφωνα με τις παρατηρήσεις και την ελάχιστη απόσταση κάτω από την οποία δεν θα πρέπει να διαχωρίζονται τα μέλη εντός της ίδια ομάδας σύμφωνα με την τυπική ακτίνα ενός σμήνους γαλαξιών. Όλα τα παραπάνω έγιναν σε κλίμακα τέτοια ώστε να λάβουμε υπόψη την αναλογία και το μέγεθος των δεδομένων της προσομοίωσης σε σχέση με τα πραγματικά δεδομένα και το πραγματικό μέγεθος του κοσμολογικού δικτύου.

ABACUS

Χρησιμοποιήσαμε τον αλγόριθμο ABACUS για την ταξινόμηση των γαλαξιών αναλόγως σε ποιες από τις μεγάλες κατηγορίες ανήκουν καθώς μας προσέφερε την δυνατότητα να αποβάλουμε τον θόρυβο (στα πλαίσια του κοσμολογικού δικτύου μπορούμε να θεωρήσουμε θόρυβο τους γαλαξίες σε κενές περιοχές). Ταυτόχρονα αυτή η συμπύκνωση πληροφορίας μας επέτρεψε να ξεχωρίσουμε κοινότητες και κυρίως μας έδωσε το έναυσμα να προχωρήσουμε σε κατάλληλο φιλτράρισμα. Ειδικότερα μέρος του αλγορίθμου είναι η συμπύκνωση των σημείων κάτω από κάποιον εκπρόσωπο και αυτός ο εκπρόσωπος έχει βάρος ίσο με το πλήθος των σημείων που αντιπροσωπεύει. Έτσι οδηγηθήκαμε σε φιλτράρισμα με βάση αυτή τη παράμετρο που μας επέτρεψε να ξεχωρίσουμε περιοχές στις οποίες υπήρχαν σμήνη ή γενικότερα μεγάλης κλίμακας κατασκευές από περιοχές κενού ή μικρής πυκνότητας. Ορμώμενοι από την εμπειρία και την γνώση μας σχετικά με την σπουδαιότητα της βαρύτητας τροποποιήσαμε τον παραπάνω αλγόριθμο ώστε κάθε εκπρόσωπος πλέον να έχει ως χαρακτηριστικό βάρος τη συνολική μάζα των γαλαξιών που αντιπροσωπεύει και στη συνέχεια φιλτράρουμε και πάλι καταλλήλως. Τα αποτελέσματα στη δεύτερη περίπτωση φάνηκαν πιο υποσχόμενα σε σχέση με αυτά της πρώτης καθώς επέτρεψαν πιο λεπτομερές φιλτράρισμα που ανέδειξε τις πιο πυκνές περιοχές έναντι των υπολοίπων ενώ παράλληλα έδωσε τη δυνατότητα να απομονωθούν περιοχές που αντιπροσωπεύουν ένα συγκεκριμένο εύρος τιμών συνισταμένης μάζας γαλαξιών.

CHAMELEON και CURE

Για λόγους πληρότητας και καθώς θεωρητικά πιστεύαμε αρχικά ότι θα οδηγήσουν σε ενδιαφέροντα αποτελέσματα εφαρμόσαμε και τους αλγορίθμους CHAMELEON και CURE. Εν τέλει παρατηρήσαμε ότι τα αποτελέσματα τους κάθε άλλο παρά υποσχόμενα ήταν αλλά για λόγους πληρότητας παρουσιάζουμε και αυτά.

Προσέγγιση με χρήση GANs

Προσεγγίσαμε το πρόβλημα με την χρήση νευρωνικών δικτύων και συγκεκριμένα GANs. Τα παραπάνω αποτελούν σημαντικό τμήμα της μηχανικής μάθησης και έχουν λάβει άνθιση τα τελευταία χρόνια. Λόγω των αποτελεσμάτων τους αλλά και της ποιότητας αυτών δοκιμάσαμε να τα χρησιμοποιήσουμε για να προβλέψουμε την μελλοντική εξέλιξη του κοσμικού δικτύου. Συγκεκριμένα εφαρμόσαμε κωδικοποίηση του αρχικού κύβου του δικτύου ΄κόβοντας΄ τον σε ΄φέτες΄ κατά μήκος και των τριών αξόνων και εικονοποιώντας τις φέτες αυτές ως είσοδο για τα νευρωνικά δίκτυα. Σκοπός μας ήταν μέσω μίας διαδικασίας μάθησης σε επίπεδα να εκπαιδεύσουμε τα νευρωνικά δίκτυα ώστε να μάθουν τον τρόπου που πραγματοποιείται η μετάβαση από το παρελθόν στο παρόν σε ότι αφορά το κοσμολογικό δικτύου. Δηλαδή αν και πολύ τολμηρό το εγχείρημα μας ήταν να μπορέσουμε να κάνουμε το νευρωνικό δίκτυο να μάθει τους νόμους της φυσικής ακόμη και κάποιους που μπορεί να μην καταλαβαίνουμε προς το παρόν ώστε να μας μεταφέρει στο μέλλον. Σε όλα τα παραπάνω υποθέσαμε ότι η μετάβαση από τη μία στην άλλη χρονική στιγμή στη διάρκεια εξέλιξης του σύμπαντος γίνεται με τρόπο ομοιόμορφο που μπορεί να τον μάθει κάποιο νευρωνικό δίκτυο. Έτσι το ζευγάρι των νευρωνικών δικτύων θα ξεκινούσε αρχικά με το ένα από αυτά να έχει εκπαιδευτεί σε κάποια χρονική στιγμή (τ1) και το άλλο να προσπαθεί γνωρίζοντας μόνο μία παρελθοντική στιγμή (τ2) να ξεγελάσει το πρώτο παράγοντας εικόνες που θα μοιάζουν με αυτές της χρονικής στιγμής τ1. Στη συνέχεια αφού γίνει αυτό οι χρονικές στιγμές θα μετατεθούν στο μέλλον και το παραπάνω θα επαναλαμβάνεται μέχρις ότου φτάσουμε στο παρόν όπου και θα ζητήσουμε από το πρώτο, πλήρως εκπαιδευμένο σε αυτό το σημείο, νευρωνικό δίκτυο να μας μεταφέρει στο μέλλον.
Η κωδικοποίηση των δεδομένων μπορεί να γίνει και με χρήση πίνακα που θα περιέχει για κάθε γαλαξία κάποιες από τις βασικές του πληροφορίες. Τέλος χρησιμοποιήσαμε το CycleGAN, ενός είδους ζευγάρι νευρωνικών δικτύων για μετάφραση εικόνων σε εικόνες, δηλαδή εικόνων ενός πεδίου (π.χ. άλογα) που μπορούν να μεταφραστούν σε εικόνες ενός άλλου παρεμφερούς πεδίου (π.χ. ζέβρες) εφόσον τα πεδία συνδέονται με κάποιον τρόπο και υπάρχει μία διαδικασία για την μετατροπή εικόνων του ενός στο άλλο (π.χ. ρίγες). Το παραπάνω ζευγάρι χρησιμοποιήθηκε για την μετάβαση ανάμεσα σε χρονικές στιγμές χωρίς να δώσει σημαντικά αποτελέσματα.

Chapter 1

Introduction

A lot of research has been conducted over the field of studying the evolution of the cosmic web. Many methods from graph approaches to even GANs have been used in order to study the evolution of the cosmic we and in order to extract useful information about its evolution and the rules which govern it.
Cosmos or universe is defined as the unity that lays around us, it is consisted or matter in different forms and some unknown substances like dark matter and dark energy. Universe began from a singularity point around 13.8 billion years ago, before that time did not exist and the physical laws were not applicable.

1.1 Previous Work

Barabási et.al. [32] have approached the problem using graphs and simple models to create a graph-like image of the network. In their approach they study some properties of the resulted networks and extract some metric graphs using knowledge about some parameters of the galaxies of the universe. The graphs resulted appear to have some properties expected from the astrophysicists. This approach uses graph theory and creates a network of the cosmic web using edges that resemble some connections between galaxies.
On the other hand the paper of Amara et.al. [43] approaches the cosmic web from the perspective of the simulations conducted in order to extract snapshots of universe's evolution. This paper uses GANs to predict possible alternatives to the evolution of the cosmic web. Its main contribution is that it may possibly save computational time from the super computers where the simulations are executed.
On the other hand astrophysicists have conducted a lot of research using either real data or simulated. They try to approach the cosmic using their previous knowledge about the physics laws that govern the universe which they attempt to integrate into simulation models and in general models that try to encapsulate universe's structure in a proper way.

1.2 Contribution

The above mentioned approaches describe either models or methods to conduct simulations faster. The field of research of the cosmic web lacks of an objective function in order to measure either the success or the failure of each model proposed. In order to do that we need to incorporate astrophysical knowledge into a function. Our main contribution is the extraction of that function which will be the main point of interest on our following research beyond this thesis.
On this thesis we aim to incorporate graph-theoretical knowledge into the field of studying the cosmic web. We will present models which will integrate both astrophysical knowledge as well as algorithmic in order to produce results about the universe. Moreover the prime advantage of our approach is that we achieve to create models that approximate cosmos well and fast. Using our proposed methods it is possible to reduce the time needed in order to extract features about galactic formations and properties of galactic constructions. The main difference from Barabási et.al. is that in our work except from constructing networks for the cosmic web we attempt to study them and perform analysis over these networks.

1.3 Graph-theoretical Results

We have used many models initiated from graph theory. Our models resulted from modifications of widely known graph tools which were altered in a proper manner so that they will be able to incorporate basic physical laws.

1.3.1 Gravity Graphons

In this approach we attempted to integrate gravity in a network-like image of the cosmic web. The intuition behind was that gravity plays a prominent role in universe's formation and in order to draw edges in the network we have taken into consideration gravitational forces between galaxies. Model runs in O(n^2) time but with a slight modification using KNN graphs it is possible to create the model in O(nlogn). Using graphon we assigned to each vertex the mass of the galaxy it represents and we draw an edge (between vertices a and b) with probability softmax(gravity(a,b)), where a, b are the galaxies that interact with each other. The results were promising and further modification could be made to use more symmetrical physical laws beyond gravity.

1.3.2 Gravity Fields

Using as intuition the paper of Barabási et.al. [32] we created a model that tries to approximate a field of influence for each galaxy. The main contribution of this model is that we do not define a constant radius for all galactic fields, rather we create a field depended on the mass and the half mass radius of each galaxy. This versatile radius depended on each galaxy's characteristics allows us to create a network for the cosmic

web that is connected and comparable to a model proposed in [32]. The complexity of the model is O(n^2) but it can once again be reduced into O(nlogn). Model works using the gravity force between pairs of galaxies and then weights it according to a metric. After that the model has a constant to control the edge density. Finally it draws an edge only if the distance between the galaxies is less that the weighted gravitational force.

1.3.3 Schwarzschild radius and Genetic Algorithm

Further attempts to create field of influence for each galaxy have been made using the Schwarzschild radius which is a physical value defined as the radius beyond which if an object is compressed then it turns into black hole. We tried to use a multiple of this value in order to represent a different radius for each galaxy but the results were not as promising as in the previous methods. Finally we implemented a genetic algorithm in order to extract a gravity like force that binds galaxies. In order to do that we created a general formula of gravity and tried to use it in a sample of the total galaxies with objective function metric to create a more power-law-like network using this formula.

1.4 Spatial Categorization Results

On this section we will present result from using modified spatial clustering algorithms.

1.4.1 Octree

Using Octree with O(nlogn) complexity we have extracted interesting results about the formation of the cosmic web. Further research along with a follow-up thesis on this topic will follow. The main results from this approach came from a smart pruning on the tree depending on the density. We have used density as it is the main parameter that separated voids and walls in the cosmic web. Further improvements of the structure are needed in order to approach the universal structure in a more compact manner.

1.4.2 Gravity Lattice

We have implemented a 3D lattice on the cosmic cube were each vertex of the lattice represents a 1kg gravitational test load. After that we inserted all galaxies one by one and calculated the forces they exerted to the test loads inside their field of influence. Finally we move the test loads with respect to the forces that have been exerted over them. The algorithm's complexity is O($C \cdot N$) as the number of test loads inside the influence of each galaxy is bounded. The algorithm was further improved into a smart filtering of the test loads according to their movement and the number of galaxies that have affected them. This filtering allowed us to examine galactic structures of determined size which is needed by the astrophysicists. The aim behind this model is to extract a gravitational footprint of the galaxies and then filter that footprint accordingly and fast.

1.4.3 HDBSCAN

Using the well known spatial clustering algorithm with parameters determined by the astrophysical data we defined the minimum cluster size, the radius beyond which we do not separate clusters and the percent of void galaxies. Using the above method and the algorithm we have extracted galactic communities and with respect to the parameters these communities represent clusters superclusters or other constructions.

1.4.4 ABACUS

We modified the spatial clustering algorithm ABACUS which results in a spinal-like structure of a point cloud. This algorithm tries to find the main constructions of a point cloud by merging and moving points. We have used this algorithm and a filtering like in the Gravity Lattice in order to extract different size constructions. We modified the algorithm so that each point that represents more that one points have also their total mass instead of having the counter of these points. Using that modification we filtered according to mass and that was helpful as it allowed us to extract galactic constructions of different sizes which is needed in the astrophysical field of research. Finally that method appeared to have eliminated almost all void galaxies and that helps us determine void regions.

1.5 Density within the Cosmic Web

In order to further extend our results over the cosmic web we have tried to determine each region's label using density as that parameter is characteristic in order to separate voids and walls. We have used the results from the HDBSCAN and in each community found from the algorithm as well as for the noise community we used convex hull to extract its volume. Taking into consideration that galaxies inside a convex hull are not able to dominate the whole hull we used a density estimator in order to find what percent of hull's volume is covered by the galaxies of the community that forms the hull. After extracting the density we compared the values among all communities. Results were promising and verified by the current model about the universe. Void regions had two orders of magnitude difference with the wall regions which is high enough to let us conclude to results that our detected communities are safe to be considered as walls while the noise detected by the algorithm can also be considered as void. Moreover we used the same approach in a deeper level so that we wanted to extract smaller communities inside the bigger ones and the results were once again promising as we have found uniform density among smaller clusters inside a bigger one.

1.6 Time Evolution of the Cosmic Web

In order to study the time evolution of the cosmic web we have used results from the filtering over ABACUS and created KNN graphs using only the filtered galaxies. That means that we have used specific sized galaxies. We have taken data from different time points of the cosmic evolution and performed the same method in order to extract the KNN graphs which we will compare. We created a map between the vertices of these KNN graphs so that vertices laying near in the cosmic cube will also be matched as we consider that it is possible to represent the same galaxy or an ancestor of the galaxy. After that matching we have compared the edges of the networks in order to determine the percentage of the edges that have changed over time. If an edge have changed that means the galaxies have declined as they are no more in the k closest neighbors of one another or some other galaxy have come closer. The model of the universe that is believed by the scientists suggests that galaxies decline as time passes. Our results agree with that model as we have observed that edges change in a fraction of 30% for large scale galactic structures and of 50% if we include smaller structures as well. That possibly can be explained as the large scale structure have enormous masses and they do not move so much while smaller ones do and that is the reason we have greater percent of edge modification.

Chapter 2

Graph Preliminaries

In this section we present some basic tools that are used during the analysis and design of the algorithms presented later in this thesis. We start by introducing the basic concepts of graph theory, scale free networks, small world networks and graphons . The above mentioned are used during the first phase of the presented thesis in order to approach the problem we have faced.

2.1 Graph Theory

Graph theory is used to study graphs, which are structures used to model pairwise relations between objects (nodes). A graph is made of vertices and edges. Vertices also known as nodes or points represent objects or agents who play an important role in our system and need to be modeled as they interact with each other. Edges known as links or lines as well are the above mentioned interactions between the nodes. Every edge represents a special relationship or interaction between two nodes. A distinction is made among graphs as they separate into two basic categories: undirected graphs and directed graphs. The first consist of edges without orientation and represent an symmetrical while the later represent asymmetrical relationships. Graphs play a very important role in modeling systems and are the prime object of study in discrete mathematics.

A restricted representation of an undirected graph is by an ordered pair. If we name the set of nodes as V and the set of edges as E a graph G is defined as G = (V, E) with:

•V the set of all nodes

•$E \in \{(x, y)|(y, x) \in V^2 \wedge x \neq y\}$

There exist graphs having multiple edges between node pairs also known as multigraphs and graphs without allowing multi-edges and self loops which are called simple graphs.

We will only be concerned with simple graphs. Bellow we show simple directed and undirected graphs.

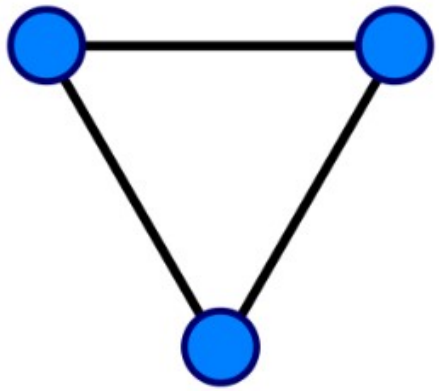

Figure 2.1: Undirected simple graph

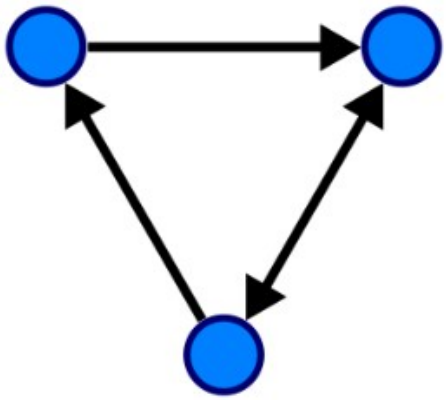

Figure 2.2: Directed simple graph

Graphs can also be divided into two categories related to the existence or absence of weight over their edges. Weight over edges represent the strength of the relationship between the nodes on the endpoints of the edge. This can represent the strength of a tight or the distance between nodes. On the other hand undirected graphs carry no weight over the edges, or as we can state it in a more proper way the edge among all edges is equal to one. In this thesis we will only use unweighted graphs whenever is needed. A common way to present graphs and their edges is through the adjacency matrix. That is a matrix of size n^2 where n is the number of nodes and the (i,j)th element of this matrix represents if there is and edge between node i and node j, and if so what is the weight of that edge. For undirected graphs the matrix is symmetrical. Also if the graph is unweighted the entries of the matrix are equal to either one or zero. [1]

2.2 Natural Complex Networks

The information mentioned in the previous section was an introduction to graph theory which is widely used as well in the network theory. A network is represented as a graph and has the same properties. We are going to depict our agents using graph nodes and the relationship between them as well as their distance using graph edges.
In the context of network theory, a network is a graph with non-trivial topological features. In order to further explain that we consider as trivial topological features the features that occur in graphs modeling of real systems such as regular lattices or random graphs.
The study of complex and big networks is an active scientific field of research which initially was inspired by empirical studies [44]. It has a wide variety of application in everyday life as networks exist around us, for example computer, biological and social networks rule

our lives.
Most of the networks that are under study tend to have connection patterns that are neither random nor regular. Such features include a heavy tail in the degree distribution, community structure and hierarchical structure [6, 10, 45]. Those three characteristics of the natural networks will be taken into examination in this thesis and using them we will extract some properties about the cosmic web. Two are the most well-known and much studied classes of complex networks: scale-free networks and small-world networks [3, 5]. The discovery of the above network classes and the extend to which they naturally appear in biology and astronomy are separated case-studies in the field. They both are characterized by specific structural features which we are going to explain. We will try to give an explanation to the above statements and inform the reader about their special traits. [4]

2.2.1 Heavy tail degree distribution

On this section we will explain the special traits of a network with heavy tail in its degree distribution [45].
Firstly, we will define the degree distribution of the network as follows. Every node is connected to a finite number of other nodes in the graph, these nodes are called neighbors. The amount of neighbors of every node is its degree. If the network is directed every node has two degrees, the in-degree and the out-degree referring to the arcs pointing to the node and to the arcs leaving from the node respectively.
Restricting the field of interest to only undirected, unweighted simple graphs the degree of a node i of the graph is the sum of all elements of the i-th row of the adjacency matrix A,

$$k_i = \sum_j a_{i,j}$$

where the sum is over all nodes in the network.

Extracting all degrees from every node of the graph results in a sequence of numbers also known as degree sequence. The probability distribution of these degrees over the degree sequence of the whole network is called the degree distribution, defined by

$$P_{deg}(k) = \mathrm{P(k)} = \text{fraction of nodes in the graph with degree k}$$

The following image depicts a graph and its degree distribution.

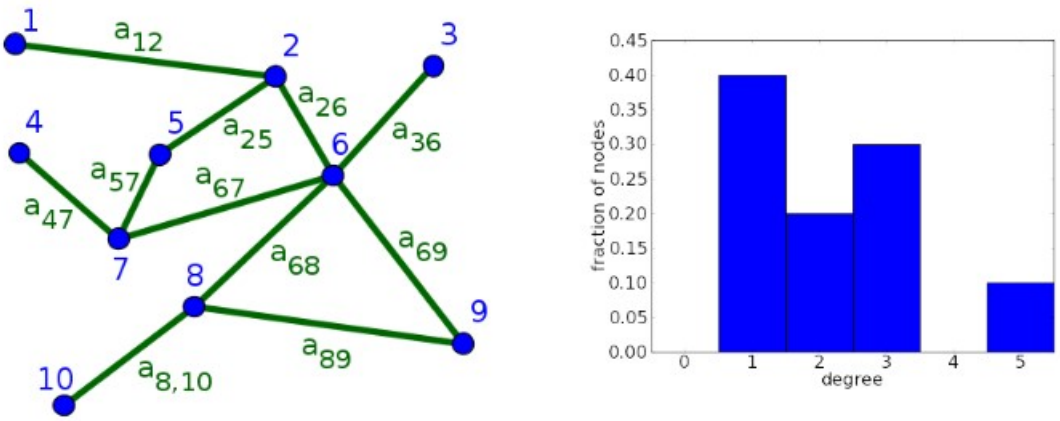

Figure 2.3: Left: Network, Right: degree distribution. Picture taken from Math Insight

As we have defined the degree distribution a heavy tailed degree distribution is a distribution having a tailed shape. In this distribution there are lots of items with tiny values and a few items with enormous values. It is also known as fat or long tailed distribution. This heavy tail of the distribution refers to the fact that a small, but not insignificant number of nodes have extremely high degree values in contradiction to the majority of the the network nodes that have small degree [46].
The reason why we are interested in such distributions is because real networks seem to have heavy-tailed degree distributions [45]. In other words, it seems that in natural networks [44] there are lots of nodes with very small degrees, but also a small number of nodes with very high degree value, known as hubs [47].

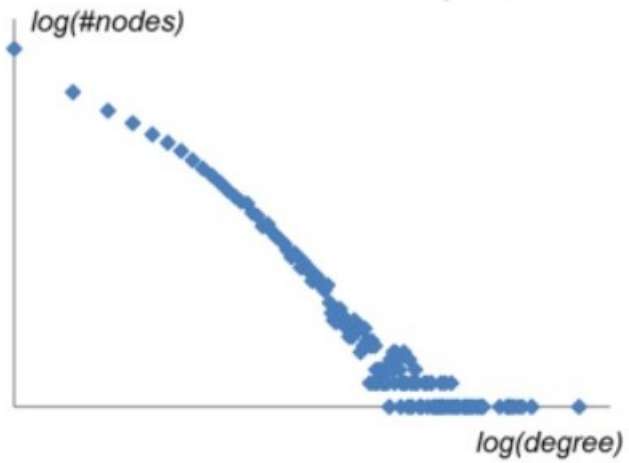

Figure 2.4: Heavy tailed distribution of a network

2.2.2 Community structure

When studying a complex network we would like to know if the nodes are placed in a way that is easily to determine possible communities among the network [48]. A network is said to have community structure if its nodes can be easily grouped into sets that are densely connected and we would also like these sets not to be highly interconnected. It is possible that these sets are also overlapping, and as a result a node might possibly belong in more than one community [49, 50]. In that case we refer as overlapping communities and the node participates in each community by a fraction/percent. On the other hand the most desired is that the network can be divided into non-overlapping communities. That would result in a natural fragmentation of the network in clusters (groups of nodes), with dense connections internally and sparse externally.[6, 7, 8]
The value of community structure of the network relies to the fact that nodes of the same group have greater probability of being connected with an edge while nodes from different groups have very small connection probability. Below we show a possible community fragmentation of the famous Zachary's Karate Club [9]. It is not the optimal community detection but we use it for demonstration reasons.

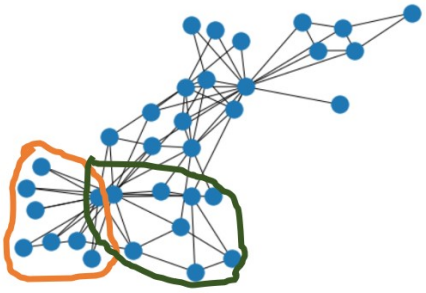

Figure 2.5: Overlapping Communities

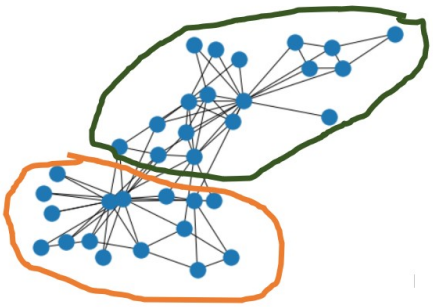

Figure 2.6: Non-Overlapping Communities

A direct result from studying theoretically the community structure of a graph is the development of community detection algorithms for every graph. It has been a lot of research on this field and many algorithms have been proposed [51]. The main idea in order to detect communities is to find group of nodes highly connected with each other and sparse connected with nodes of other groups.

2.2.3 Hierarchical Structure

Hierarchy (from the Greek word 'ιεραρχία') is the arrangement of items in a way in which there exist levels 'above' and 'below' of the level of resolution we are currently in. In order to depict that in more direct manner the reader could imagine a camera looking at a fractal and every time zooming in. The fractal is the ultimate item of hierarchy because

in every zoom the reader will see the exact same image but the point we would like to highlight here is that hierarchy depicts an uniformity in the way we do something (e.g. zooming-in) [52, 53].
Hierarchy is important in a vast variety of tasks as well as in the way we manipulate networks and try to extract their features. Hierarchy links levels of analysis directly either vertically or horizontally. Every upper level is superior to its direct descendant and there is a function that projects each level to its direct descendants and ancestor.
In order to depict that we can imagine the problem of community partition of a graph and the way we can use hierarchy in order to fulfill the task. Using a hierarchical clustering algorithm we show the dendrogram [54] of communities of the Zachary's Karate club [9]. As we can see reading it from bottom to top first we have many small communities but because we use hierarchy we treat them as nodes in every stem until we reach the final two communities on the top of the graph. [10, 11]

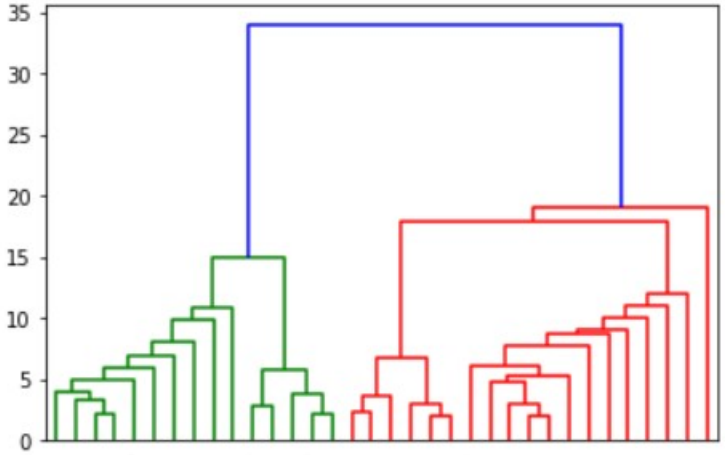

Figure 2.7: Dendrogram of Zachary's Karate Club

With the above example the reader is able to understand the importance of the hierarchy and its existence in scale free networks.

2.3 Scale Free Networks

A scale-free network is very common structure in natural networks [44] from biological to even social networks. In this type of network the degree distribution follows a power law, or at least asymptotically, and can be approximated very well from a power law function.
To be more specific, the fraction of P(k) of nodes in the network that have degree k to the total number of nodes of the network goes for large values of k as:

$$\mathrm{P(k)} \sim k^{-\gamma}$$

where γ is a parameter of the network. Its value varies between networks but typically is in the range $2 < \gamma < 3$.
A crucial property of the scale-free networks [3] is that the second moment of their degree distribution is infinite while the first moment is finite [1]. That allows them to have theoretically a degree distribution that has infinite deviation from the average degree. That points out a very important property of this type of networks. It is possible for them to have a large amount of nodes with small degrees and a small fraction of nodes with very large degrees [47].
The main reason this networks occur naturally is the preferential attachment of the nodes and the growth of the network [55, 56]. The reader can imagine that property using the following example:
Imagine there is a party with guests that are not friends with each other, at the beginning only few guests (nodes) have arrived to the party and they start talking to each other (edges). When new guests (nodes) arrive (growth) to the party they have greater probability to talk with the early arrived individuals because these people have already made many friends. [3]
The main characteristics of this type of networks is the variety of node degrees and the fact that most nodes connect with nodes of higher degree.

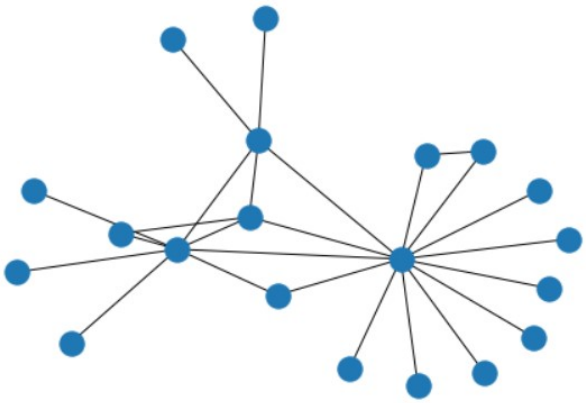

Figure 2.8: Scale-free network of 20 nodes

As expected a scale free [3] example is shown above depicting the mentioned properties.
Scale free networks have also the important property of being tolerant to faults [57]. If a fail happens uniformly at random then the possibility that it will affect a hub tends to zero. On the other hand they are not tolerant to targeted attacks because if a hub is destroyed then a great percentage of the network will vanish as well.
Furthermore it is needed to refer to scale-free property of having clustering coefficient scale-free [58, 59]. The definition of the local clustering coefficient of a node is given as the fraction of its neighbors that are neighbors with each other, or in order to make it more

clear is the fraction of its friends who are also friends with one another. Mathematically this is stated as:

$$C_i = \frac{|\{e_{jk}:u_j,u_k \in N_i, e_{jk} \in E\}|}{k_i(k_i-1)}$$

while the global clustering coefficient of the network is defined as:

$$\mathrm{C} = \frac{\text{number of closed triplets}}{\text{number of all triplets (open and closed)}} = \frac{\text{3 x number of triangles}}{\text{number of all triplets}}$$

There are many other examples of scale free networks [3] as they appear in everyday life. In order to provide to the reader full information about this type of networks and as we have already mentioned the preferential attachment [55] we define it through the principles of rich gets richer and mathematically is defined as:

$$p_i = \frac{k_i}{\sum\limits_j k_j}$$

where p_i is the probability that a new node when inserting the network to be connected to node i. Also k_i according to the previous notation is the degree of node i and the sum is used over all nodes that already pre-existed in the network. [12]
As a direct result from the mathematical relationship one can suggest that hubs tend to be early arrived nodes and are more likely to accumulate even more links as time passes. On the other hand nodes with only few links are unlikely to be chosen from a new node.

2.4 Small World Networks

Small world networks appear as well as scale-free networks [3]. They are as defined a graph in which most nodes are neighbors of one another. The special property that discriminates this type of networks is that neighbors of any given node are more likely to be neighbors of each other [60]. This fact directly points to the the main property of the small world networks, most nodes can be reached from every other node by a small number of hops/steps.
In order to make it clear to the reader choosing arbitrary two distinct nodes from the network of totaling N nodes then the distance between them (calculated in hops) is logarithmic proportional to the size N, that is:

$$L \propto \log N$$

It is also important to mention that although the average distance between two nodes is very small the clustering coefficient is not small. [5, 2]
A well studied problem over this type of networks is the 'six degrees of separation' experiment. With this experiment Travers and Milgram [13] demonstrated that individuals, nodes of the network, that look like they lay in a large distance are in fact really close, only six hops needed in order to reach from one node the other. Small world networks tend to have densely connected groups which they approach cliques or almost-cliques. Moreover almost all node pairs are connected through a relative small path.
In order to explain these properties one should take under consideration the great amount of hubs [47] that exist in this type of networks. These hubs (nodes of very large degree) serve as intermediates of the paths between node pairs, as they appear in many paths and because of them the path length stays so small.
Although small world networks share some common facts with scale free networks they tend to have fat-tailed distribution which favors more the large hubs and creates a large number of them [61].
There has been some methods for quantifying the small world property in a network but we will present the most popular which is the small-coefficient, σ. This factor is a result from comparison of clustering coefficient and path length in the given network to a random network with the same degree (on average).

$$\sigma = \frac{\frac{C}{C_r}}{\frac{L}{L_r}}$$

if $\sigma > 1$ then the network is small-world.

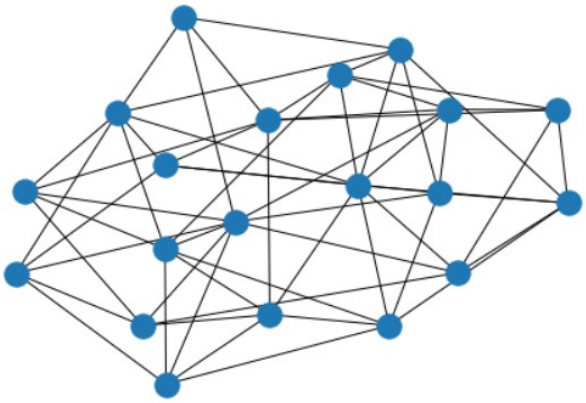

Figure 2.9: Small world network of 20 nodes, each connected with 6 neighbors and rewiring probability p = 0.7

In order to create the above graph we have used the Watts - Strogatz model. Their work provided a simple model that will result in a small world network if it is altered under specific mechanisms [62].

In order to construct this model we start with n nodes placed in a ring, then each node is connected with k other nodes which lay clockwise of it. After that for each node we take every edge that connects it to its k/2 rightmost neighbors and rewire it with probability p. Rewiring is done by replacing the end point of the edge with another node, chosen uniformly at random, from all possible nodes in the network. We perform the above task without allowing self loops nor multi-edges.
The value of the probability p is crucial to the resulted graph because for p = 0 we have complete order in the network with no rewiring, for $0 < p < 1$ we have small world network and for p = 1 we have random network. As we can see increasing p increases randomness. In order to demonstrate the above statement we present three graphs and the respected p values.

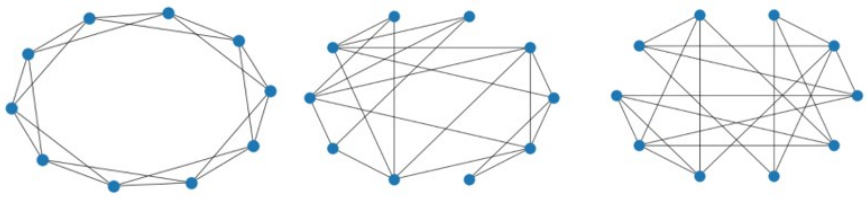

Figure 2.10: Small world network of 10 nodes, each connected with 4 neighbors. Left: p = 0, Middle: p = 0.5, Right: p = 1

2.5 Graphons

Nowadays it is common for very large graphs to emerge naturally, the world wide web is one example of this type of networks which are extremely large and tend to grow even more. Such networks and their manipulation is not a trivial task for anyone. Applying the standard graph theory to the whole network is time consuming and sampling a small fraction of nodes of the network might result in inconsistent conclusions about the network.
Taking into consideration these and the fact that large graphs are ubiquitous in mathematics and describing correctly their properties is an active field of research as well as an important role of modern combinatorics there have been proposed many methods. The method we are going to use in this thesis is the graphon method.
To this direction a way of studying large but finite networks is through a sequence of graphs. Staring to move from larger to larger such objects the idea is to go through this graph sequence and ideally limiting these objects. If this is done properly then the properties of this limiting object of the sequence of graphs will approximate the large network and will have its properties.
Graphons, kind of functions, are these limiting objects of the graph sequences, for sequences of large bu finite graphs with respect to the cut metric. Graphons were introduced by C.Borgs, J.T.Chayes, L.Lovász, V.T.Sós, B.Szegedy and K.Vesztergombi. Graphons

appear naturally when sequence of large graphs exist like in external graph theory, quasi - random graphs and others. [14, 15, 16]
To further extend the knowledge of the reader about graphons we present them as a symmetric measurable function:

$$\mathrm{W}\colon [0,1]^2 \to [0,1]$$

They are important in studying dense graphs and arise both naturally [44] as the limit of a graph sequence and as the fundamental defining objects of exchangeable random graph models. They are closely connected with dense graphs as the researchers have observed for the following reasons:

- Random graph models defined by graphons give rise to dense graphs almost surely.
- Using the regularity lemma graphon are able to encapsulate the structure of an arbitrary large and dense graph.

To statistically formulate the above theory in order to be more clear how we are going to apply it in the context of the cosmic web we present some ideas reformulated into a more applied context.
A graphon is a symmetric measurable function W as defined above. It is better understood by defining a random graph model with the following scheme:

1. Every vertex i of the graph is assigned with an independent random value chosen uniformly, in essence $u_i \sim U[0,1]$

2. Every edge (i, j) is independently added to the graph with probability W(u_i, u_j)

A random graph model is an exchangeable random graph model if and only if it can be determined using the terms of a graphon in this way.
Models based on a graphon W are denoted as G(n, W) and a graph created from a graphon W is called W - random graph.

In order to point the importance of the graphons and their relevance to well known case - studies of other network type we will present their direct application in creating a random graph model that results in the Erdös - Rényi model of random graphs G(n, p) [63].
We shall create the simplest graphon which is the constant graphon of W(x, y) = p for some constant p, where p $\in [0,1]$. Using this format we will result in a random graph

having for each edge probability p of including it.

Many random graphs can be interpreted through graphons.
During this thesis in the first phase we will use them in our approach of comsic web modeling.

Chapter 3

Cosmological Preliminaries

In this section we present the main ideas of the cosmology and its importance. We will also present the data that were used for our experiments, the reasons why we have picked them, the way they have been produced and some alternatives available for the same applications. Also we present Gadget tool, its applications and the Illustris Simulation Project. Finally we make a reference to clustering which is being used in the second phase of the problem approach.

3.1 Cosmology

Cosmology (from the Greek word κόσμος, kosmos meaning the world and the Greek word -λογία, -logia meaning study of) is the branch of astronomy concerned with the studies of the origin of the universe. The term was first used in English in 1656 by Thomas Blount.
This active field of research is also concerned about the evolution of the cosmos and the reasons and natural laws that have ruled and led to that specific evolution among all other infinite possible paths that the universe could have taken among its long lasting evolution. The study begins from the Big Bang and reaches to this day and will keep track of cosmos evolution.
Big Bang is a cosmological theory and model of the universe in order to describe how everything started and why the universe has had such a vast expansion, especially at the beginning of time. Universe expanded from an initial state of very high density and high temperature through a massive explosion that took place before circa 13.8 billion years [64, 65].
Big Bang theory is the currently most accepted theory about the origins of the universe and has been used to explain a broad variety of phenomena such that of the fact that the farther away galaxies are the faster they are moving away from Earth. This theory has also given ground for the development of other theories such that of 'primeval atom' from Georges Lemaïntre.

With cosmology itself a wide variety of studies have flourished, physical cosmology is the scientific study which is interested in the universe's origin, its large - scale structures and dynamics as well as the ultimate fate and the natural laws that govern the cosmos. Cosmology, although used at most in sciences, a field in which it has helped to evolve, it has many extends to even philosophy [17, 18].

3.2 Timeline of Cosmos

According to the aforementioned Big Bang theory the universe at the beginning was very hot and small which means the density in the early universe was very high. As the expansion took place and the universe started to grow it began to cool down. [19, 20]
Using general relativity to go back in time at the origins of the universe point out to infinite density and temperature at a finite time in the past. This irregular behavior which is not humanly understandable is known as gravitational singularity. In that point of time general relativity and the whole set of natural laws as we know them can not be applied. Models based on general relativity cannot explain this period also known as Planck epoch. This singularity is some times called the Big Bang because after that event time began, universe was born and entered a regime in which laws of physics as we understand them work.

After the singularity era universe entered its earliest phase where inflation and baryogenesis took place. Astronomical data about this epoch are not available so we can only speculate about the universe's formation. Models suggest that the universe was filled homogeneously and isotropically with a very high energy density while temperature and pressure were extremely high. This era took place for the time period of 0 to 10^{-43} seconds where the universe rapidly expanded and the four fundamental forces were unified as one. This epoch was succeeded by the grand unification epoch. At that point of time gravity separated from the other forces as the universe temperature fell, of course the universe were too hot and no particles existed.

At approximately 10^{-37} seconds from the beginning the expansion had a phase transition resulting in the so called cosmic inflation. At that period the universe grew exponentially fast, it is worth to mention that the speed of this expansion was so large that exceeded the speed of light. That resulted in a huge temperature drop.
Quantum fluctuation were able to occur because of the Heisenberg's uncertainty principle which were in fact amplified into the seeds that would lately result into the large universal structures.

At around 10^{-36} seconds the Electroweak epoch began in which the strong nuclear force separated as well leaving only the electromagnetic and the weak nuclear force unified. Inflation stopped at around 10^{-33} to 10^{-32} seconds with the universe's size to have enormously increased. To be more specific the size through the inflation has increased by

a factor of 10^{78}.
Then reheating happened until the universe reached the needed temperatures for the production of quark - gluon particles as well as all the other elementary particles. In order to enlighten the reader more about the temperature values, at that point random motions of particles were at relativistic speeds and particle - anti-particle pairs of all kinds were continuously created and annihilated into pure energy from collisions.
At some point an unknown reaction took place which is the main reason that led the universe to its form. This reaction called baryogenesis violated the conservation of baryon number and resulted in a very small excess of matter over anti-matter, the order of that little excess is of one part in thirty million. This resulted in the predominance of the matter over antimatter in the present universe.

Universe's expansion passed to the new era of cooling. Density and temperature of the universe fell and symmetry breaking phase transitions resulted into the fundamental forces of physics and the parameters of elementary particles into the form they have until now. The two before united forces (weak nuclear and electromagnetic force) were separated at about 10^{-12}.

At about 10^{-6} seconds after the energies of some particles have fallen to values compatible for particle accelerators, quarks and gluons combined together to form baryons, i.e. protons and neutrons. At that point of time the temperature was not low enough to not let any other generation of matter - antimatter pairs so we had many pairs created which again annihilated leaving a small faction of matter over antimatter in the universe, just one in 10^{10} particles were left. The same process happened at about 1 second for the electrons and positrons resulting in the final particles of the universe as we know it until now. Particles were no more moving relativistically the energy density was dominated by photons and neutrinos.

After a few seconds since the expansion the temperature was about a billion Kelvin and the density of matter was as low as to be comparable with matter density in Earth's atmosphere today. These particles mentioned before were combined and tight together with strong and weak nuclear power to create the first nuclei in the universe (deuterium and helium). This whole process of creating the nuclei is called Big Bang nucleosynthesis.

As the universe cooled down the rest of the energy density of matter came to gravitationally dominate that of the photon radiation. After 379,000 years the first chemical elements appear in the universe. The electrons and nuclei combined to atoms (hydrogen at the beginning and then fusion of hydrogen molecules to heavier elements) which were able to emit radiation. This radiation from that period which was emitted in the universe and spreaded without obstruct is the well known today microwave background radiation. The chemistry needed for life as we know it today begun during a habitable epoch when

cosmos were roughly 10 to 17 million years old.

The new era of the universe which is the epoch we are interested in and use its information throughout this thesis is the structure formation epoch.
Over a very long time period the slightly denser areas of the cosmos, where matter gravitationally attracted nearby other matter particles and as a result it increased its density. These regions were latter on the regions of formulation of gas clouds, stars, galaxies and other astronomical structure observable until today (galaxy clusters).

More information about this whole process of formation is dependent to the amount and type of matter that existed in the universe at that specific period. There are four types of matter know as cold dark matter, warm dark matter, hot dark matter and baryonic matter and the best estimation about their existence are available through the Wilkinson Microwave Anisotropy Prob (WMAP) show that the data can be fitted by a Lambda - CDM model in which dark matter is assumed to be cold and is about 23% of the matter/energy of the universe, while baryonic matter represents only 4.6%.

Figure 3.1: Left: Abell 2744 galaxy cluster - Hubble Frontier Fields, Right: Artistic depiction of WMAP, Source: Wikipedia

Moving to the final era of the cosmic evolution we reach the today's universe which is dominated by a mysterious form of energy known as dark energy which infiltrates all of space. Many research has been conducted about its nature and formation because it covers almost 73% of the total energy density of the universe. Researches suggest that is was that type of energy that was the fuel for the expansion of the universe to keep going on. Without it, it is possible that the gravity will have reverse the expansion.

All of the cosmic evolution after the inflation era can be well displayed and fitted by a Lambda - CDM model. Before the inflation researchers have not yet discovered models to describe what has happened exactly.

3.3 Cosmological Structures

Universe as we have mentioned above have passed through an era in which structures of different size have been created. From star to galaxies and galaxy clusters cosmos is

not so uniform as one might have expected.
The structure of the universe could be divided into components that can help describe the characteristics of individual regions of the universe. Each of those regions will be characterized by the structures located in there. There are two main structural components in the cosmic web [21, 22, 23, 24, 25]:

- Voids: vast ,largely spherical regions with very low cosmic mean densities, up to 100 megaparsecs (Mpc) in diameter.
- Walls: regions that contain the typical cosmic mean density of matter abundance. Walls can be further divided into two smaller structural categories:
  - Clusters: highly concentrated regions where walls meet and intersect with each other, adding to the effective size of the local wall.
  - Filaments: the branching arms of walls that can stretch for tens of Megaparsecs.

3.3.1 Voids

Cosmic voids are large spaces between filaments (the largest in scale known structures in the cosmos). They contain only a few or no galaxies. Voids typically have diameter of 10 to 100 megaparsecs. Further researches have found particularly large voids, defined by the absence of rich superclusters which are known as supervoids [66, 67].
They have less density than the average density matter abundance that is typical for the observable universe by a factor of ten. Voids were discovered in 1978 by S.Gregory and L.A.Thompson. It is widely believed that they were constructed by three main events that took place in the cosmos. These are the baryon acoustic oscillations in the Big Bang, the collapses of mass followed by implosions of the compressed baryonic matter.
These empty regions started an the early universe as small anisotropies which grew really fast as the universe expanded and evolved. This is uttered in the nature of gravity itself. Dense regions or in general regions of higher density than voids collapsed more rapidly under gravity and eventually they created large scale structures of even higher densities and left other regions of the universe almost empty. This procedure resulted in the large scale, foam-like structure of the cosmic web of voids and galaxy filaments as we observe it today.
Of course there are voids located in highly density regions but their size is significantly smaller than the size of voids located in low density spaces of the cosmos. Moreover voids appear to correlate with the observed temperature of the cosmic microwave background because of the Sachs - Wolfe effect. In order to state it in a more proper way for the reader to understand it we shall declare that colder regions correlate with voids while hotter region correlate with filaments because of the gravitational redshifting. Finally voids play a prominent role in our understanding of the universe and their existence provides physical

evidence for dark energy.

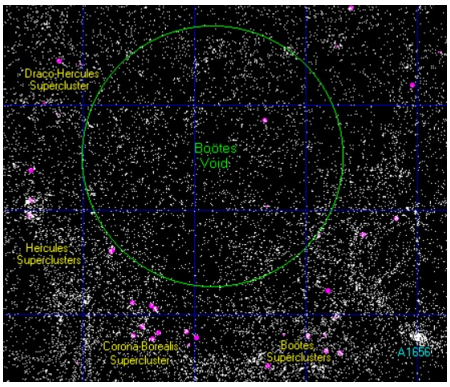

Figure 3.2: Map of Boötes Void, (The Great Nothing), Source: Wikipedia

3.3.2 Galaxy Clusters

A galaxy cluster or a cluster of galaxies is one of the sub-types of the wall structures. A typical galaxy cluster consists of a number of galaxies from hundreds to thousands of galaxies that are bound together under gravitational force [68]. They typical mass of this type o structure varies between 10^{14} to 10^{15} solar masses.
They are the largest known gravitationally bound structures of the known universe after the superclusters which were discovered in 1980s. One of the main features of the cluster is the intracluster medium (ICM). The ICM as studies have shown consists of heated gas that is located between galaxies and has temperature an most some value from 2 to 15 keV, which is dependent to the total mass of the cluster.
Galaxy clusters are not the same as star clusters that are clusters of stars and located within a galaxy. Small groups of galaxies are known as galaxy groups and if they cluster themselves they can form a galaxy cluster which in fact can also be grouped with other galaxy clusters to form superclusters [69].
Galaxy clusters typically have the following properties:

- Number of galaxies contained in the cluster are between 100 and 1000.
- The distribution of galaxies, intergalactic gas ans dark matter is almost the same in the cluster.
- They have total masses of 10^{14} to 10^{15} solar masses.
- They have diameter from 2 to 10 Mpc.
- The components of the galaxy cluster are: Galaxies in 1%, Intergalactic gas in intracluster medium in 9% and Dark matter in 90%.

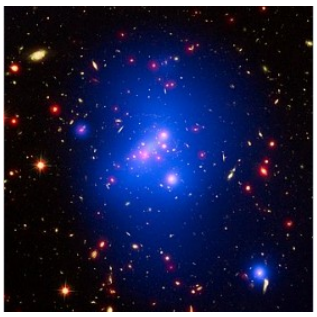

Figure 3.3: Galaxy Cluster IDCS J1426, Source: Wikipedia

3.3.3 Galaxy Filaments

Galaxy filaments play an important role in physical cosmology as they are the largest known structures in the universe. They are massive, thread - like structures having typical length of 50 to 80 megaparsecs $\cdot h^{-1}$ which is hundreds of million of light years long. These structures form the boundaries between large voids of the cosmos. Filaments consist of gravitationally bounded galaxies, it is also possible for a filament to have inside a super-cluster as well [70, 71].

According to the standard model of the evolution of the cosmos filaments of galaxies form along and follow a web - like shape. They also follow strings of dark matter. It is believed through the model that dark matter plays the most important role in the structure of the universe on the greatest of its scales. Galactic filaments have almost the same major and minor axes in cross - section, along the lengthwise axis.

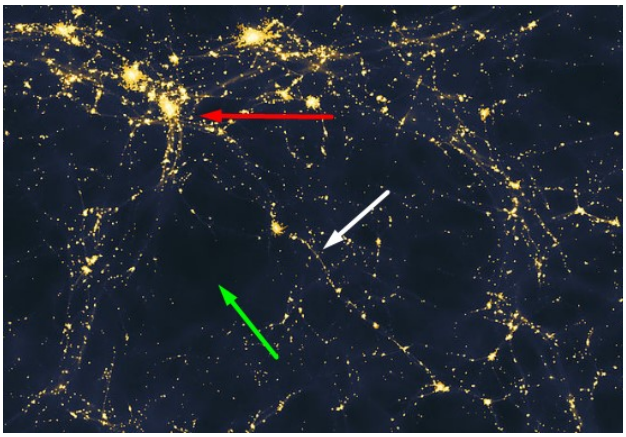

Figure 3.4: Computer simulated image, Red arrow: Cluster, Green arrow: Void, White arrow: Filament, Source: Wikipedia

On the figure above we present a typical image of the cosmos and spot the main structures on it. In this thesis we are going to be concerned mostly with the clusters and their difference from voids. We will try to determine clusters' region and evaluate the effectiveness of algorithms in finding such areas. We will also try to find void areas with little or no galactic presence.

3.4 GADGET-2 & Millennium Simulation

GADGET - 2 [26] is a code for cosmological simulations of structure formation, it is referred to the main document that is an acronym of GAlaxies with Dark matter and Gas intEracT. GADGET is an available code to the public for cosmological N-body/SPH simulations written by Volker Springel at the Max Planck Institute for Astrophysics.
GADGET computes gravitational forces as gravity is the main force that laws the universe in its greatest scales of galaxies and clusters. In order to do that GADGET uses a hierarchical tree algorithm [52], with the option to use it in combination with a particle - mesh scheme for long-range gravitational forces, and represents fluids using means of smoothed-particle hydrodynamics (SPH). The project is versatile allowing to be used for both studies of isolated systems and simulations, with or without periodic boundary conditions, included the cosmological simulations about the universe's expansion. In all of the above mentioned types of simulation GADGET follows the evolution of a self - gravitating collision-less N-body system, with also allowing gas dynamics to optionally be included in the simulation.
The force computation and the time stepping of GADGET are totally adaptive with a dynamic range which is unlimited. Therefore it is suitable to address a wide variety of astrophysical interesting problems, from galaxies' collisions to the formation of the universe itself. Moreover GADGET's variety of parameters that can be modeled and included in the simulation gives the ability to study many specific problems, i.e. dynamics of the gaseous intergalactic medium, star formation. GADGET - 2 was published in 2005 and is an improved version of GADGET - 1 as it contains a new time integration model, a new tree-code module, a new communication scheme for gravitational and SPH forces, a new domain decomposition strategy, a novel SPH formulation and the TreePM functionality [72, 73].

In the Millennium simulation project [27] an adapted version of GADGET is being used. Published in 2005 it was the largest simulation of the structure of the cosmos within the Lambda-CDM model. It uses 10^{10} particles to follow the dark matter distribution in a cubic region of $500h^{-1}$ Mpc on each side, with resolution of $5h^{-1}$ kpc. This simulation allows the formation of roughly 10^7 galaxies. During this simulation a special database

has been created in order to save and make public the results about the galaxies.
Later on 2008 a more famous simulation conducted known as Millennium Simulation II. It happened in a much smaller cube in order to have greater resolution about the mass distribution. This second model combines the multiple simulations of differing mass resolution with improved treatments of many of the underlying astrophysical processes in order to represent observed galaxies over a wider range of galaxy mass and redshift than previous models.
Millennium run kept busy the computer on which it was conducted for almost a month in order to produce the needed results. It is worth to mention that the results of the simulation were compared with observational data and surveys in order to clarify the processes underlying the buildup of real galaxies and black holes.

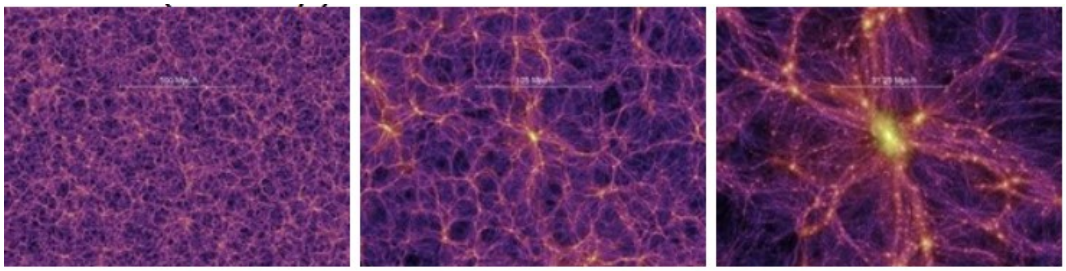

Figure 3.5: Computer simulated image for z = 0.0,Subsequent panels of zoom in by a factor of 4 with respect to the left panel, Source: The Millennium Simulation Project

The above image presents some of the results of the simulation in order to specify the importance and the quality of the simulation. In this thesis we have used the small catalog of galaxies of this simulation in order to access data about some galaxies spread into the simulated universe. In order to do that we have accessed only the catalogs without the images and without using the full catalog because we used it for demonstration reasons and because of its size.

3.5 Illustris Simulation

The Illustris project [28, 29] is a series of astrophysical simulations run by an international collaboration of scientists, it was carried out by Mark Vogelsberger and a collaboration of scientists using the V.Springel's Apero code. Prime goal for this project is the same as the Millennium simulation, studying the galaxy formation and evolution of the universe using a comprehensive model to describe the natural laws and conduct the simulations.
Illustris includes large scale cosmological simulations staring from the Big Bang using some initial conditions and going on until today, 13.8 billion years of simulation. The modeling used by this project is based on the most precise data and calculation available from the

scientists today about the cosmos and the main parameters that rule it. Its results are compared with actual observed findings including galaxy formation dark matter and dark energy.
Model is so precise that it contains physical processes thought to be very important in the formation and structure of the galaxy such as the formation of the stars and super-massive black holes.
Having as main idea the creation of precise representation of galaxies in the conducted simulations Illustris project tries to track the expansion of the universe , the gravitational pull of matter from itself and the motion of the cosmic gas that exists between galaxies in the intragalactic medium [74] and plays important role in clustering formation of galaxies.
Having described before the time line of cosmos it is obvious for the reader to expect the simulation not to start exactly at the Big Bang because the first moments of universe's evolution cannot be modeled precise nor enclosed by some specific set of laws, such as general relativity. Moreover taking into consideration that in the first thousand years there were no significant matter structure in the cosmos the Illustris project begins the simulation from a very young universe of 300,000 years after the Big Bang.
Simulation contain thousand of galaxies captures in a very detailed way covering a great variety of masses, rates of star formation, sizes and other properties. Prime goal of every simulation as well as of this one is to compare the resulted data of the simulation with the observable universe because if there is a direct connection between them that would be a step towards understanding how universe was created and eventually what are the laws that ruled this creation.
As we have mentioned before Lambda - CDM modeling of cosmos [75] and using it in the simulation was very famous in other projects and widely used to predict the evolution of the universe, even to a great extend of using one trillion particles. This model suggests that the cosmos is filled with three distinct components as we have stated in the timeline evolution these are baryonic matter, dark matter and ark energy. Trying to solve the equations given for that model without simulating it will not result in significant results as it restrain the study to some simplified problems. The Illustris main difference in the approach of the cosmos is to directly account for the baryonic matter in the simulation in order to calculate the motion of the gasses as well as they gravitational forces between them. From that point of view Illustris is able to provide a link between babyronic masses, their influence in the evolution and as a result a self-consistent and predictive method for conducting simulations.

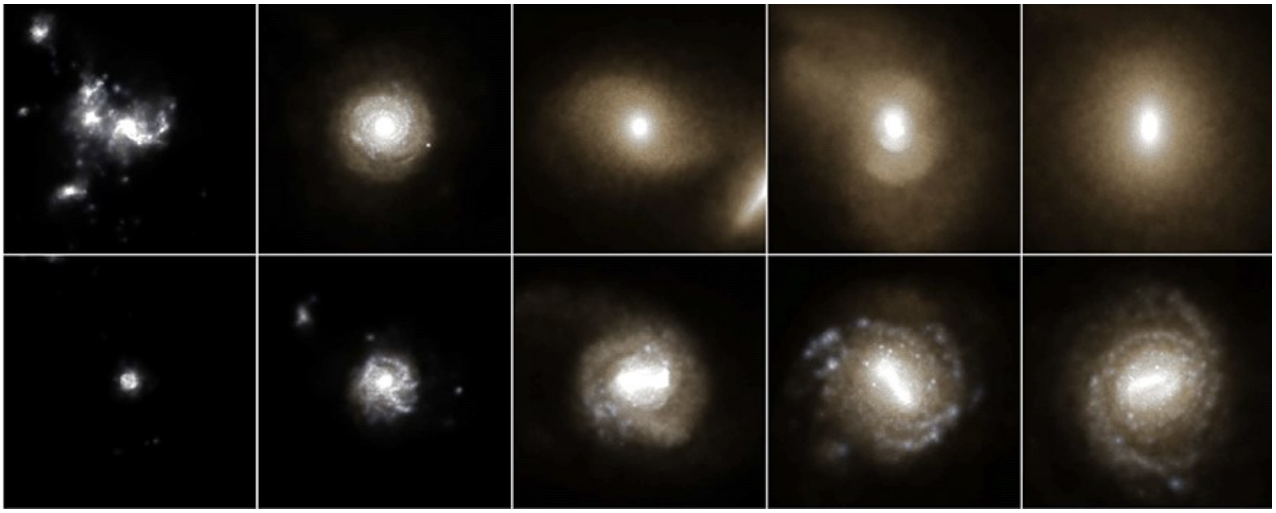

Figure 3.6: Showing two galaxies formation and evolution using Illustris method, Source: The Illustris Project

In order to further extend the reason why we have chosen Illustris data for the prime part of this thesis is because over the past decades computer simulations of the cosmos included only gravitational force as the main force that ruled the universe in this large scale. Although these simulation have given important results they have reached to a point of accuracy, the above mentioned example of the Millennium simulation is characteristic as being one of the best representatives of that school of experiments.

Illustris tries to approach the evolution doing something harder than Millennium as it includes gas treatment using one of the two main ideas either with SPH simulations, as we have explained them before, or with the approach of the mesh - based method using adaptive mesh refinement (AMR).
Illustris main idea is to use the APERO code [76] in which the treatment for gas is implemented using a moving unstructured mesh, in that method APERO tries to combine the above two mentioned in one unified model of the gas movements in the cosmos. In addition to further support this claim, recent researches have shown that this new approach used in Illustris has significant advantages in large scale simulations. As a result the reader can now understand that this simulation has many advantages over others conducted in the past, of course the results have finite resolution because the problem is as complicated and need to avoid capturing detailed some phenomena such as star formation withing the galaxies. The main goal which is in fact achieved is to simulate the galaxy formation in a detailed as possible way.
We also present the main runs that have been made form the Illustris project and the reference to more specific details about the whole project.

name	volume $[(\mathrm{Mpc})^3]$	DM particles / hydro cells / MC tracers
Illustris-1	106.5^3	$3 \times 1,820^3 \cong 18.1 \times 10^9$
Illustris-2	106.5^3	$3 \times 910^3 \cong 2.3 \times 10^9$
Illustris-3	106.5^3	$3 \times 455^3 \cong 0.3 \times 10^9$
Illustris-Dark-1	106.5^3	$1 \times 1,820^3$
Illustris-Dark-2	106.5^3	1×910^3
Illustris-Dark-3	106.5^3	1×455^3

Figure 3.7: Table of the available Illustris runs, Source: The Illustris Project

In this thesis we have used the Illustris-3 data. As we can see the Illustris simulation have a specific amount of resolution proportional to the number of particles they use in the initial state. We wanted to demonstrate our ideas so we have chosen the less detailed simulation in order to have a limited number of galaxies in the 106.5^3 $(\mathrm{Mpc})^2$ cube of the simulation.
We have used mainly the data produced at redshift z = 0.0 meaning the present time, after 13.8 billion years. In some points of this thesis we also used data from previous epochs of redshifts 0.1, 0.2 and 0.3 in order to compare the proposed methods over time.

3.6 Clustering

Cluster analysis or clustering [30] is and active field of research which was developed decades ago but it keeps growing using the new available data and their properties in order to suggest better ways to classify items. The main task of clustering is to group a set of objects in such a way that the objects in the same group, also known as cluster, to be more similar than with objects in other clusters. Clustering is used in a wide variety of tasks such as statistical analysis bio-informatics and of course astronomy in order to extract cosmological results.
Clustering is not just a specific algorithm used in every case to solve a problem but a general task needed to be solved. There have been evolved many clustering algorithms that differ in the way they understand the nature of a cluster and how to efficiently find them.
A significant part of clustering algorithms are based on the small distances that are supposed to exist between objects belonging to the same group, according to one metric distance. Another important way of detecting clusters is the through the dense areas of data space. As a result one may say that clustering is a multi objective optimization problem and the parameters used vary between algorithms and within the context of the same algorithm we have variations depending on the nature of the data needed to be clustered.
As a result of the above mentioned information it is clear to declare that clustering is not

an automatic task rather it is an iterative process of knowledge discovery through objectives optimization or even with modifications over the data in order to fulfill the needs of the algorithm.

On the context of this thesis we will only refer to density - based clustering algorithms. They have been used extensively in the second part of our approach to the cosmic web and the way we can interpret it.
In density clustering, clusters are defined as areas of higher density than the density of the remainder of the data set. On the other hand objects located in almost empty or in general sparse areas are considered as noise points which do not belong in any cluster. That has a direct connection with the approach we are trying to make throughout this thesis as one can see by the analysis of the cosmic web. In order to enlighten even more the reader we can see voids regions as empty regions in a similar clustering problem and galaxy clusters as dense regions of galaxies which is proportional to dense clusters in density - based clustering.
The most popular density based based clustering algorithm is the DBSCAN [35] which in contrast to many new methods it provides a well-defined cluster model called 'density-reachability'. It is based on connecting points within a distance threshold that satisfy a density criterion. A cluster is defined as the the density - connected points, which form a cluster of arbitrary shape, plus the points exist inside the cluster. It has low runtime, linear to the number of points. There are many other density based clustering algorithms which generalize the main idea of DBSCAN or use some extra features of the points in the dataset.
Another well known bu beyond the study of this thesis algorithm is the OPTICS [77] which is a generalization of DBSCAN that eliminates one of DBSCAN's parameters. Both of the above mentioned algorithms have a drawback as they expect some kind of density drop in order to detect cluster's boarders. That is not always the case as in many datasets there are overlapping sets and it is hard for this kind of algorithms to detect clusters.
There has been extensively research on that field and many new algorithms have risen. We shall present them by name as we do not use all of them in this thesis and we will analyze only the methods we have used on the following chapters.
Famous density clustering algorithms, not referenced before, by name (randomly): BIRCH [78], CLARANS [79], SPARCL [80], CURE [39], ROCK [81], CHAMELEON [38], DeBaCl [82], ABACUS [37].

Chapter 4

Cosmic Web Approach using Graph Tools

In this section we will present the main ideas we have used in order to extract some features about the cosmic web and its properties. As far as we know the main ideas are novel on the field. We will present gravity graphons, a novel approach using the above presented graphons, on the field of the cosmic web. We will also present the Gamma Hierarchical Analysis of the cosmic web as we have designed and used it. Moreover we will refer to the 'Gravity Fields' and 'Schwarzschild Radius Model' which were created by us in order to interpret the web behind the cosmos. Finally we will make a notion to the creation of a genetic algorithm which was used in order to guess the the way we are supposed to add edges in the graph approach of the cosmic web.

4.1 Gravity Graphons

During this chapter we will present a novel approach to the problem of the understanding of the cosmic web using graphons, that have already been defined. Throughout that approach we determine that the scope of understanding the cosmic web is through edges between galaxies (nodes). These edges might be interpreted as connections because of natural forces or as interactions between astronomical objects.

In stead of using the standard graphon theory about drawing uniformly random values for each node we have decided to test a novel approach which was initiated from the related to the cosmos bibliography [32]. On the limits of understanding of the cosmic web and its properties we have promoted gravity as the primary factor that governs the relationships between galaxies.
Playing such a prominent role on the universe it self as well as in the interactions we tried to model this force using graphons [14, 15]. In order to do that we have proceed to some specific changes to the main theory related to graphons.
The main idea, discussed below related to gravity graphons was initialized by the core

scores which is a common approach in core - periphery structures [83]. Core scores refer to a strategy roughly described as follows: each node has a score assigned to it and the possibility of including an edge between two nodes is depended on that scores. To be more specific a simple type of that scores application would be just a simple distance metric proportional to core scores and the probability of edge would be inversely proportional to that distance (e.g. euclidean distance).
Having studied both bibliographies about these topics we tried to combine them under a new approach to graphons, that take into consideration a kind of core - scores of the galaxies. On the same direction this novel approach will also contain a kind of randomness which in fact will lead to different realizations of the final proposed cosmic web. During this method we tried to encapsulate the gravity force.

- Firstly every node (galaxy) instead of taking a random parameter as its personal value is going to take its mass. It is optional but not recommended unless someone has full information about the magnitude of the masses, to regularize the masses between 0 and 1. This method will bring the proposed strategy closer to the original graphons but might result in galaxies of different masses to be projected close which will possibly bring some bias to the system.
- Staying to the direction of applying gravity force in a more 'graphic' way we suggest that every node should calculate the gravitational force of interaction between itself and every other galaxy (node). In order to do that we will use the formula from the law of gravity given as:

$$G \cdot \frac{m_1 \cdot m_2}{r^2}$$

where m_1 and m_2 are the masses of the two interacting galaxies, r is their euclidean distance and G is known as the gravitational constant (also known as Newtonian constant of gravitation) which is an empirical physical constant involved in the calculation of gravitational effects and the values of G are equal to (with some uncertainty to the final decimals):

  - G = 6.67430 x 10^{-11} $m^3 \cdot kg^{-1} \cdot s^{-2}$
  - G = 4.3009 x 10^{-3} pc$\cdot M_{sun}^{-1} \cdot (km/s)^2$

Theory that explains the above values is beyond the topic of this thesis.

- Having calculated the gravitational force of each pair of galaxies we proceed to a normalization of these values because it is possible to be extremely high or extremely low. In order to perform that normalization we have tried various approaches. A commonly used approach is that of softmax over all gravitational forces. We have chosen over other that type of normalization because it is widely applicable and it is used for soft approximation of the arg max. It is a function $\sigma : R^K \to R^K$ defined as follow:

$$\sigma(\mathrm{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \text{ for i = 1,...,K and z = } (z_1,...,z_K \in R^K)$$

  In other words this type of normalization uses exponential function in every element of the given set in order to transform the whole vector into a probability vector. Taking into consideration the properties and the slope of the exponential function one may see that the normalization is done so that to avoid present close enough values that are actually far apart. The above phenomenon is happening if there is a very large maximum in the data and one may try to normalize by dividing all values with it.
  On the other hand if the values of gravity appear to be very large (that will depend on the units over which gravity is being calculated) we advice you to use the logarithmic function in the softmax instead of the exponential. That will result in easy-to-handle values from a regular computer. In this thesis we have used both of the above methods as well as a built in normalization function of python's module sklearn.

- Having calculated the vector of the softmax of all gravitational forces of every node of the cosmic web we use it as a probability vector for the way we include edges. All these values among the vector are characteristic probabilities of connection of the node with another node represented through that element of the final vector. Drawing a uniformly random value for every possible edge we compare it with the softmax value of the gravity related with that edge in order to decide if we will include it to the final graph.

Using this approach we are able to incorporate to our system the gravity factor which plays and important role.

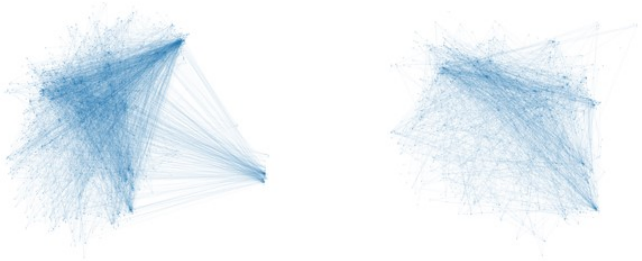

Figure 4.1: Results from the Gravity Graphon approach using Millennium Simulation data (small catalog of galaxies). Left: 3981 galaxies Right 1990 galaxies

We will also propose some variation of the algorithm that might result in interesting depictions of the cosmic web. It is possible before calculating the gravitational force between each pair in order to reduce the time taken for the task and in order to be more realistic to the expected from the astrophysicists to create a KNN graph [84] from the original data using a small K in order to find dense areas. After that one should calculate the distance of every node with its K neighbors. We arrange all these distances in descending order and erase the top 5% of them as we shall call them outliers. From the rest 95% we find the average value, which call radius R_{GG}. Finally we calculate gravitational force and continue the algorithm to its final steps using only the nodes inside a sphere having as center the current node and radius equal to R_{GG}. That results in a great time reduction and makes the algorithm versatile to the level of density we want to analyze the system. In the three dimensions of the cosmic web where KD-Trees [85] can me applied in O(nlogn) (needed for the creation of the KNN graph) the above algorithm will run in O(nlogn) where n is the number of galaxies.

Another approach that is possible to lead to interesting results uses more astrophysical theory and tries as well to incorporate it into the system we are trying to produce. In order to conduct this type or approach strong astrophysical background is needed and deeper understanding of the laws of the universe. In that alternative we shall try to incorporate for example a number (Ω) of laws/forces that govern the cosmos and we will do that with analogous way as with gravity.
Taking the intuition from the above algorithm we modify the values of each node and upgrade it from a single value to a vector of values. We also need these laws to be gravity-like forces that interact between two galaxies only. Every galaxy now is binded with a vector of the form z = (z_1,...,z_Ω) where every z_i represent a characteristic of the galaxy like mass. That could be luminosity or star formation rate or some other parameter beyond

the reach of our knowledge that is also important for an astrophysicist. Having laws of the form: Law(i, j) for i and j being the above mentioned parameters and law defines some kind of interaction between the two objects.
Taking all these into consideration the proposed method is straight forward. We first define the set of Laws (L, where $|L| = \Omega$) and calculate its values between pair of galaxies (either using the KNN approach or the all-pair approach). Using the astrophysical knowledge about the importance of the forces/interactions of the Laws in the relation between two objects we may define a set of weights W = (w_1,...,w_Ω) and $\sum W = 1$ with every weight be respectively chosen for every law in L. Finally the probability of including an edge between galaxies i and j is given as:

$$\mathrm{P(i,\ j)} = \mathrm{softmax}(\sum_{k=1}^{k=\Omega} w_k \cdot L_k(z_i, z_j))$$

And one may calculate the above probability for every pair of galaxy needed. It is obvious that the gravitational approach is a sub-case of the above generalization where gravity is the only law and its weight equals one.
Finally it is clear that this method demands strong astrophysical background and the more related is the reader with the cosmology the better he/she can use the method. Also possible is to let an algorithm determine the weights according to some metric or heuristic or even let it try to learn the weights by itself.

4.2 Gamma Hierarchical Analysis

Using as intuition the previous knowledge about scale free graph [3] as we have stated and taking into consideration that this type of networks appear very often in nature [44] we propose a new method of analysis of the cosmic web. The proposed method takes as input an already constructed graph in general but on the topic of this thesis the graph should be constructed using the astronomical data.

The main idea and contribution of this novel method is to suggest a compression of the information needed to describe a network in general, especially the cosmic web. Using the proposed method it is possible to keep track of the evolution of the network hierarchically and gives the ability to choose the level of zooming in which you study the network. It is clear to the reader that these properties are important for a large scale network such as the cosmic web, we refer to the phrase cosmic web on this topic as a graph as we have defined it.

Having as preliminaries that scale free networks appear naturally and the network under study is a natural network we hypothesized that this kind of network can be divided into separate levels and the components of each level are kind of uniform. We tried to approach the problem by contracting a general model of zooming out while keeping information [52].

We supposed that every one of these levels of zooming out are in fact scale free networks and are characterized by a γ factor. Taking into consideration the fact that the network might not be completely uniform and that it might be fragmented into connected components which are also scale free networks with different γs with each other we defined as the γ factor of each level to be the average of the γs of all components on this level. In general the intuition behind this algorithm is that the universe has similar characteristics in every level of zooming-in which are also networks that follow power law distributions [53].
The algorithm begins with the whole network and acts divisively by erasing edges in every level of zooming in. In every level it tries to fit a different γ parameter, where γ is the power law distribution's exponent [31]. The algorithm has the following steps:

- On the level we currently are we run over all edges and for every edge we check weather or not is needed to be removed in order to zoom in further. To do so we temporally remove the edge.
- We estimate all connected components of the resulted graph. Over all these components we estimate their γs using vanilla MLE estimator [31] defined as:

$$\gamma_{vanilla} = \frac{n}{\sum_{i=1}^{n} ln(\frac{d(X_i)}{k_{min}})} + 1$$

  where the above formula uses information from uniform sampling of the nodes of the network $X_1, ..., X_n$ with degrees d(X_1),...,d(X_n).
- We calculate the variation of the γs of the components of this level.
- If the removal of the edge result in isolated vertices or no variation of the γs we repair the edge else we check the new γ variation with the old one and if it is smaller we keep the edge removed and update the new γ variation.

The above algorithm could be run in an arbitrary number of times depending on the zoom we would like to result into the network. As it extract the γ sequence in every level one may keep only the average per level and use this information to reconstruct the network. We have applied the above mentioned method only to relatively small networks

in which it appeared to do well.
In order to reconstruct we start from bottom to top by using the final gamma of the sequence an with it we create a random power law tree sequence that is being used in order to configure a network model. When we need to upgrade to the above level we create a new sequence which we compare with the old one. After sorting both sequences we compare them element-wise and for every element of the second sequence that is greater that its associate in first we add edges in order to approach that new degree.

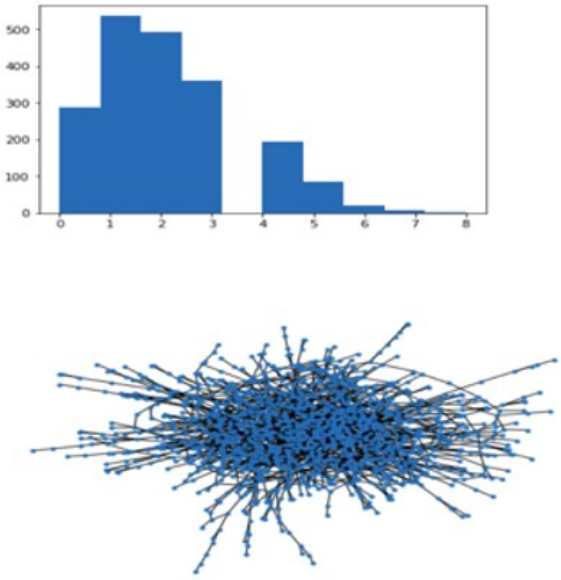

Figure 4.2: Results from the reconstruction of a 2k galaxy network from gravity graphons, Degree sequence of that reconstruction, Millennium Simulation data (small catalog of galaxies).

The above method is heuristic and appeared to reconstruct well the cosmic web produced using data from the Millennium simulation for about 2,000 galaxies. It is needed to be further developed or possibly altered to the original idea about this approach.

Although in might seem not to look like the original network presented on the previous section it has almost the same characteristics. We used a heuristic approach by calculating and comparing average shortest path, degree sequences and mean betweenness centrality over both networks.

The main idea which was to work from bottom to top. We would like to find clusters [8] at the beginning and fit a gamma parameter within these first clusters and extract the gamma average as representative of this level. Later on we would like to collapse these

clusters into super-nodes and create a new graph to which we will repeat this process until we reach one final vertex, containing the whole universe.
The main problem with that point of view lays on the fact that in order to define clusters of every level a metric is needed. The problem is rooted to the fact that we planned to extract clusters by creating a dendrogram [54] and then cutting to an optimal point. The only propose as far as we understand the cosmos we can make here is using a metric defined as:

$$\min(\sum_{i=0}^{C} (\text{cluster radius}[i] - \text{expected cluster radius}^2))$$

where the sum is over all clusters. Extracting the appropriate metric in order to define clusters on this problem is not trivial.
One may understand the proposed method as a way of going from ultra small world networks [5], which are the galactic neighbors and the clusters where the distance between galaxies is relatively small and we have many nodes close to one another, to a network of small world [60], which may be defined as the extended galactic regions. Finally as zooming out one may see a less 'small world' network which depends on the level of zoom. The understanding of the universe as it was proposed by Lambda - CDM models [75] could be related to directly with this method but in order to improve the produced results and determine a general approach we need further knowledge about the cosmos and some metrics of it that will help in the creation and reconstruction of the cosmic web.

4.3 Gravity Fields approach

It is widely known the paper of Barabási et.al. [32] about the structure of the cosmic web. This paper was the initiate about this thesis and we tried to propose new models and compare them with the models proposed on that paper. The idea is to find a heuristic way in order to construct networks of the cosmic web that have some desired properties and are meaningful for the astrophysicists.
For this experiment and for the rest of the thesis we will use data from the Illustris cosmological simulation [29]. We drew these data using a custom data crawler and the free API that Illustris provides available for public. Also we will use the catalog for subhalos (galaxies) of different redshifts but mainly for z = 0 which refers to the current age of the universe.

Many ideas have been proposed in the study of the structure behind of the cosmic web. On the referred paper there have been seven new models proposed with some of them being kind of similar or generalization of one another. The main idea was to define a radius of effect for every galaxy and using this radius to draw edges over the network. Simplest ideas suggested the construction of a network using a fixed length radius which

might not result in expected results as fixed radius fails to capture the nature of the cosmic web and the differences between galaxies. Setting this radius to very small value is possible to result in disconnected network with numerous connected components and some disconnected vertices. On the other hand setting this parameter to high values results in an ultra dense network which is possible to fail to provide the appropriate information.

Another proposed method is of creating a directed network using a widely known method of the graph and network theory. In order to create that network it is proposed to connect every galaxy with its k nearest neighbors. This results in a directed knn network [84]. Approaching galaxies in that way is useful because for small k's and by eliminating some outliers it is possible to detect dense areas and with large k's one may extract more features about the large distant neighbors of the network. Moreover with this method the average intergalactic distance is effectively calculated.
Another proposed model by the above mentioned study which had great success used a kind or arbitrary radius which seemed to did well on the context of the cosmic web. Half-mass radius ($R_i^{1/2}$) is defined as the radius containing half of the total mass of the Subhalo, meaning containing half of the total mass of the galaxy. Taking into consideration the nature of matter in universe (baryonics) and the effect of gravity over matter which will collapse to itself we could safely suppose that the value of half-mass radius will be kind of small. In the proposed method the authors suggested to use as radius of every galaxy the product of its half-mass radius with a constant 'a'. The prime advantage of this approach which might be the reason why it performed so well is deep rooted in the physics model, as far as we are able to explain it. In simple terms half mass radius allows each galaxy to have distinct radius of effect. As far as we can explain it the half mass radius is depended on the size and the mass of the galaxy. With that formation it is possible for heavier galaxies to have more connections than lighter and these connections to be meaningful for the interpretation of the cosmos. Keeping the notation of the paper this model will be referred as M4 in the following context.

Finally the study suggests two more models as well which in fact need more information about galaxies. Taken into consideration that these information might not be accurate it is possible to have pure results. In order to construct these models the velocity of the galaxies is needed.

In general all of these models require simple information about the galaxies in order to extract the final network. All of them demand the position of the galaxy, and some demand the size and velocity as well. The main problem from these approaches is to find a way to determine what model is better and how to evaluate its superiority over other models and over itself but with the use of other parameters.
In order to further depict that and enlighten the reader about the difficulty of the task of finding a metric of success we will provide some images from different models realization of the galaxies, as they are presented in the above mentioned study.

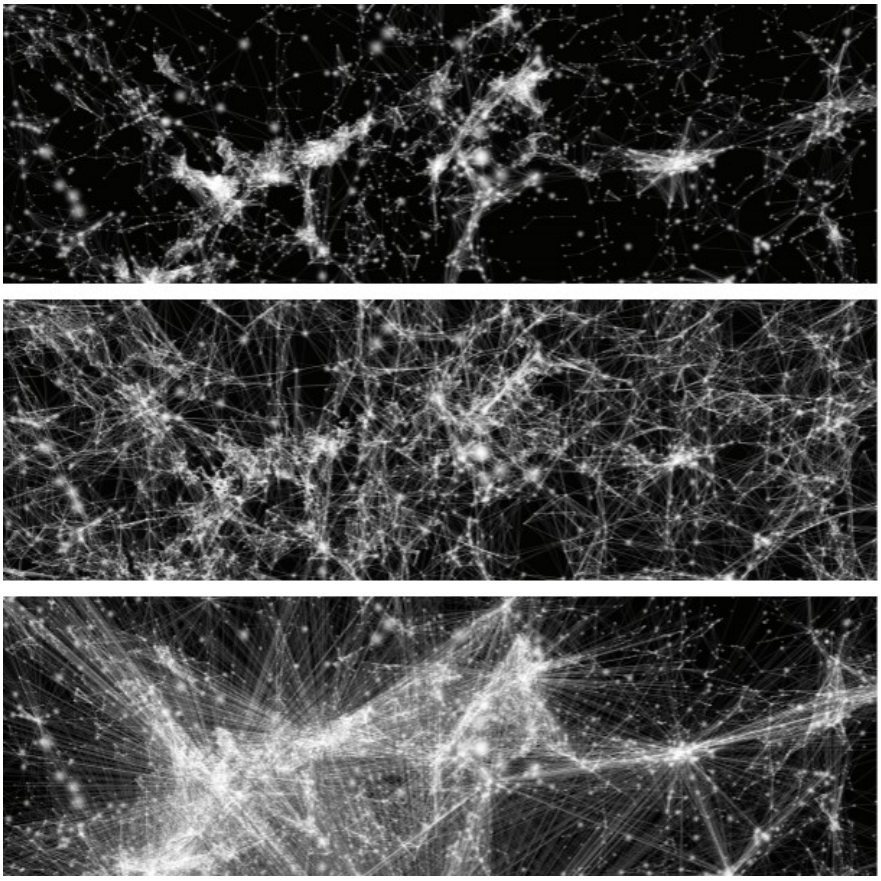

Figure 4.3: Realizations of three models proposed by the study using Illustris project data. Source: 'The Network Behind the Cosmic Web'

As it is obvious for the reader at that point that determining the optimal network between these models for so large networks is, or even for smaller ones, is a very difficult task which has little or no point of reference. Barabási et.al. [32] used some correlations among each of the model and some astronomical features of the galaxies connected in these models. These features were purely astrophysical such as peculiar speed, star metallicity, star formation rate and others. Also they have used some of the graph theory to extract features about connected components, clustering coefficients and average degree of the resulting realization of the universe [58, 59, 46].

Initializing our thought with the above applied methods and their promising results we tried to combine once again the laws of nature with graph theory in order to create a new model ourselves. We shall present the intuition behind our model and some methods to compare it with a related proposed above model as far as we could understand the physics theory behind the structure of the cosmos.

Taking again into consideration the importance of gravity it is a straight forward once again to try and find a way to incorporate it into the network theory. We wanted to establish a radius that would be directly proportional to the gravity created by a galaxy. That would be defined as a kind of gravity field and the aim to find the effective radius of this gravity field. The intuition behind this approach is based on general relativity and the way suggests and explains the space-time fabric as a lattice over which lay the cosmological objects and bend it. This inclination of space-time is in fact the gravity that some objects 'feel' from other objects. The above explanation is given for simplicity and for clarity about how the reader could imagine the time-space in order to understand the reasons for proposing the so called 'Gravity Fields' model.

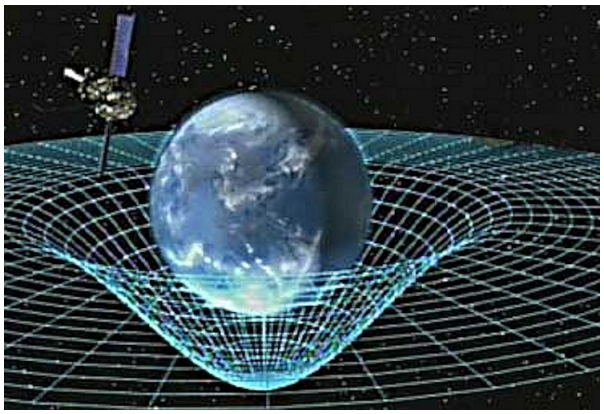

Figure 4.4: Space-time as it is believed to bend under a heavy object (Earth) Source: Wikipedia

In order to establish our method we will need information about the position and mass of the galaxies. Having these we will proceed to the method of constructing the network. We run through each pair of galaxy (unordered pair) and compute the gravitational force among the galaxies on that pair.
We compute the distance of the galaxies and if that distance is less or equal to the gravitational force multiplied by a constant factor, for scaling, we draw an edge between the galaxies. We mention that the resulted graph is undirected as all edges are symmetrical. Symmetry is natural as the gravity as a force which is applied to both objects that take part in its calculation. Formally the algorithm is:

Algorithm 2 Gravity Fields

Result: Gravity Field Network's Edges

```
M = all masses
P = all positions
HM = all half mass radii
edges =[]
a = 1e-32
for pair in all pairs do
    i, j = pair
    dist = euclidean(i, j)
    rad sum = M[i] * HM[i] + M[j] * HM[j]
    grav = gravity force(i, j)
    gravity = a * grav * rad sum
    if gravity * rad sum >= dist then
        edges.append([i, j])
    end
end
return edges
```

The algorithm defined above uses except of the gravity in order to introduce this force into network construction a constant parameter tuned by us. More over in order to further enhance the more massive galaxies of obtaining more edges we have used the rad sum metric which is a distance weighted with the galaxy mass in order to further enhance gravity. We have also tried approaches without that parameter and tuned constant 'a' to a new value. Throughout this section we will present results from the above algorithm.

For demonstration reasons we will use only a small fraction (3,800) from the $\sim$ 120,000 galaxies produced by the Illustris simulation at redshift z = 0.0. Moreover we will compare our results with the model proposed by Barabási et.al. that is related with the half mass radius of each galaxy. We believe that this comparison is solid as we have explained the way we understand the half mass radius of a galaxy and its relationship with matter and gravity.
In order to create networks with similar number of edges in order to make the comparison we tuned both constant parameters in models to values: $a_{GravFields}$ = 1e-32, a = 15 (referred to the proposed model). The main difference of our approach is the following: we are going to suppose that every galaxy is a source of a gravitational field and if two fields intersect then the sources should be connected. Parameter $a_{GravFields}$ is used to

tune the density of edges.

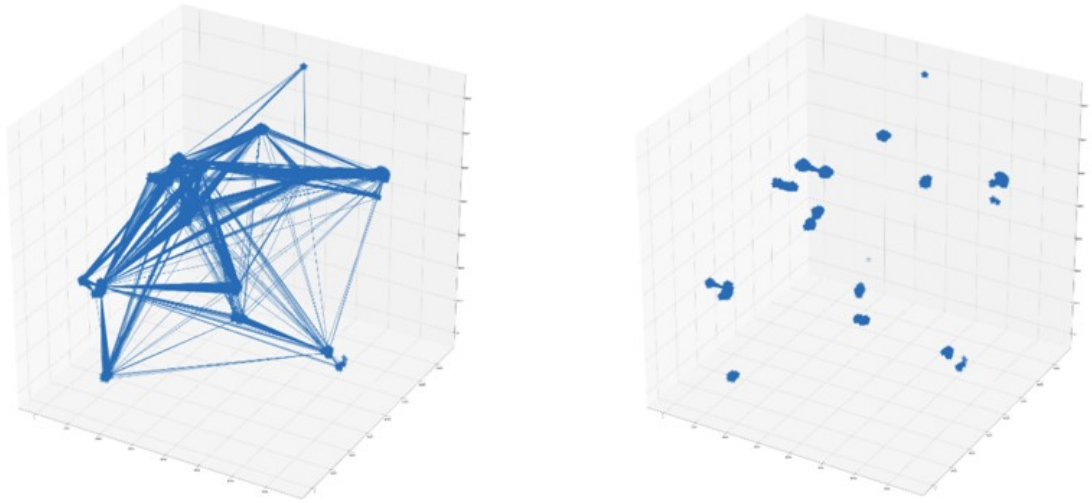

Figure 4.5: Left: Network from Gravity Fields, Right: Network from M4, Data: 3,800 galaxies from Illustris-3, z = 0.0

Both networks have circa 18,000 edges. In order to compare these two realization we have used some heuristic measures drawn from graph theory. These are the moments of the degrees (1st, 2nd, 3rd), average path lengths (if the graph is disconnected we compute it over the largest connected component) and betweenness centrality. The results are the following:

Average Path length of Gravity Fields: ~ 2.43
Average Path length of largest connected component of M4: ~ 2.01

Betweenness centrality of Gravity Fields: $\sim$ 4e-4
Betweenness centrality of largest connected component of M4: $\sim$13e-4

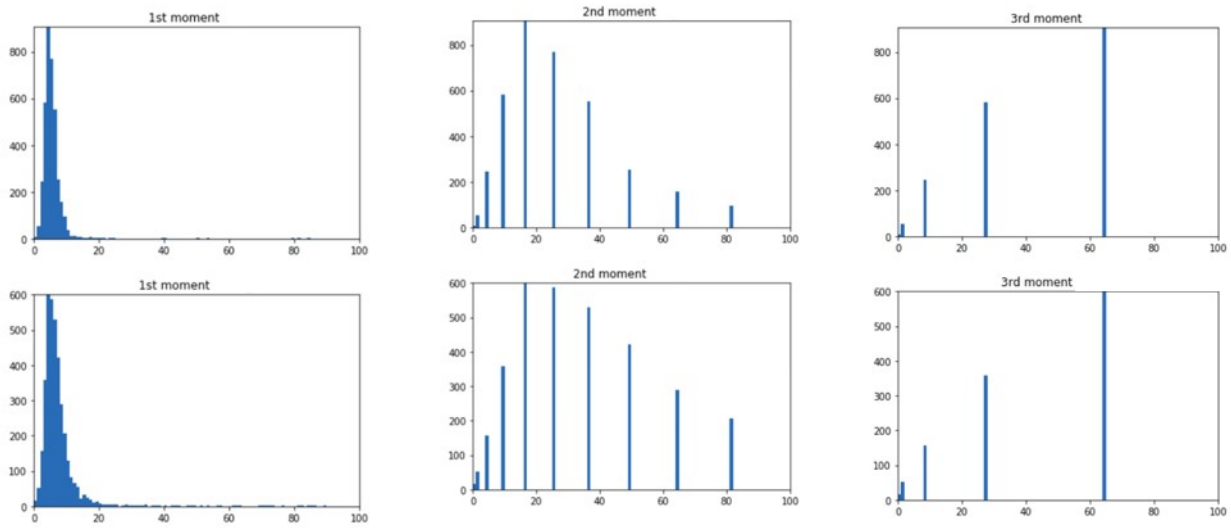

Figure 4.6: Top row: Gravity Fields, Bottom row: M4, Data: 3,800 galaxies from Illustris-3, z = 0.0

Differences of the two metrics may be rooted to the fact that M4 [32] results in disconnected graph. As far as we can use the the degree moments we can state that both networks have similar moments. Differences is possible to exist through but in general we observe similar behavior regarding the degrees.

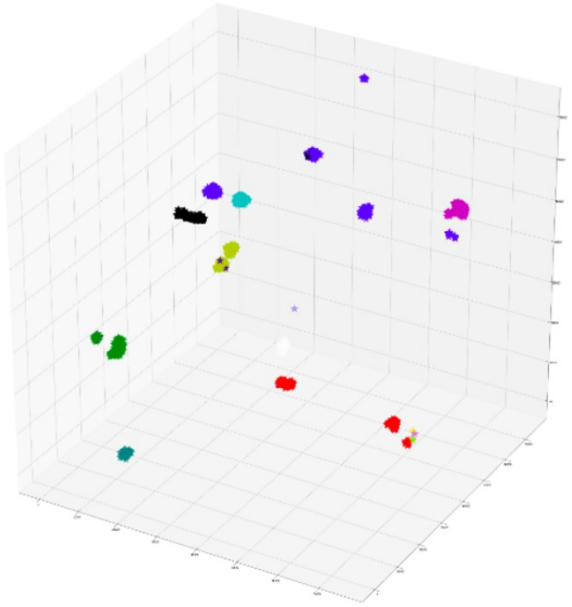

Figure 4.7: Maximum Modularity community detection of Gravity Fields' network, Data: 3,800 galaxies from Illustris-3, z = 0.0

Finally in order to finish the demonstration of this method and using the fact that

these 3,800 galaxy appear to be in communities we tried to find communities from our network using maximum modularity. The results as presented above indicate that the communities formed from M4 as connected components appear to be the results of applying community detection to our method. Of course that might not be the case in more dense network with many galaxies where the tuning of parameter $a_{GravFields}$ needs to be re-evaluated.

4.4 Schwarzschild Radius approach

On this section we will present another method used in order to model the creation of the cosmic web. Studying about the properties of the galaxies and in general of the matter in universe and using the intuition about the need of existence of a radius, similar thought as on M4 model we will make use of a well studied theoretical measure known as Schwarzschild radius [33].
Schwarzschild radius also known as gravitational radius is a physical parameter which derived during the solution of the Einstein's field equations by Schwarzschild. This radius is defined as the radius of the event horizon of a Schwarzschild black hole. It is so characteristic as a measure that is associated with every quantity of mass. It was named after K.Schwarzschild who calculated the exact solution of the theory of the general relativity. This parameter is estimated for many astronomical objects such as Earth, Milky way and others. It represents the ability of mass to make the space time to curve. In other literature is stated as the radius below which any object turn into a black hole if it is compressed to a sphere with that or less radius.
Determined as r_s Schwarzschild radius is given by the formula:

$$r_s = \frac{2 \cdot G \cdot M}{c^2}$$

where G is the gravitational constant as defined in previous chapters and c is the speed of light.

The intuition behind using this radius is straight forward. Instead of using the half mass radius that is not so strongly related with mass and gravity we have chosen to use a multiple of the r_s in order to create the model. Method of construction of the model is the same as in M4 using again a constant in order to scale this radius to an appropriate value, constant needed to be tuned as in the previous models. The main reasons we used this radius is because it is directly dependent on the mass of the object and not its shape, which is the case in the half mass radius, because using half mass radius in two galaxies of same masses but one of them being spiral and the other ellipsoid will probably result in two different radius of effect. On the other hand r_s is a parameter well studied and

based deeply on the general relativity that governs the universe. Moreover the small size of this value allows easier scale and tune in contrast of having to tune larger values (i.e. half mass radii) by using a constant.

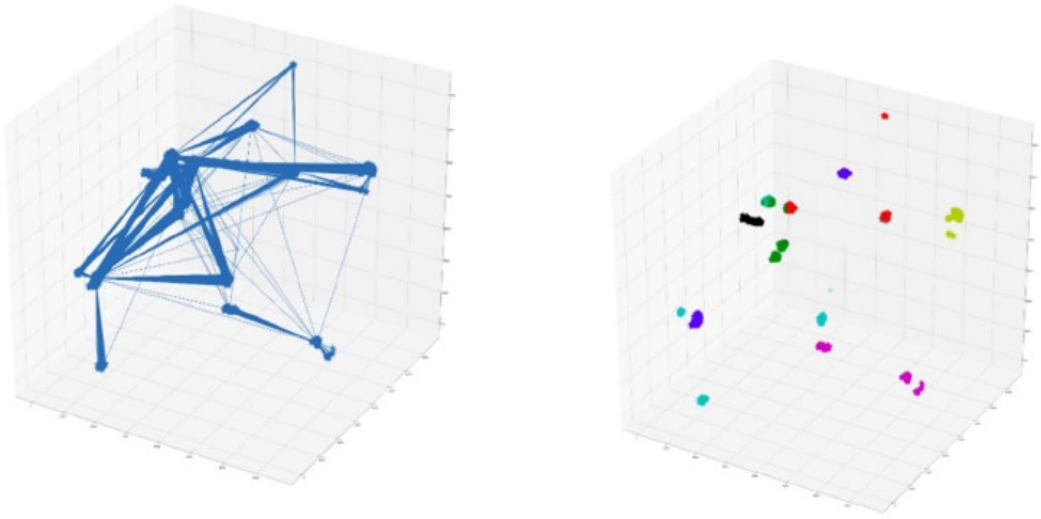

Figure 4.8: Left: Netowork from applying Schwarzschild radius model, Right: Maximum Modularity community detection, Data: 3,800 galaxies from Illustris-3, z = 0.0

Again we used the same 3,800 galaxies and the resulted network on the left has circa 18,000 edges as well.From the above graph even if we expected interesting results looking on the communities formation we may declare that this method need further optimization or should also be completed by incorporating more data.

4.5 Genetic algorithm attempt

Genetic algorithms is a meta-heuristic process which development is based on the natural genetic evolution (natural selection). Process followed by a genetic algorithm has some basic steps which can be summarized into: initialization of the population, crossover between individuals of the population, mutation of the individuals, selection of individuals of the next generation using a defined metric (fitness function) and a terminating condition.

In this section we will present a genetic process we followed in order to determine a more specific effective radius in the context of the previous discuss. The intuition behind this method is to find a formula which relates galaxy pairs and using that formula instead of the Gravity Fields or instead of the M4 model to achieve improved results.

In order to model this need into a genetic terms we had to first define how a typical chromosome of the population will look like. A chromosome will be a vector of four values: a, b, c and d. These values will be interpreted, after their final values found, into the

following formula:

$$f_{eff}(i,j) = a \cdot \frac{m_i^b \cdot m_j^c}{r^d}$$

As we can see the above formula has the formation of the gravity in a more general terms. The intuition behind it is to optimally find gravity as the best force or find some other formula that relates galaxies in such an effective manner that can be used in order to determine if their distance in comparison with f_{eff} will result in including or not the edge on the final network. In order to create the population of the genetic algorithm we selected randomly 100 galaxies and applied to this miniature of the system the formula which was represented in every chromosome. We chose that number of galaxies for execution-time reason. Moreover we did not use any tuning constant on this attempt as we add edge to the network only if the distance between the two galaxies was less than f_{eff}. Results from the best chromosome for the setup described above as well as with after community detection are:

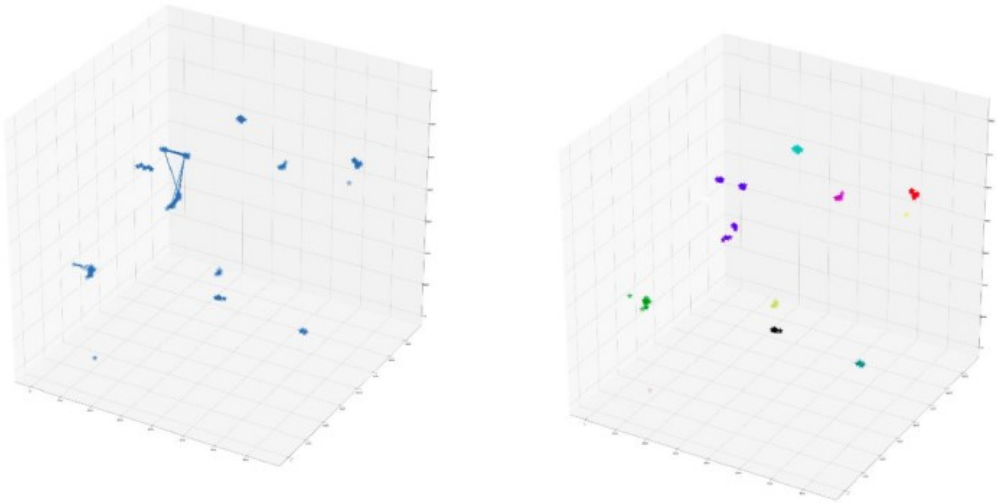

Figure 4.9: Left: Network from applying best chromosome of GA, Right: Maximum Modularity community detection, Data: 100 galaxies from Illustris-3, z = 0.0

The parameter set for the above best chromosome after executing the GA is: a ∼1e15, b = 2, c = 1, d = 5. Also we declare that the parameters of the GA are: population = 50, cross probability = 0.9, mutation probability = 0.01 and elitism = 1.
Moreover we ran the algorithm to a predetermined number of iterations (30).

Regarding the fitness function used we have applied two functions but both of them resulted in similar values for the best chromosome. The first was related with the average degree of the resulting graph which we ultimately wanted to be the same as in the other graphs we have used from the previous models (i.e. Gravity Fields, M4 etc.), circa 4.6. The second approach was related to the degree distribution of the graph. We wanted to be more close to a power law degree distribution rather than an exponential distribution (characteristic for Poisson like distribution).
Using these two approaches we tried to utilize a GA which will promote to next generation chromosomes that will result in a f_{eff} that when applied in the construction of the final network will either resemble the networks coming from the studied models or scale free networks as we believe that characterize the cosmic web. Examining the produced degree sequence we find out that the produced network tries to approach a scale free network but the resulted values about the formula do not seem promising because as we can see they are far from the gravity. That might be the result of an over-fit over the small randomly selected population. In order to improve the performance of the GA proposed we would need to conduct experiments with larger chromosome populations and perhaps altering all the other parameters of the algorithm. Of course it is needed to increase drastically the number of galaxies used to apply the formula to some thousands in order to have a more representative sample of the total set of galaxies.

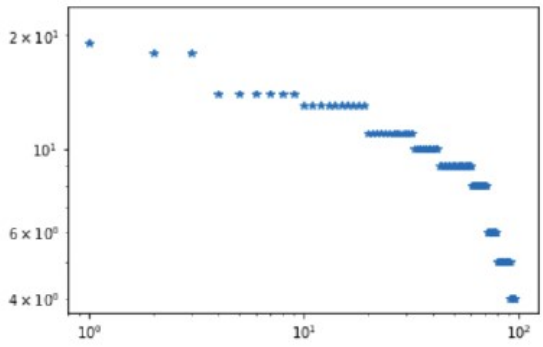

Figure 4.10: Loglog graph of degree sequence, Data: 100 galaxies from Illustris-3, z = 0.0

Chapter 5

Cosmic Web Approach using Spatial Clustering Tools

In this section we will present the main ideas we have used in order to extract some features about the cosmic web and its properties. In order to do that we have used the main ideas of the spatial clustering as they were defined in previous section. We have moved from graphical representation of the cosmic web to its analysis without using edges to extract its feature. We will only take advantage of the topology in order to extract the structure and make use of the mass as a filter. We will present the use of octrees over the cosmic web, a novel (as far as we know) approach called 'Gravity Lattice'. We will also present and apply some spatial algorithms with or without altering them over the data of the galaxies.

5.1 Octree

An octree is a tree-like data structure that is used for spatial clustering and in order to model 3D data in an appropriate way that allows the user to perform insertion and search really fast without having to look every point in the dataset. Every internal node of an octree consists of eight children. This type of structures are the most often used in 3D space. In order to construct them we will have to recursively divide the space into eight octants, as they called [34].
Octrees are the direct expansion of the quadtrees in 2D space.

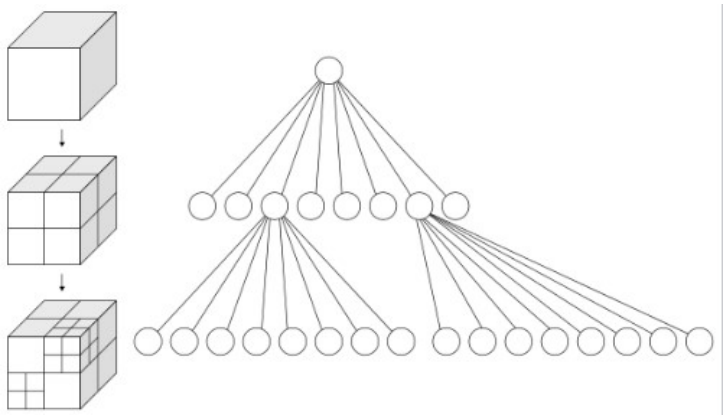

Figure 5.1: Recursive subdivision of space by an octree and its tree-like representation, Source: Wikipedia

We have made a small modification over the constructed octree in order to fulfill the needs of this thesis. As we can see in the above image after inserting all the galaxies into the octree we will result in the division of the first, giant, cube into many small sub-cubes. First of all we have tuned the number of recursive calls of the function that creates the levels of the tree. Increasing that number lead to smaller cubes with perhaps greater density.
At the beginning of that approach we thought that the more we divide the initial cube the denser the leaf-cubes will be. That was a false prediction rooted on optical depiction of the initial cube in 3D which did not allow to examine the correct properties of the data.

After tuning we concluded that the best fit for the used data in octree formation was only three levels in the data structure. The branching factor of the octree is so large that even with little levels we end up into many final cubes.
Also we made some other modification to extract even more data on the context of the cosmological approach. We model the cube as an object of mass and size. After we have inserted all galaxies used at this section we have resulted in final cubes which had their positions, sizes and masses defined.

Moreover we implemented a density based pruning [86] of the tree, which was possible since we knew the mass and size of each leaf. In order to conduct the tuning operation we have set a threshold which is global for the whole tree. We started from the root and directed to the leaves. If a leaf was empty, meaning zero density, we erase that. On the other hand if that leaf was below the threshold density we collapsed it to its father as well as all other leaves with the same father.
The above pruning method was not so effective as it was possible to have many small children that led to collapsing a whole level to the previous and that might trigger even

more collisions. The problem about the pruning and this procedure is found on the cubic shape of the leaves which are kind of strict in the context of what we are trying to study. We have observed formations that could have been described with and ellipsoid in a great success to be on this method purely encapsulated into more than one cubic leaves that are also of small density. The above mentioned observation can be interpreted well using astrophysical theory as the galaxy formations, walls, filaments, clusters, is not usually of cubic shape as we have studied and presented them previously. They are kind of either long formations or really tight and concentrated clusters.

We have eliminated the possibility of using spheres instead of cubes in order to model the parts of the cosmic web through octree but that would result in leaving many areas of the initial cube without any representative. Also we have thought and tried to model the initial cube through a sphere and then tried to fit into the initial sphere more smaller but that had the same results of not fully coverage. We refer to spheres as they appear to be more natural in order to model the galactic structure.

Another improvement that would probably work if it was combined with something else, from the computational geometry that is beyond this thesis, is using cylinders of arbitrary radius and height placed over the initial cube. Like an octree this structure could also be divided into smaller sub-cylinders. The above division should haven not in a uniform way, meaning that not all children of a cylinder would have the characteristics. This approach is possible to let us use small radii but long height cylinders to model filaments and large radii and small height cylinders for clusters. That would be the ultimate way of determining structures in the cosmic web and use their density to categorize them following the criteria about density described before. We have tried to model this process but the results were not correct as there were problems related to geometry and the way we would choose the cylinders characteristics.

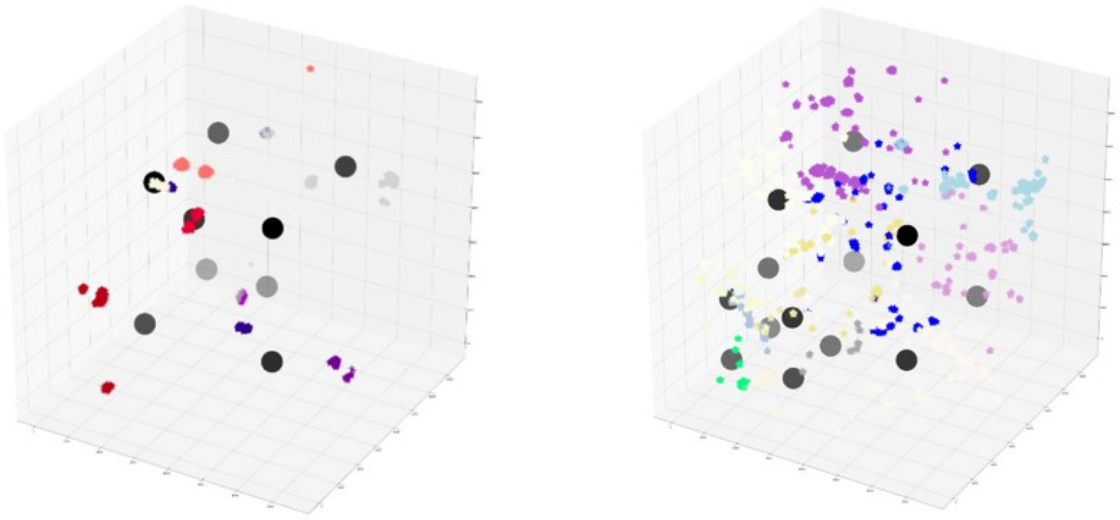

Figure 5.2: Left: Octree results with 3,800 galaxies, Right Octree results with 12,160 galaxies, Black big dots are the centers of the final cubic leaves, Data: Illustris-3, z = 0.0

Above we present the results from the usage of the modified octree as we have described int before. As we have stated the results are not so promising as we scale up with the number of the galaxies. At the beginning with small number of galaxies appear to approximate the structures in some case really well as the center of the cubic leaves appear over the clusters. On the other hand there are many clusters that wrongly assigned to a cube whose center is far from the galaxies of that cluster.
As we increase the number of galaxies and increase the depth of the octree we can see that some levels of the tree have collapsed and that resulted in assigning almost a quarter of all galaxies to one cube. Also worth to mention that during the process we have resulted into cubes with galaxies count from 100 to 900 which seems promising. Finally it is needed to mention that the visual results presented above may not be clear as the 3D data presented in a 2D image might result into inconveniences. Taking into consideration the above results and the strictness of the octree related to the cubic shaped we could summarize that this method needs improvements, as the one mentioned with the cylinders, in order to provide interesting features about the structure of the universe.

5.2 Gravity Lattice

Inspired once again, as we did in some previous chapters of this thesis by the gravity and its dominance over the universe we propose a novel approach in order to detect structures of the cosmos. We will first give an very simple example as an intuition and then explain it in a more proper way using this example.

Imagine we have an area filled with a substance like honey or yogurt. If we spread among this area some marbles of arbitrary size and weight and then remove them their signature will remain. By the word signature we mean that even if they are not there an outside observer can find out that something spherical existed and according to the curvature of the substance make an estimation about the weight of that object.

Although this example might be very simple the intuition behind it is strong and can be applied to the universe and the galaxy formation. The reader can make a direct association between galaxies and marbles, honey-like substance and cube over which all galaxies lay. The main problem of that is the way of filling the cube with that type of 'substance' and how to define the way of affecting that substance with our galaxy-marble. The answer is once again gravity, the force that rules the cosmos.

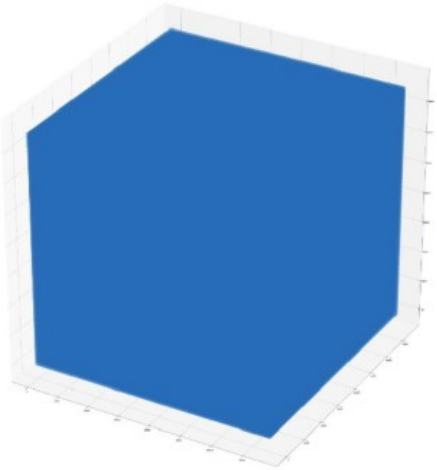

Figure 5.3: Representation after filling the cube with small spheres of 1kg

The prime aim of that model is to make the galaxies leave their gravitational signature of a substance and then remove them as their signature will remain. After that we will try to study only the signature. Like the intuition that signature will offer to an outside observer information about its owner.
In order to know specify the way of applying this method we proceed to more technical matters on which the results are dependent. Firstly, we create a 3D regular lattice without the edges. The vertices of that lattice are masses of 1kg, the intuition behind that is based on electromagnetism where in order to determine the electric force we use test loads. On this method we want to estimate gravity so we will create many 'gravity test loads'. The main parameter that rules this model is the distance of these spheres. The smaller the distance the more spheres we have and as a result better resolution of the gravitational

signature.
After having the honey-like 'substance' defined we will use it in order to put the galaxies on the system and extract their signature. Each galaxy is assigned a radius, that value could be anything defined in previous chapters as well, for the purpose of demonstration we have used as radius of each galaxy a value proportional to its half-mass radius, by a small factor. That results in creating a sphere around galaxy which encapsulates some of the gravity test loads. Even with the greater of the radius the amount of load test spheres enclosed will be finite, based on the way we have constructed the lattice.
Following we compute the gravitational force between the galaxy and each one of these spheres among the three main axes. After inserting all galaxies the each sphere will or will not have been affected by one or many of them. Each sphere will end up with a vector of three entries representing the total gravitational force that is exerted on the sphere. At that point we suppose that the movement, if the test loads were supposed to move, under the effect of the galaxies is proportional to the final gravitational force exerted on them. The intuition behind is based on Newton's second law about the total power exerted over an object and its resulting acceleration only with the difference that here we suppose the total force is proportional to the total moving.
After we have finished the insertion of the galaxies and removed them we will in fact let the spheres move. As we have expected some of the spheres have been affected among some or all axes and some others remained in their positions. So at the end we have a modified lattice where we can detect empty spaces as well as galactic formations. In order to do that we would have to go through the final lattice and detect the amount of movement of the spheres that has been made and towards what direction. The above method will also be useful in studying time evolution of the universe where galaxies may have moved, destroyed or created but the signature they have left in the past will remain and following signatures could lead us to track down galactic behaviors and patterns of the universe.

Algorithm 3 How to produce the final Gravity Lattice

Result: New 3D lattice

```
start_lattice = create_lattice(inter_dist, CubeSide)
 for gal in galaxies do
    rad = 3 * half_mass_radius
     x_neighbors = [range(x_center - rad, x_center + rad, inter_dist)]
     y_neighbors = [range(y_center - rad, y_center + rad, inter_dist)]
     z_neighbors = [range(z_center - rad, z_center + rad, inter_dist)]
     neighbors = product(round(x_neighbors), round(y_neighbors), round(z_neighbors))
     for neig in neighbors do
        neig[0] += gravity_x(gal, neig)
         neig[1] += gravity_y(gal, neig)
         neig[2] += gravity_z(gal, neig)

    end
end
return start_lattice // as it has changed through the previous procedure
```

In order to analyze the complexity of the proposed method we have presented above the algorithm used to construct the new lattice. Although it might seem at first glance that it is quadratic this is not the case. Taking into account that the maximum radius of effect of a galaxy can include finite number of spheres, C, then the nested for loop is not 'big' as a result the above algorithm run in O(C * N), where N is the total number of galaxies in the system. C is not going to be always negligible as if we wanted higher resolution meaning more spheres and less distance between them that would result in large values of C. But for the purpose of that thesis the inter-distance is kept relatively large.

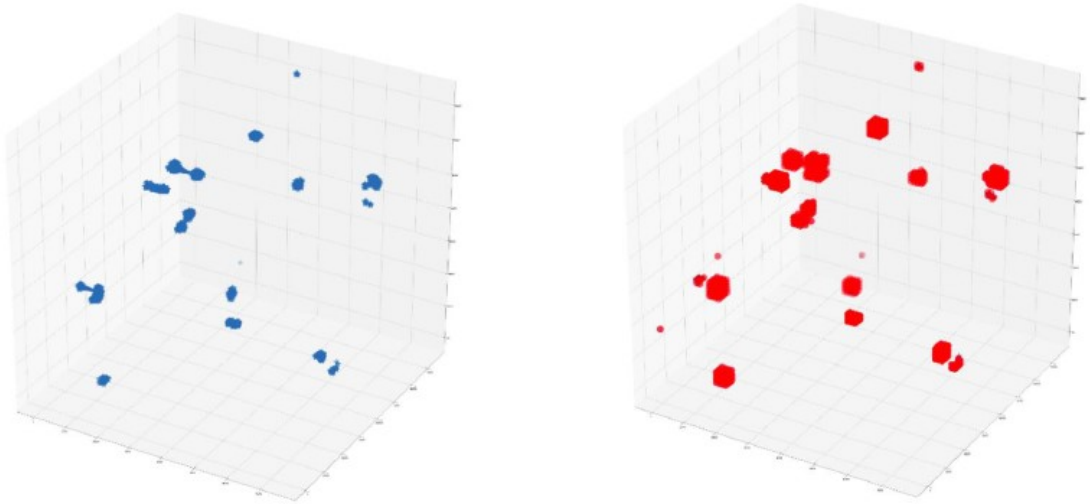

Figure 5.4: Left: Model M4 results, Right: Gravity Lattice results, Data: 3,800 galaxies from Illustris-3, z = 0.0

For the results presented above we have used inter-distance equal to 1 Mpc (but we have also tried it in 0.5 Mpc and it was still fast in execution, less than 1 sec) and as effective distance we use three times the half mass radius. After calculating the total gravitational force over all spheres we multiplied the result by a factor of 1e10 and updated the position of the spheres. Spheres beyond the limits of the cube, if exist, will be placed on limits. After that we present only the spheres that have changed their positions. Comparing the results on the above figure we can see that Gravity Lattice encloses all the information about the galaxies and their formation. Although it might seem shaped kind of cubic this model is able to fit over any shape of the galaxy formation. Even if the galaxies are formed like an arm the test loads will detect them their movement towards the galaxies will create a kind of rectangular solid.
Even without the background about this method the above figure offers visual credentials about its validity as they seem to depict vividly the galactic structure. We can detect clusters on the regions where moved test loads are more compact and voids where spheres are intact.
Another suggestion in order to improve this model would be to firstly to reduce the distance between the test loads. Another way to approach the model and its improvements is to find a proper radius of effect. For example as we use only gravity and masses it would be interesting to find a radius depended on mass. It is possible that these modification will have better results.

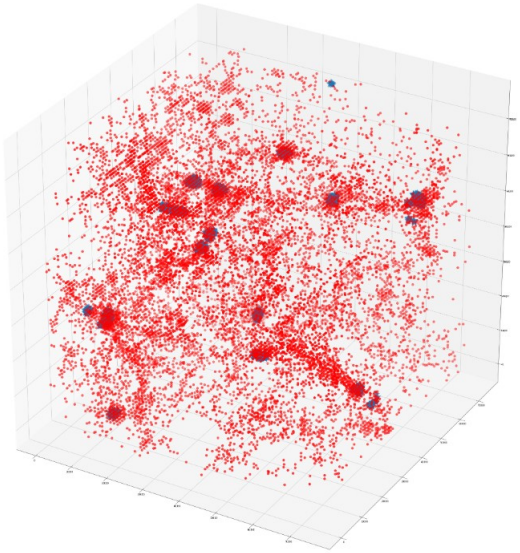

Figure 5.5: Results from Gravity Lattice, Data: 40,470 galaxies from Illustris-3, z = 0.0

Above we have presented the results for 40,470 galaxies in order to illustrate some

weaknesses of the proposed model. It is obvious that as we scale up to the number of galaxies whose signatures we want to depict the resulted cube turn out to be kind of noisy. Although the proposed method captures the nature of the galaxy formation it is difficult to extract details and results about the cluster regions of the above image as there have been many test loads that were moved. That might be because of the real position of the galaxies or because some galaxies had extremely large half mass radius which gave them falsely increased radius of effect. We can suppose that denser red areas show regions where most of the test loads have moved and as a result are clusters while other region where only a few test loads have moved are voids. In order to achieve that we have tried to filter moved test loads using as criterion their movement and if they are about to move in a short distance to cancel their movement in order to move only spheres that have been affected massively by galaxies and that would probably point out a movement towards a galaxy cluster. Although it might seem reasonable the above tactic had pure results because of the way the gravity lattice is set. Even if a galaxy belong in a void region it will have some spheres in its radius of effect that will be under massive gravity force and we cant find a way to negate the movement of such spheres.

Using the above problematic situation and the initiative of the algorithm we proposed and tested another version of the algorithm which one may say is an improved based on its results. In stead of keeping the total force exerted over a sphere we will keep the sub-forces exerted by every galaxy on that sphere as well as the number of galaxies affected each of the test loads.

From this point it is straight forward the method we are going to use in order to produce interesting results. Void regions are areas where there are only a few galaxies and the affected spheres are affected by only one or two galaxies while within a cluster the affected test loads are under the effect of much more galaxies depending on the density of the cluster and the position of the galaxies and the sphere.

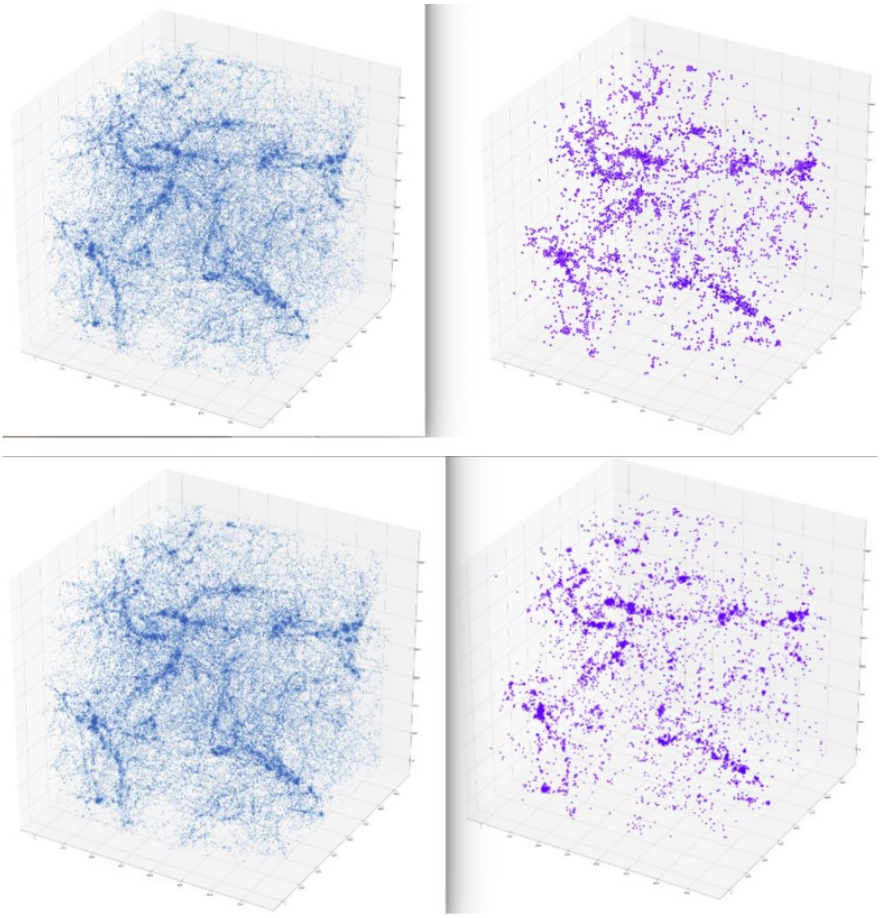

Figure 5.6: Upper row: Inter dist = 500, Lower row: Inter dist = 1000, Filter and eliminate point affected by less than 4 galaxies (Left: before filtering, Right: after filtering), Data: 100,700 galaxies from Illustris-3, z = 0.0

In order to extract information about galaxy clusters and voids taking into consideration the above mentioned setup as well as the theoretic background about the differences between void and cluster galaxies we filter the moved test loads by the number of galaxies that affected them and resulted that movement. Tuning this threshold we have finally set it to 4 by taking into account the number of such test loads and their positions. Above the reader can see the results of filtering in two different inter-distances. The less the distance the better the quality as we have stated. Depending on the needs this method can adapt to the needed resolution value.

As we can see this method is able to filter out void galaxies (noise) that is the void regions and keep only the great structures which are respectively the walls. Characteristic example is the 'Smile' on the up and left side of the cube that is also visible with our proposed method, of course visually we can see that all great structures of the left cube (all galaxies of the system) have representatives on the right cube (after applying Gravity

Lattice and filtering).

5.3 HDBSCAN in the Cosmic Web

Density based clustering is a straight forward approach to the problem of identifying the regions of the cosmic web which could be labeled as either void or cluster areas. Although there are many algorithms that could fit the demands of this thesis among the wide bibliography on this topic we have chosen to use only a small fraction of them because their description and problem application matches the problem of the cosmic web.
DBSCAN [35] is a density based spatial clustering algorithm which was at first proposed by M.Ester et.al. The main difference of this algorithm in relation with other spatial clustering algorithms is that its applications are over data with noise. The definition of noise could vary among different datasets but the way it appears when we plot the datasets is common. In order to enlighten more the reader and taking into consideration the image of the universe from the used data one can easily see that on the void regions of the cube there are only a few galaxies (dot points in the cube) which are barely visible while on the filament and cluster areas there are plenty of galaxies and that is visible through the formation of the clusters and their shaped.
If we decide to characterize the void galaxies as noise and the clusters as useful data clusters then the application of the DBSCAN seems to be appropriate and ideal as it will directly point out which regions are the empty ones, meaning that these regions are voids and which regions are covered with clusters and determining those clusters among the other as we would also like to know where a cluster is located, which galaxies are part of it and what is the shape of the cluster.
This algorithm is non parametric and given a set of data points it groups them together in cluster and noise. In order to group them the algorithm takes into account the distance between points and groups together points that are closely packed with each other into clusters as it believes that these points have higher probability of belonging into the same cluster. On the same way points that lay on sparse areas, meaning areas of less density, are marked as noise because their nearest neighbors are too far away. DBSCAN is one of the most prominent and common spatial clustering algorithm.
In order to further explain DBSCAN we present its main ideas in the way it categorizes the points of the dataset. At the beginning we set a radius ε of a neighborhood with respect to some point. Then all points will be clustered as either core or outlier points as follows:

- A point p is a core point if at least there are min_points points within distance ε of it (including p).
- A point q is said to be directly reachable from a point p (p is core point) if it lays

within a distance of ε from p.

- A point q is said to be reachable from a point p if there is a path where each point is directly reachable from the previous point of the path, that means that all points among the path except (maybe) the last point must be core points.
- All points that are not reachable from any other point are characterized as outliers or noise points.

In order to illustrate that if a point p is a core point then it forms a cluster with all the points that are reachable from it, each cluster consists at least form one core point. Non core points can be part of the cluster but if they are alone then they are characterized as noise. This formation as we can see is applicable in our setup of the cosmic web as from our point of few core points are galaxies laying into dense areas of clusters of filaments and outliers are galaxies in void areas. Using this method it is not needed from us to know the exact shape of the cluster is it may be spherical, elliptic or any other shape and the algorithm will still be able to find it out if we set the appropriate parameter ε about the radius of the neighborhood of the core points.
The above defined reachability of points within the context of the DBSCAN algorithm as we can see is not symmetric as core points can reach non core points but the opposite is not true. So the algorithm has defined a notion about connectedness that is defined as follow: two points are density - connected if there is a third one (p) such that both of them are reachable from p.
Finally DBSCAN's clusters satisfy two properties:

1. All points within a cluster are mutually density - connected

2. If a point is density - reachable from a point of the cluster then it is part of the cluster as well.

Strategy behind the algorithm is obvious as it finds the ε of the neighborhood in order to identify the core points and then finds the connected components of the core points ignoring all non - core points in the neighbor graph. Finally from the rest of the points it either assigns it to a cluster if it is reachable directly or through a path to a core point or in assigns it to noise.
The intuition and the usage of the algorithm in the cosmic web is natural as without knowing its existence someone would probably have tried to do the same in order to classify regions according to the galaxies lay inside them.
The main problem of this algorithm lays over its complexity. As we have described it above it the algorithm each time tracks down a point in order to classify it as noise of cluster point. In order to do that it will need time O(N * classification), where classification time is depended on the structure and dimension of the points. If points are not

sorted in a spatial way in order to define which points lay inside the ε radius then the worst case scenario needs O(N^2) time, while in other cases time of O(NlogN) is needed if accelerating index structures are used.
It is worth to mention that the algorithm is appropriate for out application as it does not need to specify from the beginning the number of clusters nor their shape or central points. It also allows us by determining the radius parameters ε to define the level of resolution we would like the algorithm to concentrate in order to find the clusters. Finally parameters are easily understandable and have an apparent relationship with the noise level we want to extract and the density of the final clusters we would like to find.
Of course the algorithm has some drawbacks such as the tuning of the parameters and the choice of the appropriate distance metric but in the context of this thesis we will disregard them.
In order to avoid the drawback of DBSCAN we will use in this problem a hierarchical version of the algorithms called HDBSCAN [36]. Hierarchy is once again appealing into this problem as we have used it in previous attempts of modeling the cosmic web. HDBSCAN proposed by D.Moulavi et. al. is a hierarchical based algorithm for spatial data that has avoided some of the main drawbacks of the common spatial clustering algorithms. It generates a complete density - based hierarchy by the most important clusters while it also offers a measure of significance of the clusters in order to hierarchically sort them. HDBSCAN is a direct expansion of the well studied and previously presented DBSCAN which we have used on our data because of its hierarchic nature and especially because of its speed of computation.
We will present the most promising results from using HDBSCAN in a Python implementation:

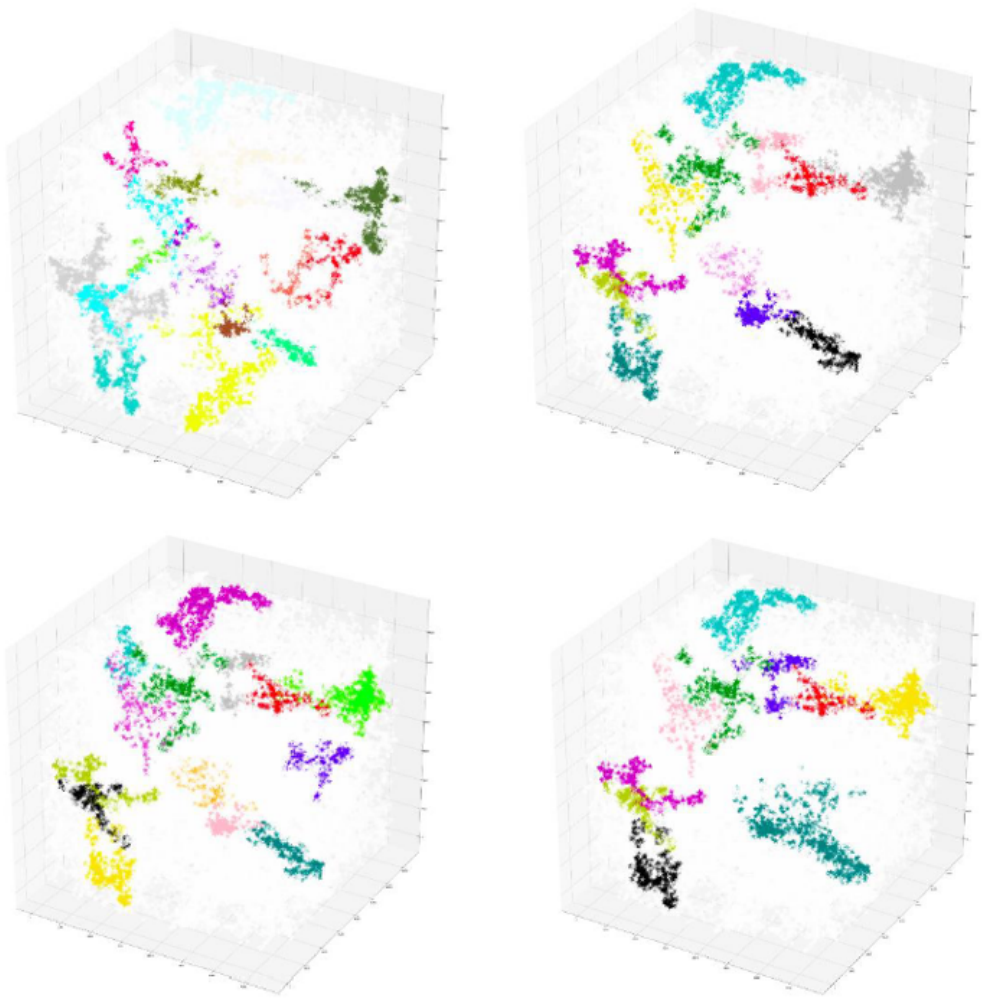

Figure 5.7: Upper row: [700, 30, 1000], [1200, 100, 1000], Lower row: [900, 100, 1000], [1500, 100, 1000], Format of the parameters [MinClusterSize, MinSamples, Epsilon] Data: 100,700 galaxies from Illustris-3, z = 0.0

In the context of this thesis we have used in order to use the above mentioned algorithms the module HDBSCAN* of the Python and tried to tune the parameters of the model in a way that it will reflect the natural world and on the same manner reflect the appropriate clusters of the cosmos. Method offered us three parameters which we could tune:

- Minimum cluster size defines the minimum number of points needed for one cluster.
- Minimum samples parameter that defines the amount of data that will be declared as noise, this parameters offer a measure to check how conservative the clustering will be.
- Cluster ε is the parameter that defines under which distance the algorithm will not

separate clusters.

Using the above parameters and knowledge of the universal parameters that are depicted on these three we have conducted experiments and resulted in the previously presented clustering of the cosmos.
In order to extract the parameters about these three values we used some of the knowledge about the galactic clusters and super clusters. Firstly we have set the parameter about the MinSamples to a value that led to a circa 70% of the galaxies to be assigned as noise (void) because the universe itself is compromised mainly of empty regions as the astrophysicists suggest.
Moreover we have used many and different cluster size as the reader can see, the less the cluster size the more clusters created by the algorithm. On the same context as previous we have decide to focus our interests on great galactic structures, clusters or even superclusters that consists of many galaxies as the theory suggested. That was the reason why we have chosen so many and different values about the MinClusterSize parameter. As we can see that resulted in different cluster formations that also mainly maintain they structure of the universe as it was visually observed in previous chapters. It is characteristic that the larger the size of the cluster the more vivid are the big galactic structures as they are colored in order to differentiate them in a way that they look like the the results after Gravity Fields filtering. By the use of look like we mean that the same structures extracted after filtering as the main structures of the universe are also extracted by spatial clustering of the HDBSCAN.
Finally we have used the epsilon selection in order to model the diameter and in general the size of the clusters and superclusters of the cosmic web. The value of 1000 or equally 1 Mpc is the regular radius of a small galactic cluster. We have chosen that parameter because the data we used from the Illustris project are simulated in a small (compared to the universe) cube and because other experiments we have done with greater of less epsilon values resulted into undesired clustering of the galaxies where the main structures were either fragmented into smaller ones or combined into enormous structures.

5.4 ABACUS in the Cosmic Web

Continuing on the same context as previous about the spatial clustering we will now move to an alternative form of spatial clustering which applications and results might be useful if applied on the cosmic web. Having study the implementation and theoretical background of the ABACUS spatial clustering algorithm proposed by V.Chaoji et. al. [37] we have implemented the algorithm using Python and applied it to the galactic data.

ABACUS is based on one main hypothesis that one can consider as valid in the cosmic

web data. This hypothesis is that spatial clusters can be generated from a set of core points. These core points will play the role of a backbone-like structure of the cluster and will give the intrinsic shape of it. For example an elliptic galactic formation like a filament will have a backbone which will be consisted of the main galaxies along its axis, the same will apply to an gamma like cluster that has many galaxies but the main core of it are the galaxies among the two segment lines.

To further illustrate the intuition and the reason why we will use this algorithm on the cosmic web and in general in spatial clustering is the equivalence-like relation between the backbone of a dataset and the dataset itself. Having the backbone of the dataset can result into the dataset or a close approximation of it through a specific process. Each point of the backbone has two main properties the spreading radius and spreading number. The first parameter is the parameter defining the radius within which a core point can generate other points and the second parameter defines the number of points that can be generated from a single core point.
As we can see if a dataset can be obtained fully or in a good approximation using the backbone points' properties of spreading then if we manage to inverse this process we will result into the backbone points of the whole dataset without losing information about it. ABACUS is the algorithm that tries to inverse this process, to do so instead of spreading each of the backbone points and 'divide' it into many other points within a given radius it tries to aggregate points of the dataset to single points that will be their representatives. These representatives should be located in an area that is appropriate so that if spreading is applied then the original points will appear with some probability. In order to succeed that ABACUS will gather together points of the main structures of the data and extract representatives of these points. The above process can take place many times dependent on the level of abstraction we would like to go and the number of compression of the given information we would like to make.

The main functions that model the above described process taking place in the ABACUS algorithm are (adjusted to the three dimensions of the cosmic web):

- Globing, which involves finding a representative of a group of points. In order to do that we first create a ball in the three dimensional space that has as its center a point of the dataset and radius r which full definition will stated afterwards. All points inside this ball will have as their representative the point laying on the center of the ball. Each point of the dataset has a weight assigned to it, that weight represents the number of points that are aggregated together in this point, so at the beginning all points have weights equal to one. As the process goes one and points are being globed to one representative the algorithm aggregates their weights and updates the weight of the representative which equals the total sum of the weights of the other

points. So at the final iteration of the globing process the representative's weight will represent (p) the amount of points that have been globed on p. Of course a representative could also be globed under another representative and the weights between them will sum up. The radius of the sphere can be understood as the spreading radius if we will follow the inverse process in the backbone points of the dataset.

- Point movement is the process where the representatives move towards the points they are going to represent. It is a gravity-like force that is exerted to points by their neighboring points. It is like a point feels the attraction of other points that lay within its neighborhood. Like gravity points tend to move towards other points under the influence of that force. Another similarity with gravity is that the magnitude of the change of the position of the points towards other points is proportional to the forces exerted to that point an the direction of the movement is is the weighted sum of the the force vector. In other words the weights play a mass-like role if we take into consideration the forces between points as gravitational forces. Of course in the context of the algorithm which is far from the gravitational model we described the point moves towards the most likely component that is responsible for generating the point if an opposite process is followed.

The process of the algorithm consists of repeating the above two steps until a convergence or a stopping condition is fulfilled. The algorithm described above is:

Algorithm 4 ABACUS with slightly changes as we implemented it

Result: Backbone resulting data points

D = all galaxy positions
$w_i = 1, \forall i \in D$
K = knn_graph(D)
r = knn_radius(D, k, K)
all_m = []
for *i in range 3* **do**
    K = glob_points(K, r, k)
    K = move_points(K, r, k)
    m_i = number_of_points_moved
    all_m.append(m_i)
end
j = 2
while $\frac{m_{j-1}}{m_{j-2}} \geq \frac{m_j}{m_{j-1}}$ **do**
    K = glob_points(K, r, k)
    K = move_points(K, r, k)
    m_j = number_of_points_moved
    all_m.append(m_j)
end
return K

Calculating the radius of the algorithms is done using the KNN graph and eliminating the top 5% of the outliers. The choice of k in the KNN graph is important as for small values of k the graph is concentrated on dense regions while for larger values the results are more general from the KNN graph. With the use of the 95% of the distances as calculated from the KNN graph radius r is defined as the average value of them.
During the globing phase all we go through all points which have formed a ball of radius r in the three dimensional space and check how many of the other points lay inside that ball. That process is important as it ensures that no outlier point or noise point will be the center of the coming globing as these points will not have many other points within their radius. At the globing process we modify the points as we erase some of the points which will from now on have a representative and at the same time we update the representative's weight. Each representative will have the summed up weight of all the points that have globed to that representative.
Needed to mention that hard-wired values to radius r might not capture the full dynamic of the system.

During the movement phase as it would happen if we have used the gravity there exists a gravity-like equation that is applied to all points that have not been globed. These points will be affected of other points that are also not globed. TO state this process more illustrative if we had a system of four galaxies and they were globed in pairs

then only the final pair will affect the position of the other pair. Furthermore in order to do that the algorithm will take into consideration only the k-nearest neighbors that are not globed an use their weights and distance from the point in order to find out the final position of the point. Although it is not directly related to gravity the formula that gives the new coordinates of the point has a distance-like factor on the denominator. For y_i the y_i^{new} is given as (i is the dimension):

$$y_i^{new} = \frac{y_i \cdot w_y + \sum\limits_{z \in R_k(y) \wedge d(y,z) > r} z_i \cdot w_z \cdot \frac{1}{dist(y,z)}}{w_y + \sum\limits_{z \in R_k(y) \wedge d(y,z) > r} w_z \cdot \frac{1}{dist(y,z)}}$$

where z is a point that has not been globed under y and lays among y k nearest neighbors.

As we can see from all described above ABACUS is a promising algorithm on the problem of the cosmic web and its application might have interesting results. Also its complexity is O(NlogN), where N is the number of points, because our dataset lays inside the three dimensions and we can make use of the KD-Trees when constructing the KNN graph.
We used ABACUS in the context of the 100,700 galaxies of the dataset starting with a 3-NN graph because we wanted to pay attention to the denser areas. Using the algorithm form different sizes of the dataset we observed that the resulting backbone points where less than the half points or in some cases almost one third of them. Below we show the results of the circa 43k galaxies from the original circa 100k galaxies.

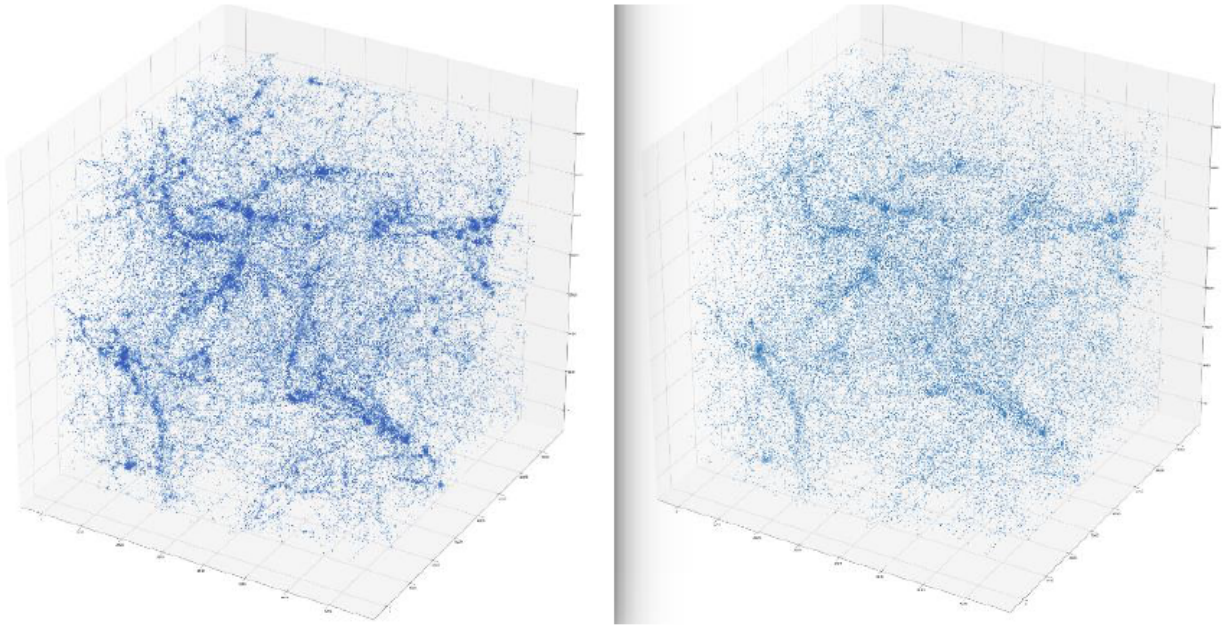

Figure 5.8: Left: All galaxies, Right: Points after ABACUS application, Data: 100,700 galaxies from Illustris-3, z = 0.0

As we observe there is not a great loss of information and we can declare that the structures on the right are more distinct in comparison to the left because many galaxies have been globed and especially inside the big structures we observe that although their shape is clear they seem more sparse because of the globing. That is the reason why they do not look so toned as they appear on the left where there are more point on them.

Taking the initiative from the Gravity Lattice and the filtering we applied on that method we tried to do the same here. As we have mentioned each of the representatives have as weights the sum of weights of the galaxies they represent, which in case is equal to the number of these weights. So we hypothesized that galaxies after the ABACUS algorithm that lay inside the big structures will have greater weights in comparison with galaxies lay in void regions. We filtered the right side image of the previous figure in order to keep only the backbone points which represent five or more galaxies, the value of five is the result of tuning that threshold.

On the right image of the figure again the initial universe and on the left the resulting galaxies or their representatives after filtering. As we can see here the great structures are apparent while the voids are almost perfectly filtered out. Of course if someone wanted it is possible to go into a greater level of resolution by filtering out even more galaxies and keep greater in size structures intact or the same can also apply in order to keep only small or between a given range structures. It is characteristic that the left image has many similarities with the results from Gravity Lattice filtering as we expected.

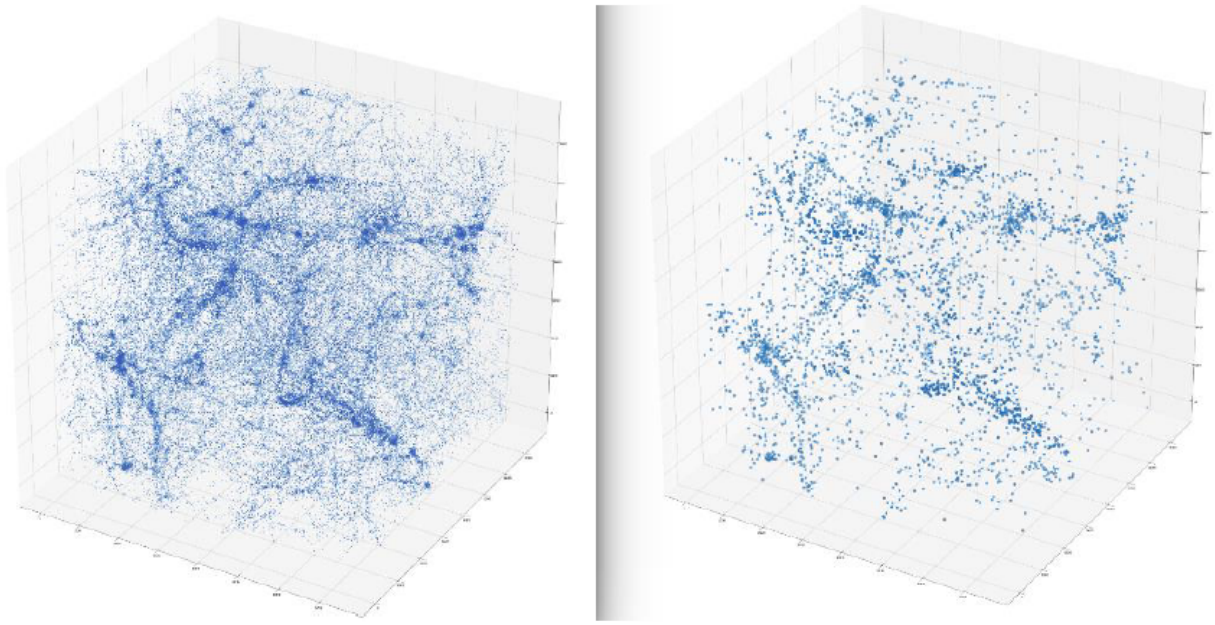

Figure 5.9: Left: Before filtering, Right: After filtering out points with weight less than 5, Data: 100,700 galaxies from Illustris-3, z = 0.0

Once again using the prime force of the universe, gravity, and the intuition about mass, its properties and the way it affects the cosmos we made a slight change to the ABACUS implementation and instead of setting the initial weight of each point to 1 we used galaxy masses to initialize each point of the dataset. The reason behind that implementation is that is allows us to differentiate the galaxies as they are now not equally weighted points which are going to be clustered based only on their positions. With that weight the movement of the representatives as well as the total weight they result in is dependent on the galaxies they have globed. So the final backbone has points whose mass is the sum of galactic masses and that allows us a more proper filtering in comparison to the previous filter. In the previous filtering only the number of globed galaxies played a role and not their total mass which is the case in the context of the universe.
We have left the equation that defines the magnitude of movement intact as it was in the ABACUS because we chose not to implement a version of gravity instead of the equation because that might have led to same results as in Gravity Lattice. Below we present the results from filtering with different mass thresholds, as the threshold increases the resulted after filter structures are the most immense and the void regions are correctly ruled out by the filter. The main point of this idea is to find the appropriate value of mass that keeps the greater structures intact without changing their shape nor their position a lot. All mass thresholds indicated below are measured in 10^{10} of sun's mass. The best of them appears to be for mass threshold around 100 maybe a value of 250 would be even better. On the other hand if we only wanted the greater structures resulted from filtering with mass filter 1000 are appropriate as they rule out more than 95% of the galaxies but the resulting points detect the formation of large scale structures (i.e. the 'Smile' on left upper of the cube).
For each of the levels of resolution they appear even less galaxies because a larger fraction of them is discarded as noise. Moreover as we low the threshold we observe the appearance of more detailed structures without the insertion of noise. In order to solidify that we have extracted only the noise (void galaxies) of the cosmic and notice that as we expected lay almost everywhere in the cube and especially between the large scale structures.

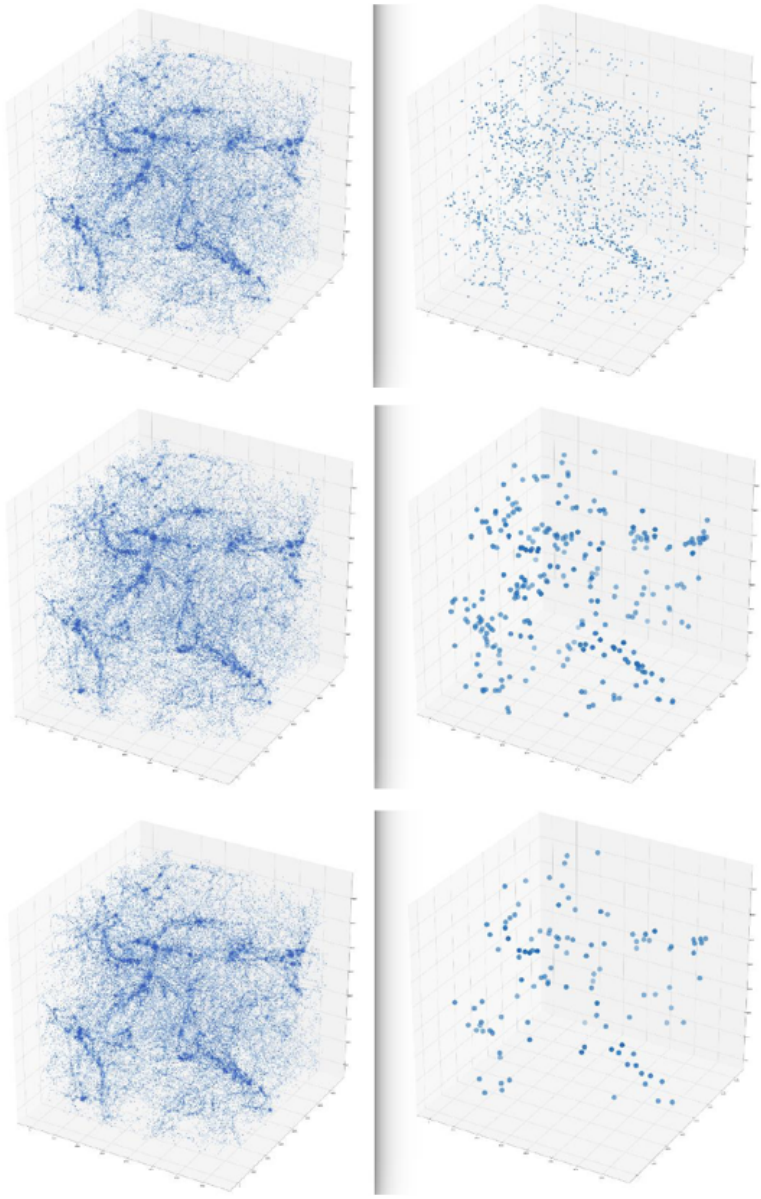

Figure 5.10: Upper row: Filtering out mass less than 100, Middle row: Filtering out mass less than 500, Lower row: Filtering out mass less than 1000, Left: Before filter, Right: After filter, Calculated in 10^{10} mass of sun, Data: 100,700 galaxies from Illustris-3, z = 0.0

Finally taking into account the speed of execution and the results after the filtering we can declare that this algorithm is promising and can extract important information about the cosmos.

5.5 Chameleon & CURE in the Cosmic Web

In the context of the previous success of the spatial clustering algorithms over the problem of regions' classification of the cosmos we were tempted to use more of these

algorithms.

The most promising of the spatial clustering algorithms according to the results he has achieved is CHAMELEON [38] which manage to perform well in two dimensional data but in higher dimensions it appears to be time consuming. We used an implementation of this algorithm and tried to approach the cosmic web. In the construction of the KNN graph we have used the ten nearest neighbors of each point and we will present the results for 5,000 and 10,000 galaxies. As the results below indicate the algorithm is not appropriate for this application as it fails to understand the nature of noise points which seems to be really important in the context of the universe as all these points assigned as noise are galaxies in void regions. As we can see on the images below it provides us with clusters that encapsulate noise points and end up in many clusters which are falsely identified as we can see galaxies once belonged in the same large scale structure using the previous mentioned algorithms to belong in different clusters as CHAMELEON result.

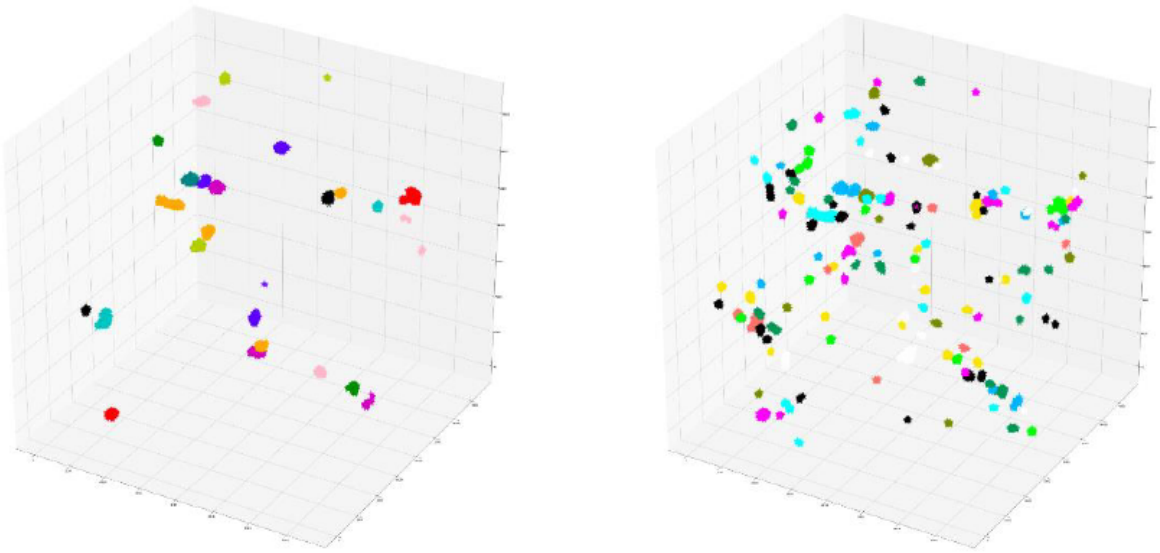

Figure 5.11: CHAMELEON results with 10-NN graph, for Left: 5k galaxies, Right: 10k galaxies, Data: galaxies from Illustris-3, z = 0.0

On the same topic we also used the CURE algorithm [39] which is a classic approach for spatial clustering than seemed to work properly for circa 10k galaxies but taking into account its quadratic time complexity it is rather pointless to try to use it for the whole data points. The image below shows CURE's result that appear to properly detect the structures of the universe but its in a larger scale that algorithm will take much time to complete which make it hard to use and tune its parameters. In the below image we asked CURE to determine ten clusters, determining the number of clusters at the beginning of

the process is rather impossible for so many data points as these of the cosmos.

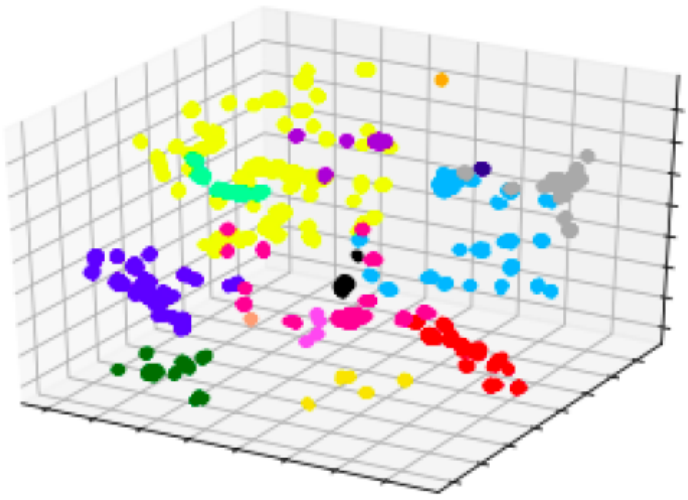

Figure 5.12: CURE results for 10k galaxies and 10 communities, Data: 10,000 galaxies from Illustris-3, z = 0.0

Chapter 6

Approaching Cosmic Web with GANs

Generative Adversarial Networks (GANs) is a class of machine learning frameworks design by I.Goodfellow et. al. in 2014 [40]. Recently because of the increase of the data and the computational power they have become really famous. On this chapter we will present an approach of that method the context of the cosmic web. Our method as far as we know is novel and its results can be improved in order to extract important features about the cosmic web.

6.1 Predicting the future using GANs

The initiative behind this approach lays on the nature of the GANs and their wide application field. We will present a novel approach of predicting the future of the cosmos based on simulated data by the Illustris project and a setup we have to propose.

It is known that different redshifts (z) represent different time frames of the simulation of the cosmos and as a result Illustris project offers us a wide variety of snapshots of different time points that could be the data for training a GAN model. In every snapshot we have different number of galaxies and different position among the same galaxies as in the pass of so many years galaxies are born and die an some of them tend to move towards other galaxies as well. The formulation of the above setup does not allow us to start from the galaxy set and try to apply a trajectory prediction model in order to predict the the future image of the universe. In order to overcome this obstacle we decided to study GANs and propose a novel method related to them. Because of its successful results in many datasets, i.e. celebrity dataset, we will propose the usage of DCGAN with slight different setup in order to meet our needs.
The main aim of this approach is that we will try to train the discriminator of the GAN with data of the present and the generator with data of the past. The generator will try using only past data and noise to create data of the present that will be so realistic that the

discriminator, which is trained on present data, will not be able to tell the difference. For example the generator is using data from z = 0.3 (past) and the discriminator is trained using data of z = 0.1 (more recent data in comparison of the data of z = 0.3). So we will ask the generator knowing data about the past to present us data about the present (z = 0.1).

In order to train the neural networks we will use the standard methods about this process. As we know GANs are really good when are used in cases where the data are represented with images. Using this and the already known success of DCGAN [41] over celebrity dataset we have to propose a novel as far as we know encoding of the data given by the Illustris simulation in a manner that is acceptable by the DCGAN. We determine an image (even though it will not be an exact image rather than an array of information but we shall call it image in the context of what it is needed by the GANs) as follows:

From the pool of galaxies we choose randomly 100 of them and for each of them we extract features like: mass, position, velocity, spin, half mass radius. In total we will keep 11 different values for each galaxy, the choice of the above mentioned values is not strict as someone may want to enclose more information of each galaxy but we propose them as the basic. Then we create an image of 25 x 50, meaning that in every row there are 2 galaxies and 3 empty spaces between them, using a single color channel. Of course we could encode using 3 color channels but we will keep it with single color for the time being. It is now clear that the images are in fact arrays of information and not real images as the values of the above mentioned parameters are huge, a scaling is not possible as we can not determine if a larger value will ever come to the future that will make the current scale invalid. Moreover when choosing galaxies we manage not to select the same galaxy twice for the same image.
Having now a big dataset that allows us to believe that all galaxies will have appeared on the frames of the dataset, with great probability, we shall now start the training. The way we will approach the training is kind of level-ish as it has many stages and levels of train between the pair of neural networks of the GAN. For example at the beginning we star with z = 10 defined as past and z = 9.5 defined as present and proceed to the training application. After the generator has learned to transform galaxies of the past to the present we then change the definition of past and present without changing to random weights the weights of the generator. The discriminator is trained to the new present and the generator tries to mimic the before process it was trained for in order to predict new images taking into account its current past. Thae procedure will continue until we reach a position where present will be defined as z = 0.0, which is the real present. After that final step then we will keep only the generator model and we will feed it with present data as past in order to predict us the future of the cosmos. The training process is like a game with levels where each level is different from the previous and the generator tries to pass it using what is has learned in all previous levels.

Among the training the two networks chose randomly from the galaxy pool but on the final application of the generator to predict the future we will use a different setup in order not to allow same galaxies to appear in different of same images more than once. We allow single appearance of the galaxies of z = 0.0 in order the future results to have each galaxy affected them only once. When the final generator has produced all images from the whole present data we combine them in order to predict the future image of the cosmos.
Of course in the above setup more data can be encoded and the results from the prediction of the future could also be used in order to further predict the future be setting these results as past and move on in time.

The main intuition behind this approach and the reason we believe that it will succeed its task in a reasonable time is that the way the universe evolves is governed by the same physical laws everywhere and the training phase after a few levels will be much faster than the beginning because the generator will have evaluated and learned the procedure needed in order to move from past to present. As we can see it is deep rooted into physical understanding of the world and that is the main reason behind the information about each galaxy we have decided to encode as these parameters form a distinctive set for each galaxy.

6.2 CycleGAN in the Cosmic Web

CycleGAN [42] is a technique that involves the automatic translation from image to image with or without image pairs, proposed by T.Park et. al. in 2017 is a quiet new application of GANs in the image translation. Its results were really tempting in the context of translating horse to zebras and especially in translating pictures to paintings. This implies that it is possible for that setup of GAN to translate images from one domain to another as long as there is a direct connection between them and a direct distinctive features that defines these two domains. As it is clear for the above mentioned method we would like to use that setup in predicting the future of the cosmos using GANs. In order to do that we would have to create images of the past and present in order to predict the GAN model. As the above method of creating a dataset would not result into a set of real images, that someone can visually observe, we used the method proposed by A.Amara et. al. [43] on their paper which as far as we know is the only GAN application in the cosmic web. They propose a way of constructing images by pixalisation the data of the cosmic web. We used the data from the Illustris project and sliced among three axes in order to take images every 0.1 Mpc, we transformed the slice of 0.1 Mpc of the cube into an image where where we either had black background and galaxies with white color reflecting the number of them that are represented on than pixel. Also we moved on a new encoding where each galaxy was a dot and its color was given with respect its mass into a five color mass range.

Using all different encoding we tried to train the CycleGAN model but we had pure results which offered no important information about predicting the future. The model was trained for only a few epochs but because of having no results we terminated it. Needed to mention is that the paper of A.Amare et.al. used for codification data from the Millennium-II simulation which might be in fact better for GAN applications and its usage in codification may result in the desired results about predicting the future of the cosmos.

Below we present some of the images used in training process in order to illustrate what was the dataset that had no results in order to help the others avoid the same mistakes we did:

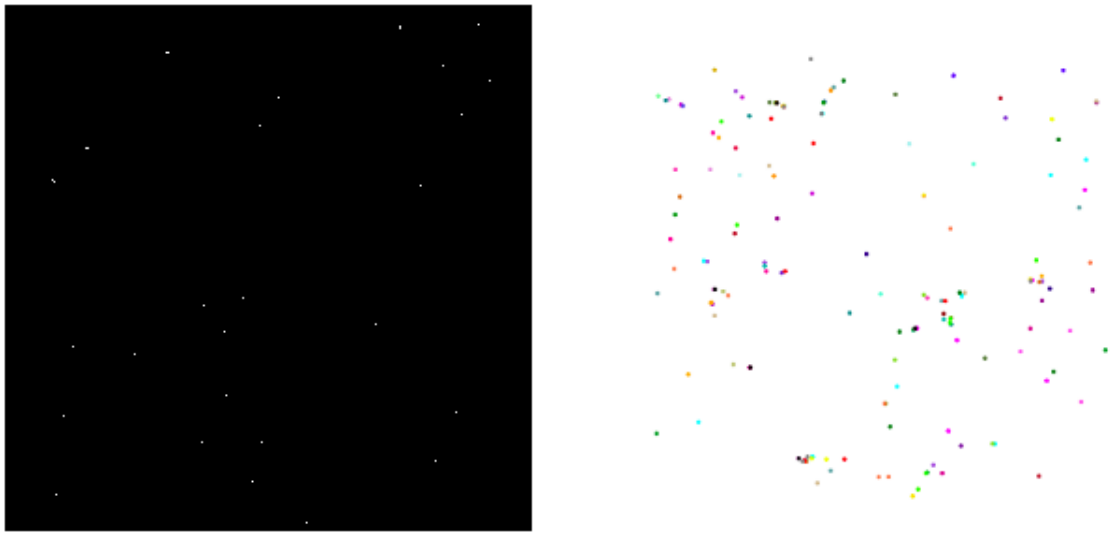

Figure 6.1: Left: Pixels proportional to the number of galaxies of the thin slice on that x,y coordinate, Right: Galaxies represented as colored dots where colors are based on mass, Data: 10,000 galaxies from Illustris-3, z = 0.0

Chapter 7

Future Work

Continuing this work would have to deal with many difficulties but it would probably result in interesting results.
Interesting direction would be to further extend the Gravity Graphons to enclose more physical laws in a generalized Graphon approach. As we discussed about it with the correct background it would be possible to result into a cosmic web depiction where edges will play an important role as they would point out directly the connection between galaxies and will also point out possible connection that astrophysicists may have miss judged.
Gravity Lattice offers a wide variety of applications not only on the direction of the cosmic web but it could also be applied in general. It is also needed to be further tested about its results when the test loads are denser or sparser populated inside the cube. Furthermore is would be interesting to further examine the filtering application on the results of the Gravity Lattice and try to make more sensitive filters whose parameters will be directly connected with the astrophysical theory.
Another important direction is towards improving the spatial clustering by applying more algorithms such as ROCK that seems to be promising, based on its applications and the way it works. Also needed to improve and further modify other spatial clustering algorithms, as we did in ABACUS, in order to fulfill our needs and be able to distinct void and cluster regions of the cosmic web. Also needed here to again reevaluate the filter we have used in order to extract more specific data about the galaxies and their masses. Stricter filters will result in pointing out greater structures and with the help of the background theory of the astrophysics it would be possible to directly point out these structures by their typical masses.
Finally really interesting and perhaps the most interesting future perspective of this work would be to further extend the usage of GANs in the cosmic web. Further encoding as we mentioned in the final chapter as well as new encoding methods of the data and using the pixalisation proposed by other along with the Cycle or other setup of GANs will probably result in interesting predictions about the future. The main idea about training in back and forth way and in a level-ish manner is promising and might result in solid prediction of the form of the cosmos in the future.

Bibliography

[1] Albert - László Barabási. Network Science.

[2] S. H. Strogatz, D. J. Watts (1998). "Collective dynamics of 'small-world' networks" In *Nature. 393 (6684): 440–442*.

[3] A. Barabási, E. Bonabeau (2003). "Scale-Free Networks". In *Scientific American: 50–59*.

[4] S. H. Strogatz (2001). "Exploring Complex Networks". In *Nature. 410 (6825): 268–276*.

[5] R. Cohen, S. Havlin, "Scale-free networks are ultrasmall" In *Phys. Rev. Lett. 90, 058701 (2003)*.

[6] M. Girvan; M. E. J. Newman (2002). "Community structure in social and biological networks". In *Proc. Natl. Acad. Sci. USA. 99 (12): 7821–7826*.

[7] F. D. Malliaros; M. Vazirgiannis (2013). "Clustering and community detection in directed networks: A survey". In *Phys. Rep. 533 (4): 95–142*.

[8] M. E. J. Newman (2004). "Detecting community structure in networks". In *Eur. Phys. J. B. 38 (2): 321–330*.

[9] Girvan, M.; Newman, M. E. J. (2002). "Community structure in social and biological networks". In *Proceedings of the National Academy of Sciences. 99: 7821–7826*.

[10] Aaron Clauset; Cristopher Moore; M.E.J. Newman (2008). "Hierarchical structure and the prediction of missing links in networks". In *Nature. 453 (7191): 98–101* .

[11] V.D. Blondel; J.-L. Guillaume; R. Lambiotte; E. Lefebvre (2008). "Fast unfolding of community hierarchies in large networks". In *J. Stat. Mech. 2008 (10): P10008* .

[12] Dorogovtsev, S.; Mendes, J.; Samukhin, A. (2000). "Structure of Growing Networks with Preferential Linking". In *Physical Review Letters. 85 (21): 4633–4636*.

[13] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. In *Social Networks, pages 179–197. Elsevier, 1977*.

[14] Justin Eldridge Mikhail Belkin Yusu Wang. Graphons, mergeons, and so on! .

[15] Daniel Glasscock. WHAT IS ... a Graphon? .

[16] Lovász, L. Large Networks and Graph Limits. In *American Mathematical Society*.

[17] Luminet, Jean-Pierre (2008). The Wraparound Universe. In *CRC Press. p. 170*.

[18] David N. Spergel (Fall 2014). "Cosmology Today". In *Daedalus. 143 (4): 125–133*.

[19] Whitney Clavin (17 March 2014). "NASA Technology Views Birth of the Universe". In *NASA. Retrieved 17 March 2014*.

[20] Dennis Overbye (19 June 2014). "Astronomers Hedge on Big Bang Detection Claim". In *The New York Times. Retrieved 20 June 2014*.

[21] Kravtsov, A. V.; Borgani, S. (2012). "Formation of Galaxy Clusters". In *Annual Review of Astronomy and Astrophysics. 50: 353–409*.

[22] Clavin, Whitney; Jenkins, Ann; Villard, Ray (7 January 2014). "NASA's Hubble and Spitzer Team up to Probe Faraway Galaxies". In *NASA*.

[23] Hubble Pinpoints Furthest Protocluster of Galaxies Ever Seen". In *ESA/Hubble Press Release. Retrieved 13 January 2012*.

[24] Freedman, R.A., & Kaufmann III, W.J. (2008). Stars and galaxies: Universe. In *New York City: W.H. Freeman and Company.* .

[25] Gott III, J. Richard; Mario Jurić; David Schlegel; Fiona Hoyle; et al. (2005). "A Map of the Universe". In *The Astrophysical Journal. 624 (2): 463–484* .

[26] Volker Springel "GADGET-2 A code for cosmological simulations of structure formation" In *https://wwwmpa.mpa-garching.mpg.de/gadget/*.

[27] "The Millennium Simulation Project" In *https://wwwmpa.mpa-garching.mpg.de/galform/virgo/millennium/*.

[28] "The Illustris Simulation Project" In *https://www.illustris-project.org/*.

[29] Nelson, D., A. Pillepich, S. Genel, M. Vogelsberger, V. Springel, P. Torrey, V. Rodriguez-Gomez, et al. 2015. "The Illustris Simulation: Public Data Release." In *Astronomy and Computing 13 (November): 12–37*.

[30] Kriegel, Hans-Peter; Kröger, Peer; Sander, Jörg; Zimek, Arthur (2011). "Density-based Clustering". In *WIREs Data Mining and Knowledge Discovery. 1 (3): 231–240*.

[31] Buddhika Nettasinghe1 and Vikram Krishnamurthy1. "Maximum likelihood estimation of power-law degree distributions using friendship paradox based sampling"..

[32] B.C. Coutinho, Sungryong Hong, Kim Albrecht, Arjun Dey, Albert-László Barabasi, Paul Torrey, Mark Vogelsberger and Lars Hernquist. "The Network Behind the Cosmic Web". .

[33] K. Schwarzschild, "Über das Gravitationsfeld eines Massenpunktes nach der Einsteinschen Theorie". In *Sitzungsberichte der Deutschen Akademie der Wissenschaften zu Berlin, Klasse fur Mathematik, Physik, und Technik (1916) pp 189* .

[34] Meagher, Donald (October 1980). "Octree Encoding: A New Technique for the Representation, Manipulation and Display of Arbitrary 3-D Objects by Computer". In *SRensselaer Polytechnic Institute (Technical Report IPL-TR-80-111)* .

[35] Daren Wang, Xinyang Lu and Alessandro Rinaldo. "DBSCAN: Optimal Rates For Density-Based Cluster Estimation".

[36] Claudia Malzer and Marcus Baum. "A Hybrid Approach To Hierarchical Density-based Cluster Selection".

[37] Vineet Chaoji, Geng Li, Hilmi Yildirim, Mohammed J. Zaki. "ABACUS: Mining Arbitrary Shaped Clusters from Large Datasets based on Backbone Identification".

[38] George Karypis, Eui-Hong (Sam) Han, Vipin Kumar. "CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling".

[39] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim. "CURE: An Efficient Clustering Algorithm for Large Databases".

[40] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. "Generative Adversarial Networks".

[41] Alec Radford, Luke Metz, Soumith Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks".

[42] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks".

[43] A.C. Rodríguez, T. Kacprzak, A. Lucchi, A. Amara, R. Sgier, J. Fluri, T. Hofmann, and A. Réfrégier. "Fast Cosmic Web Simulations with Generative Adversarial Networks".

[44] G Caldarelli. "Large scale structure and dynamics of complex networks: from information technology to finance and natural science".

[45] Olivier Cappe, Eric Moulines, Jean-Christophe Pesquet, Athina Petropulu, Xueshi Yang. "Long-range dependence and heavy-tail modeling for teletraffic data".

[46] SN Dorogovtsev, JFF Mendes, AN Samukhin. "Size-dependent degree distribution of a scale-free growing network".

[47] MP Rombach, MA Porter, JH Fowler, PJ Mucha. "Core-periphery structure in networks".

[48] MA Porter, JP Onnela, PJ Mucha. "Communities in networks".

[49] S Gregory. "Finding overlapping communities in networks by label propagation".

[50] V. Nicosia, G. Mangioni, V. Carchiolo, M. Malgeri. "Extending the definition of modularity to directed graphs with overlapping communities".

[51] J Leskovec, KJ Lang, M Mahoney. "Empirical comparison of algorithms for network community detection".

[52] HW Marsh, R Shavelson. "Self-concept: Its multifaceted, hierarchical structure".

[53] J Kleinberg. "Bursty and hierarchical structure in streams".

[54] A Clauset, MEJ Newman, C Moore. "Finding community structure in very large networks".

[55] MEJ Newman. "Clustering and preferential attachment in growing networks".

[56] A Vázquez. "Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations".

[57] W Najjar, JL Gaudiot. "Network resilience: A measure of network fault tolerance".

[58] K Klemm, VM Eguiluz. "Highly clustered scale-free networks".

[59] P Holme, BJ Kim. "Growing scale-free networks with tunable clustering".

[60] AL Barabási. "Scale-free networks: a decade and beyond".

[61] ML Weitzman. "Fat-tailed uncertainty in the economics of catastrophic climate change".

[62] Y Xia, J Fan, D Hill. "Cascading failure in Watts–Strogatz small-world networks".

[63] Marco Avella-Medina, Francesca Parise, Michael T. Schaub, and Santiago Segarra. "Centrality measures for graphons: Accounting for uncertainty in networks".

[64] S Dodelson. "Modern cosmology".

[65] AK Raychaudhuri. "Theoretical cosmology".

[66] YB Zeldovich, J Einasto, SF Shandarin. "Giant voids in the universe".

[67] Jörg M. Colberg, Ravi K. Sheth, Antonaldo Diaferio, Liang Gao, Naoki Yoshida. "Voids in a ΛCDM universe".

[68] SDM White, CS Frenk, M Davis, G Efstathiou. "Clusters, filaments and voids in a universe dominated by cold dark matter".

[69] NA Bahcall. "Large-scale structure in the universe indicated by galaxy clusters".

[70] Jörg M. Colberg, K. Simon Krughoff, Andrew J. Connolly. "Intercluster filaments in a ΛCDM Universe".

[71] S Bharadwaj, SP Bhavsar, JV Sheth. "The size of the longest filaments in the universe".

[72] JJ Monaghan, JB Kajtar. "SPH particle boundary forces for arbitrary boundaries".

[73] JS Bagla. "TreePM: A code for cosmological N-body simulations".

[74] BT Draine. "Physics of the interstellar and intergalactic medium".

[75] V Sahni, AA Sen. "A new recipe for CDM".

[76] MP Deseilligny, I Cléry. "Apero, an open source bundle adjusment software for automatic calibration and orientation of set of images".

[77] M Ankerst, MM Breunig, HP Kriegel, J Sander. "OPTICS: ordering points to identify the clustering structure".

[78] T Zhang, R Ramakrishnan, M Livny. "BIRCH: an efficient data clustering method for very large databases".

[79] RT Ng, J Han. "CLARANS: A method for clustering objects for spatial data mining".

[80] V Chaoji, M Al Hasan, S Salem, MJ Zaki. "SPARCL: an effective and efficient algorithm for mining arbitrary shape-based clusters".

[81] S Guha, R Rastogi, K Shim. "ROCK: A robust clustering algorithm for categorical attributes".

[82] BP Kent, A Rinaldo, T Verstynen. "Debacl: A python package for interactive density-based clustering".

[83] SP Borgatti, MG Everett. "Models of core/periphery structures".

[84] YM Zhang, K Huang, G Geng, CL Liu. "Fast kNN Graph Construction with Locality Sensitive Hashing".

[85] C Silpa-Anan, R Hartley. "Optimised KD-trees for fast image descriptor matching".

[86] J Hu, J Xu. "Density based pruning for identification of differentially expressed genes from microarray data".