A scaling relationship between the WHIM density and the luminosity density along filament regions inside of the Illustris TNG simulation

Benjamin Byrd
A SCALING RELATIONSHIP BETWEEN
THE WHIM DENSITY AND THE
LUMINOSITY DENSITY ALONG FILAMENT
REGIONS INSIDE OF THE ILLUSTRIS TNG
SIMULATION

Benjamin Byrd

A THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science
in
The Department of Physics and Astronomy
to
The Graduate School
of
The University of Alabama in Huntsville
August 2023

Approved by:
Dr. Massimiliano Bonamente, Research Advisor/Committee Chair
Dr. Stephen Walker, Committee Member
Dr. Richard Lieu, Committee Member
Dr. James Miller, Department Chair
Dr. Rainer Steinwandt, College Dean
Dr. Jon Hakkila, Graduate Dean
A SCALING RELATIONSHIP BETWEEN THE WHIM DENSITY AND THE LUMINOSITY DENSITY ALONG FILAMENT REGIONS INSIDE OF THE ILLUSTRIS TNG SIMULATION

Benjamin Byrd

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

Physics and Astronomy
The University of Alabama in Huntsville
August 2023

This thesis presents the current progress towards deriving a scaling relation between the Luminosity density and the density of the Warm Hot Intergalactic Medium (WHIM) located inside of filaments of the Illustris The Next Generation simulation 300-1 at redshift z=0. The filaments are located using the DisPerSE algorithm. This paper also presents a memory-efficient manner of performing nearest neighbor searches for large data sets for small dimensionality via the K Dimensional Tree, which enabled the limitation of WHIM exterior to halos of the full simulation. The limited WHIM was binned into $600^3$ cells with a side length of .505 Mpc. The same was done for the LD with a Gaussian smoothing parameter of 1.2. The current best linear fit has a slope of $0.75688 \pm 0.00659$ and an intercept of $-0.41329 \pm 0.00641$. This is a preliminary result that is still undergoing refinement.
Acknowledgements

Often times this project felt not dissimilar to the endeavor of Sisyphus, just as cresting the tip of the summit of what was thought to be sure footing would prove to be deceptive gravel, sending both the boulder back down to the base of the mountain and us to the drawing board to devise new methods to approach our problem.

I would like to thank Dr. Massimiliano Bonamente for allowing me the opportunity to work on and contribute to this exciting project and for acting as my advisor. Further, I want to thank Dr. Bonamente for always picking up my motivation and morale when the boulder did thunder down that hill and suggesting different ways to approach a difficult problem. I am also grateful to Dr. Patrick Holt and his invaluable insight on all matters and his experience working with simulations, without whom the pace at which this project progressed would have been greatly hindered. I would also like to acknowledge and thank Dr. Daniela Galárraga-Espinosa, whose continued contribution to this project has been vital to understanding and refining results, providing insight into both the way data is stored with Illustrs TNG and efficient ways to manage said data.

I would also express my gratitude to the members of my committee for their consideration and for their time. Lastly, I would like to thank The University of Alabama in Huntsville (UAH) and The UAH Physics Department for both allowing me to attend as a master’s student and providing funding by way of a Graduate Teaching Assistantship.

I dedicate this work to my family. To my Mother, Christina, whose belief that I could achieve anything inspired me to try. To my Father, Shell, whose perseverance convinced me to stay the course and appear strong when I felt weak. To my brothers; Matthew, whose eager passion for learning has me racing to keep up, Jon,
ferocity and dedication have kept me modeling the same intensity and Kristian, who is a reminder that I need to be better than what I have been. Finally, I would dedicate this work to Danielle Zaros, who has been a constant source of unconditional support and love, reminding me to take care of myself when I only wanted to take care of work. I couldn’t have completed any of this without the help and love of any of these individuals.
# Table of Contents

Abstract ................................................................. ii

Acknowledgements ....................................................... iv

Table of Contents ......................................................... vii

List of Figures ............................................................. viii

List of Tables .............................................................. x

List of Symbols ............................................................ xi

Epigraph ................................................................. xii

Chapter 1. Introduction .................................................. 1

1.1 Missing Baryon Problem ............................................. 1

1.2 Motivation of Study ................................................ 3

1.3 Illustris Simulations ................................................ 6

1.4 Filling Out the Filaments ......................................... 9

Chapter 2. Methodology .................................................. 16

2.1 Luminosity Over-Density Grid Construction ................. 16
2.1.1 Limiting Subhalo Information .................................. 16
2.1.2 Gaussian Filter and Smoothing Parameter ............. 18

2.2 Excluding WHIM Interior to Halos and Grid Construction .. 21
  2.2.1 K-Dimensional Tree ....................................... 22
  2.2.2 Implementation of the KD-Tree ........................... 27
  2.2.3 WHIM Grid Construction ................................. 30

2.3 Filament Grid and the Scaling Relationship ................. 31
  2.3.1 Filament Grid Construction .............................. 31
  2.3.2 The Scaling Relationship ................................ 32
  2.3.3 Cumulative Mass Distribution ............................ 37
  2.3.4 Linear Fit via Linear Regression ......................... 40
  2.3.5 Linear Fit via Monte Carlo Markov Chain .............. 44

Chapter 3. Conclusions and Future Work .......................... 52

3.1 Discussion of Results .......................................... 52
3.2 Systematic Error ............................................... 53
3.3 Recommendations for Future Study ............................ 55
3.4 Conclusion ..................................................... 56

References ............................................................ 58
List of Figures

1.1 Universe Energy Budget ........................................ 2
1.2 Example of Cosmic Structure ................................. 10

2.1 Linear Correlation Coefficient As a Function Of The Smoothing Parameter ........................................ 19
2.2 Luminosity Density Grid Slice ............................... 20
2.3 KD-Tree Visual Representation ............................... 23
2.4 KD-Tree Graphical Representation ......................... 25
2.5 KD-Tree FOF Search Example ............................... 26
2.6 Overplot of the Luminosity Density and WHIM Density Grid ........................................ 31
2.7 Full Range of log space of Luminosity over-density vs WHIM over-density with bias ....................... 34
2.8 Full Range of log space of Luminosity over-density vs WHIM over-density bias excluded ................... 35
2.9 LD vs WHIM over-density with the luminosity range limited ........................................ 36
2.10 WHIM Cumulative Mass Distribution ...................... 38
2.11 WHIM Cumulative Mass Distribution Compared To The Full Simulation ........................................ 39
2.12 First Iteration of Linear Fit .................................. 40
2.13 Linear Fit With High LD Excluded ......................... 42
2.14 Linear Fit With High LD Excluded and Low WHIM over-density Excluded ........................................ 43
2.15 Parameter a’s ”walk” ........................................... 46
2.16 Parameter b’s ”walk” ........................................... 47
2.17 Distribution of Parameter a ................................. 48

viii
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.18</td>
<td>Distribution of Parameter b</td>
<td>49</td>
</tr>
<tr>
<td>2.19</td>
<td>Comparison of MCMC results and the original fit</td>
<td>51</td>
</tr>
</tbody>
</table>
## List of Tables

1.1 Simulation Parameters ........................................ 7
1.2 Filament Parameters ........................................... 15

2.1 Linear Regression Parameters ................................. 44
2.2 MCMC Parameters ............................................... 50

3.1 Systematic Error ............................................... 54
## List of Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Omega_\Lambda$</td>
<td>Dark Energy Density</td>
</tr>
<tr>
<td>$\Omega_b$</td>
<td>Baryon Density</td>
</tr>
<tr>
<td>$\Omega_m$</td>
<td>Mass Density</td>
</tr>
<tr>
<td>$\bar{\rho}_b$</td>
<td>Mean Baryon Density</td>
</tr>
<tr>
<td>BLA</td>
<td>Broad Ly$\alpha$ Absorber</td>
</tr>
<tr>
<td>COS</td>
<td>Cosmic Origins Spectrograph</td>
</tr>
<tr>
<td>FUV</td>
<td>Far Ultra Violet</td>
</tr>
<tr>
<td>HST</td>
<td>Hubble Space Telescope</td>
</tr>
<tr>
<td>KD-Tree</td>
<td>K-Dimensional Tree</td>
</tr>
<tr>
<td>$r_{200}$</td>
<td>The Virial Radius</td>
</tr>
<tr>
<td>WHIM</td>
<td>Warm Hot Intergalactic Medium</td>
</tr>
</tbody>
</table>
Like I said I thought I was focused
I thought I had it all figured out
I had to organize my words good
Before they’d fall right out my mouth

- The Front Bottoms, "West Virginia"
Chapter 1. Introduction

1.1 Missing Baryon Problem

The energy budget of the universe is comprised of three contributors: dark matter, dark energy, and ordinary matter. The contributions of dark matter and dark energy account for a combined 95% of the total energy budget, dark energy comprising 68% and dark matter accounting for 27%. This leaves approximately 5% for ordinary matter, also known as baryonic matter. Baryons are a class of particles called a Hadron, meaning that baryons interact with the strong nuclear force. They are also Fermions, having the attribute of spin 1/2, which requires them to follow the Pauli Exclusion Principle. They are made up of three quarks and a baryon number of B=1, which is always a conserved quantity. Common examples of baryons include protons and neutrons, while more exotic examples include lambda and sigma particles.

The Wilkinson Microwave Anisotropy Probe found that baryonic matter accounts for 4.58% of the energy density of the universe [17]. Of that percentage, there is a popular consensus that there is $\approx 30 \pm 10\%$ left unaccounted [27]. The results of the Hubble Space Telescope (HST) Cosmic Origin Spectrograph (COS) estimate that there is missing baryonic content of $\approx 50 - 60\%$. This estimation is derived from traces due to O VI and H I being substantially smaller than assumed.
Figure 1.1: Visual representations of contributions of the universe energy budget (pictured left) and the contributors to the total baryonic content of the universe. As discussed below there is a suspected overlap between the O VI contributions and the Lyman-α contributions leading to closer to 50-60% of baryonic matter unaccounted for [33]. Values presented in this figure are obtained from Komatsu et al. (2011) [17] and Tuominen et al. (2021) [33].

by Shull et al. (2012) [11]. There are numerous uncertainties associated with the HST/COS traces of O VI and H I in the far ultraviolet (FUV). The first is that for the baryon content, \(i.e.\) the associated hydrogen mass, to be determined from an O VI line measurement, one needs the metallicity and the temperature of the gas. Temperature and metallicity cannot be obtained from only the measurement of the O VI emission line [33]. Danforth et al. (2016) [11] overcame this problem by utilizing cosmological simulations and the relationships between the metallicity and ion fraction and column density. The marriage between the observed O VI
measurements and the simulation parameters muddles the resulting conclusion of the hydrogen abundance as observed baryonic matter.

Secondly, the uncertainties for the temperature and metal abundance are underestimates. This is a direct result of the aforementioned combination of observation and simulation. The simulation accounts for the cosmic variation affecting the correlation between the ion fraction, column density, and temperature. However, as no observation data for this effect exists, the variation is not propagated to the final uncertainty [33]. Third, the temperature range of $\log T(\text{K}) = 5.0 - 5.5$ overlaps both O VI measurements and Broad Ly$\alpha$ (BLA) measurements [33]. The temperature range is associated with the low end of the WHIM temperature range. Observations directly made via BLA are treated as independent of the hydrogen abundance traced from the O VI measurement. Since these measurements are treated as independent, double counting of the respective populations occurs [33]. Hence the O VI baryon fraction ranges from 0%, total overlap with BLA, to 10% being wholly independent of the measurements of the BLA [33]. This leaves a sizable 50-60% of total missing baryons depending on the amount that is double counted. This unaccounted percentage is the crux of the "missing baryon problem."

1.2 Motivation of Study

The majority of baryonic matter at all redshifts is located within the diffuse Intergalactic Medium (IGM). At high redshifts, $z \geq 1.6$, baryons are expected to be photoionized and detectable via the Ly$\alpha$ forest [25]. This is useful for
examining chemical evolution, halos, and of course the high redshift components of the IGM. As for low redshift, the majority of baryonic content is going to be located as part of the Warm Hot Intergalactic Medium (WHIM) [4], which is defined with a temperature range of $\log T(K) = 5 - 7$ [4]. At this low redshift and temperature range, only a portion of the expected baryonic content can actually be observed. This is done through far ultraviolet observations with the Hubble Space Telescope and the Far Ultraviolet Spectroscopic Explorer (FUSE) [4]. High-temperature portions of the WHIM at $\log T(K) = 6 - 7$ have been observed as H I BLA and O VIII $k_\alpha$ absorbers [5]. These observations are made in the x-ray but are limited by the resolution of the detectors (Hubble Space Telescope (HST), Chandra, Sloan Digital Sky Survey (SDSS)) [5][1].

Due to the limitations of observation methods, it is useful to consider clever ways to indirectly observe the WHIM. The method used here is by using luminosity density to determine WHIM density. By comparison, the optical luminosity density is an observable, which is simpler to obtain than the x-ray lines of the various absorbers. The motivation for this relation is based on the distribution of dark matter that is centered around filament spines, the cores of the filament, that act as the primary source of gravitational potential. Baryons inside of this potential accrete, and experience adiabatic compression heating as well as shock heating [16]. This shock heating is stronger towards the center of the gravitational potential, the spine of the filament. This is as expected, with Tuominen et al. (2021) [33] showing with the Evolution and Assembly of GaLaxies and their Environments (EAGLE) simulation data that within $\approx 1$ Mpc of a spine,
the temperature ranges from log $T(K) = 6-7$. Further, Galárraga-Espinosa et al. (2021) [14] presented preliminary results for gas phases inside of filaments simulated with Illustris The Next Generation (TNG) showing similar results; however, the filaments inside of IllustrisTNG yield a temperature range at $r < 1.5$ Mpc of log $T(K) = 5-7$, depending on the length of the filament [14].

Using the C12 simulation developed by Cui et al. (2012) [10], a scaling relationship between the Luminosity density and the WHIM density was developed by Nevalainen et al. (2015) [24]. This relation was applied to both the Sculptor Wall and Pisces-Cetus supercluster. Both of these formations are observed at low redshifts with the Sculptor Wall at $z \approx 0.03$ and the supercluster at $z \approx 0.06$. The scaling relationship proved to be in agreement with observations by x-ray absorbers which provides justification for seeking a scaling relation regarding the luminosity density and the WHIM density. Holt et al. (2022) refined this scaling relation using the more recent EAGLE simulation and applied it to low redshift galaxies ($0.02 \leq z \leq 0.05$) SDSS DR12 data [16]. They found that their scaling relation was able to account for approximately $\Omega_{b,LD} = 31 \pm 10\%$ of the baryon density [16], concluding that it is consistent with observations. This study aims to develop a scaling relation between the luminosity density and WHIM density using the IllustrisTNG simulation and apply the scaling relation to observations.

In this work, the temperature range of the WHIM is considered to be log $T(k) = 5-7$. Due to the location of the filaments, the WHIM that is located within $r_{200}$ of any halo is excluded from the data set. This is discussed more in section 2.2. The subhalos are limited based on their luminosity, and if they are correlated
with the Flag filter. The Flag filter determines if a given subhalo is an artifact of the simulation or a subhalo, further discussed in section 2.1. The definitions presented here are used when finding the scaling relation.

1.3 Illustris Simulations

The IllustrisTNG simulation was generated as a way to model the galactic evolution of black holes, dark matter, gas, and stars. The simulation incorporates relevant physical processes that alter galactic structure such as radiation, diffuse gas, gravitational, hydrodynamic, and magnetic field effects [23]. The simulation has a redshift range of $z = 0 - 127$. IllustrisTNG is simulated assuming the ΛCDM (Lambda Cold Dark Matter) model and makes use of the cosmology parameters of the Planck Collaboration XIII (2016) [26]. The cosmological parameters used in the IllustrisTNG simulation and the parameters of the simulation used for this paper, TNG 300-1, can be seen in Table 1.1. IllustrisTNG makes use of adaptive mesh grids to handle its simulations. This process is called AREPO, and is known as moving unstructured mesh [30]. By utilizing AREPO, Illustris is able to mitigate some of the errors that can be present in other simulations. Two of the most common methods for cosmological simulations are done either using the Adaptive Mesh Refinement (AMR) with Eulerian hydrodynamics or Lagrangian Smoothed Particle Hydrodynamics (SPH), which is used in the EAGLE simulation [32]. Both of these are viable and commonly used techniques for cosmological simulations, however, they do have drawbacks that lead to inaccuracies in specific circumstances. For example, SPH can lead to suppression of fluid
Table 1.1: Simulation Parameters. This table provides the parameters that the IllustrisTNG simulation uses from Planck Collaboration XIII (2016) [26], as well as specific parameters of TNG 300-1 at z=0.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dark Energy Density (From Planck)</td>
<td>$\Omega_{\Lambda} = 0.69011$</td>
</tr>
<tr>
<td>Mass Density (From Planck)</td>
<td>$\Omega_{m} = 0.3089$</td>
</tr>
<tr>
<td>Baryon Density (From Planck)</td>
<td>$\Omega_{b} = 0.0486$</td>
</tr>
<tr>
<td>Mean Baryon Density (From Planck)</td>
<td>$\bar{\rho}<em>{b} = 6.18 \times 10^{10}M</em>{\odot} \text{Mpc}^{-3}$</td>
</tr>
<tr>
<td>Dimensionless Hubble Constant</td>
<td>$h = 0.6774$</td>
</tr>
<tr>
<td>Scalar Spectral Index (From Planck)</td>
<td>$n_s = 0.9667$</td>
</tr>
<tr>
<td>Number of Cells</td>
<td>216 000 000</td>
</tr>
<tr>
<td>Total Volume</td>
<td>$302.6^3 \text{ Mpc}^3$</td>
</tr>
<tr>
<td>Individual Cell Volume</td>
<td>$0.505^3 \text{ Mpc}^3$</td>
</tr>
<tr>
<td>Dark Matter Resolution</td>
<td>$m_{DM} = 4 \times 10^7 M_{\odot} h^{-1}$</td>
</tr>
<tr>
<td>Number of Particles</td>
<td>$N_{DM} = 2500^3$</td>
</tr>
<tr>
<td>Average Luminosity Density</td>
<td>$4.9 \times 10^7 L_{\odot} \text{Mpc}^{-3}$</td>
</tr>
</tbody>
</table>

instabilities, and AMR lacks Galilean-invariance [30]. As described in Springel 2010 [30], AREPO is a set of discrete points defined by Voronoi tessellation to
create an unstructured moving mesh. Voronoi tessellation is a method of partitioning space where each cell contains the region of space that is closest to that point, and each cell border is equidistant to the neighboring point. As each point is allowed to move freely, the cells of the mesh grid are able to adjust their location and volume without overlapping with neighboring cells [30]. This allows for smooth cell evolution over time. A Riemann solver along with the unstructured mesh then solves the ideal hydrodynamic hyperbolic conservation laws based on a second-order unsplit Godunov scheme with finite volume (see Springle 2010). This method also allows for the points to be treated as either stationary or moving. If treated as moving, AREPO is very similar to SPH and other Lagrangian smoothing techniques; however, due to the adaptability of Voronoi cells, AERPO lacks the limitations related to mesh distortion that occurs when neighboring cells begin to overlap and tangle [30]. While the center points are treated as moving the code is completely Galilean-invariant. In the context of IllustrisTNG, AERPO is used to solve the coupled evolution of gravity and ideal continuous magnetohydrodynamics (MHD) [23]. Gravity is handled by a Tree-PM (particle mesh) approach, which is described by Bagla 2002 [2] as combining the high resolution of Tree codes with the periodic boundary conditions of particle mesh. The MHD is simulated through the previously discussed Voronoi cells with finite volume [23]. Through these methods, IllustrisTNG is able to accurately simulate a broad range of physical phenomena including [23]:

1. Radiative gas processes
2. Stellar formation inside of regions of the dense interstellar medium
3. Stellar population and chemical enrichment following supernovae

4. Outflows on the galactic scale due to supernovae

5. Supermassive black hole formation and growth

6. Dual-mode black hole feedback at high and low accretion rates

IllustrisTNG does not include a treatment of cosmic rays or radiation. This is a notable drawback as both of these could affect galactic evolution. Nevertheless, the TNG model is still calibrated based on observational data to match observed attributes and statistics of galaxies [23].

IllustrisTNG produced three different primary runs: TNG50, TNG100, and TNG300. All three correspond to different simulation sizes and resolutions. The one used in this work is TNG300-1, which contains both the largest simulation volume of $302.6^3\text{Mpc}^3$ and the highest mass resolution. This simulation catalog is also at a redshift of $z=0$. TNG300-1 is selected based on the advantage of allowing examining cosmic structures, such as filaments and their gas contents, on smaller scales at higher resolutions.

1.4 Filling Out the Filaments

Galactic structure on the cosmic scale does not present itself in a uniform manner. Instead, the matter that comprises the universe coalesces into defined structures: galaxy clusters (Figure 1.2 A), filaments (Figure 1.2 B), walls (massive filament structures), and voids (large swaths of space that are not home to particles seen in Figure 1.2 C). All of these components form the cosmic web, the
Figure 1.2: Pictured is a slice of 100 kpc of the TNG300-1 at z=0. The color gradient pictured is representative of the gas column density. Location A, in the center of the image is a galaxy cluster. The largest in TNG300-1. Location B is an example of a filament structure. The focus region of this paper. Location C is an example of a void. The unmodified version of this image is provided by the TNG Collaboration [31].

large-scale structure of the universe, named as such due to its resemblance to a spider web [8]. An example of this structure can be seen in Figure 1.2 and is
provided by the TNG Collaboration [31]. The first observation of the cosmic web occurred in 1986 by de Lepparent et al. [12] as part of the Center for Astrophysics Galaxy examining the Coma cluster, making the comparison that the examined slice of the universe looks like a soapy kitchen sink with the galaxies appearing to sit firmly on the surface of bubbles. The bubbles are the void here and the film exterior of the bubble the filaments. Physically, the expansive structure that is the cosmic web is comprised of dark matter and gas, which formed these structures due to the gravitational instability caused by the anisotropic collapse of the initial perturbations of the density field [35], leading to matter "falling" into this galactic structure.

Filaments, the primary region of interest in this study (section 1.2), still have a broad definition and are challenging to categorize. This is not due to a lack of effort by the scientific community. It is simply a matter of variety that the filament structure can form in. In reference to the Coma cluster, Malavasi et al. 2020 [19] examined the filament network in a 75 Mpc radius of the Coma cluster. They used data from SDSS and the Discrete Persistent Structure Extractor (DisPerSE) to locate and analyze filament structures. They found that there are three filaments that connect the Coma cluster to other clusters in the same region such as Abell 1367, which agrees with x-ray observations [19]. Another study conducted by Bonjean et al. (2018) [9] examined the intra-cluster region between the cluster pair A399-A401 as well as the A21-PSZ2 cluster and G114.09034.34 cluster. They reported that for the A399-A401 clusters, the region expected to be filament was home to passive, early type, and "red and dead" galaxies as well that the filament
was full of hot dense gas [9]. The A399-A401 intra-cluster region was further examined through radio emission with the Low-Frequency Array (LOFAR) due to the large scale, throughout the volume of the filament, and weak shock [15]. This is notable as most studies related to filaments and large-scale galaxy structures are conducted using x-ray observation. The absence of a standardized definition of the filament has led to the creation of multiple different algorithms that locate and identify filament structures given either dark matter particle information or galaxy distribution catalogs.

As mentioned there are a number of different ways to locate and identify filaments. This investigation used the DisPerSE search algorithm [29], based on topology. Other algorithms include the Bisous model, the NEXUS algorithm, and the SpineWeb. The DisPerSE search algorithm uses the discrete distribution of particles to identify all aspects of the cosmic web, the voids, filaments, and clusters. This process is by no means simple. DisPerSE applies Delaunay tessellation (DT) to the sample population [29]. Delaunay tessellation is another method of partitioning space, however, unlike Voronoi tessellation, DT takes the points density distribution and spreads the distribution radially outwards while also partitioning the space into a tetrahedron where the vertices of the tetrahedron make up the points of the given distribution. In the context of DisPerSE, the DT follows discrete Morse Theory, a form of computational topology which has been thoroughly applied to astronomical data sets and is able to be similarly applied to numerical data sets that mirror observational data. Which allows for structures to be easily identified [29]. The identification of filaments specifically
is done so by using discrete Morse Theory to find critical points of the density field. The critical points are located at the saddles of the field where the gradient of the field is zero, meaning that these points are located at either the minima or maxima of the field in a given region. Physically the maxima of the density saddle correspond to the region of the filament structure that is near a halo, whereas the low-density region of the saddle corresponds to the middle, or region of the filament that is located in-between two galaxy clusters in the intra-cluster medium. The filaments are treated as the edges or ridges of the triangulated DT cells. They are segmented points connecting a filed maxima to a saddle point. In an effort to separate the noise from the true critical points, the persistence value of a topological feature is used to consider the significance of different critical points. Topological persistence is a method of identifying the significance of different topological features and based on their significance limiting the features to those that are relevant. In general, the persistence of a topological feature is a measure of how long maxima and minima survive inside of an excursion set [29]. That is to say in a defined threshold, the excursion set contains values that exceed this threshold. When the threshold crosses a critical point, the excursion set is modified to account for this new point. In the case of either a positive or negative critical point that has no opposing value equal in magnitude, it is added to the set. If there is an opposing point of equal magnitude, then that feature is destroyed or removed from the excursion set. The difference between a pair of points is the persistence of that topological feature. The persistence is also considered to be the lifetime of the feature or its robustness against changes [29].
For this work, a persistence threshold of $2\sigma$ was chosen to be applied to the DT density field. This persistence threshold was chosen based on the abundance of features that were likely to be attributed to noise. Persistence values that are lower than $2\sigma$ allowed for smaller-scale features to be retained in the final filament skeleton. These smaller features are likely noise. On the other end, for persistence values greater than $2\sigma$, is the range at which topological features that are smaller filament features begin to be excluded. As a higher persistence threshold, only more prominent features are kept as they are deemed the most reliable. At $2\sigma$ the resulting filament skeleton had the highest correlation between the critical density points and the over-dense regions of the input galaxy set, meaning that the filament skeleton aligned well with the locations of the halos. This persistence threshold resulted in a detected 12,244 filaments, the attributes of which can be seen in Table 1.2. The radii of the filaments are as of yet undefined as DisPerSE only returns the spine location of the filaments. The discussion related to the filament grid construction is seen in Section 2.3.1.
Table 1.2: Filament Parameters. A few key attributes related to the skeleton of the filaments that were located by DisPerSE. Inside of the TNG300-1 Simulation at z=0. The skeleton was generated with a persistence threshold of 2σ.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Filaments</td>
<td>$n_F = 12,244$</td>
</tr>
<tr>
<td>Max Filament Length</td>
<td>$L_{max} = 86.2$ Mpc</td>
</tr>
<tr>
<td>Minimum Filament Length</td>
<td>$L_{min} = .3$ Mpc</td>
</tr>
<tr>
<td>Mean Filament length</td>
<td>$\mu_F = 12.3$ Mpc</td>
</tr>
<tr>
<td>Median Filament Length</td>
<td>$\text{med}_F = 9.7$ Mpc</td>
</tr>
</tbody>
</table>
Chapter 2. Methodology

In order to obtain a scaling relation for the results of the TNG300-1, the data needed to be slightly modified. These modifications were imposed to ensure the validity of the results. Such as removing artifacts of the simulation or limiting the data to the filament regions. The first part of the data limitation was applied to the subhalo information and is discussed in section 2.1. Following that constraints were applied to the WHIM (Section 2.2) and then the locations of the filaments were applied to the data set to determine a scaling relationship, which is discussed in depth in section 2.3.

2.1 Luminosity Over-Density Grid Construction

2.1.1 Limiting Subhalo Information

As discussed in section 1.2 it is necessary to limit subhalo information that is used for both the luminosity over-density grid (LD grid) and the filtering of WHIM. There are two filters that are applied to the subhalo information. The first of which is a Flag filter. This filter is based on a parameter that is provided in the Illustris data which tell whether a subhalo is suitable for analysis or if it should be excluded. It is a boolean value of either True or False. If the value is True then the subhalo is marked for exclusion and the subhalo is not of cosmological origin.
and is most likely either a baryonic fragment or a formation inside of another halo [22]. If the subhalo is marked as False then it is considered good and is included in the analysis.

The second filter applied to the subhalos is the luminosity filter. Following both Tuominen et al.[33] and Holt emhet al. [16] method to constructing a Luminosity over-density grid the filter is limiting subhalos with an r-band luminosity value below a certain $M_r$. For the Evolution and Assembly of Galaxies and their Environments (EAGLE) simulation the limit is $M_r \leq -18.4$ [33]. This r-band limit is chosen in an effort to match the galaxy number density of the Sloan Digital Sky Survey (SDSS). SDSS has a redshift of $z = .05$ and a limiting magnitude of $M_r = -19$ [33]. Fainter galaxies need to be included in the sample population of both EAGLE and Illustris to match the number density of SDSS. This method for EAGLE yields an average luminosity density of $<\text{LD}> = 7.1 \cdot 10^7 L_\odot \text{Mpc}^{-3}$ [33]. The r-band magnitude for IllustrisTNG to match SDSS and EAGLE number density is $M_r = -18.014$. This conformity was made so that the scaling relation between can be applied to SDSS and other observations. As well as be compared to the scaling relation found in EAGLE by Holt et al.[16]. As such the subhalos in the simulation were filtered based upon $M_r < -18.04$. This yields an average luminosity density of $<\text{LD}> = 4.9 \cdot 10^7 L_\odot \text{Mpc}^{-3}$. This average luminosity density is lower than EAGLE due to the inclusion of dimmer galaxies to match the number density of SDSS.

Following the application of filters to the luminosity values, the values are divided by the average luminosity value (see section 2.3) and binned into a grid
to create the luminosity over-density. This grid is comprised of $600^3$ cells with a side length of $0.505$ Mpc. This dimensionality was chosen based on the resolution and limited by the computational cost. At a higher cell count the side length of the cells decreases, which increases the resolution of the grid. However, the more cells used to create the grid the more computationally expensive the grid is to construct and manipulate. The side length of $0.505$ Mpc and count of $600^3$ were determined to be an adequate balance between resolution and computational cost. As choosing a higher resolution exponentially increased the run time of scripts to times that were unfeasible.

### 2.1.2 Gaussian Filter and Smoothing Parameter

After the grid is constructed, it is necessary to apply a Gaussian filter to smooth the luminosity values to nearby cells. This is done so that the luminosity values related to a cell are "spread", smoothed, to nearby cells so that the corresponding luminosity value is better representative of the filament. The optimal smoothing parameter is chosen based on which smoothing parameter yields the strongest correlation coefficient. This process was done by creating $30$ LD grids that were smoothed using the Gaussian filter with varying smoothing parameters in the range $a = 0.1 - 3$ Mpc [16]. Then using the WHIM grid (Section 2.2) and each LD grid calculated the Pearson linear correlation coefficient as a function of the smoothing parameter. The Pearson linear correlation coefficient is calculated based on the slopes. The slopes for each LD grid relation to the WHIM grid are calculated via simple linear regression. The Pearson Linear correlation coefficient
is found as \( r^2 = b \cdot b' \) [6], where \( b \) is the slope of the linear regression of \( x \) on \( y \) and \( b' \) is the slope of the linear regression of \( y \) on \( x \) [6]. The null hypothesis for the Pearson linear correlation coefficient is that if \( r^2 = 0 \) then there is no correlation between \( X \) and \( Y \), alternatively, for true linear correlation \( r^2 = 1 \) [6], as the slope of \( x \) on \( y \) will be the inverse of \( y \) on \( x \). The results of this process can be seen in Figure 2.1. The smoothing parameter with the highest linear correlation coefficient is 1.2 Mpc. With a smoothing coefficient less than 1.2 Mpc

**Figure 2.1:** The linear correlation coefficient of the logarithm of the WHIM overdensity and logarithm of the LD as a function of the Gaussian filter smoothing parameter. The smoothing parameter of 1.2 associated with the maximum correlation is denoted by the green dashed line.
the majority of WHIM that populates the filaments would not correspond to a luminosity value as the LD would remain closer to a point value. Which would yield a smaller correlation coefficient and thus a worse fit. Whereas at smoothing parameters greater than 1.2 the LD begins to exceed the radius that is populated by the WHIM, meaning luminosity cells that have no corresponding WHIM value would appear. Similarly leading to a smaller correlation coefficient. Both of these behaviors can be seen in Figure 2.1

**Figure 2.2:** A 2 Mpc thick slice of the luminosity over-density grid. The color bar represents the smoothed luminosity over-density values of the Galaxy clusters and filaments with a smoothing of \( a=1.2 \).
It should be noted that to match the cell length used in Holt et al.[16] of .2 Mpc the grid would need to be made of $1500^3$ cells. This was found to be too computationally expensive for the hardware used for this study. Further, here a Gaussian Filter was applied instead of a B$_3$ spline kernel. However, these differences are assumed to impact the results of the smoothing in a minor way, as the Gaussian smoothing is a reasonable approximation to the B$_3$ spline distribution [18]. An example slice of the Luminosity over-density grid can be seen in Figure 2.2. The most luminous regions are expected to be the galaxy clusters and filaments, which the luminosity is taking the shape of. Figure 2.2 is not exclusive to regions that include filaments. Filament grid construction and application are discussed in section 2.3.

2.2 Excluding WHIM Interior to Halos and Grid Construction

It is necessary to limit WHIM to that exterior of $R_{200}$ of the halos, as the focus of this study is focused on the filament region of intracluster space. This process was assumed to be computationally expensive and the data set very large so the first iteration of this process involved an under-sampled data set of the WHIM which only contained 1 in 400 particles. This undersampled set still contained information regarding temperature, pressure, mass, density, and three-dimensional coordinates. This data set had 17,383,565 particles. Post masks applied the remaining halos, which comprised the entire simulation, had 11,504,190 halos that were suitable for analysis. The number of WHIM particles and halos is especially important when considering the run time of Python code. The first iteration of
limiting WHIM exterior to $r_{200}$ was a rather crude method attempted through nested for loops. This was incredibly computationally expensive. A for loop has an expected run complexity of $O(n)$ where $n$ is the number of iterations the loop performs. A nested for loop has a quadratic time complexity of $O(n^2)$. As the first attempt had to loop over both the halo positions and the WHIM positions it had a time complexity of $O(1.99984 \cdot 10^{14})$. This naturally led to a very generous estimated run time of $\approx 6,736$ days, completely unacceptable.

### 2.2.1 K-Dimensional Tree

The next iteration involved a more sophisticated approach. The implementation of a K-dimensional tree (KD-Tree), which is a method for partitioning data structures into binary trees, where every data point is considered a node and stored in the tree. Every point in the tree separates further into either another node or null point [3]. Consider a two-dimensional data set, comprised of x-y coordinates. This data set is of the following points:

\[ A = (1, 5), (4, 10), (3, 19), (8, 3), (12, 5), (3, 9), (16, 9) \] (2.1)

\[ \text{Median}_{x-axis} = (4, 10). \] (2.2)

For a balanced tree which is what was implemented, the method for selecting the primary node is to first calculate the median along the 0th order axis or the x-axis [3]. The parent node is seen in the first row of Figure 2.3 in blue. The subsequent nodes are then organized by separating the remaining points based upon $x \geq 4$ or
Figure 2.3: Completed 2 dimensional KD-Tree. Where each node represents the mean of the points within a partitioned space. The dark cyan nodes represent mean values along the x-axis, and the orange nodes are the mean value along the y-axis.

$x \leq 4$, which partitions the space into two areas based on the parent node. These partitions can best be thought of as hyperplanes, and in higher $k$ dimensions, hypercubes:

$$A_1 = (1, 5), (3, 9), (3, 19) \quad (2.3)$$

$$Median_{A_1} = (3, 9) \quad (2.4)$$

$$A_2 = (8, 3), (12, 5), (16, 9) \quad (2.5)$$

$$Median_{A_2} = (12, 5). \quad (2.6)$$

Every subsequent node is calculated cyclically based on axis, $A_1$ and $A_2$ are calculated based on the median of the y-axis or the 1st order axis [3]. These
child or branch nodes can be seen in Figure 2.3 on the second row in orange. These two branches further partition the space into four areas. As there are only two remaining data points for each node. The mean for each partition is simply the remaining points. The final sort is based on the x-axis. The completed tree is seen in Figure 2.3. The final row of nodes is considered to be the leaves of the branches [3]. While Figure 2.3 is a useful visualization of the KD-tree showing how the branches flow from the root node down into the leaf nodes, it lacks any representation of the partitioned space. Figure 2.4 shows the graph space representation of how the KD-Tree partitions the space. The ordering results in 8 partitioned regions. Figure 2.3 can be visually confusing, however, upon examination with the creation steps in mind the pattern is clear. From the parent node (4,10), there are two child nodes both in red (3,9) and (12,5), whose selection was previously discussed. From the child nodes, two more branches are formed to the last remaining nodes on the leaves of the tree (see the third row of Figure 2.3).

In the context of our work, the KD-Tree is used to perform a nearest neighbor search in order to find nodes of the KD-Tree that lie within a sphere. The nearest neighbor search algorithm is a recursive one, meaning that it will repeat the process of searching until the best match is found. Continuing with the prior 2D example consider two circles. The first, $\alpha$, with its origin located at (10,15) with a radius of 3, and the second, $\beta$, with an origin at (5,3) and a radius of 4. Figure 2.5 presents both the circles as well as the hyperplanes and nodes of the KD-Tree. The algorithm for the nearest neighbor starts similarly by adding
Figure 2.4: Graphical representation of the KD-tree. Nodes that were found along the 0th axis are represented by the blue circle and the 1st axis via red circles. This tree partitions the data into 8 distinct regions.

another node to the KD-Tree. The input point is filtered along the branches of the tree, as previously discussed until a leaf node is reached [21]. This means that the current best value for circle $\alpha$ is (16,9) and the current best value for $\beta$ is (8,3). The next segment of the process is for the distance to the leaf node to be calculated, and determine if the node lies within the input radius [20]. For $\alpha$ the leaf node (16,9) does not lie interior to its radius. The next step is to return to the parent node, (12,5), and determine if it lies interior to the input radius. As $\alpha$
does not intersect any other hyperplanes, see Figure 2.5, the alternate branch of node (12,5) is not considered at all and is immediately eliminated from the search [20]. The algorithm then moves to the root node (4,10) and repeats the distance calculation, and then returns that there are no interior points for circle \( \alpha \). \( \beta \) is a much more interesting example. Evident from Figure 2.5, \( \beta \) has one point interior as well as an intersection with three hyperplanes so the algorithm behaves slightly different. As previously mentioned the current best value is (8,3), the leaf node

**Figure 2.5:** Graphical representation of the KD-tree with the input circles represented in green. Point \( \alpha \), the upper circle, intersects no nodes of the tree. Point \( \beta \), the lower circle, intersects one point (8,3) as well as 3 hyperplanes.
(see Figure 2.3). Similarly, the next step is the algorithm recursively checking node (12,5) and performing the distance calculation. However, since the circle overlaps along the horizontal hyperplane of node (12,5) the other branch cannot be eliminated so the algorithm will also check the other branch, being (16,9) [20]. These steps occur regardless of if nodes (8,3) or (12,5) fall within the radius as the intersection and possible overlap of the hyperplanes cause the algorithm to check the corresponding nodes to determine the nearest point. Again the next step is to return to the root node (4,10) to determine if it falls within the radius. Since $\beta$ intersects the vertical hyperplane of node (4,10), the other side of the tree cannot be excluded. The algorithm goes through the branches starting at node (3,9) until reaching a leaf, and recursively checks each node until returning to the parent node again [20]. Then would return that there are no points interior to $\alpha$ and that there is one node, (8,3), interior to $\beta$.

2.2.2 Implementation of the KD-Tree

Why? Why try to implement this complex process? The run-time complexity of both the construction and the nearest neighbor search! As previously mentioned the first attempt of the nested for loops had a time complexity of $O(n^2)$, where $n$ is the number of iterations. In contrast the construction of the KD-Tree has a time complexity of Equation 2.7 [13]:

$$O(knlog(n))$$  \hspace{1cm} (2.7)

$$O(log(n)),$$  \hspace{1cm} (2.8)
where $k$ is the number of dimensions of the data set. As is evident, the only time the run times are comparable to the for loops is for high dimensionality of data while the nearest neighbor search has a time complexity that can be seen in Equation 2.8 [28]. As these processes are not nested the time complexity is added. So in general the total time complexity is as follows:

\[
O(kn\log(n) + \log(n)) \quad (2.9)
\]

\[
O(3.7 \cdot 10^8 + 7.06). \quad (2.10)
\]

The KD-Tree was constructed using the SciPy package for Python. The tree was made of the three-dimensional coordinates of the undersampled WHIM data set using the following function [34]:

\[
\text{scipy.spatial.KDTree}. \quad (2.11)
\]

This function takes the arguments of k-dimension arrays as well as other specifications such as leaf size and returns a KD-Tree with the input specifications. For our purpose, all other arguments were kept as default. The nearest neighbor search was also easily performed using the SciPy package with [34]:

\[
\text{scipy.spatial.KDTree.query_ball_point}(), \quad (2.12)
\]

which takes arguments for the location and radii of spheres/circles depending on the dimension of the data. The halo positions and their corresponding $r_{200}$
radii were used. All other arguments were kept as the default value except for
workers which were set to -1 to maximize parallel processing which allows for
the computer to use all of the cores of the processor to complete different parts
of the computation simultaneously [34]. Constructed and queried in this manner
Equation 2.10 represents the time complexity. This translated to approximately a
20-minute run time as the tree could be manually saved and loaded for future use.
However, this was for an under-sampled WHIM population using only every 1 in
every 400 particles. Thankfully, Daniela Galárraga-Espinosa, a postdoc working
at The Max Planck Institute for Physics and a colleague has access to IlustrisTNG.
Due to how the halo and WHIM data are stored, it requires a more complex
code. The location information of the halos and WHIM is not stored as a large
list. Instead, the volume of the simulation is broken into smaller cubes where
the position information is then stored. Therefore, the filter method had to be
modified to a structure that could manage this storage structure. The approach
is:

1. Call the WHIM location for the entire simulation and generate the tree.

2. Call the halo location information and the $r_{200}$ information for that iter-
tion’s sub-volume.

3. Perform the nearest neighbor search and store/exclude WHIM information
   interior to the halos.

4. Save the limited WHIM location and parameter information.

5. Repeat this process for each sub-volume inside of the simulation.
We further broke down the total regions into three separate scripts and ran them simultaneously. With each iteration taking 5-6 minutes the final run time for all WHIM inside of the IllustrisTNG simulation took approximately 6 hours. This was a pivotal breakthrough for the project as it allowed for the use of the entire simulation data in a time and memory-efficient manner.

### 2.2.3 WHIM Grid Construction

The remaining WHIM is used to construct a grid. This is done identically to Luminosity over-density grid [16]. Meaning that the WHIM grid is constructed of $600^3$ cells with a cell length of 0.505 Mpc. Further, to transform the WHIM grid to the over-density grid each cell is divided by the mean cosmic baryon density $\bar{\rho}_b = 0.618 \cdot 10^{10} \text{M}_\odot \text{Mpc}^{-3}$ [26]. Overlaying the WHIM and luminosity over-density grids show very close alignment and agreement (Figure 2.6). The slice presented is the same as in Figure 2.2. There are two locations of notable overlap between Figure 2.2 and Figure 2.6. The first is the structure at the top of the figure, and the second appears at $Y \approx 50$ Mpc. These structures are most likely galaxy clusters and are only appearing here because the data have not been limited to just the filaments.
Figure 2.6: A 2 Mpc thick slice of the luminosity over-density grid with the WHIM over-density grid also overlaid. The red color bar represents the smoothed luminosity while the blue color bar is the WHIM density value. Features of Figure 2.2 can be easily recognized (0,50) and (150,275). This highlights the agreement between the Luminosity and WHIM over-density grids.

2.3 Filament Grid and the Scaling Relationship

2.3.1 Filament Grid Construction

The skeleton of the filaments is found with DisPerSE (see section 1.4). As the filament grid is to be made up of 600$^3$ cells with Boolean values, True or False. In this case True corresponds to the cell containing a filament and False is a cell...
that does not. The filaments are binned into these cells based on their locations. Due to the filament information just being point information a couple of issues arise. The first is that for segments that are close to each other, they have the possibility of being binned into the same cell. Conversely for points that are far apart have the chance that the information between them is counted as False and not being represented as connected. To combat the cells being incorrectly represented as False we added more intermediate points to the initial filament locations given by the skeleton. Between each point 20 sub-points were added, and this allowed for mitigation of both errors as it spread out nearby filament points and helped to define the space between distant points. As mentioned in section 1.4 we intend to consider filaments of radius 0.5 Mpc to 3 Mpc. The varying radii are accounted for simply by changing the nearby cells from False to True based on the distance away from the initial True cell. The filament grid is then applied to the LD grid and the WHIM over-density grid to limit those grids to exclusively information pertaining to the filaments.

2.3.2 The Scaling Relationship

The relationship between the WHIM density and luminosity density is sought after similarly to Nevalainen et al. 2015 [24] and Holt et al. 2022 [16]. As discussed in section 2.1 the Luminosity over-density is defined as:

$$\delta_{LD} = \frac{LD}{<LD>}$$ 

(2.13)
where LD is the luminosity density of the limited regions and as mentioned in section 2.1 \(<LD>\) is the average luminosity density with a value of \(<LD>=4.9 \cdot 10^7 L_\odot \text{Mpc}^{-3}\). This is the mean luminosity density for the entirety of the simulation and not just the limited data ranges discussed previously. The WHIM over-density is similarly defined as:

\[
\delta_\rho = \frac{\rho_{WHIM}}{\rho_b},
\]  

(2.14)

where \(\rho_b\) is the mean cosmic baryon density [26]. Nevalainen et al. [24] modeled the LD-WHIM scaling relationship as a power law, expressing the WHIM over-density as a function of the luminosity over-density:

\[
\delta_\rho = A\delta_{LD}^B
\]  

(2.15)

with the best-fit parameters being A and B. These values were reported as \(A = .7 \pm .1\) and \(B = .9 \pm .2\). Instead of a power law, it was decided to work with a logarithmic scale [16]. The expected scaling relationship then becomes:

\[
\log\delta_\rho = \log A + B \log \delta_{LD}.
\]  

(2.16)

Using this logarithmic scaling relation necessitates that the Luminosity over-density grid and WHIM over-density grid be transformed into logarithmic scales. Before this transformation could take place, the zero values of the grids had to be removed as zero is undefined in log space. Luminosity over-density values that
correspond to $\delta_{LD} > 1$ are the values of interest. As LD values that are smaller than 1 correspond to dim and uninteresting regions. So removing zeros from the data set is not considered to impact the scaling relation in a significant manner. Finally, the data are plotted as $\log \delta_{\rho}$ as a function of $\log \delta_{LD}$. This result can be seen in Figure 2.7. While Figure 2.7 shows a positive correlation between the

![Figure 2.7: $\log \delta_{\rho}$ as a function of $\log \delta_{LD}$ with a filament radius of 1Mpc. The brighter yellow colors indicate a higher occurrence/count of that individual value. This figure highlights the fang feature, which spans high to low WHIM over-density in the high LD. This feature is a result of a bias in the WHIM over-density grid.](image)

LD over-density and the WHIM over-density, there is the emergence of the fang feature, seen as the vertical feature in the high LD region that spans low to high
WHIM over-density. This peculiar feature is a result of an unaccounted-for bias in the WHIM over-density grid. The origin of this bias is due to the underestimation of the WHIM densities during the construction of the over-density grid. When the cells located inside of the halos were discarded, the remaining mass exterior to the halos was used to calculate the density. However, the same treatment for the volume of the discarded cells was neglected, causing the resulting bias and underestimation of WHIM density values. The method for correcting the bias of

\[
\log \delta \rho \text{ as a function of } \log \delta_{LD} \text{ with a filament radius of 1Mpc. Again the brighter yellow colors indicate a higher occurrence/count of that individual value. Without analysis, there is a visual correlation between the values.}
\]
Figure 2.9: \( \log \delta_p \) as a function of \( \log \delta_{LD} \) with a filament radius of 1Mpc. This plot excludes dimmer values of luminosity. Further, it highlights the shift in the densest region with the bias excluded.

the WHIM over-density grid is rather straightforward. Due to the high resolution of the LD and WHIM over-density grids, we are able to identify pixels that fall within halos and discard them outright. This was accomplished by creating a Boolean mask, where True corresponded to being exterior to a halo and False being interior to a halo. This mask was applied to the WHIM over-density grid. This revealed that there was 1.294% of WHIM cells that were interior to halos still included in the over-density grid. These were promptly removed. After these
modifications, the relation was plotted again yielding Figure 2.8. Limiting the Luminosity over-density grid to the brighter region $\log \delta_{LD} > -0$ shows the significance of this correction. As indicated by Figure 2.9 the densest region of log space shifted a significant portion. Moving from the highest over-density regions (Figure 2.7) to the middle region. Of course, this is a result of the removal of the bias.

2.3.3 Cumulative Mass Distribution

The cumulative mass distribution (CMD) for the WHIM is created in an effort to examine the different regions of the luminosity over-density and the percentage of mass located in those the regions. The WHIM CMD for the mass located in the filaments can be seen in Figure 2.10. The cumulative mass distribution for the WHIM is calculated from the density values of the WHIM over-density grid and the volume of the cell. The mass calculation is done by using the cell volume instead of the particle volume. This choice or limitation was made based on the format of the data. The decision to move forward using the every whim particle inside of the simulation meant that the only information that is stored locally inside of the over-density grid can be used which is the cell location, and the related density for that cell. Figure 2.10 shows the WHIM for the simulations filament regions with 95% of the mass of the WHIM located inside of the filaments between -5.4 and 2.5. This is notable, and if correct, 40% of the mass located inside of the filaments are less than $\log \delta_{LD} = 0$, which is the average over-
Figure 2.10: The cumulative mass distribution of the WHIM as a function of the $\log \delta_{LD}$. This figure shows the LD range where 95% of the mass is located for filament regions.

density value, meaning that almost half of the mass interior to filaments are under-dense.

To further investigate the mass distribution of the WHIM the filament regions were over-plotted with the WHIM CMD of the full simulation. This can be seen in Figure 2.11. The 95% range presented in Figure 2.11 corresponds to the region that contains 95% of the mass that is located inside of the filaments, not the full simulation. The behavior of the WHIM CMD of the full simulation
The cumulative mass distribution of the WHIM as a function of the log$\delta_{LD}$. This figure shows the LD range where 95% of the mass is located for filament regions as well as the Full simulation. The 95% range here corresponds to the filament regions, that there is 95% of the filament mass contained in that range.

Compared to the filament regions is as expected. This is to say that the magenta line of Figure 2.11 has more mass located at higher log$\delta_{LD}$ regions physically corresponding to the dense and luminous regions of the filaments. However, for the full simulations there is a larger spread of mass at all luminosity regions, as it includes both the filaments and the less populated not belonging to either the filaments or halos.
2.3.4 Linear Fit via Linear Regression

With the removal of the bias, a preliminary attempt at determining the scaling relation was made. This involved simply using simple weighted linear regression. The weights correspond to the count of appearance of each data point in log space. This method of fitting the data was the first and most simple effort made to fit the data. This yielded a slope of $a = 0.757 \pm 0.007$ and an intercept of $b = -0.413 \pm 0.006$. This fit can be seen in Figure 2.12. Interestingly the fit

![Figure 2.12: \( \log \delta_p \) as a function of \( \log \delta_{LD} \) with a filament radius of 1Mpc. This figure shows the linear fit traced with the red line, with a slope of $a = 0.757 \pm 0.007$ and an intercept of $b = -0.413 \pm 0.006$.](image)
produced from this data does not follow the expected path. The expected path is to go through the highest count region of medium density and follow in such a manner as to cross the sparse high-density region. Instead the behavior that is exhibited, visually, the trend line seems to follow the shape of the high count medium density range, which completely bypasses the high-density region.

In an effort to refine the scaling relationship, more data limitations were imposed. This was done by manually excluding the remaining particles that belonged to the region that initially held the bias. The results of this can be seen in Figure 2.13. The motivation for this was to determine if these data points were playing a significant role in the behavior of the slope of the best-fit line. The limitations imposed were a simple one; if there was a data point belonging to the region of \( \log\delta_{LD} > 2.3 \), the data point was excluded. This modification can be seen in Figure 2.13. The slope of this fit is \( a = .757 \pm 0.007 \) and has an intercept of \( b = -0.413 \pm 0.006 \). Evidently, the exclusion of the remnants of the bias plays little effect against the concentration of data points in the medium density range.

A third attempt to model the data using simple linear regression looked to further artificially restrict the data. This was done by keeping the imposed limit that excludes the biased region and implementing a new restriction. This time on the WHIM over-density specifically. The motivation behind this exclusion was to examine if the low-density region (\( \log\delta_\rho < -1.2 \)) had a significant impact on the scaling relation that has been determined thus far. This can be seen in Figure 2.14. Again this fit was calculated through a simple linear regression where the weights of each value were determined based on the occurrence of the data point.
Figure 2.13: log$\delta_\rho$ as a function of log$\delta_{LD}$ with a filament radius of 1Mpc. This figure is to highlight the minimal change in the trend line when excluding the remaining data points in the bias region. The fit parameters are a slope of $a = 0.757 \pm 0.007$ and an intercept of $b = -0.413 \pm 0.006$.

This again yielded a minimal change in both the slope and the intercept of the best-fit trend line. With the log$\delta_\rho < -1.2$ data limitation applied, the slope is calculated as $a = 0.649 \pm 0.006$ and has an intercept of $b = -0.236 \pm 0.005$. This instance of limited data does have the largest difference from the initial fit. However, it is also the largest limitation of data. Interestingly it lowers the slope instead of the expected increase. All of the fit parameters can be seen in Table
Figure 2.14: $\log \delta_\rho$ as a function of $\log \delta_{LD}$ with a filament radius of 1Mpc. Again this figure highlights the significance of the medium-density region and its significant influence on the best-fit line. With the data exclusions the parameters are as follows: $a = 0.649 \pm 0.006$ and has an intercept of $b = -0.236 \pm 0.005$

2.1, along with the percent difference from the initial results of the log plots, Figure 2.12.

The only notable difference comes from limiting the lower range of the WHIM over-density. However, the changes that are present are not desired, and the rationale for performing these cuts are only curiosity in nature and have no physical meaning. Further, what this implies is that the densely populated (See Figure 2.12 region accounts for the majority of the WHIM and Luminosity
Table 2.1: Linear Regression Parameters. The Slope and intercepts of all of the above-discussed data ranges. The percent difference row corresponds to the percent difference between that column and that of the Original Fit, the unmodified range. This provides insight into the effect that the cut ranges have on the linear fit, that is to say very little.

<table>
<thead>
<tr>
<th>Name</th>
<th>Original Fit</th>
<th>High LD Limited</th>
<th>Low WHIM Limited</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope (a)</td>
<td>0.757 ± 0.007</td>
<td>0.757 ± 0.007</td>
<td>0.649 ± 0.005</td>
</tr>
<tr>
<td>Intercept (b)</td>
<td>−0.413 ± 0.006</td>
<td>−0.413 ± 0.006</td>
<td>−0.236 ± 0.005</td>
</tr>
</tbody>
</table>

\[ a_{\%\text{diff}} \] 0% 0% 14.2%
\[ b_{\%\text{diff}} \] 0% 0% 42.8%

Information, and holds the most sway in terms of the linear regression as the regression is weighted based on the count. This result inspired a different method of approach in regards to fitting the data.

2.3.5 Linear Fit via Monte Carlo Markov Chain

The new approach to fitting the data is a Monte Carlo Markov Chain (MCMC) with Metropolis-Hastings sampling method, which is a useful approach to obtaining best-fit parameters for both analytic and non-analytic functions. The methodology of the MCMC will be discussed alongside of its implementation. The first step is choosing the initial parameters, which are automatically added into the Markov chain. As there are 2 free parameters in the case of the linear fit.
(slope and intercept) one must be chosen for each with \(a\) and \(b = 3\). This choice was completely arbitrary.

Step two is to select the candidate chain member. This is drawn from a uniform distribution that has a fixed width. The width of the distribution is based upon the "step" at which the chain will converge to the parameters that correspond to maximum likelihood. Here, the step size was chosen to be ±.01 from the current parameter location, and the proposal step was drawn from a uniform distribution of that length. As such the prior distribution is also chosen to be uniform. This naturally leads to the acceptance or rejection phase of the proposed parameters. This is done through the acceptance probability \(\alpha\) the cornerstone of the Metropolis-Hastings method [7]. The acceptance probability is a value between 0-1 that describes the probability of the proposal parameters \((\theta')\) being accepted into the chain. \(\theta_n\) corresponds to the parameters at the chain location \(n\):

\[
\alpha \left( \frac{\theta'}{\theta_n} \right) = \min \left( \frac{e^{\frac{\chi^2(\theta_n) - \chi^2(\theta')}{2}}}{1}, 1 \right).
\]

(2.17)

The acceptance probability is calculated using the difference in the \(\chi^2\) distributions, which inherently assumes that that data are drawn from a normal distribution. If the calculated acceptance probability is \(\alpha = 1\) then the proposal parameters are added to the chain as this is a 100% probability of acceptance. The more interesting case is if \(\alpha < 1\). Then a random number \((u)\) is drawn, again from a uniform distribution, that has a width of 0-1. If \(u > \alpha\) then the proposed parameters are rejected and the prior parameters are used at the current chain
Otherwise, if \( u < \alpha \) then the proposed parameters are accepted into the chain and used to determine the next chain index.

\[ \chi^2 \]

Before the results of the MCMC are discussed some limitations and caveats need to be stated. The first is that the distribution of the parameters is assumed to be normal distributions. The second is that by using the \( \chi^2 \) distribution it is assumed that a linear fit with 2 free parameters, \( a \) and \( b \), is the correct model of the data. Further, as there is a lack of variance associated with the data points.
for the Luminosity or the WHIM a uniform variance is assumed as well. That being said the results of the MCMC are as follows.

**Figure 2.16:** The "walk" of parameter b. Similarly to parameter a there is a burn-in period until approximately the 2000 iteration. The total chain length is 10000.

The process described above was repeated over 10000 iterations. This yielded the parameter "walks" for both a and b. The expected behavior of a MCMC is seen here as both parameters start at 3 and quickly converge to a closer region that is related to the most likely value. The chains can be seen in Figure 2.15 for parameter a and Figure 2.16 for parameter b. Very clearly the so-called "burn in" period, the region that can be excluded due to it simply
Figure 2.17: The distribution for parameter a. It follows a normal distribution and is limited to after the burn-in period of 2000 iterations. The parameters of this figure can be seen in Table 2.2.

being the steps taken to enter the region of the true parameter value, occurs at approximately iteration 2000. As such the burn in region is excluded. The parameter values can be seen in Table 2.2. The distribution of parameter a can be seen in Figure 2.17 and parameter b in Figure 2.18. The distributions are limited to iterations after 2000 to exclude the burn in period. The distributions of both parameter a and parameter b follow normal distributions. Out of 10000 iterations, there was an acceptance rate of 41.35%. This means that there were
Figure 2.18: The distribution for parameter b. It follows a normal distribution, and is limited to after the burn in period of 2000 iterations. The parameters of this figure can be seen in Table 2.2.

4135 proposed values that were accepted into the chain that were not reused prior values. Further approximately half of those values can be attributed to the burn in region. Both Figure 2.15 and Figure 2.16 show clear direction to the point of maximum likelihood. Table 2.2 shows the results of this MCMC.

In comparison to the original fit using the polyfit function of Python (see Table 2.1) the parameters are in agreement. The original fit had a slope of $a=0.757$ and an intercept of $b=-0.413$ whereas the MCMC centered around a
Table 2.2: MCMC Parameters. This table shows the results of the MCMC after excluding the burn in period. This table includes the slope and intercept as well as 1 $\sigma$ confidence level.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Median</th>
<th>Standard Deviation</th>
<th>68% confidence level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope (a)</td>
<td>0.763</td>
<td>0.763</td>
<td>0.008</td>
<td>0.756 $\leq a \leq$ 0.771</td>
</tr>
<tr>
<td>Intercept (b)</td>
<td>-0.42</td>
<td>-0.416</td>
<td>0.007</td>
<td>-0.423 $\leq b \leq -0.041$</td>
</tr>
</tbody>
</table>

slope of a=0.763 and an intercept of b=-0.416. The values of the original fit fall within the 68% confidence interval of the MCMC, which indicate that the MCMC and the original fit are in agreement with each other. Figure 2.19 shows both the fit of the results of the MCMC and the fit due to polyfit. As seen in Figure 2.19 the different results of linear fitting methods are in such an agreement that they overlap and are indistinguishable from each other which is a double-edged sword. IN other words, it is good that the different forms of linear regression agree with each other. It further implies that the implemented Monte Carlo Markov Chain and the Metropolis-Hastings sampling method are most likely correct and an accurate representation of the model. However, this also implies that this is the correct fit of the data. This means this is not what was expected of the results to appear as. Examining Figure 2.19 the best-fit line almost seems to have too low of a slope. Shooting underneath the densest and most luminous region of the data. While it is evident that the region that contains the highest count of data points influences the fit, it is unexpected to have such a significant impact on the
Figure 2.19: This figure shows an over plot of three different methods of linear fits. The SciPy package [34], poly-fit, and the results of the MCMC. The only line that is visible is that of the MCMC, this is due to the overlap between the lines as the results are almost identical.

line of best fit. There are more attempts presently in progress to further refine the line of best fit that will be discussed in Section 3.2.
Chapter 3. Conclusions and Future Work

3.1 Discussion of Results

After many hours and many, many more error logs, it is time to discuss the results of this endeavor. This work investigated the relationship between the luminosity and WHIM of filament regions. This was attempted through the use of the IllustrisTNG simulation, which simulated a $300^3 \text{ Mpc}^3$ space. The filament spines were identified using the DisPerSE algorithm using advanced topology techniques in an effort to determine a scaling relationship between the luminosity and the WHIM density in the filaments between galaxy clusters. While the initial scaling relationship between the WHIM over-density and the luminosity over-density indicates that there may be a strong correlation, there are unique challenges that still need to be addressed. The initial linear fit does not follow the initial expectation and appears to have a slope that is too small. However, as discussed, multiple other methodologies were applied to the data such as imposing limits on the data set to determine the significance of certain regions, and further implementing a Monte Carlo Markov Chain to determine the best-fit parameters. The MCMC agreed with the previous linear fit results, being within a 68% confidence interval for both the slope and intercept. However, this still yielded the surprising line of
best fit, which misses the densest WHIM and most luminous region of the data. This still needs to be investigated further.

This paper also presents a novel approach to the traditional nearest-neighbor search: through the k-dimensional tree [3]. This approach to storing and searching data yields many benefits. The primary one is the quick run time of \( O(\log(n)) \), when in lower k dimensions. That is \( k < 20 \). The application of this useful computer science technique allowed for results of much higher quality than what would have been presented without it. The KD-Tree allowed for the use of the entire simulation of IllustrisTNG instead of the initial under-sampled 1 in 400 WHIM particles.

### 3.2 Systematic Error

Sources of error are always one of the most important and most dreaded topics of discussion, as it allows the humble art of self scrutiny to become a group activity. The sources of systematic error in this paper include: the range limitation of the luminosity over-density, the future implementation of the different radii for filaments, and the choice of the smoothing parameter. Considering Figure 2.12 as the "correct" or baseline fit, the errors can be seen in Table 3.1. As the investigation is still in progress these error can change, ideally, shrink in magnitude; however, in the same vein, other sources of error may arise, potentially as a result of the choice of the LD range. As presented in Figure 2.12, the LD range is from 0-2.5. This choice of LD range has a few errors associated with it. Figure 2.11 shows that the 95% of mass contained in filaments is associated with a LD
range of -4-1.8. The percent difference between the fit parameters of these two ranges can be seen in Table 3.1. Further, as a result of the reduced LD range, there is a -40% mass exclusion of the WHIM for a range of 0-2.5. The smoothing parameter choice is also an impacting aspect as it influences which cells have a non-zero luminosity value and would allow for the inclusion of otherwise zero valued cells. Figure 2.1 shows the chosen smoothing parameter of 1.2. The peak

<table>
<thead>
<tr>
<th>Source of error</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choice of LD range $a_{diff}$</td>
<td>51.23%</td>
</tr>
<tr>
<td>Choice of LD range $b_{diff}$</td>
<td>105%</td>
</tr>
<tr>
<td>Choice of 1.2 smoothing parameter $a_{diff}$</td>
<td>12.29%</td>
</tr>
<tr>
<td>Choice of 1.2 smoothing parameter $b_{diff}$</td>
<td>20.54%</td>
</tr>
<tr>
<td>Radii of filaments</td>
<td>To be determined</td>
</tr>
<tr>
<td>WHIM Mass Exclusion (As presented)</td>
<td>-40%</td>
</tr>
</tbody>
</table>

Table 3.1: Systematic Error. Possible contributors of error and their contributions to impact the scaling relation. The choice of LD range is a percent difference between the ranges of -4-2.5 (95% of WHIM mass contained see Figure 2.10) and 0-2.5 (the presented range). The choice between a smoothing parameter of 1.2 and 1.4 as the best smoothing parameter is a range see Figure 2.1. The contribution is a percent difference from what is considered to be the true or correct scaling relation as presented in Figure 2.12. The WHIM exclusion is found based on the CMD of Figure 2.10.
smoothing parameter is a range rather than a point, so in an effort to see the impact of the smoothing parameter on the fit, a smoothing parameter of 1.4 was chosen to compare. The percent difference of the fit parameters can be seen in Table 3.1. The topic of systematic errors and ideas to mitigate the effects of them is an ongoing discussion.

3.3 Recommendations for Future Study

Currently, the linear relationship is still being refined to ensure that the figures presented in this paper are accurate. The most recent investigation is comparing the linear fits between the "red" and "blue" galaxies of the IllustrisTNG simulation. The approach involves first separating the red-shifted galaxies and the blue-shifted galaxies and replicating the steps discussed in sections 2.1-2.3.2. Then compare the resulting linear relations to each other and also to the original fit that was the result of combining the red and blue galaxies. Further, the creation process for both the WHIM over-density grid and the Luminosity over-density grid is undergoing inspection to ensure that there are no other biases present. The motivation for this is due to the bias that was created from the underestimation of the WHIM density during the construction of the over-density grid (see section 2.3.2 for this discussion).

Further, all results presented in this paper are for filaments that have a radius of 1 Mpc. It would be of interest to extend this radius to 3 Mpc in increments of 0.1 Mpc, then compare the resulting linear relations. This would provide insight into how the density of the WHIM and its corresponding luminosity evolves
as the radius moves towards the less populated regions that are further from the
spine of the filament. Further, by looking at the cumulative mass distributions
of the WHIM inside of filaments with different radii could also provide insight as
to where the majority of the WHIM is located inside of filament regions. The
cumulative mass distribution of the WHIM could also provide insight as to what
region of luminosity values the majority of the mass of the WHIM is located, al-
lowing for further refinement of the scaling relation, possibly sectioning the data
into two regions of high luminosity values and low luminosity values.

Once the linear relationship is sufficiently refined the next goal would be to
apply it to observation. This would be applied to SDSS to see if our results would
be both accurate and useful. The motivation for this comparison is the ability to
accurately and reliably make predictions for the WHIM density of filaments that
have been detected while only using the more easily acquired optical information
about a region of the sky, to both encourage and identify interesting regions of
the sky to observe in the far ultraviolet and x-ray for sight lines that correspond
to high column density of the WHIM.

3.4 Conclusion

The primary goal of this project was to determine a scaling relationship
between the luminosity and the density of the WHIM located in filament regions
inside of the IllustrisTNG 300-1 simulation at $z=0$. The systematic process de-
vised to meet this goal first started by limiting the luminosity values, removing
the artifacts of the simulation via the Flag filter and limiting values based on
$M_r = -18.014$ (see Section 2.1). Then, removing the WHIM interior to the halos by utilizing the KD-Tree in section 2.2. Finally, constructing the filament grids to further limit the LD and WHIM to filament regions. Arriving at this limited data set allowed for the investigation of the CMD of the WHIM as a function of the LD. Figure 2.10 shows that 95% of the WHIM mass correspond to a LD range of -4 to 1.8. Further, after eliminating the bias created by the construction of the WHIM grid (Section 2.3.2), the first scaling relation attempt was made and compared against different ranges of the LD and WHIM density (Table 2.1). This showed very little impact excluding higher LD values. Following this to verify the results of the linear regression, the parameters were calculated via Monte Carlo Markov Chain (Section 2.3.5), which yielded parameters that were nearly identical to those of the linear regression. The parameters found via the MCMC include the original fits parameters inside of the 68% confidence level (See Table 2.2).

While this is the extent of the work presented in this thesis, the project is still in progress, and different areas are being investigated as discussed in section 3.3. To conclude, the linear fit presented with this thesis is solid foundation for future investigation and is currently undergoing refinements to determine if it is an accurate prediction method for WHIM density, with plans to apply the refined scaling relation to both the IllustrisTNG simulation and SDSS observations.
References


