Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **485,** L126–L130 (2019)
Advance Access publication 2019 March 30

doi:10.1093/mnrasl/slz045

A new approach to finding galaxy groups using Markov Clustering

L. Stothert,[1,2] P. Norberg[1,2]★ and C. M. Baugh[1]

[1] *Institute for Computational Cosmology, Department of Physics, Durham University, South Road, Durham DH1 3LE, UK*
[2] *Centre for Extragalactic Astronomy, Department of Physics, Durham University, South Road, Durham DH1 3LE, UK*

Accepted 2019 March 23. Received 2019 March 18; in original form 2018 December 17

ABSTRACT

We present a proof of concept of a new galaxy group finder method, Markov graph CLustering (MCL) that naturally handles probabilistic linking criteria. We introduce a new figure of merit, the variation of information (VI) statistic, used to optimize the free parameter(s) of the MCL algorithm. We explain that the common friends-of-friends (FoF) method is a subset of MCL. We test MCL in real space on a realistic mock galaxy catalogue constructed from an *N*-body simulation using the GALFORM model. With a fixed linking length FoF produces the best group catalogues as quantified by the VI statistic. By making the linking length sensitive to the local galaxy density, the quality of the FoF and MCL group catalogues improve significantly, with MCL being preferred over FoF due to a smaller VI value. The MCL group catalogue recovers accurately the underlying halo multiplicity function at all multiplicities. MCL provides better and more consistent group purity and halo completeness values at all multiplicities than FoF. As MCL allows for probabilistic pairwise connections, it is a promising algorithm to find galaxy groups in photometric surveys.

Key words: methods: statistical – galaxies: groups: general – galaxies: haloes.

1 INTRODUCTION

The fundamental assumption behind galaxy formation theory is that galaxies form inside dark matter haloes (White & Rees 1978). The hierarchical assembly of haloes and the time-scale for galaxy mergers means that haloes often have a main or central galaxy, accompanied by distinct satellite galaxies. There are clear predictions for the properties of the galactic content of haloes that can be tested if we can identify a high fidelity sample of galaxy groups from galaxy surveys that retains a connection to the underlying dark matter haloes (Eke et al. 2004, 2005; van den Bosch et al. 2005; Yang et al. 2005a).

The identification of a galaxy group requires an algorithm to associate galaxies with a common, unique dark matter halo. Many ways have been explored to do this, with the most common being Friends of Friends (FoF; e.g. Huchra & Geller 1982; Zeldovich, Einasto & Shandarin 1982). For example, Eke et al. (2004) and Robotham et al. (2011) created FoF galaxy group catalogues from the 2dF Galaxy Redshift Survey (Colless et al. 2001) and the Galaxy And Mass Assembly survey (GAMA; Driver et al. 2011). Liu et al. (2008) extended FoF for galaxies with photometric redshifts, which was then applied to the Pan-STARRS1 medium deep survey (Jian et al. 2014). Yang et al. (2005b) developed a halo-based group finder that was used to construct a group catalogue using Sloan Digital Sky Survey (SDSS) galaxies (Yang et al. 2007).

However, despite the success of FoF-based methods they are far from perfect and struggle when applied to low-density samples as is the case with galaxy catalogues. This should be contrasted with their application to numerical simulations where the particle distribution is thousands of times denser (if not more) than a typical galaxy distribution. When applied to galaxy catalogues, FoF tends to create either too many low multiplicity groups (by fragmentation of the larger ones) or groups that are too big (by spuriously joining smaller groups to bigger ones). Measures of purity and completeness are then used to rate the quality of the group catalogue and these statistics tend to be combined in some way, to create a statistic that should be minimized to ensure an ‘optimal’ set of groups (see, for example, Eke et al. 2004). It is worth noting that FoF does not use all of the available pairwise information, nor can it be extended naturally to handle probabilistic positional information, as is the case with e.g. photometric redshifts.

Here we show that the FoF approach to galaxy group finding is just one solution to the graph clustering problem (e.g. Schaeffer 2007). Graph clustering aims to find clusters of points given all pairwise connection amplitudes between them. It is a problem that occurs in many situations, such as detecting communities in social networks (e.g. Liu et al. 2014). We explain, in Section 2, how the FoF algorithm is a subset of the Markov graph CLustering (MCL) algorithm (Van Dongen 2000), which we apply to the problem of galaxy group detection. MCL has been widely used in the field of bioinformatics in detecting groups of proteins based on their pairwise interactions (e.g. Vlasblom & Wodak 2009).

Our overall aim is to construct a group catalogue using the narrow band PAU Survey (PAUS; e.g. Stothert et al. 2018; Eriksen et al.

★ E-mail: peder.norberg@durham.ac.uk

2019). A PAUS group catalogue would probe significantly fainter galaxies than one built using SDSS or GAMA, and would cover a larger area with better completeness in both sampling and redshift than a group catalogue constructed using similar depth surveys such as zCOSMOS (Lilly et al. 2007) or VIPERS (Guzzo et al. 2014). Hence a PAUS group catalogue would provide a better probe of the redshift evolution of haloes as traced by galaxy groups and better sampling of low-mass haloes. The challenge with finding galaxy groups in PAUS lies in the varying accuracy of the PAUS photometric redshifts. MCL is a promising approach as it allows probabilistic pairwise connections (see also Tempel et al. 2018, for another approach), something that could be useful for PAUS where it is more natural to frame pairwise connections as probabilities than as binary links.

Section 2 presents the MCL algorithm and explains its relation to the standard FoF algorithm. Section 3 presents the mock catalogue which is used to test the algorithm. Section 4 summarizes the metrics we use to assess the group finding performance. Section 5 presents the results in real space. We provide our conclusions and future prospects of the MCL algorithm in Section 6. Hereafter we refer to a 'clustering' of galaxies interchangeably with a 'grouping' of galaxies. Throughout we assume a flat Λ cold dark matter cosmology, with parameters $\Omega_{\rm m} = 0.272$, $\sigma_8 = 0.81$, and $h = 0.704$, consistent with those used to create the mocks (as described in Section 3). We refer the reader to Stothert (2018) for additional details regarding the algorithm, the mocks, and some of the additional tests performed (and not reported here).

2 MARKOV CLUSTERING

The MCL algorithm was developed as a fast, scalable approach to graph clustering[1] (Van Dongen 2000). Graph clustering (e.g. Schaeffer 2007) is a solution to the problem of finding clusters of points given their pairwise connection amplitudes. One obvious and instructive example of a graph clustering problem is detecting communities within a social network (Liu et al. 2014). Here users are 'friends' with other users. The entire friendship network can be represented by a (symmetric) binary matrix, which we call the pairwise connection matrix W, with elements w_{ij}. If users i and j are friends, w_{ij} is 1 and is 0 otherwise. A graph clustering algorithm detects communities within this structure. MCL was chosen for two key reasons: (1) in one of its limits it tends to the standard FoF algorithm as explained later; (2) it supports probabilistic pairwise connections rather than just fixed binary links, which is essential for finding galaxy groups with photometric redshifts.

The MCL algorithm has one free parameter, the inflation parameter Γ, which has to be greater than or equal to unity. The algorithm takes the initial pairwise connection matrix, W_0 (specified by its elements w^0_{ij}), as an input and assigns points to clusters following an iterative process, where W_k is the pairwise connection matrix after k steps:

(i) Normalize w^0_{ij} column-wise such that $\sum_j w^0_{ij} = 1$.

(ii) At step k, create W_k by squaring the pairwise connection matrix W_{k-1}, i.e. $W_k = W^2_{k-1}$.

(iii) Raise every element of w^k_{ij} to the power of Γ, i.e. $(w^k_{ij})^\Gamma$.

(iv) Renormalize w^k_{ij} column-wise such that $\sum_j w^k_{ij} = 1$.

(v) Repeat from (ii) until all elements of W_k have converged individually to within a specified tolerance.

(vi) Rearrange the converged cleaned W_k matrix into a block diagonal matrix and read off the groups.

We now explain each step in turn. The initial column-wise normalization in step (i) above – and those that follow in step (iv) – are necessary to ensure that the pairwise connection elements relating to point i can be treated as probabilities. By squaring the pairwise connection matrix W_{k-1} to create a new pairwise connection matrix, W_k, the MCL algorithm approximately simulates a random walk on the graph by using the elements w^k_{ij} as transition probabilities to determine which pairs are more bound than others.[2] Step (iii), raising the elements of w^k_{ij} to the power Γ, is designed to boost the more travelled connections and reduce the less travelled inter-cluster ones. This process of matrix multiplication (here assumed to be squaring), element inflation (to the power of Γ) and column-wise normalization is repeated until a predefined convergence criteria is met by the pairwise connection matrix W_k. The convergence criterion is that the final matrix becomes idempotent, i.e. invariant under expansion and inflation. The exact criterion is expressed in terms of the maximum over all columns of the difference between the maximum value in a column and the sum of all elements squared of that column. Once converged, the matrix W_k is cleaned (by setting to zero all w^k_{ij} elements below a pruning value of 10^{-4}) and then rearranged with row replacement into a block diagonal matrix, with members of each group defined by the matrix blocks.

At face value MCL is an iterative N^2 process as all links between N points need to be defined at each iteration. The larger the value of the inflation parameter, Γ, the more rapidly the pairwise connection tends towards zero during the iterations and the faster the MCL algorithm will split structures into smaller components. A structure that is split by inflation parameter Γ_1 will always be split by any $\Gamma > \Gamma_1$. In principle Γ has no maximum value but there will be a value of Γ above which the catalogue stops splitting, as all clusters become fully connected sub-graphs with equal pairwise connections, i.e. all points in every cluster are connected only to all other points within the same cluster with the same w_{ij} value (and such clusters are not split by MCL). We note that a Γ value of unity will connect any structure that has any path connecting it. In that case MCL tends to converge extremely slowly as no links are ever trimmed from the matrix (see Section 4 for a practical application).

In the astrophysical case we first have a connection criterion that sets the values of w_{ij} between galaxies i and j. This is normally based on a distance criterion between two galaxies, setting w_{ij} to 1 if the galaxies are closer to each other than some specified linking length and 0 otherwise. The standard FoF algorithm connects all points that could be reached via a succession of links between points. This outcome is exactly the same as that for MCL with the inflation parameter Γ set to unity. Therefore the FoF algorithm should be considered as the limit towards which MCL converges when Γ tends to unity, i.e. formally FoF is a subset of MCL. An advantage of MCL over FoF is that, even though MCL like FoF uses all pairwise links, MCL gives higher priority to points that are more connected than those with fewer connections, unlike FoF. By carefully using the inflation parameter, the less well connected points (or less important pairwise links) can be broken up. Only through detailed tests on mocks (see Section 5) can the accuracy of the MCL algorithm be assessed against e.g. FoF.

[1] The MCL code is publicly available at http://micans.org/mcl/.

[2] See e.g. Van Dongen (2000) for a discussion of why this approach produces a similar result to a standard random walk, while strictly speaking it is not a random walk.

3 MOCK CATALOGUE

To test the MCL approach to galaxy group finding we apply it to a realistic real space galaxy mock catalogue. We use real space rather than redshift space to better understand the impact of changing the clustering algorithm. We use a $z = 0$ snapshot of the GALFORM model presented in Gonzalez-Perez et al. (2018), implemented in the 125 h^{-1} Mpc per side MilliGas simulation cube. Note that this simulation has the same cosmology and number of snapshots as the 500 h^{-1} Mpc MR7 simulation (Guo et al. 2013). We use a smaller simulation to speed up the calculations, as deciding between methods of linking galaxies and optimization of free parameters requires running the algorithm many times. The catalogue is limited in the rest-frame r band to $M_r - 5\log h < -20.0$ and contains $\sim$20 000 galaxies, corresponding to a galaxy density of $\sim 10^{-2}$ $(h^{-1}\,\mathrm{Mpc})^{-3}$, comparable to the GAMA survey at $z \sim 0.15$ (Driver et al. 2011; Liske et al. 2015; Baldry et al. 2018). By construction each galaxy belongs to a unique dark matter halo and each halo contains one or more galaxies. See Stothert (2018) for further details of how the mock catalogue was constructed.

4 GOODNESS OF FIT METRICS

We assess the quality of group finding using the measures of purity and completeness. Group purity, P, quantifies the extent to which galaxies in the same group are actually in the same halo (e.g. Manning, Raghavan & Schütze 2008; Wu, Xiong & Chen 2009):

$$P = \frac{1}{\sum_{i=1}^{N_G}\sum_{j=1}^{N_H} n_{ij}} \sum_{i=1}^{N_G} \max{}_j(n_{ij}), \tag{1}$$

where n_{ij} is the number of galaxies in group i *and* halo j, N_G is the total number of groups, and N_H is the total number of haloes. Similarly, we define the halo completeness, C, which quantifies the extent to which galaxies in the same halo are placed in the same galaxy group:

$$C = \frac{1}{\sum_{i=1}^{N_G}\sum_{j=1}^{N_H} n_{ij}} \sum_{j=1}^{N_H} \max{}_i(n_{ij})\,. \tag{2}$$

We also use the associated cumulative measures C ($\geq M$) and P ($\geq M$) defined, respectively, as the completeness of haloes and the purity of groups with multiplicity (i.e. number of members) greater than or equal to M. For the cumulative measures, the multiplicity cut is only applied to the haloes for C ($\geq M$) and groups for P ($\geq M$).

To optimize the parameters of the MCL algorithm a single statistic is desirable. Here we would like a problem agnostic measure to build an 'optimal' group catalogue. Most astrophysical applications invoke combinations of bijective measures of completeness and purity (Gerke et al. 2005; Robotham et al. 2011; Knobel et al. 2012; Jian et al. 2014). Instead we follow Wu et al. (2009) who tested multiple goodness of fit metrics in a statistical context and choose to use the variation of information (VI; Meilă 2003).

VI, also called the shared information distance, quantifies the distance between two clusterings by looking at the amount of information in each that cannot be inferred using the other clustering. A smaller value of VI means a better clustering, so we minimize this metric to determine the best MCL parameters. Using a definition of entropy from statistical physics, VI is formally written as

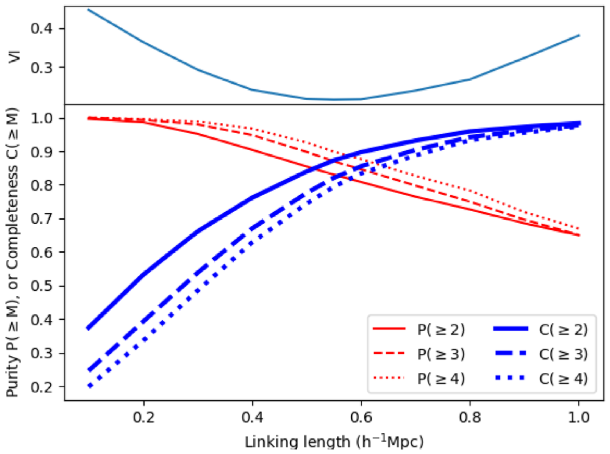

Figure 1. Variation of information (VI; top), group purity, and halo completeness [P ($\geq M$) and C ($\geq M$); main panel] as a function of linking length in a standard FoF approach to galaxy group finding for three values of minimum group/halo multiplicity M. The minimum of VI provides a good compromise between group purity and halo completeness at all multiplicities.

$$\begin{aligned}\mathrm{VI} = &-\sum_{j=1}^{N_H} p_{\Sigma j}\ln(p_{\Sigma j}) - \sum_{i=1}^{N_G} p_{i\Sigma}\ln(p_{i\Sigma})\\ &-2\sum_{i=1}^{N_G}\sum_{j=1}^{N_H} p_{ij}\ln\left(\frac{p_{ij}}{p_{i\Sigma}\,p_{\Sigma j}}\right),\end{aligned} \tag{3}$$

where $p_{xy} = n_{xy}/n_{\Sigma\Sigma}$ for any x or y. This includes the special case of $x = \Sigma$ (or $y = \Sigma$ or $x = y = \Sigma$) for which we define $n_{\Sigma j} = \sum_{i=1}^{N_G} n_{ij}$, $n_{i\Sigma} = \sum_{j=1}^{N_H} n_{ij}$, and $n_{\Sigma\Sigma} = \sum_{i=1}^{N_G}\sum_{j=1}^{N_H} n_{ij}$, corresponding to the number of galaxies in group j, the number of galaxies in halo i, and the total number of galaxies, respectively.

We validate the use of VI by testing how it relates to the more familiar measures of halo completeness and group purity (equations 1 and 2). Fig. 1 shows the VI and three values of P ($\geq M$) and C ($\geq M$) as a function of the assumed fixed linking length for a standard FoF algorithm applied to our mock galaxy catalogue. The minimum value of VI gives a catalogue that is well balanced between completeness and purity. The minimum value of VI also agrees with the value of the linking length relative to the mean galaxy separation found in e.g. Eke et al. (2004). This shows that our choice of optimization statistic is sensible, and that using it in standard FoF produces results comparable to those found in previous work.

5 RESULTS

We compare the results of applying two different clustering methods (MCL and FoF) to the mock galaxy catalogue. In each case the free parameters are found by minimizing VI (equation 3). All models set the binary connection between galaxies i and j, w_{ij}, to unity if the pairwise separation r_{ij} is smaller than the linking length L_{ij}, and 0 otherwise.

In our first groupings we adopt a constant linking length, i.e. L_{ij} is fixed. For FoF this is the only free parameter. Fig. 1 indicates that the optimal value is $L_{ij} = 0.55\,h^{-1}$ Mpc. The optimal solution with MCL using a fixed linking length is achieved, according to VI, when the inflation, Γ, tends to unity, indicating that the FoF algorithm is

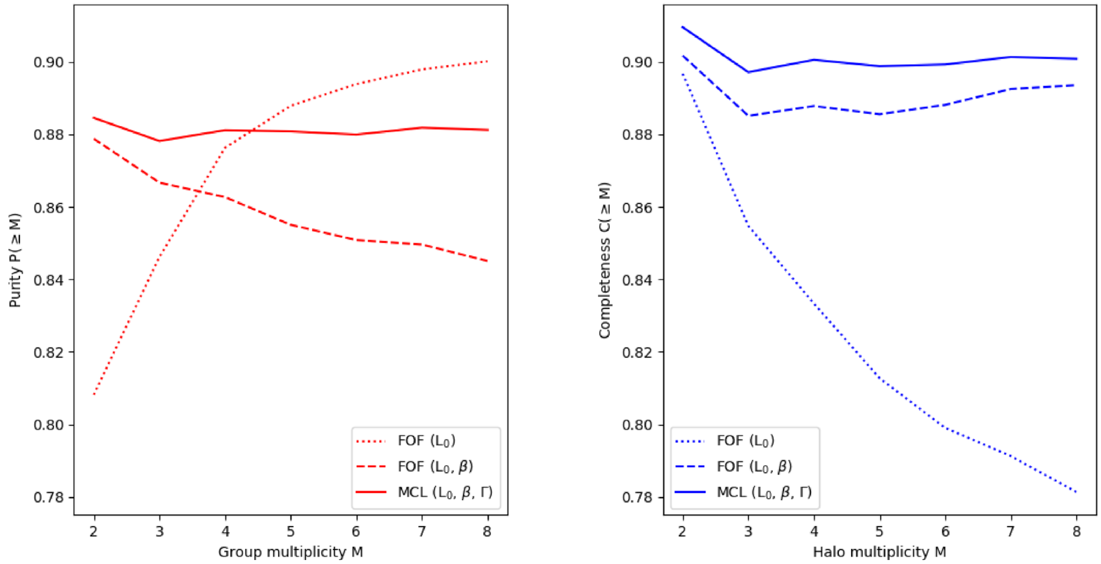

Figure 2. Group purity, P ($\geq M$) (left-hand panel), and halo completeness, C ($\geq M$) (right-hand panel), as a function of minimum multiplicity, M, for different VI minimized galaxy group catalogues: simple FoF (dotted), FoF with density enhancement (dashed), and MCL with density enhancement group catalogues (solid). The purity and completeness of the MCL group catalogue is the most consistent as a function of multiplicity, and has undoubtedly the best halo completeness.

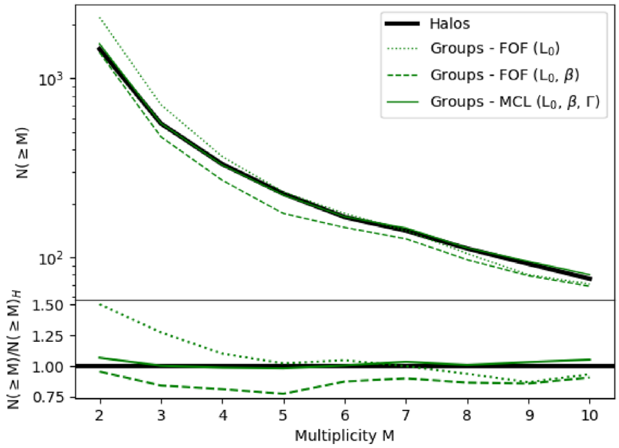

Figure 3. The cumulative multiplicity function, N ($\geq M$), for haloes (truth, thick black line) and three groups catalogues (green lines), which are the simple FoF (dotted), FoF with density enhancement (dashed), and MCL with density enhancement (solid). The bottom panel shows the ratio of the multiplicity functions to the truth, $N(\geq M)_H$. MCL with density enhancement recovers the true halo multiplicity function extremely well (to better than 7 per cent at all multiplicities).

preferred over MCL in this fixed linking length scenario. This is because with a fixed linking length small structures have poor purity and large structures have poor completeness, and increasing Γ only splits the larger structures further, worsening the situation. Hence, hereafter we only show the FoF results for fixed linking length.

The second set of models use a variable linking length set by the geometric mean of the local galaxy density in an attempt to include the known scale dependence of the clustering, as was done in e.g. Eke et al. (2004) and Robotham et al. (2011). We calculate the local density, ρ_i, at the position of galaxy i using a 3D Gaussian kernel with $\sigma = 1\,h^{-1}$ Mpc truncated at 4σ. Other reasonable values of the smoothing scale were tested with no significant improvement found. L_{ij} is now given by

$$L_{ij} = L_0 \left(\frac{\sqrt{\rho_i \rho_j}}{\langle \rho \rangle \, (r_{ij})} \right)^{\beta}. \tag{4}$$

L_0 and β are free parameters and $\langle \rho \rangle \, (r)$ is the mean value of the geometric mean of the pairwise local densities at separation r

$$\langle \rho \rangle \, (r) \equiv \frac{\sum_i \sum_j \sqrt{\rho_i \rho_j}}{\sum_i \sum_j}, \tag{5}$$

where the sums are over all galaxies separated by r. This process extends the linking length for galaxy pairs in overdense region relative to those in underdense ones. A scale-dependent normalization is necessary because, for pairs of galaxies at small separations, the product of their local galaxy densities will on average be larger than that of galaxy pairs at larger separations.

The first density enhanced model connects groups using the FoF algorithm and has two free parameters, β and L_0. The best value of VI is at $\beta\ = 0.6$ and $L_0\ = 0.9\ h^{-1}$ Mpc. From its VI value, this best FoF density enhanced model is preferred over the best model with a constant linking length.

The second density enhanced model uses MCL, so adds the inflation Γ as a third free parameter. The minimum value of VI is now given by $\Gamma = 1.6$, $\beta\ = 0.6$, and $L_0\ = 1.1\ h^{-1}$ Mpc. From its VI value, this optimal MCL density enhanced algorithm produces the best catalogue of the four algorithms considered (FoF and MCL, with and without density enhanced linking lengths).

Fig. 2 shows the group purity P ($\geq M$) and halo completeness C ($\geq M$) as functions of group and halo multiplicity respectively for the optimal catalogue produced by each of the three models. FoF has low purity for small groups and poor completeness for large haloes. FoF with density enhancement performs significantly better, but still tends to overjoin some larger groups, explaining the fall in purity with increasing multiplicity. The density-enhanced MCL algorithm improves on both aspects and produces a group purity and halo completeness that are largely independent of multiplicity. A catalogue with high purity and completeness that are only mildly dependent on multiplicity is preferable. This MCL also produces a catalogue that has higher halo completeness for all multiplicities considered here than the corresponding FoF algorithm with density enhancement. We note that the purity of high multiplicity groups is larger for the simple FoF case, but this is at the expense of a very poor halo completeness.

Fig. 3 shows the cumulative multiplicity function, $N\,(\geq M)$, for the underlying haloes and the three galaxy group catalogues. By including density enhancement, FoF provides a better estimate of the number of small groups, but the number of large groups remains underestimated. MCL with density enhancement impressively recovers the correct numbers of groups at all multiplicities tested here to better than 7 per cent, and often to better than 3 per cent. This is to be compared to the best FoF performance which underestimates the number of haloes by as much as 25 per cent from the truth at $N \geq 5$ and ∼15 per cent for most multiplicities. Note these results were not used to identify the optimal group finder, which is determined by minimizing the VI for each clustering model.

Our results show that MCL can better address the stochastic nature of 'bridges' connecting structure that appear with FoF. FoF needs to be more cautious about the connection criterion as there is a large penalty if even a single link is found between two large structures, whereas MCL reduces this penalty by using inflation to break loosely connected structures. These 'bridges' cause the number of high multiplicity FoF groups to be underestimated (see Fig. 3), and their group purity to be low (see Fig. 2). Both aspects are improved significantly upon using MCL.

6 CONCLUSIONS

For the first time in an astronomical context we apply the MCL algorithm (Van Dongen 2000), which is part of the more general graph clustering algorithms, to identify galaxy groups. MCL has one free parameter, inflation, Γ. We show that the widely used FoF algorithm is a subset of MCL; with $\Gamma = 1$, MCL produces the same result as the deterministic FoF algorithm. We apply MCL to detect galaxy groups in a real space galaxy mock catalogue. We minimize the VI (Meilă 2003) to compare group catalogues to real haloes. We validate this choice by showing that the minimum value of VI for a simple FoF approach is found at linking lengths that are in good agreement with previous values (e.g. Eke et al. 2004).

For a constant linking length FoF produces the best group catalogue. Nevertheless, FoF returns too many spurious small groups and too few large groups: increasing inflation away from unity only makes this discrepancy worse. Using a linking length sensitive to the local density to account for the scale dependence of the grouping, MCL is superior to FoF (i.e. VI is minimized with $\Gamma > 1$). In both cases the group purity and halo completeness are improved over a fixed linking length FoF for all multiplicities. The MCL group catalogue has better halo completeness and group purity than the comparable FoF catalogues, with a completeness and purity that is approximately independent of multiplicity. As a result, MCL provides a better estimate of the number of groups of a given multiplicity than either of the two FoF models considered. In particular, compared to the best FoF approach (as measured by VI), it significantly improves the purity of, and the estimate of the number of, high multiplicity groups. This is most likely because MCL addresses better, through its inflation parameter, the problem of bridges linking large structures together, a common limitation of FoF.

MCL allows pairwise connection amplitudes that are not just ones and zeros, which may prove useful in catalogues with mixed redshift measurement precision, such as those from the PAUS (e.g. Eriksen et al. 2019). Even in real space, where pairwise connections are not probabilistic, MCL produces better group catalogues than FoF. Future work will test MCL on more detailed mock galaxy catalogues in redshift space with photometric errors.

ACKNOWLEDGEMENTS

We thank the referee, Elmo Tempel, for insightful comments. This work was supported by the Science and Technology Facilities Council [ST/J501013/1, ST/L00075X/1, ST/P000541/1]. PN acknowledges the receipt of a Royal Society University Research Fellowship. We acknowledge support from the Royal Society international exchange programme. This work used the DiRAC Data Centric system at Durham University, operated by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility www.dirac.ac.uk. This equipment was funded by BIS National E-infrastructure capital grant ST/K00042X/1, STFC capital grant ST/H008519/1, and STFC DiRAC Operations grant ST/K003267/1, and Durham University. DiRAC is part of the National E-Infrastructure.

REFERENCES

Baldry I. K. et al., 2018, MNRAS, 474, 3875
Colless M. et al., 2001, MNRAS, 328, 1039
Driver S. P. et al., 2011, MNRAS, 413, 971
Eke V. R., Baugh C. M., Cole S., Frenk C. S., King H. M., Peacock J. A., 2005, MNRAS, 362, 1233
Eke V. R. et al., 2004, MNRAS, 348, 866
Eriksen M. et al., 2019, MNRAS, 484, 4200
Gerke B. F. et al., 2005, ApJ, 625, 6
Gonzalez-Perez V. et al., 2018, MNRAS, 474, 4024
Guo Q., White S., Angulo R. E., Henriques B., Lemson G., Boylan-Kolchin M., Thomas P., Short C., 2013, MNRAS, 428, 1351
Guzzo L. et al., 2014, A&A, 566, A108
Huchra J. P., Geller M. J., 1982, ApJ, 257, 423
Jian H.-Y. et al., 2014, ApJ, 788, 109
Knobel C. et al., 2012, ApJ, 753, 121
Lilly S. J. et al., 2007, ApJS, 172, 70
Liske J. et al., 2015, MNRAS, 452, 2087
Liu H. B., Hsieh B. C., Ho P. T. P., Lin L., Yan R., 2008, ApJ, 681, 1046
Liu R., Feng S., Shi R., Guo W., 2014, Procedia Comput. Sci., 31, 85
Manning C. D., Raghavan R., Schütze H., 2008, Introduction to Information Retrieval. Cambridge Univ. Press, Cambridge
Meilă M., 2003, in Schölkopf B., Warmuth M. K., eds, Learning Theory and Kernel Machines. Springer, Springer-Verlag Berlin Heidelberg, p. 173
Robotham A. S. G. et al., 2011, MNRAS, 416, 2640
Schaeffer S. E., 2007, Comput. Sci. Rev., 1, 27
Stothert L. et al., 2018, MNRAS, 481, 4221
Stothert L. J., 2018, PhD thesis, Durham University
Tempel E., Kruuse M., Kipper R., Tuvikene T., Sorce J. G., Stoica R. S., 2018, A&A, 618, A81
van den Bosch F. C., Yang X., Mo H. J., Norberg P., 2005, MNRAS, 356, 1233
Van Dongen S., 2000, PhD thesis, University of Utrecht
Vlasblom J., Wodak S. J., 2009, BMC Bioinformatics, 10, 99
White S. D. M., Rees M. J., 1978, MNRAS, 183, 341
Wu J., Xiong H., Chen J., 2009, in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '09. ACM, New York, NY, USA, p. 877
Yang X., Mo H. J., van den Bosch F. C., Jing Y. P., 2005a, MNRAS, 356, 1293
Yang X., Mo H. J., van den Bosch F. C., Jing Y. P., 2005b, MNRAS, 356, 1293
Yang X., Mo H. J., van den Bosch F. C., Pasquali A., Li C., Barden M., 2007, ApJ, 671, 153
Zeldovich I. B., Einasto J., Shandarin S. F., 1982, Nature, 300, 407

This paper has been typeset from a TEX/LATEX file prepared by the author.