# Machine Learning Structure Formation: Data Access

Parimah Safarian, Saba Etezad Razavi and Erfan Abbasgholinejad
Supervisor: prof. Raiesi

March 2020

## Contents



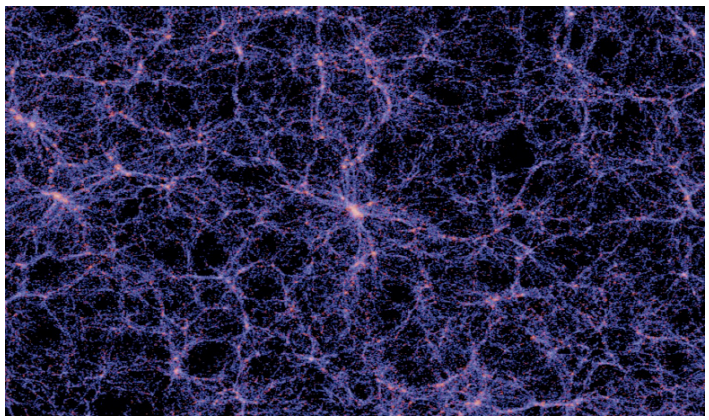

Figure 1: Cosmic Web

# 1 Introduction

Population of dark matter halos and their formation initiated from early universe density fluctuations have received attention because of their important role in cosmological structure formation specifically galaxies and galactic clusters. This is because these halos acts as a potential wells for capturing baryonic matter to form massive structures. Our observation about redshift of structure formation zone confirms our idea about dark matter halos role.
Press-Schechter [4] method and more novel methods like Excursion Set Theory [6] are most succeeded semi-analyctical models for describing halo collapse. Despite their successful results they have some problems:

- Extremely dependance on cosmological model
- They are useful in the regime we are allowed to apply Newtonian gravitation. In this regime density threshold depends on scale factor linearly and general relativity effects are neglecated.
- Press-Schechter mass function overestimates halos whith lower masses and underestimates massive halos.

The more numerical method to investigating structure formation is to run large simulations at each time, but it has heavy computational expenses.

In this project we are going to approximately repeat what people have done in two papers [2] & [**paper2018**] applying machine learning methods and then use them for solving our problems.
The mos important advantage compared with older methods is that it is not model base which make us able to even test the models and it can be applied in all regimes, also there is no need to run a heavy simulation for each initial condition. We train the machine once and we use it for all times!
It is important to mention this method is sensitive to the simulation assumptions.
The main goal is to train machine to find out a correspondence between initial conditions and final halos. In the paper Random Forest algorithm has been used and the problem has been mentioned as a supervised classification problem.

# 2 Data

## 2.1 Source of Data

For our purpose for analysing the conditions and features of particles which end up into a halo dark matter at $z = 0$, we needed the data of particles in different snapshots.

Thus for this purpose we used the "Virgo-Millennium" database.The simulation was carried out using a modified version of the publicly available code "GADGET-2" (Springel 2005). The algorithm "SUBFIND" (Springel et al.

2001) was used to identify all self-bound halos containing at least 20 particles and all self-bound subhalos within these halos down to the same mass limit.

A millennium run uses $10^10$ particles each mass $8.6108h^{-1}M_{\odot}$, to follow the evolution of the dark matter distribution within a cubic region of side $500h^1Mpc$ from redshift $z = 127$ (snapNumber = 67) until $z = 0$ (snapNumber = 0). The Benchmark model [1] is assumed for this simulation which the cosmological parameters in it are $\Omega_m = \Omega_D M + \Omega_b = 0.25$, $\Omega_b = 0.045$, $\Omega_\Lambda = 0.75$, $h = 0.73$, $\sigma 8 = 0.9$ and $n = 1$ with standard definitions for all quantities.

## 2.2 Access Data

In order to access the data we used the standard "Structural Query Language" (SQL). We derived our demanded data in two ways:
1. A data set from merger trees database [3] on the history of a given halo at a given redshift. Which in this context we started to classify our data in two groups in a special way, "In-halo-particle" at any redshift is a sub-halo with mass between 20 to 100 $m_(particle)$[1] which will end up into a larger halo with the minimum mass of 1600 $m_(particle)$ at $z = 0$ and "out-halo-particle" is visa versa a sub-halo with mass between 20 and 100 $m_(particle)$, which do not end up in a halo with a mass larger than the above limit at $z = 0$.
2. The particles data which contain all the data of each particle at a given redshift, but as this data were quite large (about 1TB on the hole millennium run) and also in order to make a fair balance between accuracy of results and run-time (even run-time for our queries! which couldn't be more than 30 minute in each run), we forced to work on the "millimil" instead of the hole run. It contains the data of a $60^3h^{-3}Mpc^3$ box with about 20 million particles and $5kpc$ resolution. Here is the general information about the related tables which millennium database contain, we used MPAHalo, MillimilSnapshotIDs and MillimiSnapshots for our purpose:

for making our "IN" and "OUT" classes, we couldn't run an exact query for getting the information straight from the database because of the limit in the gate's openning time, so we collected the information about all the particles which exists in halos with $np > 1600$ at $z = 0$ and we traced these particles into $z = 127$ by running a python code, which you can find it here

## 2.3 Our Data

Now we have collected about 30MB data in 4 different snapshots, equivalently different redshifts, of the first type. including velocity vectors, spatial indices, halo mass, and the last progenitors(see the merger trees) IDs.
Also we have collected about 4GB data from "Millimil" [5] for all the particles in the box, which includes the position of each particles, it's velocity vectors, particle ID and the Peano-Hilbert key in three snapshots which are in redshifs 0, 16.7 and 127. The Peano-Hilbert key corresponds to the position of the particle.

[1] Remember that $m_(particle) = 8.610^8h^{-1}M_{\odot}$


Figure 2: Caption

Based on a $32^3$ grid. You can see a further description and Data samples in our Github repository[5].

# 3 A Quick Analysis

## 3.1 Particles Distribution and Velocity

In order to visualize our data we have plotted the particles distribution in space for different redshifts. They are shown in figure 5 to 11. In the figure 11 homogenous density is well shown and it is because redshift is large enough to not see any sign of structures yet. As redshift decreases by time we can see more heterogeneity in figure 12 and finally in figure 5 as $z = 0$ corresponds to $t = t_0$ and cosmic structures are now quite formed it is the plot which shows highest difference between voids and dense regions.

## 3.2 Dark Matter Halos

We defined our halo class in a way that the minimum mass for a halo is equal to 1600 particles. as can be seen in the plots below, the number of halos with mass less than 10000 particles is much more than the bigger ones, as we expected from our cosmology. Total number of halos with np more than 1600 is also 706, while the total number of halos in the box (including the lighter ones) is around 37000. As it is shown in figure 6 the halos mass distribution histogram shows a narrow pick in low masses.

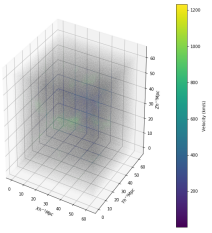

Figure 3: $z = 127$

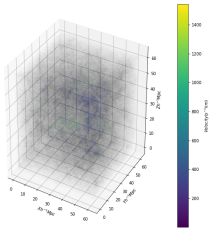

Figure 4: $z = 16.7$

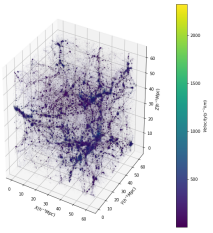

Figure 5: $z = 0$

Particles' position in the a $60^3 h^{-3} Mpc^3$ box.

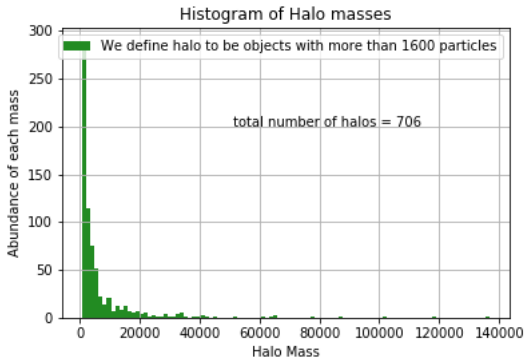


Figure 6: Dark matter halos mass distribution

# 4 What We Want to Train

Eventually we need the initial condition of each particle and wether it is in a halo or not at z=0 (with our definition of halo), we should have derived this data from the both halo and particle data and consequently classify our data in two classes "IN" and "OUT". For this purpose we wrote a code[2] to build a matrix which includes the information of all the particles in halos and then apply it to our particle data to find our class "OUT" objects as well. we made a column named "$In_halo$" at the end of our data frame which will take one of the two values 0 or 1 for each particle. 0 for "OUT" class with the number of 19 million particles and 1 for "IN" class with the number of 5 million particles. In the next step we are going to train initial conditions and "In" or "Out" categories. Machine will help us to understand a new particle with a given initial condition is going to bound in a halo or not. Figure 13 to 15 illustrate particles.

[2]You can find the code in Github repository

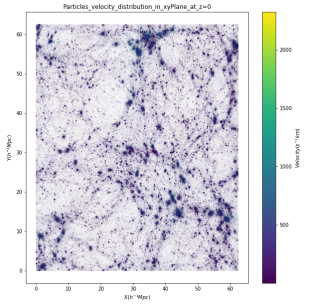


Figure 7: $z = 0$

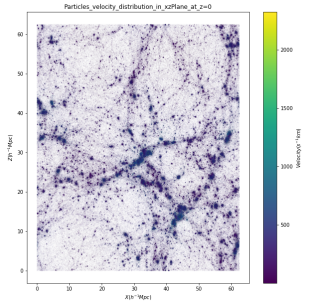


Figure 8: $z = 0$

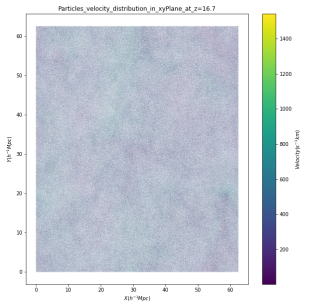


Figure 9: $z = 16.7$

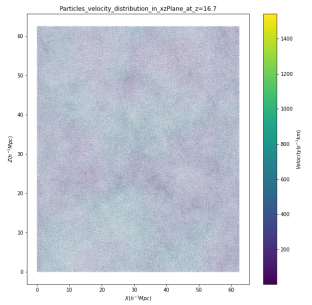


Figure 10: $z = 16.7$

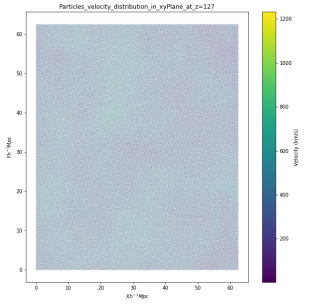


Figure 11: $z = 127$

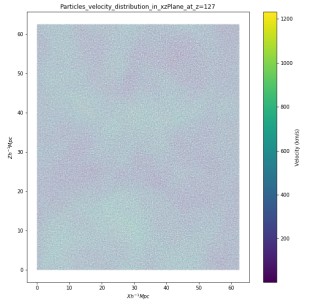


Figure 12: $z = 127$

Particles' position projected in xy and xz plane.

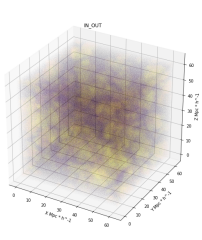

Figure 13: $z = 127$

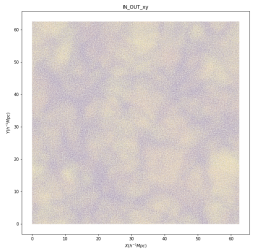

Figure 14: $z = 0$

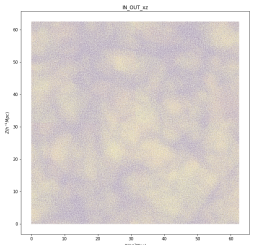

Figure 15: $z = 0$

The purple particles are "In" halos and the yellow ones are "Out"

# 5 Conclusion

We collect the data from "Millimil" and produced the parameters we are going to train in next steps. For a better knowledge about the data we plotted some things and checked them to be compatible with our knowledge and also to have a view from the physical properties of the problem. We learned how to work with large databases, using SQL, terminal and Github and then pushed all the codes and plots on our project's repository and now the way and next steps are much clearer! For finding codes, sample data, plots and further description check out our Github repository:
https://github.com/Machine-Learning-in-Structure-formation/NLSFML

# References


[1] E. Macaulay et al. "First Cosmological Results using Type Ia Supernovae". In: (2019). URL: https://arxiv.org/pdf/1811.02376.pdf.

[2] Andrew Pontzen Luisa Lucie-Smith Hiranya V. Peiris. "An interpretable machine learning framework for dark matter halo formation". In: (2019). URL: https://arxiv.org/abs/1906.06339.

[3] "Merger Trees Database". In: (). URL: http://gavo.mpa-garching.mpg.de/MyMillennium/Help?page=mergertrees.

[4] P. Press W. H. Schechter. "Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation". In: (1974). URL: http://articles.adsabs.harvard.edu/full/1974ApJ...187..425P.

[5] Parimah Safarian Saba Etezad Razavi Erfan Abbasgholinejad. "Source codes". In: (2020). URL: https://https://github.com/Machine-Learning-in-Structure-formation/NLSFML.

[6] Andrew R. Zentner. "The Excursion Set Theory of Halo Mass Functions, Halo Clustering, and Halo Growth". In: (2006). URL: https://arxiv.org/abs/astro-ph/0611454.