THE CLUSTERING OF ALFALFA GALAXIES: DEPENDENCE ON H I MASS, RELATIONSHIP WITH OPTICAL SAMPLES, AND CLUES OF HOST HALO PROPERTIES

Emmanuil Papastergis1, Riccardo Giovanelli1, Martha P. Haynes1, Aldo Rodríguez-Puebla2, and Michael G. Jones1

1 Center for Radiophysics and Space Research, Space Sciences Building, Cornell University, Ithaca, NY 14853, USA; papastergis@astro.cornell.edu, riccardo@astro.cornell.edu, haynes@astro.cornell.edu, jonesmg@astro.cornell.edu
2 Instituto de Astronomía, Universidad Nacional Autónoma de México, A. P. 70-264, 04510 México, D.F., Mexico; apuebla@astro.unam.mx

Received 2013 June 16; accepted 2013 August 12; published 2013 September 24

ABSTRACT

We use a sample of ≈6000 galaxies detected by the Arecibo Legacy Fast ALFA (ALFALFA) 21 cm survey to measure the clustering properties of H i-selected galaxies. We find no convincing evidence for a dependence of clustering on galactic atomic hydrogen (H i) mass, over the range $M_{\text{HI}} \approx 10^{8.5} - 10^{10.5} M_\odot$. We show that previously reported results of weaker clustering for low H i mass galaxies are probably due to finite-volume effects. In addition, we compare the clustering of ALFALFA galaxies with optically selected samples drawn from the Sloan Digital Sky Survey (SDSS). We find that H i-selected galaxies cluster more weakly than even relatively optically faint galaxies, when no color selection is applied. Conversely, when SDSS galaxies are split based on their color, we find that the correlation function of blue optical galaxies is practically indistinguishable from that of H i-selected galaxies. At the same time, SDSS galaxies with red colors are found to cluster significantly more than H i-selected galaxies, a fact that is evident in both the projected as well as the full two-dimensional correlation function. A cross-correlation analysis further reveals that gas-rich galaxies “avoid” being located within ≈3 Mpc of optical galaxies with red colors. Next, we consider the clustering properties of halo samples selected from the Bolshoi CDM simulation. A comparison with the clustering of ALFALFA galaxies suggests that galactic H i mass is not tightly related to host halo mass and that a sizable fraction of subhalos do not host H i galaxies. Lastly, we find that we can recover fairly well the correlation function of H i galaxies by just excluding halos with low spin parameter. This finding lends support to the hypothesis that halo spin plays a key role in determining the gas content of galaxies.

Key words: galaxies: evolution – galaxies: halos – galaxies: statistics – large-scale structure of universe – radio lines: galaxies – surveys

Online-only material: color figures

1. INTRODUCTION

In the currently accepted hierarchical theory of structure formation, the clustering of galaxies is jointly determined by the large-scale structure of dark matter (DM) in the universe as well as the way in which baryons trace DM through the formation of galaxies (see, e.g., Cooray & Sheth 2002). As a result, the quantitative study of galaxy clustering through the correlation function, $\xi(r)$, has been instrumental both for constraining cosmological models (e.g., Jenkins et al. 1998) as well as for furthering our understanding of galaxy formation and evolution (e.g., Zheng et al. 2007).

Given a cosmological model, galaxy clustering can be used to constrain the galaxy–halo connection, thus testing and informing models of galaxy formation. A large quantity of work has been devoted to studying the correlation function of galaxies as a function of their optical properties, such as luminosity, color, and morphological and spectral type (e.g., Zehavi et al. 2005, 2011; Li et al. 2012b). These studies have established with increasing precision a number of fundamental clustering phenomena, such as the trend for stronger clustering with increasing luminosity and the fact that galaxies with blue colors, late-type morphologies, and elevated star formation activity cluster significantly less than red, early-type, quiescent galaxies. Moreover, several studies have used the halo occupation distribution formalism to make quantitative predictions of the properties of halos hosting a certain class of galaxies (e.g., Zehavi et al. 2005). These analyses have suggested that more luminous galaxies inhabit more massive halos on average and that red galaxies have a higher chance of being hosted by a subhalo compared with blue galaxies. These results are in agreement with theoretical expectations and are supported by a number of other observational methods (e.g., abundance matching: Guo et al. 2010; Behroozi et al. 2010; Moster et al. 2010; Leauthaud et al. 2012, and galaxy–galaxy weak lensing: Mandelbaum et al. 2006; Dutton et al. 2010; Reyes et al. 2012). These clustering-based galaxy occupation models then feed back into cosmological studies, since they provide the necessary link between the measured distribution of galaxies and the distribution of matter that is determined by cosmological parameters (e.g., Reddick et al. 2013).

Until recently, similarly detailed studies of the clustering characteristics of galaxies selected by their atomic hydrogen content (H i-selected) were not feasible, due to the lack of large-volume blind 21 cm surveys. In recent years, however, the H i Parkes All Sky Survey (HIPASS; Meyer et al. 2004) and the Arecibo Legacy Fast ALFA1 (ALFALFA; Giovanelli et al. 2005) survey have provided adequate samples for this purpose. Basilakos et al. (2007) and Meyer et al. (2007) have both analyzed the HIPASS data set, establishing the fact that H i-selected galaxies are among the most weakly clustered galaxy populations known. In addition, both of these works investigated the dependence of clustering strength on galaxy H i mass ($M_{\text{HI}}$), arriving at different conclusions. More recently, Martin et al. (2012) used $\approx 10,000$ galaxies from the 40% ALFALFA catalog ($^{\alpha} \approx 40^\circ$ catalog; Haynes et al. 2011) to

1 The Arecibo L-band Feed Array (ALFA) is a seven-feed receiver operating in the L band ($\approx 1420$ MHz), installed at the Arecibo Observatory.
measure the correlation function of gas-rich galaxies. Among their main findings were that H\textsc{i}-selected galaxies show a markedly anisotropic clustering pattern (their Figure 3; see also Figure 14 in this paper) and that they are anti-biased with respect to DM on scales $\lesssim 5$ Mpc (see also Passmoor et al. 2011).

In this work, we take advantage of the large H\textsc{i} data set provided by ALFALFA to conduct a detailed investigation of the clustering properties of gas-rich galaxies. We furthermore draw samples from the spectroscopic database of the seventh data release of the SDSS (Abazajian et al. 2009) spanning the same volume as the ALFALFA sample, to make comparisons with the clustering properties of optically selected galaxies. The fact that the ALFALFA and SDSS samples are drawn from the same volume allows for a further cross-correlation analysis, measuring the spatial relationship between H\textsc{i} from the same volume allows for a further cross-correlation analysis, measuring the spatial relationship between H\textsc{i} and optical galaxies. Lastly, we select halos from the Bolshoi $\Lambda$CDM simulation (Klypin et al. 2011), to investigate what halo properties are associated with weak clustering, providing evidence of the characteristics of halos hosting gas-rich galaxies.

This paper is organized as follows. In Section 2, we present the ALFALFA and SDSS samples used to measure the clustering properties of gas-rich galaxies. We furthermore present our halo samples selected from the Bolshoi simulation and study their clustering as a function of their properties (mass, spin, etc.). We conclude in Section 5 by summarizing the main findings of this work. We forewarn the reader that—unlike most correlation function studies—all distances in this work assume a Hubble constant of $H_0 = 70$ $h_{70}$ km s$^{-1}$ Mpc$^{-1}$. In order to facilitate comparisons with the literature, however, the upper $x$-axis of the figures is expressed in terms of $h \equiv H_0/100$ km s$^{-1}$ Mpc$^{-1}$, when appropriate.

2. DATA AND METHODS

2.1. ALFALFA Sample

The ALFALFA survey is a wide-area, blind 21 cm emission-line survey performed with the 305 m radio telescope at the Arecibo Observatory (Giovanelli et al. 2005). The survey has recently completed data acquisition and a source catalog covering $\approx 40\%$ of the final survey area has been publicly released—the $\alpha$.40 catalog. ALFALFA has greater sensitivity, finer spectral resolution, and better centroiding accuracy than previous blind H\textsc{i} surveys of comparable sky coverage (e.g., HIPASS; Barnes et al. 2001) and $\alpha$.40 already represents the largest H\textsc{i}-selected galaxy sample to date.

In this article, we use a parent sample of 6123 H\textsc{i}-selected galaxies detected by the ALFALFA survey. In particular, we select galaxies over a contiguous rectangular sky region of $\approx 1700$ deg$^2$ ($135° < R.A. < 230°$ and $0° < decl. < 18°$) in the redshift range $z \approx 0.0023$–0.05 ($v_{\text{CMB}} = 700$–15,000 km s$^{-1}$). The parent sample has significant overlap with the publicly available $\alpha$.40 sample, but has been supplemented by newly processed ALFALFA regions covering the declination ranges $0° < decl. < 4°$ and $16° < decl. < 18°$. The sample is restricted to “Code 1” ALFALFA detections, i.e., it is comprised only of confidently detected extragalactic sources ($S/N_{H\text{\textsc{i}}} > 6.5$). In addition, parent sample sources have a combination of observed 21 cm flux ($S_{H\text{\textsc{i}}}$) and 21 cm line profile width ($W_{50}$) that places them in the region of the ($S_{H\text{\textsc{i}}}$, $W_{50}$) plane where the completeness of the ALFALFA survey is at least 50% (see Section 6 and Figure 12 in Haynes et al. 2011).

Lastly, the sample is limited to line widths $W_{50} > 18$ km s$^{-1}$ and H\textsc{i} masses\footnote{Atomic hydrogen (H\textsc{i}) masses for ALFALFA galaxies are calculated from their 21 cm line flux though the relation $M_{H\text{\textsc{i}}} = 2.356 \times 10^7 S_{H\text{\textsc{i}}} d^2$. In this formula, $M_{H\text{\textsc{i}}}$ is measured in $M_{\odot}$ and the flux $S_{H\text{\textsc{i}}}$ is in Jy km s$^{-1}$. The distance $d$ is measured in Mpc and calculated from the galaxy’s recessional velocity in the cosmic microwave background (CMB) frame as $d = v_{\text{CMB}}/H_0$.} $M_{H\text{\textsc{i}}} > 10^5 M_{\odot}$.

From the parent sample described above, we select a number of subsamples such that their H\textsc{i} masses are above a specified limit (H\textsc{i} mass thresholds) or within a specific range (H\textsc{i} mass bins). Figure 1 shows a histogram of $M_{H\text{\textsc{i}}}$ for the parent sample, with a graphical representation of the H\textsc{i} mass-threshold and the H\textsc{i} mass-binned subsamples used in this work.

2.2. SDSS Sample

We select an optical sample of galaxies from the spectroscopic database of the seventh data release of the Sloan Digital Sky Survey (SDSS DR7; Abazajian et al. 2009). This optically selected parent sample is restricted to the same volume as the H\textsc{i}-selected sample used in this work: $135° < R.A. < 230°$, $0° < decl. < 18°$ and $z \approx 0.0023$–0.05 ($v_{\text{CMB}} = 700$–15,000 km s$^{-1}$). We only select SDSS galaxies that are spectroscopically classified as galaxies (SpecClass = 2) and...
that have an apparent magnitude in the $r$ band brighter than 17.6, after correction for Milky Way (MW) extinction ($m_r < 17.6$).

In addition, we impose a color cut on our spectroscopic sources, $(i - z)_{\text{model}} > -0.25$, which excludes a small number of objects; the vast majority of excluded objects are cases where star-forming knots and structures in nearby extended spirals or dwarf irregular galaxies are erroneously classified as separate galaxies by the SDSS pipeline. Lastly, our SDSS parent sample is limited to MW extinction-corrected absolute magnitudes in the $r$ band brighter than $-17$ ($M_r < -17$) and is comprised of a total of 18,516 galaxies.

From this optical parent sample, we create subsamples, selected based on specifying their faintest $r$-band absolute magnitude (magnitude thresholds) or their range of $r$-band absolute magnitudes (magnitude bins). Figure 2 shows a histogram of $M_r$ for the parent sample, with a graphical representation of the magnitude-threshold and magnitude-binned subsamples used in this work. Furthermore, we define three color-based subsamples according to the position of galaxies in a color–magnitude diagram (CMD), as shown in Figure 3. The “red” subsample is composed of red sequence galaxies, the “blue” subsample is composed of blue cloud galaxies, and the green subsample is composed of galaxies with intermediate locations on the CMD, sometimes referred to as “green valley” galaxies.

2.3. Sample Selection Functions and Random Catalogs

Measuring the clustering of galaxies with certain properties involves comparing the spatial distribution of an observed galactic sample with the spatial distribution of a catalog of random points, which reflects the galactic sample’s selection function (see, e.g., Peebles 1980). The selection function, $\varphi(d)$, describes the fraction of a hypothetical volume-limited sample of galaxies with the desired properties that is included in an observational sample at a distance $d$ from the observer. For example, Figure 4 shows $\varphi(d)$ for the H I mass-threshold samples used in this work; samples restricted to more massive galaxies are complete (i.e., $\varphi = 1$) out to larger distances.

Deriving the selection function for a sample is not straightforward and necessitates two inputs: (1) the cuts used to define an observational sample and (2) the intrinsic distribution of the galaxy properties that determine the inclusion of a galaxy in the observational sample. The SDSS sample described in Section 2.2 is mostly flux-limited, because included galaxies satisfy an apparent $r$-band magnitude cut, $m_r < 17.6$. We thus need to calculate the intrinsic distribution of $r$-band luminosity for SDSS galaxies, most commonly referred to as the galactic luminosity function. Then, $\varphi(d)$ can be calculated in terms of...
the galaxy luminosity function, \( n(M_r) \), as

\[
\varphi(d) = \frac{\int_{M_{\text{r, min}}}^{M_{\text{r, max}}} n(M_r) \, dM_r}{\int_{M_{\text{r, min}}}^{M_{\text{r, max}}} n(M_r) \, dM_r}.
\]  

\( M_{\text{r, max}} \) and \( M_{\text{r, min}} \) are the faint and bright absolute magnitude limits defining a specific subsample, respectively, while \( M_{\text{r, lim}}(d) \) is the faintest absolute magnitude that a galaxy at a distance \( d \) can have and still have an apparent magnitude brighter than \( m_r = 17.6 \). The luminosity function, \( n(M_r) \), is the volume-limited number density of galaxies within a bin of magnitude centered on \( M_r \), and has units of Mpc\(^{-3}\) mag\(^{-1}\). In informal terms, the denominator in Equation (1) represents the volume-limited number density of a specific subsample, while the numerator represents the number density of galaxies in the subsample that are detectable at a distance \( d \).

On the other hand, the ALFALFA sample described in Section 2.1 is not a purely flux-limited sample, but it is mostly defined by a flux-width-dependent cut (Equations (4) and (5) in Section 6 of Haynes et al. 2011). We therefore need to know the intrinsic two-dimensional distribution of H\(_i\) mass and line width, \( n(m_{H_1}, w_{50}) \), of ALFALFA galaxies. The mass-width function is customarily expressed in logarithmic intervals of mass and width, so \( m_{H_1} = \log(M_{H_1}/M_\odot) \) and \( w_{50} = \log(W_{50}/\text{km s}^{-1}) \). The selection function for any ALFALFA subsample is then given by the expression:

\[
\varphi(d) = \frac{\int_{w_{50, \text{min}}}^{w_{50, \text{max}}} \int_{m_{H_1, \text{min}}}^{m_{H_1, \text{max}}} n(m_{H_1}, w_{50}) \, dm_{H_1} \, dw_{50}}{\int_{w_{50, \text{min}}}^{w_{50, \text{max}}} \int_{m_{H_1, \text{min}}}^{m_{H_1, \text{max}}} n(m_{H_1}, w_{50}) \, dm_{H_1} \, dw_{50}}.
\]  

Again, \( m_{H_1, \text{min}}, m_{H_1, \text{max}}, w_{50, \text{min}}, \) and \( w_{50, \text{max}} \) are the H\(_i\) mass and line width limits defining a specific ALFALFA subsample, while \( m_{H_1, \text{lim}}(d, w_{50}) \) is the minimum H\(_i\) mass detectable at a distance \( d \) for a source of line width \( w_{50} \), as dictated by the ALFALFA 50% completeness limit. \( n(m_{H_1}, w_{50}) \) is again the number density of galaxies within a logarithmic bin of H\(_i\) mass centered on \( m_{H_1} \) and a logarithmic bin of width centered on \( w_{50} \), with units of Mpc\(^{-3}\) dex\(^{-2}\).

The \( r\)-band luminosity function for SDSS galaxies and the mass-width function for ALFALFA galaxies used in Equations (1) and (2) are shown in Figure 5. We calculate these quantities by applying appropriate volume correction factors to the sample histograms of \( M_r \) and \( (m_{H_1}, w_{50}) \), respectively. We use the maximum-likelihood, non-parametric “1/V\(_{\text{eff}}\)” method (Zwaan et al. 2005) to calculate the volume weights, on a galaxy-by-galaxy basis. For a more detailed description of the application of the method to the SDSS and ALFALFA data sets; see Section 3.1 in Papastergis et al. (2012) and references therein.

Once a data sample selection function is known, it is straightforward to construct a random catalog suitable for the calculation of the sample correlation function. Initially, random points are created within the subsample volume with a constant expected number density throughout, \( (dN_{\text{rand}}/dV) = \text{const} \). Random points are therefore randomly distributed in R.A., \( \sin(\text{decl}) \), and \( d^2 \). Subsequently, each random point is kept with a probability \( \varphi(d) \) (where \( d \) is the random point’s distance from the observer), in order to reproduce the subsample’s selection function. In the case of an ALFALFA sample, an additional step is necessary: accounting for the effects of radio frequency interference (RFI). RFI disrupts ALFALFA’s performance in the frequency bands where it occurs, resulting in galaxies with certain heliocentric velocities having a lower chance of being detected. Figure 6 shows the fractional ALFALFA volume lost to RFI as a function of heliocentric velocity. In order to reproduce the effects of RFI on the spatial distribution of ALFALFA samples, points in the random catalog are kept with a probability \( f_{\text{RFI}}(v_r) \), where \( v_r \) is the heliocentric velocity of the random point.

Figure 7 compares the distribution of data and random points for the \( M_{H_1} = 10^{9.5} - 10^{10} M_\odot \) ALFALFA sample. Panel (a) displays the coneplot (i.e., a projection of R.A. and \( d \) in polar coordinates) and the sky distribution of the data sample, while panel (b) displays the same distributions for the corresponding random catalog. Panel (c) compares the distance histograms of the two samples. The non-uniform distribution of the data set and the large-scale structure in the survey volume are readily visible in all three panels.

2.4. Clustering Measures

The galaxy correlation function at a given length scale, \( \xi(r) \), is defined as the excess probability of finding a pair of galaxies separated by a distance \( r \) compared with the case of a randomly distributed set of points. It then follows that a positive value of \( \xi(r) \) means that the sample under consideration tends

\footnote{The heliocentric velocity of a random point is calculated by first considering its velocity relative to the CMB, \( v_{\text{CMB}} = H_0 \, d \), and then by converting the velocity from the CMB to the heliocentric frame according to the random point’s position in the sky.}
to cluster on length scales \( r \), while a negative value means that the sample avoids clustering on this scale (a randomly distributed sample would have \( \xi(r) = 0 \) for all \( r \)). In formal terms, the correlation function is defined through the relation \( \langle d^2 N_{\text{pair}} \rangle = \bar{n}_{\text{gal}}^2 (1 + \xi(r)) \, dV_1 \, dV_2 \). Here, \( dV_1 \) and \( dV_2 \) are two volume elements separated by a distance \( r \) and \( \bar{n}_{\text{gal}} \) is the average galaxy number density. \( \langle d^2 N_{\text{pair}} \rangle \) is the average number of galaxy pairs within those volume elements, which would be just \( \bar{n}_{\text{gal}}^2 \, dV_1 \, dV_2 \) in the absence of clustering. In practice, when a galactic sample and its corresponding random catalog are available, the correlation function is calculated in terms of the number of data–data, random–random, and data–random pairs whose separation falls in the bin \( r \pm \Delta r/2 \). These pair counts are denoted by \( P_{\text{DD}}(r) \), \( P_{\text{RR}}(r) \), and \( P_{\text{DR}}(r) \), respectively. If the data sample contains \( N_D \) objects and the random sample contains \( N_R \) objects, we can compute the normalized counts

\[
\text{DD}(r) = P_{\text{DD}}(r)/(N_D(N_D - 1)/2) \\
\text{RR}(r) = P_{\text{RR}}(r)/(N_R(N_R - 1)/2) \\
\text{DR}(r) = P_{\text{DR}}(r)/(N_D N_R),
\]

where, in all three cases, the denominator represents the total number of available pairs.

The most intuitive estimator for the correlation function is then \( \hat{\xi}(r) = \text{DD}(r)/\text{RR}(r) - 1 \), which just computes the ratio of the fraction of data–data pairs and random–random pairs separated by a distance \( r \) and compares this quantity with unity. However, Landy & Szalay (1993) have shown that an alternative estimator, \( \hat{\xi}_{\text{LS}}(r) = (\text{DD}(r) - 2\text{DR}(r) + \text{RR}(r))/\text{RR}(r) \), has better statistical performance; for volume-limited, weakly clustered samples equipped with large random catalogs (\( N_R \gg N_D \)),
the \( E_{LS} \) estimator is unbiased and its variance is determined just by the counting noise associated with the number of data–data pairs. In this paper, we adopt the Landy–Szalay (LS) estimator throughout, dropping from now on the excess notation:
\[
\hat{E}_{LS}(r) \rightarrow \hat{E}(r).
\]

Despite the fact that the "real space" correlation function, \( \hat{E}(r) \), is the fundamental quantity related to galaxy clustering, physical separation is not generally available for extragalactic objects. The measurable quantities in a spectroscopic galaxy survey are position on the sky (R.A., decl.) and recessional velocity (\( v_{\text{CMB}} = c z \)). As a result, we consider in this paper the "redshift space" separation between two objects, \( s \), given by
\[
s = \sqrt{(v_1^2 + v_2^2 - 2v_1v_2 \cos \theta)/H_0},
\]
where \( v_1 \) and \( v_2 \) are the recessional velocities of galaxies 1 and 2, respectively, in km s\(^{-1}\), \( \theta \) is the angle between them on the sky, and \( H_0 \) is the Hubble constant (recall that \( H_0 = 70 \) km s\(^{-1}\) Mpc\(^{-1}\) in this paper). In addition, we can consider separately the components of the separation along the line of sight (\( \pi \)) and on the plane of the sky (\( \sigma \)) defined as
\[
\pi = |v_1 - v_2|/H_0,
\]
\[
\sigma = \sqrt{s^2 - \pi^2}.
\]
We can therefore calculate the redshift space correlation function \( \xi(s) \) by counting the number of pairs whose separation is within \( s \pm \Delta s/2 \). Similarly, we can calculate the two-dimensional correlation function \( \xi(\sigma, \pi) \) by counting pairs separated by \( \sigma \pm \Delta \sigma/2 \) in the tangential plane and \( \pi \pm \Delta \pi/2 \) along the line of sight. Note that in the absence of galaxy peculiar velocities, (1) \( \xi(s) \) would be identical to \( \hat{E}(r) \) and (2) \( \xi(\sigma, \pi) \) would contain no additional information compared with \( \xi(s) \), since galaxy clustering is expected to be intrinsically isotropic. However, due to "redshift-space distortions" (RSDs; Kaiser 1987), the two-dimensional correlation function has a characteristic non-isotropic shape (see Figure 14) and contains non-trivial cosmological information (see, e.g., Reid et al. 2012; Contreras et al. 2013).

Lastly, we can measure the "projected correlation function," which is denoted by \( \Xi(\sigma)/\sigma \) and is defined as
\[
\Xi(\sigma)/\sigma = 2 \int_{\pi = 0}^{\pi_{\text{max}}} \xi(\sigma, \pi) d\pi,
\]
where \( \pi_{\text{max}} = 45 h_{70}^{-1} \) Mpc is used in this work. \( \Xi(\sigma)/\sigma \) is a correlation measure that is integrated over the line-of-sight direction. As a result, it is not affected by RSDs and therefore is the most closely related to \( \hat{E}(r) \). In fact, if the real space correlation function follows a power-law form, parameterized as \( \xi(r) = (r/r_0)^{-\gamma} \), then
\[
\Xi(\sigma)/\sigma = \frac{r_0}{\Gamma(1/2)\Gamma((\gamma - 1)/2)} \Gamma(\gamma/2) \sigma^{-\gamma}.
\]
In other words, a power-law projected correlation function has the same exponent \( \gamma \) as the real space correlation function, while

\[\text{Table 1}
\begin{tabular}{|c|c|c|c|}
\hline
Sample & \( N_{\text{gal}} \) & \( \bar{n} \)/\( h_{70}^3 \) Mpc\(^{-3}\) & \( r_0 \)/\( h_{70} \) Mpc \\
\hline
H1 mass-threshold samples & & & \\
7.5 & 6123 & 0.05562 & 4.74 ± 0.20 & 1.604 ± 0.049 \\
8.0 & 5980 & 0.03561 & 4.85 ± 0.20 & 1.642 ± 0.055 \\
8.5 & 5650 & 0.02147 & 4.75 ± 0.24 & 1.613 ± 0.058 \\
9.0 & 5160 & 0.01123 & 4.53 ± 0.23 & 1.624 ± 0.062 \\
9.5 & 3789 & 0.004135 & 4.77 ± 0.22 & 1.607 ± 0.058 \\
10.0 & 1211 & 0.0007632 & 4.83 ± 0.54 & 1.606 ± 0.129 \\
\hline
H1 mass-binned samples & & & \\
8.5–9.5 & 1861 & 0.01733 & 3.49 ± 0.31 & 2.190 ± 0.266 \\
9.5–10.5 & 2578 & 0.003372 & 5.12 ± 0.36 & 1.554 ± 0.069 \\
10–10.5 & 1181 & 0.0007377 & 5.14 ± 0.49 & 1.598 ± 0.109 \\
\hline
Magnitude-threshold samples & & & \\
−17 & 18,516 & 0.02029 & 6.925 ± 0.046 & 1.772 ± 0.013 \\
−18 & 17,025 & 0.01288 & 6.945 ± 0.045 & 1.775 ± 0.012 \\
−19 & 12,914 & 0.007952 & 7.144 ± 0.061 & 1.781 ± 0.017 \\
−20 & 7026 & 0.004287 & 7.505 ± 0.073 & 1.759 ± 0.019 \\
−21 & 2571 & 0.0015576 & 8.011 ± 0.168 & 1.701 ± 0.039 \\
−22 & 386 & 0.0002320 & 9.311 ± 0.661 & 1.816 ± 0.145 \\
\hline
Optical color samples & & & \\
Blue & 9031 & 0.009896 & 4.767 ± 0.094 & 1.589 ± 0.025 \\
Green & 2622 & 0.002873 & 6.061 ± 0.204 & 1.719 ± 0.057 \\
Red & 6863 & 0.007520 & 9.892 ± 0.104 & 1.982 ± 0.021 \\
\hline
Velocity-binned halo samples (halos and subhalos) & & & \\
60–82 & 53,858 & 0.01847 & 5.142 ± 0.028 & 1.701 ± 0.009 \\
82–114 & 24,769 & 0.008494 & 5.668 ± 0.039 & 1.720 ± 0.013 \\
114–157 & 9190 & 0.003151 & 5.874 ± 0.068 & 1.715 ± 0.021 \\
157–217 & 3634 & 0.001246 & 5.952 ± 0.134 & 1.746 ± 0.038 \\
>217 & 2143 & 0.0007349 & 6.982 ± 0.194 & 1.780 ± 0.055 \\
\hline
Spin parameter halo samples (halos and subhalos) & & & \\
60–82 & 41,459 & 0.01421 & 3.848 ± 0.043 & 1.567 ± 0.011 \\
82–114 & 19,213 & 0.006588 & 4.274 ± 0.066 & 1.615 ± 0.017 \\
114–157 & 7332 & 0.002513 & 4.535 ± 0.110 & 1.605 ± 0.026 \\
157–217 & 3041 & 0.001042 & 5.083 ± 0.229 & 1.668 ± 0.058 \\
>217 & 1912 & 0.0006553 & 6.340 ± 0.257 & 1.708 ± 0.058 \\
\hline
Notes. The fits reported above assume a power-law real space correlation function of the form \( \xi(r) = (r/r_0)^{-\gamma} \). The parameters are derived by fitting the measured projected correlation functions according to Equation (8), over the separations \( \sigma = 0.7–14 h_{70}^{-1} \) Mpc. The parameters and their errors were calculated with the mpfit\footnote{Affected by finite-volume effects.} least-squares fitting procedure, written in IDL. 

\[\text{at the same time its normalization can be used to determine the clustering scale-length parameter, } r_0. \text{ Table 1 contains fitted values of } r_0 \text{ and } \gamma \text{ parameters for all the projected correlation functions appearing in this work.}\]
2.5. Pair-weighting

In order to increase the effective volume probed by the ALFALFA and SDSS samples, we weight each pair roughly inversely to the product of the individual selection function values for the two constituent objects. This weighting aims at taking into account the large number of pairs that remain undetected at large distances. More specifically, each data–data, random–random, and data–random pair is counted toward \( P_{\text{DD}}, P_{\text{RR}}, \) and \( P_{\text{DR}} \) with a weight \( w_{ij} \) given by

\[
 w_{ij} = w_i \times w_j, \quad \text{where} \quad w_i = \frac{1}{1 + 4\pi \bar{n} J_3 \psi(d_i)}. \tag{10}
\]

In the expression above, \( w_i \) and \( w_j \) are the weights of object \( i \) and \( j \), respectively, while \( \psi(d_i) \) is the selection function at the distance of object \( i \). \( \bar{n} \) is the average volume-limited number density of the sample, while \( J_3 \) is a shorthand notation for 

\[
 J_3(s = 30\,\text{Mpc}) = \int_{s=0}^{s=30\,\text{Mpc}} s^2 \xi(s) \, ds.
\]

The results are not sensitive to the exact value of \( J_3 \), so a value of \( J_3 = 2962\,\text{Mpc}^3 \) is used here, corresponding to a fiducial \( \xi_{\text{fid}}(s) = (s/5\,\text{Mpc})^{-1.5} \).

In essence, the weight in Equation (10) reduces to \( w_{ij} \propto (1/\psi_i) \times (1/\psi_j) \) when the selection function is relatively large \( (\psi(d) \gg 1/(4\pi \bar{n} J_3)) \), while \( w_{ij} \approx 1 \) when the selection function is small \( (\psi(d) \ll 1/(4\pi \bar{n} J_3)) \).

In the case of SDSS data–data pair counts, an additional weight is applied to correct for SDSS “fiber collisions.” SDSS spectroscopic fibers cannot generally be placed closer than 55″, with the value of some correlation measure in a separation bin \( i \), while the superscript \( (b) \) denotes the \( b \)th bootstrap sample realization. Note that, as is usual practice, we use in this paper random catalogs with more objects than our data samples \( (N_R > N_D) \). This methodology ensures that the contribution of the counting noise of random–random pairs to the overall error budget is subdominant to the error from data–data and data–random pairs. Specifically, each random catalog is 10 times the length of the corresponding data catalog (up to a maximum of 100,000) and 25 bootstrap realizations are used to estimate the mean, variance, and covariance. We note here that the number of bootstrap realizations is highly inadequate to compute accurate covariance matrices for the two-dimensional correlation functions presented in this work. As a result, we do not attempt to perform a rigorous quantitative analysis of these two-dimensional clustering measurements; we rather present them in order to offer better intuitive understanding of the results concerning the one-dimensional correlation functions.

Here, \( \xi_i \) denotes generically the value of some correlation measure in a separation bin \( i \), while the superscript \( (b) \) denotes the \( b \)th bootstrap sample realization. Note that, as is usual practice, we use in this paper random catalogs with more objects than our data samples \( (N_R > N_D) \). This methodology ensures that the contribution of the counting noise of random–random pairs to the overall error budget is subdominant to the error from data–data and data–random pairs. Specifically, each random catalog is 10 times the length of the corresponding data catalog (up to a maximum of 100,000) and 25 bootstrap realizations are used to estimate the mean, variance, and covariance. We note here that the number of bootstrap realizations is highly inadequate to compute accurate covariance matrices for the two-dimensional correlation functions presented in this work. As a result, we do not attempt to perform a rigorous quantitative analysis of these two-dimensional clustering measurements; we rather present them in order to offer better intuitive understanding of the results concerning the one-dimensional correlation functions. Finally, to ease the computational workload, the normalized random–random counts are only computed once, while the data–data and data–random pairs are computed for each realization.

2.6. Error Estimation

We calculate errors on our clustering measurements by bootstrapping our data sample. If a data sample has \( N_D \) elements, bootstrap resampling involves forming sample realizations by randomly extracting \( N_R \) elements from the original data set, with replacement. If \( k = 1, \ldots, K \) sample realizations are produced in this way, we can calculate statistical properties of the measured correlations such as the average, variance, and covariance matrix:

\[
\langle \xi_i \rangle = \frac{1}{K} \sum_{k=1}^{K} \xi_i^{(k)} \tag{12}
\]

\[
\sigma_{\xi_i}^2 = \frac{1}{K-1} \sum_{k=1}^{K} (\xi_i^{(k)} - \langle \xi_i \rangle)^2 \tag{13}
\]

\[
\text{Cov}(\xi_i, \xi_j) = \frac{1}{K-1} \sum_{k=1}^{K} (\xi_i^{(k)} - \langle \xi_i \rangle)(\xi_j^{(k)} - \langle \xi_j \rangle). \tag{14}
\]

3. RESULTS

3.1. Dependence of Clustering on \( H_1 \) Mass

Figure 8 shows the measured projected correlation function, \( \Xi(\sigma)/\sigma \), for the \( H_1 \) mass-threshold samples shown in the upper panel of Figure 1 and described in Section 2.1. Darker shades of blue represent samples with a lower \( H_1 \) mass threshold, as listed in the figure legend. Note that some error bars for the \( M_{200} > 10^{10} M_\odot \) sample extend below the legend. Bottom panel: the same projected correlation functions as above, normalized to the correlation function of the \( M_{200} > 10^{10} M_\odot \) sample. The unity line is also plotted for reference. (A color version of this figure is available in the online journal.)
functions seem to deviate from a simple power-law form at separations larger than this characteristic value. This behavior has been noted in multiple studies and seems to hold for both optically selected and H\textsc{i}-selected samples (e.g., Li et al. 2012b; Norberg et al. 2002; Zehavi et al. 2011 for optical samples and Martin et al. 2012; Basilakos et al. 2007 for H\textsc{i} samples).

Most importantly, however, the correlation functions of the H\textsc{i} mass-threshold samples show no significant differences among one another, within the errors of the present analysis. Note that Figure 8 shows no evidence for enhanced clustering for the samples with the highest H\textsc{i} masses; this result is in stark contrast to the strong clustering displayed by galaxies with high stellar mass or optical luminosity (e.g., Zehavi et al. 2011; Beutler et al. 2013 to name a few; see also Figure 11 in this work). Figure 8 is not ideal, however, for assessing the clustering properties of low H\textsc{i} mass galaxies, as the upper panel of Figure 1 shows that even the samples with the lowest H\textsc{i} mass thresholds (as low as $M_{\text{H}\textsc{i}} = 10^{7.5} \, M_\odot$) are still dominated by fairly H\textsc{i} massive galaxies. We therefore display in Figure 9 the $\xi(\sigma)/\sigma$ measurements for the three non-overlapping H\textsc{i} mass binned samples shown in the lower panel of Figure 1. Galaxies with intermediate and high masses ($M_{\text{H}\textsc{i}} = 10^{9.5} - 10^{10} \, M_\odot$ and $M_{\text{H}\textsc{i}} = 10^{10} - 10^{10.5} \, M_\odot$ bins) show no significant differences in their clustering properties in Figure 9, in accordance with the results in Figure 8. Interestingly enough, however, galaxies with low H\textsc{i} mass ($M_{\text{H}\textsc{i}} = 10^{8.5} - 10^{9.5} \, M_\odot$ bin) seem to be much more weakly clustered than their more massive counterparts.

The H\textsc{i} mass dependence of the clustering properties of H\textsc{i}-selected galaxies has remained a controversial issue in the literature. For instance, Basilakos et al. (2007) and Meyer et al. (2007) have both analyzed data sets from HIPASS, but came to different conclusions regarding the issue. On one hand, Basilakos et al. (2007) found that HIPASS galaxies with $M_{\text{H}\textsc{i}} < 10^{9.4} \, M_\odot$ have a significantly lower clustering amplitude than galaxies with H\textsc{i} masses larger than this value (see their Figure 5). On the other hand, Meyer et al. (2007) dissected the HIPASS sample at a similar H\textsc{i} mass, $M_{\text{H}\textsc{i}} = 10^{9.25} \, M_\odot$, but found no convincing differences in the correlation function of the low-mass and high-mass subsamples (see their Figure 12). At face value, the ALFALFA measurements shown in Figure 9 seem to lend support to the Basilakos et al. (2007) claim. One complication arises, however, due to the fact that the volume probed by the $M_{\text{H}\textsc{i}} = 10^{9.5} - 10^{10} \, M_\odot$ sample is $\approx 6$ times smaller than the volume probed by the $M_{\text{H}\textsc{i}} = 10^{8.5} - 10^{9} \, M_\odot$ sample. As a result, the observed discrepancy could be caused by finite volume effects. We therefore re-calculate the projected correlation function of the $M_{\text{H}\textsc{i}} = 10^{8.5} - 10^{9} \, M_\odot$ sample, but restricting it to the smaller volume available to the $M_{\text{H}\textsc{i}} = 10^{8.5} - 10^{9} \, M_\odot$ sample. Figure 10 shows the result: even though the correlation functions of the two samples are very different from one another when both are calculated over their full volumes (dark blue and blue solid lines), they show no significant differences when the two samples are restricted to a common volume (dark blue solid line and blue dash-dotted line).

Overall, we find no conclusive evidence for a dependence of the clustering properties of H\textsc{i}-selected galaxies on their H\textsc{i} mass, over the mass range $M_{\text{H}\textsc{i}} \approx 10^{8.5} - 10^{10.5} \, M_\odot$. Despite the fact that Figure 9 displays a weak correlation function for low H\textsc{i} mass galaxies ($M_{\text{H}\textsc{i}} = 10^{8.5} - 10^{9.5} \, M_\odot$), Figure 10 suggests that this behavior could be entirely due to finite volume effects. An extension of this work to both higher and lower masses will necessitate the next generation of H\textsc{i} surveys, such as the planned WALLABY survey with the ASKAP array (Koribalski 2012) and the H\textsc{i} surveys to be performed with the APERTIF instrument on the Westerbork Synthesis Radio Telescope interferometer, which will probe a much larger volume than ALFALFA and are expected to detect $\approx 10 \times$ more sources. These surveys will also provide more accurate measurements of the correlation function over the H\textsc{i} mass.
range probed in this work, potentially uncovering trends that cannot be detected with the current precision.

3.2. Bias Relative to Optical Galaxies

Several studies have found that H\textsc{i}-rich galaxies are among the most weakly clustered galactic populations known (e.g., Basilakos et al. 2007; Meyer et al. 2007; Martin et al. 2012). This fact can be clearly seen in Figure 11, which compares the projected correlation function of one representative ALFALFA sample with the luminosity-threshold SDSS samples (as depicted in the upper panel of Figure 2 and described in Section 2.2). The figure shows that a “typical” H\textsc{i}-selected sample is significantly less clustered than a “typical” optically selected sample, regardless of the optical sample’s limiting luminosity. Furthermore, we can compare the correlation function of the same ALFALFA sample with the correlation function of the SDSS luminosity-binned samples (as shown in the lower panel of Figure 2). The result is shown in Figure 12, and we note the following points. First, the optical samples display a clear trend of stronger clustering with increasing luminosity, unlike H\textsc{i}-selected samples. When no color cuts are applied to the latter sample. This statement is valid even for relatively faint galaxies (at least as faint as $M_r \approx -18$). In addition, optically selected samples of all luminosities display slightly steeper correlation functions compared with H\textsc{i}-selected samples.

2. The correlation function of H\textsc{i}-selected galaxies is practically indistinguishable from the correlation function of optical galaxies with blue colors. The relative bias of the two samples is $b_{\text{rel}} \approx 1$, over almost the whole range of separations probed.

3. Red galaxies show much stronger clustering than H\textsc{i}-selected galaxies, with the relative bias reaching values $b_{\text{rel}} > 3$ at small separations ($\sigma \lesssim 1$ Mpc). Moreover, they seem to display a steeper correlation function regardless of luminosity.\footnote{Note that the $M_r = -18$ to $-17$ sample has not been taken into account when making these statements. The correlation function of this sample is probably affected significantly by finite-volume effects, similarly to the case of the $M_{\text{HI}} = 10^8.5 - 10^9$ M$_\odot$ ALFALFA sample (see Figure 10).}

Figure 13, on the other hand, compares the clustering of the same $M_{\text{HI}} > 10^9$ M$_\odot$ ALFALFA sample with the clustering of three optical subsamples split by color (see Figure 3 and Section 2.2). By combining the information in Figures 11–13, we arrive at the following conclusions.

1. H\textsc{i}-selected galaxies cluster less than optically selected galaxies, when no color cuts are applied to the latter sample. This statement is valid even for relatively faint galaxies (at least as faint as $M_r \approx -18$). In addition, optically selected samples of all luminosities display slightly steeper correlation functions compared with H\textsc{i}-selected samples.

2. The correlation function of H\textsc{i}-selected galaxies is practically indistinguishable from the correlation function of optical galaxies with blue colors. The relative bias of the two samples is $b_{\text{rel}} \approx 1$, over almost the whole range of separations probed.

3. Red galaxies show much stronger clustering than H\textsc{i}-selected galaxies, with the relative bias reaching values $b_{\text{rel}} > 3$ at small separations ($\sigma \lesssim 1$ Mpc). Moreover,
the projected correlation function of red optical galaxies is significantly steeper than that of HI-selected galaxies.

Points 1–3 above hold also for the full two-dimensional correlation function, $\xi(\sigma, \pi)$: Figure 14 shows $\xi(\sigma, \pi)$ for the parent ALFALFA sample ($M_{HI} > 10^{7.5} M_\odot$), which can be compared with $\xi(\sigma, \pi)$ for the blue and red SDSS subsamples (Figures 15 and 16, respectively). Note that common contour levels are used in Figures 14–16. The two-dimensional correlation functions for the ALFALFA and blue SDSS galaxies are very similar in amplitude and shape. In particular, both samples display a characteristic “flattening” of $\xi(\sigma, \pi)$ along the $\pi$-axis on intermediate scales ($\pi \gtrsim 10$ Mpc), as well as a weak “finger-of-god” effect (i.e., the elongated structure along the $\pi$-axis at $\sigma \approx 0$ Mpc). In contrast, the red SDSS subsample shows a $\xi(\sigma, \pi)$ with much larger overall amplitude, as well as a very distinct finger-of-god feature. In addition, the $\xi(\sigma, \pi)$ contours for the red SDSS sample display a more symmetric, “round,” shape on intermediate scales ($\gtrsim 10$ Mpc). We remind the reader that a rigorous, quantitative analysis of the two-dimensional correlation functions in Figures 14–16 is beyond the scope of this paper; the two-dimensional correlation functions are presented here in order to provide some intuitive understanding of the results observed in the one-dimensional correlation functions.

These results are not unexpected; it is well established that gas-rich galaxies are associated with late-type morphology, blue optical colors, and elevated specific star formation rates (e.g., Huang et al. 2012; Catinella et al. 2010; Li et al. 2012a). For example, Huang et al. (2012) show that the ALFALFA sample is heavily biased against red-sequence galaxies, while it samples very well the less luminous and more actively star-forming galaxies in the “blue cloud” (their Figure 11). The main conclusions summarized in points 1–3, therefore, are a direct consequence of the fact that blue galaxies are significantly less clustered than red galaxies, irrespective of luminosity (see, e.g., Figure 16 in Zehavi et al. 2011). The bias of blind HI surveys against red-sequence galaxies also helps explain the marked difference in the shape of $\xi(\sigma, \pi)$ between the ALFALFA and SDSS red samples (Figures 14 and 16, respectively). Red galaxies are usually found in high-density environments, such as clusters of galaxies and compact groups, and their clustering bears the signs of large and incoherent peculiar motions that are characteristic of these environments. In particular, the red sample has an increased number of galaxy pairs that have small physical but large velocity separations; these pairs produce the strong finger-of-god feature in $\xi(\sigma, \pi)$ at $\sigma \approx 0$. On the other
hand, galaxies with blue colors and H\textsc{i} galaxies tend to inhabit the lower density "field." As a result, they trace the ordered flow toward matter overdensities without significant noise from peculiar motions. This fact is why the characteristic asymmetric shape of $\xi(\sigma, \pi)$ at separations $\gtrsim 10$ Mpc, which is caused by these systematic motions, is more pronounced in the blue and H\textsc{i} samples.

### 3.3. Cross-correlation between H\textsc{i}-selected and Optically Selected Samples

The results above can be used to compare the clustering properties of H\textsc{i} and optical galaxies, but these results do not contain information about the spatial relationship among the samples under consideration. In particular, they cannot address questions such as whether or not H\textsc{i} galaxies inhabit the same environments as a given class of optical galaxies. It is already known based on studies of individual clusters (Giovanelli & Haynes 1985; Haynes & Giovanelli 1986; Solanes et al. 2002) that galaxies in high-density environments tend to have lower gas fractions than their counterparts in the field, and thus have a lower probability of being included in an H\textsc{i}-selected sample. Since optical galaxies with red colors are found preferentially in dense environments, we expect H\textsc{i}-selected samples to show some degree of "segregation" with respect to red galaxies. Here we use the large galaxy samples provided by the ALFALFA and SDSS surveys to obtain a statistical measurement of this effect and to determine the length scale over which environment can affect the gaseous contents of galaxies.

The spatial relationship of two galactic samples is encoded in their cross-correlation function. In this work, we calculate cross-correlation functions using a modified version of the LS estimator (following Zehavi et al. 2011):

$$\hat{\xi}_{\text{cross}} = (DD_1DD_2 - DD_1RR_2 - DD_2RR_1 + RR_1RR_2)/RR_1RR_2.$$  \hspace{1cm} (15)

Here, DD, RR, and DR are the normalized data–data, random–random, and data–random pair counts, and the subscripts 1 and 2 are used to denote the two samples. Generally, the information present in the cross-correlation function is most intuitively presented in terms of its ratio with the geometric mean of the correlation functions of the two constituent samples, $R(r) = \xi_{\text{cross}}(r)/\sqrt{\xi_1(r)\xi_2(r)}$. In essence, $R$ measures the degree to which two samples are spatially "aware" of one another: Two spatially independent samples have $R = 1$ for all $r$ (i.e., $\xi_{\text{cross}}(r) = \sqrt{\xi_1(r)\xi_2(r)}$). Conversely, a ratio of $R < 1$ at some separation $r$ means that the two samples "avoid" each other on the length scale under consideration.

Figure 17 shows the cross-correlation function between the ALFALFA parent sample ($M_{H_1} > 10^{10} M_\odot$) and the blue SDSS sample. The lower panel shows that $R \approx 1$ on all probed scales, meaning that H\textsc{i}-selected galaxies and optical galaxies with blue colors have no special spatial relationship. In other words, detecting a blue SDSS galaxy at a given location in space does not influence our chance of finding an ALFALFA-detected galaxy in its vicinity, beyond what is expected from the clustering of the samples. The situation is very different in Figure 18, which shows that the cross-correlation function between H\textsc{i}-selected galaxies and optical galaxies with red colors is systematically lower than their geometric mean at small separations (i.e., $R < 1$ at $\sigma \lesssim 3$ Mpc). This fact means that the existence of a red SDSS galaxy at a given position in space lowers the chances that an H\textsc{i}-rich galaxy is positioned within $\approx 3$ Mpc of it.

These results also hold for the two-dimensional cross-correlation functions between the H\textsc{i}-selected ALFALFA sample and the color-based SDSS samples. Figure 19, for example, shows a two-dimensional map of $R(\sigma, \pi)$ calculated from the cross-correlation function between the ALFALFA parent sample and the SDSS blue sample. Regions that are enclosed by solid contours are those where $R$ deviates significantly from unity ($R < 0.85$ or $R > 1.15$). We can clearly see that, barring the large fluctuations at the outskirts of the map caused by noise,
Figure 18. Top panel: the projected correlation function, $\Xi(\sigma)/\sigma$, for the red SDSS sample (red solid line) and the parent ALFALFA sample (light blue solid line), compared with their projected cross-correlation function (orange solid line). Bottom panel: the ratio of the cross-correlation function to the geometric mean of the correlation functions of the two constituent samples, $R(\sigma)$. Note the clear tendency for SDSS galaxies with red colors and H$_i$-rich ALFALFA galaxies to avoid each other at separations $\lesssim 3$ Mpc. (A color version of this figure is available in the online journal.)

$R(\sigma, \pi) \approx 1$ over most of the map. The situation is very different when the cross-correlation between the ALFALFA parent sample and the SDSS red sample is considered. Figure 20 shows that regions corresponding to $\sigma \lesssim 3$ Mpc have systematically low values of $R$, over the whole range of $\pi$-axis separations. This characteristic shape demonstrates graphically that H$_i$-selected galaxies avoid regions of space where the finger-of-god effect is large, corresponding mostly to galaxy clusters and rich groups. Once again, we note that in this paper we do not aim to perform a quantitative analysis of the two-dimensional distributions in Figures 19 and 20. We present these distributions in order to offer some intuition regarding the results of the one-dimensional cross-correlation analysis.

In general, measurements of the cross-correlation properties of H$_i$ galaxies with respect to various optical samples are especially important in the context of cosmological studies with next generation H$_i$ surveys (e.g., Beutler et al. 2011) and 21 cm intensity-mapping experiments at moderate redshift (e.g., Masui et al. 2013). In particular, Figures 11–20 show that an H$_i$-selected sample traces the cosmic large-scale structure differently than most optical surveys. For example, due to the very different clustering properties of H$_i$-rich galaxies and galaxies with red colors, a 21 cm survey would provide a very different view of the large-scale structure compared with a survey of, e.g., luminous red galaxies (as in, e.g., Eisenstein et al. 2005). On the other hand, a survey targeting actively star-forming galaxies (such as the UV-selected WiggleZ survey; Drinkwater et al. 2010) will be a much closer match in terms of clustering properties and spatial distribution. The considerations above have an effect on the potential of future 21 cm surveys for cosmological studies: for example, the low clustering amplitude of H$_i$ galaxies may lower our sensitivity for detecting the baryon acoustic oscillation feature in the large-scale galaxy correlation function. On the other hand, measurements that are based on the anisotropy of $\xi(\sigma, \pi)$ may benefit considerably from the low levels of peculiar motion “noise” achieved with an H$_i$-selected sample. In addition, the low bias of H$_i$ samples leads to smaller deviations from linearity compared with other highly biased
samples, potentially aiding the cosmological interpretation of clustering measurements. As a result, HI surveys may prove advantageous in measuring RSDs and the growth of structure (“fσ8” measurements, e.g., Reid et al. 2012; de la Torre et al. 2013; Contreras et al. 2013).

4. WHICH HALOS HOST GAS-RICH GALAXIES?

4.1. ΛCDM Halo Sample

We select a sample of DM halos from the Bolshoi ΛCDM simulation (Klypin et al. 2011). The Bolshoi simulation is a high-resolution dissipationless simulation, run for a set of cosmological parameters consistent with the seven year results of the Wilkinson Microwave Anisotropy Probe (Jarosik et al. 2011) and other recent cosmological studies. We use the halo catalogs9 extracted with the Bound-Density-Maxima (BDM) halo finder algorithm (Klypin & Holtzman 1997). A detailed description of the BDM halo finding method can be found in Appendix A of Klypin et al. (2011). Here, we briefly summarize the method’s main characteristics: the algorithm identifies density peaks in the survey volume and associates neighboring particles into halos. The virial radius of each halo is defined as the radius enclosing an average overdensity of about 360 times the cosmic matter density (at z = 0). Unbound particles are removed iteratively, and a number of properties are then recorded for each halo (e.g., virial mass, maximum circular velocity, angular momentum, etc.). In the case of subhalos (i.e., density maxima lying within the virial radius of a larger halo), the extent is confined to particles that are bound to the substructure. The Bolshoi simulation is one of the highest resolution cosmological simulations available today and is complete down to a maximum circular velocity of 50 km s$^{-1}$ (or $M_{\text{vir}} \approx 2 \times 10^{10}$ $h^{-1} M_{\odot}$).

In particular, we select a box region of the Bolshoi simulation of size $\approx 140$ $h^{-1}$ Mpc on a side, such that the volume of the halo sample is comparable to the ALFALFA volume. In addition, we restrict ourselves to halos with maximum circular velocities $v_{\text{halo}} > 60$ km s$^{-1}$, in order to exclude halos that are close to the resolution limit of the simulation. In total, our halo sample consists of 94,671 halos, including both distinct halos as well as subhalos. From this parent halo sample, we create subsamples by specifying five $v_{\text{halo}}$ ranges, as shown in Figure 21: 60–82 km s$^{-1}$, 82–114 km s$^{-1}$, 114–157 km s$^{-1}$, 157–217 km s$^{-1}$, and >217 km s$^{-1}$.

Furthermore, we consider halo subsamples split by their spin parameter, $\lambda_K$, defined as

$$\lambda_K = \frac{J K^{1/2}}{G M_{\text{vir}}^{5/2}}. \quad (16)$$

$J$ is the total angular momentum of the particles within the virial radius of the halo, $K$ is the total kinetic energy of these particles (not including the energy associated with the bulk motion of the halo through space), and $M_{\text{vir}}$ is the halo virial mass. Note that the BDM database for the Bolshoi simulation reports spin parameters defined in terms of the halo kinetic energy ($K$), instead of the more common definition based on total energy ($\lambda = J |E_{\text{tot}}|^{1/2} / GM_{\text{vir}}^{5/2}$). However, the two definitions yield very similar results for well-virialized halos, since in this case $K \approx |E_{\text{tot}}|$. Figure 22 displays graphically the three spin-based subsamples, referred to as the “low spin,” “average spin,” and “high spin” samples, respectively.

The halo samples described above are volume-limited, meaning that inclusion in some specific sample does not depend on their position in the simulation box. As a result, all halo samples share the same random catalog, which is straightforwardly created by a set of points with uniformly distributed $x, y, z$ coordinates. In addition, no pair-weighting (see Section 2.5) is necessary, since all halo pairs in the simulation can be accounted for. Lastly, the separations between pairs of halos are readily available in terms of physical length, and are not affected by RSDs. It follows that for these halo samples $\xi(r) \equiv \xi(s)$, while $\Xi(\sigma)/\sigma$ can be calculated by projecting separations on any arbitrary axis.

---

9 www.multidark.org/MultiDark/Help?page=databases/bolshoi/database
Abundance matching is a simple yet powerful statistical approach to connect galaxy properties (such as the luminosity, stellar or baryonic mass, velocity, etc.) to host DM (sub)halos (e.g., Shankar et al. 2006; Conroy et al. 2006; Guo et al. 2010; Reddick et al. 2013; Rodríguez-Puebla et al. 2012; Papastergis et al. 2012). In its most simple form, the observed abundances of galaxies at a given property are matched against the theoretical halo plus subhalo abundances. The result is a galaxy property versus halo mass empirical relation. In reality, galaxy properties are not determined solely by the mass of the halo in which they reside but, due the complexity of the galaxy formation process, a dependence on other halo and/or environmental properties is expected. To take this fact into account, recent works have extended the abundance matching technique to include a scatter around the mean relation (e.g., Behroozi et al. 2010; Moster et al. 2010; Hearin et al. 2013; Rodríguez-Puebla et al. 2013).

Here, we use the abundance matching technique in order to obtain an average relation between galaxy H\(_i\) mass and host (sub)halo mass (\(M_{\text{H}i}=M_h\) relation). To do so, we employ the observed H\(_i\) mass function of the H\(_i\)-selected sample of Papastergis et al. (2012; see their Table 1) and the halo plus subhalo mass function obtained from the Bolshoi simulation (see Section 4.1). Reddick et al. (2013) have shown that the measure of halo mass that is most tightly related to the stellar properties of galaxies (e.g., stellar mass or optical luminosity) is the maximum mass reached along the entire merger history of the (sub)halo, \(M_{\text{peak}}\); we therefore perform our abundance matching analysis using \(M_{\text{peak}}\) as the halo mass, dropping from now on the excess notation (\(M_{\text{peak}}\)). Note that \(M_{\text{peak}}\) is approximately equal to the present-day mass for distinct halos, but in the case of subhalos it can be significantly larger than \(M_h\). We furthermore assume that the distribution of H\(_i\) mass at a given (sub)halo mass is drawn from a lognormal distribution with a mean \(M_{\text{H}i}(M_h)\) and a scatter of \(\sigma_{\text{H}i}\). Here, we will assume that \(\sigma_{\text{H}i}\) is independent of halo mass. While the scatter around the

\[
\begin{align*}
\text{Figure 23.} & \quad \text{Top panel: the projected correlation function for the velocity-binned Bolshoi halo samples (shown in Figure 21 and described in Section 4.1), compared with the correlation function of the } M_{\text{H}i} > 10^9 M_\odot \text{ ALFALFA sample. Both halos and subhalos are included in the computation of the halo } \Xi(\sigma) \text{ distribution, as listed in the figure legend. Bottom panel: the same projected correlation functions as above, normalized to the correlation function of the ALFALFA sample. The unity line is also plotted for reference.} \\
\text{Figure 24.} & \quad \text{Same as in Figure 23, but for distinct halos only. Note that some error bars for the } v_{\text{halo}} > 217 \text{ km s}^{-1} \text{ sample extend below the legend. (A color version of this figure is available in the online journal.)}
\end{align*}
\]
average stellar mass–halo mass relation (the $M_\ast(M_h)$ relation) has been discussed extensively in the literature (e.g., Cacciato et al. 2009; More et al. 2009, 2011; Yang et al. 2009; Leauthaud et al. 2012; Rodríguez-Puebla et al. 2013), $\sigma_{H_\text{i}}$ has not been discussed previously. Here we opt to use two different values of $\sigma_{H_\text{i}}$ in order to gauge the uncertainty introduced from our lack of knowledge on its value and mass dependence: $\sigma_{H_\text{i}} = 0.1$ dex and $\sigma_{H_\text{i}} = 0.4$ dex. A more thorough exploration of this scatter is deferred to a future publication.

The left-hand panel of Figure 25 shows the resulting average $M_{H_\text{i}}(M_h)$ relations for both values of $\sigma_{H_\text{i}}$. Note that, while for the $\sigma_{H_\text{i}} = 0.1$ dex case the $M_{H_\text{i}}(M_h)$ relation is clearly monotonic at all masses, in the $\sigma_{H_\text{i}} = 0.4$ dex case $M_{H_\text{i}}$ is nearly independent of halo mass (except perhaps for $M_h < 10^9 M_\odot$). According to the abundance matching result, galaxies with $M_{H_\text{i}} > 10^9 M_\odot$ are hosted by halos with $M_h > 10^{11.3} M_\odot$ or $v_{\text{halo}} > 100 \text{ km s}^{-1}$, in reasonable agreement with our claims based on Figures 23 and 24. The right-hand panel of Figure 25 shows the average $M_{H_\text{i}}/M_h$ ratio as a function of $M_h$. The maximum H$_\text{i}$-to-halo mass ratio is obtained at $M_h \approx 10^{11.3} M_\odot$ in the case of $\sigma_{H_\text{i}} = 0.1$ dex and at $M_h \approx 10^{11.1} M_\odot$ in the case of $\sigma_{H_\text{i}} = 0.4$ dex. Note that both values are lower than the values commonly obtained for the location of the peak of the $M_\ast/M_h$ ratio, $M_h \approx 10^{12} M_\odot$.

Once we have assigned a value of $M_{H_\text{i}}$ to each (sub)halo of the Bolshoi simulation, we can compute the correlation functions of modeled samples with any range of H$_\text{i}$ mass and compare them with the ALFALFA results. For example, in Figure 26 we compare the projected correlation functions of three ALFALFA H$_\text{i}$ mass-threshold samples with the correlation functions of modeled samples with the same H$_\text{i}$ thresholds. Note that, due to the $v_{\text{halo}} > 60 \text{ km s}^{-1}$ cut that we have imposed on the Bolshoi halos, our parent halo sample is complete only down to $M_h \approx 10^{11} M_\odot$; in order to avoid biases related to this selection criterion, we conservatively restrict our abundance matching analysis to galaxies with $M_{H_\text{i}} > 10^9 M_\odot$. Overall, we find that the clustering dependence on H$_\text{i}$ mass is rather weak. Nevertheless, in the $\sigma_{H_\text{i}} = 0.1$ dex case, modeled galaxies with large H$_\text{i}$ masses ($M_{H_\text{i}} > 10^{10} M_\odot$) show stronger clustering than their lower H$_\text{i}$ mass counterparts. On the other hand, in the $\sigma_{H_\text{i}} = 0.4$ dex case, the clustering amplitude of modeled galaxies is almost independent of H$_\text{i}$ mass. This latter case is therefore in better agreement with the observational results of Figure 8. We conclude that the clustering properties of ALFALFA galaxies favor an $M_{H_\text{i}}$–$M_h$ relation with a large scatter and a very weak dependence on halo mass (at least for galaxies with $M_{H_\text{i}} \gtrsim 10^{9.5} M_\odot$).

Moreover, Figure 26 shows that all modeled samples display consistently stronger clustering than the actual ALFALFA samples. This is because our abundance matching analysis assumes that all subhalos host an H$_\text{i}$ galaxy. If we repeat the analysis considering only distinct halos, we find the opposite result: all modeled samples consistently underestimate the clustering amplitude of the actual ALFALFA samples (figure not shown). Our second conclusion is therefore that only a subset of subhalos host H$_\text{i}$ galaxies, with the rest presumably hosting gas-poor galaxies that are not detected by ALFALFA. In view of the results above, it is important to ask whether halo properties other than mass and halo/subhalo status may be playing a major role in determining the gas content of galaxies.

### 4.3. Halo Spin Parameter

The spin parameter of the host halo has been suggested to be a decisive factor in setting a number of galaxy properties such as the galaxy’s stellar and gas surface density. In fact, low surface brightness (LSB) galaxies are currently believed to be hosted by halos with higher-than-average spin parameters (e.g., Boissier & Prantzos 2000). Several lines of evidence also
suggest that halo spin may be closely related to the overall gas-to-stellar mass ratio of a galaxy, in the sense that halos with higher spin parameters host more gas-rich systems at a fixed stellar mass. First, gas-rich galaxies are known to be of relatively LSB and low stellar mass density (Catinella et al. 2010; Zhang et al. 2009; Huang et al. 2012), properties that are typically associated with high spin halos. Huang et al. (2012) have furthermore directly estimated the galactic spin parameter of the entire ensemble of ALFALFA galaxies, obtaining the result that measured galaxy spin increases with increasing gas fraction (their Figure 14). Lastly, recent hydrodynamical simulations by Kim & Lee (2013) have shown that, at fixed halo mass, halos with higher spin parameters have more extended gaseous disks and larger overall gas-to-stellar mass ratios (J.-h. Kim 2013, private communication).

Here, we aim to investigate the spin–gas content relation by comparing the clustering properties of the gas-rich ALFALFA galaxies and halos with "low," "average," and "high" spin parameters. Figure 27 shows the projected correlation functions for the three halo samples and compares them with the projected correlation function of the $M_{\text{HI}} > 10^9 M_\odot$ ALFALFA sample (solid cyan line). Note that both distinct halos and subhalos are included in the computation of $\Xi(\sigma)/\sigma$ for these halo samples and no cuts are performed based on halo velocity. The figure clearly shows that halos with low spin parameters display a markedly stronger correlation function compared with halos with average and high spin parameters. This behavior can be attributed to two main causes. First, the low spin sample contains proportionally more subhalos than the average and high spin samples (see Figure 22), which tend to be a highly clustered population. Second, low spin parameter halos tend to cluster significantly more than their higher spin counterparts, even when the sample is restricted to distinct halos only (figure not shown). Note that the results hold also when spins are measured according to the “classical” definition, $\lambda = J |E_{\text{tot}}|^{1/2} / G M_h^{3/2}$.

We caution the reader, however, that the results presented above are sensitive to the algorithm used to identify halos and extract their properties. We have carried out an alternative
In contrast with BDM halos, ROCKSTAR halos do not display clustering variations when split based on their spin parameter. At the same time, ROCKSTAR halos have a different distribution of spin parameters than BDM halos, with the most conspicuous difference being the lack of the low-spin tail in the distribution of ROCKSTAR subhalos (cf. Figure 22). The difference in the spin measurements between the two halo finders seems to be caused by the fact that BDM measurements are made in spherical shells, while ROCKSTAR measurements can be performed over significantly asymmetric domains, both in real and velocity space (A. Klypin 2013, private communication). In this paper, we choose to focus on results obtained with the BDM halo finder. The BDM algorithm has been used extensive in the literature and several aspects of the performance of the algorithm have been studied in detail.

Aside from a potential sensitivity to the details of halo finding, the result above is remarkable since it shows that a crude selection of halos based solely on their spin parameter can reproduce fairly well the clustering of ALFALFA galaxies. This finding lends further support to the hypothesis that halo spin—and perhaps not halo mass—is the main property setting the gas content of galaxies. The result could also have important implications for the modeling of gas-rich galaxies, which will be an integral part of the scientific interpretation of near-future 21 cm large-scale surveys. In particular, it may be necessary for semi-analytic models of H I galaxies (e.g., Obreschkow et al. 2009; Lu et al. 2012) to consider halo spin—in addition to mass—in their implementation. Empirical approaches such as abundance matching may also need to be revised and may potentially need to consider matching H I galaxy abundances with the joint distribution of halo mass and spin (see, e.g., Hearin & Watson 2013 for a similar approach pertaining to galaxy optical colors).

5. CONCLUSIONS

We use a sample of galaxies detected by the ALFALFA blind 21 cm survey to study the clustering characteristics of H I-selected galaxies (i.e., galaxies selected based on their atomic hydrogen content). In particular, we divide the ALFALFA galaxies into subsamples based on their H I mass, creating six H I mass-threshold and three H I mass-binned samples, spanning the entire $M_{HI} = 10^{7.5} - 10^{11} M\odot$ range. We measure the projected correlation function for each of the samples and find no compelling evidence for a dependence of clustering on H I mass. The data yield a lower amplitude correlation function for the least massive H I mass-binned sample ($M_{HI} = 10^{7.5} - 10^{9.5} M\odot$), but we attribute this effect to the small volume probed by the specific sample (see Figures 9 and 10).

Moreover, we compare the clustering characteristics of the H I sample with those of optically selected galaxies drawn from the SDSS spectroscopic database. We follow a similar procedure as described above and divide the optical galaxies into subsamples based on thresholds and ranges in their $r$-band absolute magnitudes (see Figure 2). In addition, we split the parent optical sample into three color-based subsamples, based on the galaxies’ position on a CMD (see Figure 3). We find that H I-selected galaxies cluster more weakly than their optical counterparts, even those at faint absolute magnitudes (at least as faint as $M_r \approx -18$). On the other hand, we find that the correlation function of ALFALFA galaxies is matched extremely well by the correlation function of SDSS galaxies with blue colors. Conversely, SDSS galaxies with red colors display much stronger clustering than the H I-selected samples, resulting in a projected correlation function with a markedly steeper slope and higher amplitude. The results above hold also for the full two-dimensional correlation function, $\xi(\sigma, \pi)$: both the ALFALFA and SDSS blue samples display a strongly anisotropic shape at scales $\gtrsim 10$ Mpc and a very weak finger-of-god feature at small on-sky separations (see Figures 14 and 15).

On the other hand, the two-dimensional correlation function of SDSS red galaxies shows a prominent finger-of-god feature and a more isotropic shape at intermediate scales (Figure 16).

In addition, we carry out a cross-correlation analysis between the ALFALFA and color-based SDSS samples. The H I × red cross-correlation function shows that the gas-rich ALFALFA galaxies “avoid” being located within $\approx 3$ Mpc of optical galaxies with red colors. In particular, they avoid environments where the finger-of-god effect is strong, presumably corresponding to clusters and rich groups. This fact amounts to a statistical measurement of the “H I deficiency” of galaxies in clusters and yields a quantitative result for the length scale over which dense environments typically affect the gas contents of galaxies.

We also measure the clustering properties of the halos in the Bolshoi ΛCDM simulation to gain insights into the characteristics of halos hosting gas-rich galaxies. By comparing the clustering of ALFALFA galaxies with that of halo samples split based on their mass and halo/subhalo status, we arrive at the conclusions that (1) H I mass is not tightly related to the mass of the host halo and (2) a sizable fraction of subhalos do not host gas-rich galaxies. We furthermore perform a more detailed modeling of the clustering of halos hosting gas-rich galaxies, based on the $M_{HI} - M_h$ relation inferred from abundance matching. The results confirm our previous findings, by favoring an $M_{HI} - M_h$ relation that has large scatter and a weak dependence of $M_{HI}$ on host halo mass (see Figures 25 and 26). Lastly, we consider the clustering of halos with different spin parameters. We find that halos with low spin parameters (as measured by the BDM halo finder algorithm) cluster more strongly than halos with higher spin parameters. Remarkably, this leads to the correlation function of ALFALFA galaxies being reproduced fairly well by simply excluding low-spin halos from the computation (Figure 27). This finding provides indirect support to the hypothesis that halo spin plays a central role in determining the gas contents of present-day galaxies.

The authors acknowledge the work of the entire ALFALFA collaboration team in observing, flagging, and extracting the catalog of galaxies used in this work. The ALFALFA team at Cornell is supported by NSF grants AST-0607007 and AST-1107390 to R.G. and M.P.H. and by grants from the Brinson Foundation. E.P. would also like to thank Anatoly Klypin and Peter Behroozi for sharing their Bolshoi halo catalogs and for helpful comments.

10 The ROCKSTAR halo finding algorithm is a phase-space (six-dimensional) friends-of-friends (FOF) algorithm with an adaptively refined linking length. The method is described in detail in Behroozi et al. (2013a), but we give here a basic overview of its operation: first, the simulation volume is split into large groups by a position-space (three-dimensional) FOF algorithm. Then, a phase-space linking length is calculated based on the position and velocity dispersions of the particles in the group. Structures are identified via a six-dimensional FOF algorithm with the predefined linking length. The process above is repeated in the newly identified structures to detect substructure. When no more substructure can be identified, unbound particles are excluded and the structure properties of all substructures are measured. Catalogs of halos extracted from the Bolshoi simulation with the ROCKSTAR algorithm can be found at www.slac.stanford.edu/~behroozi/Bolshoi_Catalogs/.
