Predicting Galaxy Interloper Fraction with Graph Neural Networks

Elena Massara
Waterloo Centre for Astrophysics, University of Waterloo
200 University Ave W, Waterloo, ON N2L 3G1, Canada
Department of Physics and Astronomy, University of Waterloo
200 University Ave W, Waterloo, ON N2L 3G1, Canada
elena.massara.cosmo@gmail.com

Francisco Villaescusa-Navarro
Center for Computational Astrophysics, Flatiron Institute,
162 5th Avenue, 10010, New York, NY, USA
fvillaescusa@flatironinstitute.org

Will Percival
Waterloo Centre for Astrophysics, University of Waterloo
200 University Ave W, Waterloo, ON N2L 3G1, Canada
Department of Physics and Astronomy, University of Waterloo
200 University Ave W, Waterloo, ON N2L 3G1, Canada
Perimeter Institute for Theoretical Physics,
31 Caroline St. North, Waterloo, ON N2L 2Y5, Canada
will.percival@uwaterloo.ca

Abstract

Upcoming emission line spectroscopic surveys, such as Euclid and the Roman Space Telescope, will be prone to systematics due to the presence of interlopers: galaxies whose redshift and distance from us are miscalculated due to line confusion in their emission spectra. Particularly pernicious are interlopers involving the confusion between two lines with close emitted wavelengths, since these interlopers correlate with the target galaxies. An interesting example is Hβ emitters confused as [O III] emitters. They introduce a particular pattern in the 3D distribution of the observed galaxy catalog that can bias the cosmological analysis performed with that sample. We present a novel method to predict the fraction of interlopers in a galaxy catalog, using simulations and halos as a proxy for galaxies. This method uses Graph Neural Networks to learn the posterior distribution of the interloper fraction while marginalizing over cosmological and astrophysics unknowns.

1 Introduction

Slitless spectroscopy used in upcoming galaxy surveys, such as Euclid and Roman, will provide emission line galaxy spectra with low signal-to-noise ratio, but will allow us to observe an unprecedented number of galaxies out to redshift $z = 3$. Measurements of the galaxy redshifts will be performed using one or two lines in each spectrum, making these surveys prone to contain interlopers. Interlopers are galaxies whose redshift has been miscalculated due to line confusion, which leads to a wrong prediction for their distance from us. The confusion happens because an emission line is wrongly assumed to be the target line: Hα in Euclid and Roman up to $z = 1.8$, and [O III] in Roman

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

at redshifts $z = 1.8 - 3$. We consider a particular type of interlopers, those that correlate with the target sample because the two confused lines have similar emitted wavelengths, and thus the distance between the true and wrongly inferred position of the interloper is small and interlopers live in the volume of the survey. In particular, we consider Hβ emitters that are confused as [O III] emitters. Their distance is underestimated by about $97\,h^{-1}$Mpc at $z = 1$ and $85\,h^{-1}$Mpc at $z = 2$. This systematic shift of a subsample of objects introduces a particular anisotropic pattern in the 3D distribution of the observed galaxies that can bias the analysis performed with that data sample [Massara et al., 2021]. In this work we train Graph Neural Networks (GNNs) [Battaglia et al., 2018] to infer the unknow fraction of Hβ-[O III] interlopers in a catalog using a likelihood-free field-level method.

2 Data Set

We use halos (clumps of dark matter) as a proxy for galaxies to test our method. Indeed, both types of objects trace the 3D distribution of the matter field with bias schemes that can be tuned. Moreover, to demonstrate the method, we do not need to build galaxy catalogs with emission line distributions that have realistic fractions of interlopers in them. We instead need to generate a sufficiently large data set with enough variation in the objects' bias, the underlying cosmology, and the fraction of interlopers, so that the training set can be used to train a flexible enough model that can predict the fraction of interlopers effectively marginalizing over the other unknowns[1]. We build our data set from the halo catalogs of a subset of the Quijote simulations [Villaescusa-Navarro et al., 2020a]:

`SET1`: 100 simulations at the so-called fiducial cosmology, a flat ΛCDM cosmology with matter density parameter $\Omega_{\rm m} = 0.3175$, baryon density parameter $\Omega_{\rm b} = 0.049$, dimensionless Hubble constant $h = 0.6711$, spectral index $n_s = 0.9624$, linear matter fluctuation amplitude $\sigma_8 = 0.834$, and sum of neutrino masses $M_\nu = 0$ eV.

`SET2`: 155 simulations selected among 2,000 simulations whose cosmological parameters are arranged in a Latin Hypercube (LH) configuration. The selected boxes have parameters $\Omega_{\rm m} \in [0.18 - 0.42]$, $\Omega_{\rm b} \in [0.038 - 0.062]$, $h \in [0.58 - 0.82]$, $n_s \in [0.88 - 1.12]$, $\sigma_8 \in [0.68 - 0.92]$.

All the simulation boxes cover a volume equal to $1\,h^{-3}{\rm Gpc}^3$, and they contain 512^3 dark matter particles from which halos have been identified using the Friends-of-Friends (FoF) algorithm. Due to constraints from the memory of the GPUs, we crop each box along the $\hat{x}$ and $\hat{y}$ directions to obtain multiple sub-boxes with fewer halos. The $\hat{z}$ axis is assumed to be the line-of-sight direction, along which we apply redshift space distortions (halos are displayed depending on their velocity) and shift by $97\,h^{-1}$Mpc randomly selected objects to mimic interlopers. The fraction of such objects is uniformly sampled within $f_i \in [0.0 - 0.2]$ and varied among different sub-boxes. Then, from the halos in each sub-box we build a graph, where halos are the nodes that are connected via edges if their distance is smaller than the linking radius $r_{\rm link}$, which is a hyperparameter that we tune. We add attributes that respect the symmetry of the problem[2] by describing the spatial distribution of halos via edge attributes only [Villanueva-Domingo et al., 2022]:

$$\mathbf{e}_{ij} = \left[r_\parallel = \frac{\mathbf{d}_{ij} \cdot \hat{z}}{r_{\rm link}}\,, r_\perp = \frac{|\mathbf{d}_{ij} \times \hat{z}|}{r_{\rm link}}\,, \cos\theta = \frac{\mathbf{v}_{i\perp} \cdot \mathbf{v}_{j\perp}}{|\mathbf{v}_{i\perp}||\mathbf{v}_{j\perp}|} \right] \tag{1}$$

where $\mathbf{v}_i$ is the vector connecting the centroid of the catalog and the node i, $\mathbf{d}_{ij} = \mathbf{v}_i - \mathbf{v}_j$ is the vector connecting the two nodes i and j at the beginning and end of the edge $\mathbf{e}_{ij}$, and $\mathbf{v}_\perp = \mathbf{v} \times \hat{z}$ is the component of $\mathbf{v}$ perpendicular to the line-of-sight. Moreover, we introduce global attributes describing the whole graph in some cases. Once the graphs are built, we divide the data into training, validation, and test sets with an 80/10/10 split ratio.

[1] However, realistic galaxy catalogs that include survey geometry and observational systematics will be needed to train such types of GNN models before using them on real data.

[2] The Universe is invariant under translations and rotations, however we observe redshifts rather than distances. The galaxy redshift is not only determined by the Hubble flow, hence its distance to us, but also by the peculiar velocity of the galaxy along the line-of-sight $\hat{z}$, causing an observed anisotropic distortion. Moreover, when converting redshift and angles into distances, a fiducial cosmology needs to be assumed. If that is different from the cosmology of the Universe (or of the simulation box considered), additional anisotropic distortions are introduced in the dataset. The observed Universe thus exhibits a cylindrical symmetry and we implement it in the attributes of the edges in our graphs.

3 Graph Neural Network

We build GNNs to determine the fraction of interlopers in a catalog via likelihood-free inference. The GNN architecture is composed of GNN blocks that take as input a graph and output the same graph with updated node, edge, and global attributes via message passing. Thus, even if the initial graph does not contain node attributes, the GNN blocks will assign and update them. Each block l is composed of the following elements:
The edge model that updates each input edge attribute $\mathbf{e}_{ij}^{(l-1)}$ to the output $\mathbf{e}_{ij}^{(l)}$,

$$\mathbf{e}_{ij}^{(l)} = \phi_l \left(\left[\mathbf{n}_i^{(l-1)}, \mathbf{n}_j^{(l-1)}, \mathbf{e}_{ij}^{(l-1)} \right] \right) , \tag{2}$$

and the node model that updates the node attributes,

$$\mathbf{n}_i^{(l)} = \psi_l \left(\left[\mathbf{n}_i^{(l-1)}, \bigoplus_{j \in \mathcal{N}_i} \mathbf{e}_{ij}^{(l)}, \mathbf{u} \right] \right) , \tag{3}$$

where ϕ and ψ are MLPs, $\mathbf{u}$ is the global attribute (when specified), and $\bigoplus = [\,\max,\ \mathrm{mean},\ \sum]$ is a permutation invariant aggregation operator applied to all edges $\mathbf{e}_{ij}$ with $j \in \mathcal{N}_i$ and $\mathcal{N}_i$ being the indexes of the nodes connected to node $\mathbf{n}_i$. The number of GNN blocks, $N_{\rm block}$, determines the number of times the message passing operation is performed and the attributes are updated. After the GNN blocks, an additional aggregation operation compresses the information of each graph, then it is concatenated to the global feature $\mathbf{u}$, when present, and passed to a final MLP τ to obtain the output vector y. All MLPs are built using two fully connected layers with the ReLU activation function. The number of GNN blocks and neurons per fully connected layer are hyperparameters that we optimize.

We train the GNNs to output the vector $\mathbf{y}(\mathcal{G}) = [\mu(\mathcal{G}), \sigma(\mathcal{G})]$ with

$$\mu(\mathcal{G}) = \int df_i\, p(f_i|\mathcal{G}) f_i\,, \qquad \sigma(\mathcal{G}) = \left[\int df_i\, p(f_i|\mathcal{G})(f_i - \mu)^2 \right]^{1/2} \tag{4}$$

being the mean and standard deviation of the marginalized posterior distribution $p(f_i|\mathcal{G})$,

$$p(f_i|\mathcal{G}) = \int d\theta_1 ... d\theta_n\, p(f_i, \theta_1, ..., \theta_n|\mathcal{G}) \tag{5}$$

where $\theta_1, ..., \theta_n$ are cosmological and/or astrophysical parameters. In order to train such a model, we implement the loss function [Jeffrey and Wandelt, 2020, Villaescusa-Navarro et al., 2022]

$$\mathcal{L} = \log \left\{ \sum_{j \in \rm batch} (f_{i,j} - \mu_j)^2 \right\} + \log \left\{ \sum_{j \in \rm batch} \left[(f_{i,j} - \mu_j)^2 - \sigma_j^2 \right]^2 \right\} \tag{6}$$

whose minimization has been shown to be equivalent to solving for the mean and standard deviation of the posterior distribution [see Villaescusa-Navarro et al., 2020b]. We minimize the loss function using the ADAM optimizer [Kingma and Ba, 2017], with values for the learning rate and weight decay that we treat as hyperparameters to be optimized. The optimization of all hyperparameters (learning rate l_r, weigh decay w_d, number of GNN blocks $N_{\rm block}$, number of neurons $N_{\rm hid}$ in MLPs, and linking radius $r_{\rm link}$) is performed using the OPTUNA package [Akiba et al., 2019] with at least 100 trials, each of those consisting in the training of a model with a specific choice for the value of hyperparameters. We select the GNN model with hyperparameters that give the best validation loss after training.

We quantify the performance of the GNN using various metrics applied to the test sets. We consider: The root mean square error $\mathrm{RMSE} = \sqrt{< (\mu - f_i)^2 >}$ with $< ... >$ indicating the mean among the test set, which quantifies the precision of the model—the lower the RMSE, the more precise the model is; The coefficient of determination $\mathrm{R}^2 = 1- < (\mu - f_i)^2 > / < (f_i - < f_i >)^2 >$ that measures the accuracy of the model (the closer it is to 1, the more accurate the model is); An estimation for the bias, $b =< \mu - f_i >$; The $\chi^2 =< [(\mu - f_i)/\sigma]^2 >$ that indicates if the standard deviations of the posterior distributions are well determined by the model (happening when $\chi^2 \sim 1$).

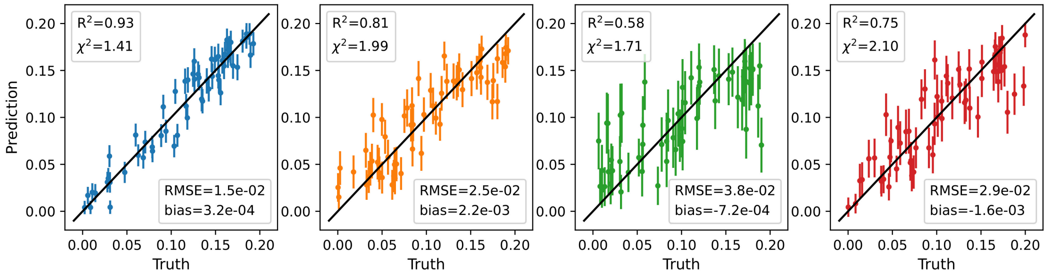

Figure 1: Likelihood-free inference of interloper fraction in the test sets. Left to right: fixed cosmology and halo bias, fixed cosmology and varied bias, varied cosmology and bias, latter with global feature.

4 Results

Fixed cosmology and halo bias We consider the case where all catalogs share the same cosmology and halo bias, and are built using SET1. In this case, the posterior distribution of the fraction of interlopers $p(f_i|\mathcal{G})$ in equation 5 is not marginalized over any other parameter. From each simulation we crop sub-boxes of size $150 \times 150 \times 1000$ $(h^{-1}\text{Mpc})^3$ along the $\hat{x}$, $\hat{y}$, and $\hat{z}$ directions, respectively. As a result, these sub-boxes contain about 4,500 objects. Depending on the sub-box, a different fraction of halos are selected to represent interlopers.

After training more than 100 models via an Optuna study, we identify the model with the best validation loss having hyperparameters $N_{\text{block}} = 1$, $N_{\text{hid}} = 35$, $l_r = 3.8 \times 10^{-4}$, $w_d = 10^{-2}$ and $r_{\text{link}} = 11.68\, h^{-1}$Mpc. The single GNN block and the low value for r_{link} indicate that the GNN is using small-scale clustering properties to determine the fraction of interlopers. As expected, there is lot of information on small scales, since the number of pairs and their spatial distribution change depending on the number of interlopers in the catalog and graph. The left panel of Figure 1 shows the inference performed on the test sets, with the x-axis indicating the true value f_i, and the y-axis denoting the prediction of the model. The black line shows the values $y = x$, points denote the predicted mean, and error bars denote the predicted standard deviation in each test catalog. As shown by the figure, there is no bias, $b \sim 0$, and $\chi^2 = 1.4$, indicating that the standard deviation is slightly under-predicted, probably because it does not take into account the epistemic error of the GNN (we measured it to be 0.002, corresponding to about 15% of the standard deviation). The model has coefficient of determination close to 1, $R^2 = 0.93$, and a precision equal to ± 0.015, corresponding to a 15% error on the mean range value $f_i = 0.1$. This precision is obtained using a volume equal to $0.0225\, h^{-3}\text{Gpc}^3$, which is a very small fraction of the [O III] survey in the Roman Space Telescope and a volume thousands of times smaller than the one considered in Foroozan et al. [2022], where the authors developed a model fit for both the baryon acoustic oscillation (BAO) position and the interloper fraction. We can compute the $RMSE$ for their BAO+f_i analysis using the results in their Figure 1 for $\Delta d = 97\, h^{-1}$Mpc. We obtained $RMSE = 3.1 \times 10^{-3}$, which rescaled to the volume considered here is equal to 0.65. Therefore, it is 40 times larger than the $RMSE$ obtained with the GNNs; in other words, the GNN models are 40 times more precise in predicting the interloper fraction than the BAO+f_i fit. Moreover, the bias from the BAO+f_i model is equal to -2.1×10^{-3}, which is an order of magnitude larger than the bias obtained with GNNs. However, we should remember that the comparison is not truly fair since the GNN has been trained at fixed cosmology, whereas the BAO+f_i fit includes variation in cosmology via the dilation parameters.

Variation of halo bias and cosmology We consider progressively more complicated tasks where the training dataset displays variation in halo bias among different catalogs but same cosmology (using SET1) and variation in both halo bias and cosmology (using SET2). In the former case, the GNN aims at learning the mean and the standard deviation of the posterior distribution $p(f_i|\mathcal{G})$ marginalized over the halo bias scheme; in the latter, the posterior distribution is marginalized over both cosmology and halo bias. In both cases, the performance of the best GNN models is degraded compared to the instance where both halo bias and cosmology are fixed, with precision equal to ± 0.025 (± 0.038) and $R^2 = 0.8$ ($R^2 = 0.6$) in the first (second) case (see second and third panel of Figure 1). The Optuna study performed in the second case suggests the need for larger linking radii (currently limited by the GPU memory). An alternative way to add information coming from large scales consists in declaring a global attribute for each graph, for example a guess for the cosmology in the form of 5 scalars describing the cosmological parameters. In this case the GNN can reach a precision equal to ± 0.029 and $R^2 = 0.75$ (right panel in Figure 1), giving a better performance similar to the case where only the halo bias is varied.

5 Conclusions

GNNs are designed to handle sparse and irregular data, such as galaxy or halo catalogs. In this paper, we showed that they can solve problems where one needs to identify a sub-sample of objects whose spatial properties differ from that of the core sample, such as interloper galaxies. In particular, we investigated if small-scale information, which is typically not used for cosmological constraints, can be used by GNNs to determine the interloper fraction. We found that GNNs perform well if the underlying cosmology is known or can be guessed with a small uncertainty, while they produce worse results if the cosmology is unknown and the large-scale information cannot be accessed. In order to improve the efficiency of the GNN, we need to include large-scale information in the graph, such as the large-scale power spectrum. However, given the small volume occupied by the graphs, their power spectrum is cosmic variance (noise) dominated on large scales. A more promising venue to include large-scale information is, for example, represented by hierarchical GNN [Sobolevsky, 2021], which would allow for larger r_{link}. Moreover, GNNs have shown to be powerful tools to constrain cosmology using the full 3D galaxy field information [Villanueva-Domingo and Villaescusa-Navarro, 2022, de Santi et al., 2023]. A further generalization of the method proposed in this paper could consist of the simultaneous inference of interloper fraction and cosmological parameters.

Acknowledgments and Disclosure of Funding

EM and WP acknowledge the support of the Canadian Space Agency. WP also acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number RGPIN-2019-03908]. Research at Perimeter Institute is supported in part by the Government of Canada through the Department of Innovation, Science and Economic Development Canada and by the Province of Ontario through the Ministry of Colleges and Universities. This research was enabled in part by support provided by Compute Ontario (computeontario.ca) and the Digital Research Alliance of Canada (alliancecan.ca).

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. *arXiv e-prints*, art. arXiv:1907.10902, July 2019. doi: 10.48550/arXiv.1907.10902.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv e-prints*, art. arXiv:1806.01261, June 2018. doi: 10.48550/arXiv.1806.01261.

Natalí S. M. de Santi, Helen Shao, Francisco Villaescusa-Navarro, L. Raul Abramo, Romain Teyssier, Pablo Villanueva-Domingo, Yueying Ni, Daniel Anglés-Alcázar, Shy Genel, Elena Hernandez-Martinez, Ulrich P. Steinwandel, Christopher C. Lovell, Klaus Dolag, Tiago Castro, and Mark Vogelsberger. Robust field-level likelihood-free inference with galaxies. *arXiv e-prints*, art. arXiv:2302.14101, February 2023. doi: 10.48550/arXiv.2302.14101.

Setareh Foroozan, Elena Massara, and Will J. Percival. Correcting for small-displacement interlopers in BAO analyses. *JCAP*, 2022(10):072, October 2022. doi: 10.1088/1475-7516/2022/10/072.

Niall Jeffrey and Benjamin D. Wandelt. Solving high-dimensional parameter inference: marginal posterior densities & Moment Networks. *arXiv e-prints*, art. arXiv:2011.05991, November 2020. doi: 10.48550/arXiv.2011.05991.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017.

Elena Massara, Shirley Ho, Christopher M. Hirata, Joseph DeRose, Risa H. Wechsler, and Xiao Fang. Line confusion in spectroscopic surveys and its possible effects: shifts in Baryon Acoustic Oscillations position. *Mon. Not. Roy. Astron. Soc.*, 508(3):4193–4201, December 2021. doi: 10.1093/mnras/stab2628.

Stanislav Sobolevsky. Hierarchical Graph Neural Networks. *arXiv e-prints*, art. arXiv:2105.03388, May 2021. doi: 10.48550/arXiv.2105.03388.

Francisco Villaescusa-Navarro, ChangHoon Hahn, Elena Massara, Arka Banerjee, Ana Maria Delgado, Doogesh Kodi Ramanah, Tom Charnock, Elena Giusarma, Yin Li, Erwan Allys, Antoine Brochard, Cora Uhlemann, Chi-Ting Chiang, Siyu He, Alice Pisani, Andrej Obuljen, Yu Feng, Emanuele Castorina, Gabriella Contardo, Christina D. Kreisch, Andrina Nicola, Justin Alsing, Roman Scoccimarro, Licia Verde, Matteo Viel, Shirley Ho, Stephane Mallat, Benjamin Wandelt, and David N. Spergel. The Quijote Simulations. *Astrophys. J. Supp.*, 250(1):2, September 2020a. doi: 10.3847/1538-4365/ab9d82.

Francisco Villaescusa-Navarro, Benjamin D. Wandelt, Daniel Anglés-Alcázar, Shy Genel, Jose Manuel Zorrilla Mantilla, Shirley Ho, and David N. Spergel. Neural networks as optimal estimators to marginalize over baryonic effects. *arXiv e-prints*, art. arXiv:2011.05992, November 2020b. doi: 10.48550/arXiv.2011.05992.

Francisco Villaescusa-Navarro, Shy Genel, Daniel Anglés-Alcázar, Leander Thiele, Romeel Dave, Desika Narayanan, Andrina Nicola, Yin Li, Pablo Villanueva-Domingo, Benjamin Wandelt, David N. Spergel, Rachel S. Somerville, Jose Manuel Zorrilla Matilla, Faizan G. Mohammad, Sultan Hassan, Helen Shao, Digvijay Wadekar, Michael Eickenberg, Kaze W. K. Wong, Gabriella Contardo, Yongseok Jo, Emily Moser, Erwin T. Lau, Luis Fernando Machado Poletti Valle, Lucia A. Perez, Daisuke Nagai, Nicholas Battaglia, and Mark Vogelsberger. The CAMELS Multifield Data Set: Learning the Universe's Fundamental Parameters with Artificial Intelligence. *Astrophys.J.Supp.*, 259(2):61, April 2022. doi: 10.3847/1538-4365/ac5ab0.

Pablo Villanueva-Domingo and Francisco Villaescusa-Navarro. Learning Cosmology and Clustering with Cosmic Graphs. *Astrophys. J.*, 937(2):115, October 2022. doi: 10.3847/1538-4357/ac8930.

Pablo Villanueva-Domingo, Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, Federico Marinacci, David N. Spergel, Lars Hernquist, Mark Vogelsberger, Romeel Dave, and Desika Narayanan. Inferring Halo Masses with Graph Neural Networks. *Astrophys. J.*, 935(1):30, August 2022. doi: 10.3847/1538-4357/ac7aa3.