

THE C4 CLUSTERING ALGORITHM: CLUSTERS OF GALAXIES IN THE SLOAN DIGITAL SKY SURVEY

CHRISTOPHER J. MILLER,^{1,2} ROBERT C. NICHOL,^{1,3} DANIEL REICHART,⁴ RISA H. WECHSLER,^{5,6} AUGUST E. EVRARD,^{7,8}
JAMES ANNIS,⁹ TIMOTHY A. MCKAY,⁷ NETA A. BAHCALL,¹⁰ MARIANGELA BERNARDI,^{1,11} HANS BOEHRINGER,¹²
ANDREW J. CONNOLLY,¹¹ TOMOTSUGU GOTO,¹³ ALEXIE KNIAZEV,^{14,15,16} DONALD LAMB,¹⁷ MARC POSTMAN,¹⁸
DONALD P. SCHNEIDER,¹⁹ RAVI K. SHETH,¹¹ AND WOLFGANG VOGES¹²

Received 2004 August 17; accepted 2005 March 15

ABSTRACT

We present the C4 Cluster Catalog, a new sample of 748 clusters of galaxies identified in the spectroscopic sample of the Second Data Release (DR2) of the Sloan Digital Sky Survey (SDSS). The C4 cluster-finding algorithm identifies clusters as overdensities in a seven-dimensional position and color space, thus minimizing projection effects that have plagued previous optical cluster selection. The present C4 catalog covers $\sim 2600 \text{ deg}^2$ of sky and ranges in redshift from $z = 0.02$ to 0.17 . The mean cluster membership is 36 galaxies (with measured redshifts) brighter than $r = 17.7$, but the catalog includes a range of systems, from groups containing 10 members to massive clusters with over 200 cluster members with measured redshifts. The catalog provides a large number of measured cluster properties including sky location, mean redshift, galaxy membership, summed r -band optical luminosity (L_r), and velocity dispersion, as well as quantitative measures of substructure and the surrounding large-scale environment. We use new, multicolor mock SDSS galaxy catalogs, empirically constructed from the Λ CDM Hubble Volume (HV) Sky Survey output, to investigate the sensitivity of the C4 catalog to the various algorithm parameters (detection threshold, choice of passbands, and search aperture), as well as to quantify the purity and completeness of the C4 cluster catalog. These mock catalogs indicate that the C4 catalog is $\simeq 90\%$ complete and 95% pure above $M_{200} = 1 \times 10^{14} h^{-1} M_\odot$ and within $0.03 \leq z \leq 0.12$. Using the SDSS DR2 data, we show that the C4 algorithm finds 98% of X-ray-identified clusters and 90% of Abell clusters within $0.03 \leq z \leq 0.12$. Using the mock galaxy catalogs and the full HV dark matter simulations, we show that the L_r of a cluster is a more robust estimator of the halo mass (M_{200}) than the galaxy line-of-sight velocity dispersion or the richness of the cluster. However, if we exclude clusters embedded in complex large-scale environments, we find that the velocity dispersion of the remaining clusters is as good an estimator of M_{200} as L_r . The final C4 catalog will contain $\simeq 2500$ clusters using the full SDSS data set and will represent one of the largest and most homogeneous samples of local clusters.

Key words: catalogs — galaxies: clusters: general

Online material: color figures, machine-readable table

1. INTRODUCTION

Catalogs of clusters and groups of galaxies are used extensively throughout extragalactic astronomy and cosmology, from constraining the cosmological parameters (e.g., Oukbir & Blanchard 1992; Henry & Arnaud 1991; Viana & Liddle 1996; Bahcall et al. 1997; Reichart et al. 1999; Miller et al. 2001a, 2001b) to magnifying the most distant galaxies in the universe (Sand et al. 2002). Considerable effort has been invested over the last half-century in constructing catalogs of clusters and groups

of galaxies (e.g., Zwicky 1952; Abell 1958; Abell et al. 1989; Gioia et al. 1990; Lumsden et al. 1992; Dalton et al. 1992; Henry et al. 1995; Postman et al. 1996; Romer et al. 2000; Böhringer et al. 2000; Gladders 2002; Postman et al. 2002). In this paper, we present one of the first catalogs of clusters and groups constructed directly from the spectroscopic data of the Sloan Digital Sky Survey (SDSS). This is now possible because of the present size of the SDSS data set (see § 7) and is complementary to the

¹ Department of Physics, Carnegie Mellon University, 7325 Wean Hall, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890.

² Currently at NOAO Cerro Tololo Inter-American Observatory, 950 North Cherry Avenue, Tucson, AZ 85719.

³ Institute of Cosmology and Gravitation, University of Portsmouth, Mercantile House, Hampshire Terrace, Portsmouth PO1 2EG, UK.

⁴ Department of Physics and Astronomy, University of North Carolina, CB 3255 Phillips Hall, Chapel Hill, NC 27599.

⁵ Center for Cosmological Physics, Department of Astronomy and Astrophysics, and Enrico Fermi Institute, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637.

⁶ Hubble Fellow.

⁷ Department of Physics, University of Michigan, 2477 Randall Laboratory, 500 East University Avenue, Ann Arbor, MI 48109.

⁸ Department of Astronomy, University of Michigan, 830 Dennison, 501 East University Avenue, Ann Arbor, MI 48109.

⁹ Fermi National Accelerator Laboratory, P.O. Box 500, 500 Wilson Road, Batavia, IL 60510.

¹⁰ Princeton University Observatory, Peyton Hall, Ivy Lane, Princeton, NJ 08544.

¹¹ Department of Physics and Astronomy, University of Pittsburgh, 100 Allen Hall, 3941 O’Hara Street, Pittsburgh, PA 15260.

¹² Max-Planck-Institut für Extraterrestrische Physik, Giessenbachstrasse Postfach 1312, Garching 85748, Germany.

¹³ Institute for Cosmic-Ray Research, University of Tokyo, 515 Kashiwanoha, Kashiwa, Chiba 277-0882, Japan.

¹⁴ Special Astrophysical Observatory, Nizhnij Arkhyz, Karachai-Circassia 369167, Russia.

¹⁵ Max-Planck-Institut für Astrophysik, Königstuhl 17, 69117 Heidelberg, Germany.

¹⁶ Isaac Newton Institute of Chile, SAO Branch, Santiago, Chile.

¹⁷ Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637.

¹⁸ Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218.

¹⁹ Department of Astronomy and Astrophysics, Pennsylvania State University, University Park, PA 16802.

SDSS cluster catalogs selected using the SDSS photometric data (e.g., Annis et al. 1999; Goto et al. 2002; Kim et al. 2002; Bahcall et al. 2003; Lee et al. 2004).

The distribution of matter in the universe is described by the statistics of overdensities. When these overdensities are small, the equations of motion that follow the evolution of matter can be linearized and solved. As gravitational clustering is amplified into the nonlinear regime, a description of the matter distribution as a point set of extended dark matter halos becomes more appropriate (e.g., Cooray & Sheth 2002). How halos are populated with galaxies of specific colors and luminosities (typically referred to as the halo occupation) is not known precisely and remains a serious challenge in cosmology and astrophysics. The details of how galaxies occupy halos will clearly have an affect on attempts to identify clusters in optical catalogs. But progress can be made on both fronts: gaining understanding about how galaxies populate clusters will be invaluable to those who study structure and galaxy formation/evolution. Vice versa, mock catalogs that are representative of the real universe will be invaluable for accurately measuring the selection function of any clustering algorithm; this specifies the contamination and completeness of a data set and is a prerequisite to many scientific analyses.

The challenge in constructing a cluster catalog from galaxy data is to minimize projection effects (or false positive detections) while maximizing completeness, i.e., controlling the selection function. Previous analyses of large optical catalogs of clusters (Lucey 1983; Sutherland 1988; Frenk et al. 1990) have claimed various levels (10%–25%) of contamination (see also Miller et al. 1999, 2002). The next generation of cluster catalogs must have very little contamination and precisely known selection functions in order to compete with the increasingly precise cosmological constraints from other methods (e.g., Perlmutter & Schmidt 2003; Mandolesi et al. 2002). Because of a concern about the presence of projection effects in optical cluster catalogs, there has been renewed emphasis over the past decade on new ways of finding clusters of galaxies in wave bands other than the optical. For example, many authors have constructed catalogs of clusters from X-ray surveys of the sky (e.g., Edge & Stewart 1991; Gioia et al. 1990; Böhringer et al. 2000; Romer et al. 2000), as this is believed to be more robust for selecting mass-limited samples than optical methods (see Ebeling et al. 1997). An X-ray-selected SDSS cluster sample has also been presented by Popesso et al. (2004). In addition, many authors have proposed the construction of catalogs of clusters using the Sunyaev-Zel'dovich effect (Carlstrom et al. 2002; Romer et al. 2004) and weak gravitational lensing (Wittman 2002). Furthermore, Kochanek et al. (2003) recently presented a new catalog of clusters derived from the Two Micron All Sky Survey infrared photometric data.

As we outline below, we have now mitigated the problem of projection effects in optical cluster catalogs by simultaneously using both SDSS photometric and spectroscopic data to find clusters. The details of our cluster-finding algorithm, which we refer to as the “C4” algorithm, are presented in § 2. The premise of this algorithm is that optical clusters and groups of galaxies are dominated, at their cores, by a single, coevolving population of galaxies that possess similar spectral energy distributions, e.g., the “E/S0 ridgeline” or “red envelope” (Baum 1959; McClure & van den Bergh 1968; Lasker 1970; Visvanathan & Sandage 1977). The evidence for such a coevolving population of galaxies in the cores of clusters has been presented by many authors; see Gladders (2002) for a detailed review of this evidence. Blakeslee et al. (2003) provide evidence that this coevolving population extends beyond a redshift of 1. As such,

Ostrander et al. (1998), Gladders (2002), and Goto et al. (2002) have all used the existence of a coevolving population of galaxies in the cores of clusters as a basis for their cluster-finding algorithms.

In Figure 1, we present the color-magnitude diagram, in all four of the SDSS colors ($u - g$, $g - r$, $r - i$, and $i - z$), for a newly discovered cluster of galaxies at $z = 0.06$ in the SDSS Early Data Release (Stoughton et al. 2002). The nearest cluster in the literature is a Zwicky group that is $\sim 9'$ to the southeast. All galaxies in this figure are within a projected $1 h^{-1}$ Mpc radius of the cluster center. (We use $h = H_0 / 100 \text{ km s}^{-1} \text{ Mpc}^{-1}$.) The figure demonstrates the existence of a tight relationship between the colors of cluster galaxies, which can have an observed scatter of ~ 0.05 (see Bower et al. 1992). The figure also illustrates that this “red sequence” of galaxies is present in all four SDSS colors and not just in the one-color system used by Gladders & Yee (2000). Note also the existence of a “blue ridgeline” in Figure 1 (top left), which has been highlighted previously by several authors (see Chester & Roberts 1964; Tully et al. 1982; Baldry et al. 2004). In this one example (at $z = 0.06$), the 4000 Å break sits between the u and g filters. Thus, the $u - g$ color-magnitude diagram can separate the blue star-forming galaxies from the older, red, passive population. As one moves toward redder colors, the star-forming and passive populations begin to overlap in color. In the reddest colors, the red sequence contains a mixture of old passive ellipticals and young star-forming galaxies. So instead of using an algorithm that attempts to model the red sequence, we simply allow galaxies in clusters to have similar colors. Our clusters will be detected via a mixture of galaxy types, both passively evolving and star forming.

In this paper we first present an outline of our algorithm based on the premise of galaxies clustering in both position and color. We then spend a significant amount of effort analyzing how the C4 algorithm’s free parameters affect the completeness and contamination of the final cluster catalog. We do this using a new breed of mock galaxy catalogs that populate N -body cosmological simulations with realistic galaxy populations. Previous authors have realized the necessity of such a detailed understanding of their clustering algorithms (e.g., Diaferio et al. 1999; Adami et al. 2000; Postman et al. 2002; Kim et al. 2002; Goto et al. 2002; Eke et al. 2004; Yang et al. 2005; Gerke et al. 2005). In many of these cases, the authors have embedded fake clusters of various forms into real or simulated data, which are then searched for to characterize the contamination and completeness of the algorithm. In this work, we have gone one step further by using mock galaxy catalogs generated from full N -body simulations. A similar technique was used in recent work by Eke et al. (2004) on Two-Degree Field groups, using N -body simulations populated with semianalytic models. The catalogs we use, developed by Wechsler (2004), embed galaxies with realistic luminosities and colors into cosmological simulations that contain all the messiness of structure formation, including merging systems, systems with lots of substructure, systems with ill-defined E/S0 ridgelines, systems that nearly overlap in redshift space, etc.

In § 2, we provide details of the C4 algorithm. In § 3, we describe the use of novel mock catalogs—based on populating large cosmological simulations with realistic galaxy properties—for calibrating the C4 algorithm and determining the completeness and purity of our SDSS C4 cluster catalog. In § 5, we introduce the observables that we measure for each cluster and then discuss the scaling relations between those observables and halo mass in § 6. The SDSS data and the SDSS C4 cluster catalog are presented in § 7, and we conclude in § 8. Where appropriate, we

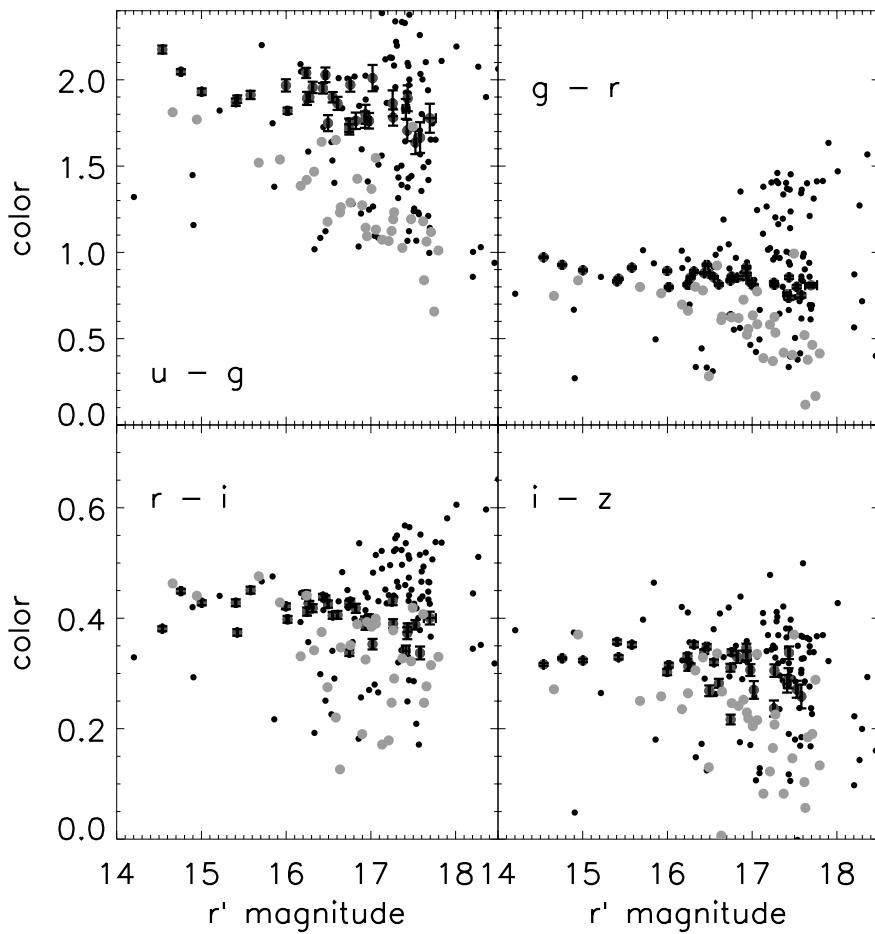


FIG. 1.—Color-magnitude relation of galaxies in all four SDSS colors for a previously unknown cluster of galaxies identified in the SDSS DR2 data set. Black circles show galaxies within a projected aperture of $1 h^{-1} \text{ Mpc}$ around the cluster center. Dark gray and light gray circles show cluster members; the dark gray circles have low $\text{H}\alpha$ emission, and the light gray circles have high $\text{H}\alpha$ emission, indicative of ongoing star formation. Error bars on the colors of the dark gray galaxies indicate the typical color errors in the spectroscopic sample. [See the electronic edition of the Journal for a color version of this figure.]

have used $h = H_0/100 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_m = 0.3$, and $\Omega_\Lambda = 0.7$ throughout this paper.

2. THE C4 ALGORITHM

In this section, we present the C4 cluster-finding algorithm. Details and tests follow. Many readers will only want a brief explanation of the algorithm, and we suggest that they examine the flowchart of the algorithm given in Figure 2 and read this overview section. For those who desire more details, each step is described in more detail throughout the rest of this section. The application of the C4 algorithm to the SDSS data is discussed in § 7.

The C4 algorithm begins by placing each galaxy in a seven-dimensional space of right ascension (R.A.), declination (decl.), redshift, and four color dimensions ($u - g$, $g - r$, $r - i$, and $i - z$). On each such target galaxy, we then perform the following steps:

1. We place an aperture around each target to only include galaxies in a specified range of right ascension, declination, and redshift. We then measure the probability that every galaxy within this spatial aperture has colors equal to the target galaxy. The probabilities are summed to obtain a “number count.”

2. Using the target galaxy’s spatial and color aperture, we then select 100 random galaxies and perform step (1). These 100 random locations provide a number count distribution for the target galaxy.

3. Using the number count distribution, we compute the probability of obtaining at least the observed number count around the original target galaxy. By definition, target galaxies with low probabilities will be in clustered regions.

4. We repeat this exercise for all (target) galaxies in our sample and then rank all the target galaxy probabilities obtained from step (3).

5. Using the false discovery rate (FDR) algorithm (Miller et al. 2001c), we determine a threshold in probability below which target galaxies are removed; our threshold choice typically results in the eradication of $\simeq 90\%$ of all galaxies. The galaxies that remain are called “C4 galaxies.” By construction, these reside in high-density regions, with neighbors that possess similar colors.

6. We determine the local surface density around all C4 galaxies, using only the C4 galaxies. We then rank order these measured densities and locate C4 cluster centers as peaks in this density field.

In summary, the C4 algorithm is a semiparametric implementation of adaptive kernel density estimation. The key difference of our approach compared to previous color-based cluster-finding algorithms is that we do not attempt to model either the colors of the cluster galaxies (e.g., Gladders & Yee 2000; Goto et al. 2002) or the properties of clusters (e.g., Kepner et al. 1999; Postman et al. 1996; Kim et al. 2002). Instead, we only demand that the colors of nearby galaxies are similar to those of the target galaxy. In this way we are sensitive to a diverse range of cluster and group

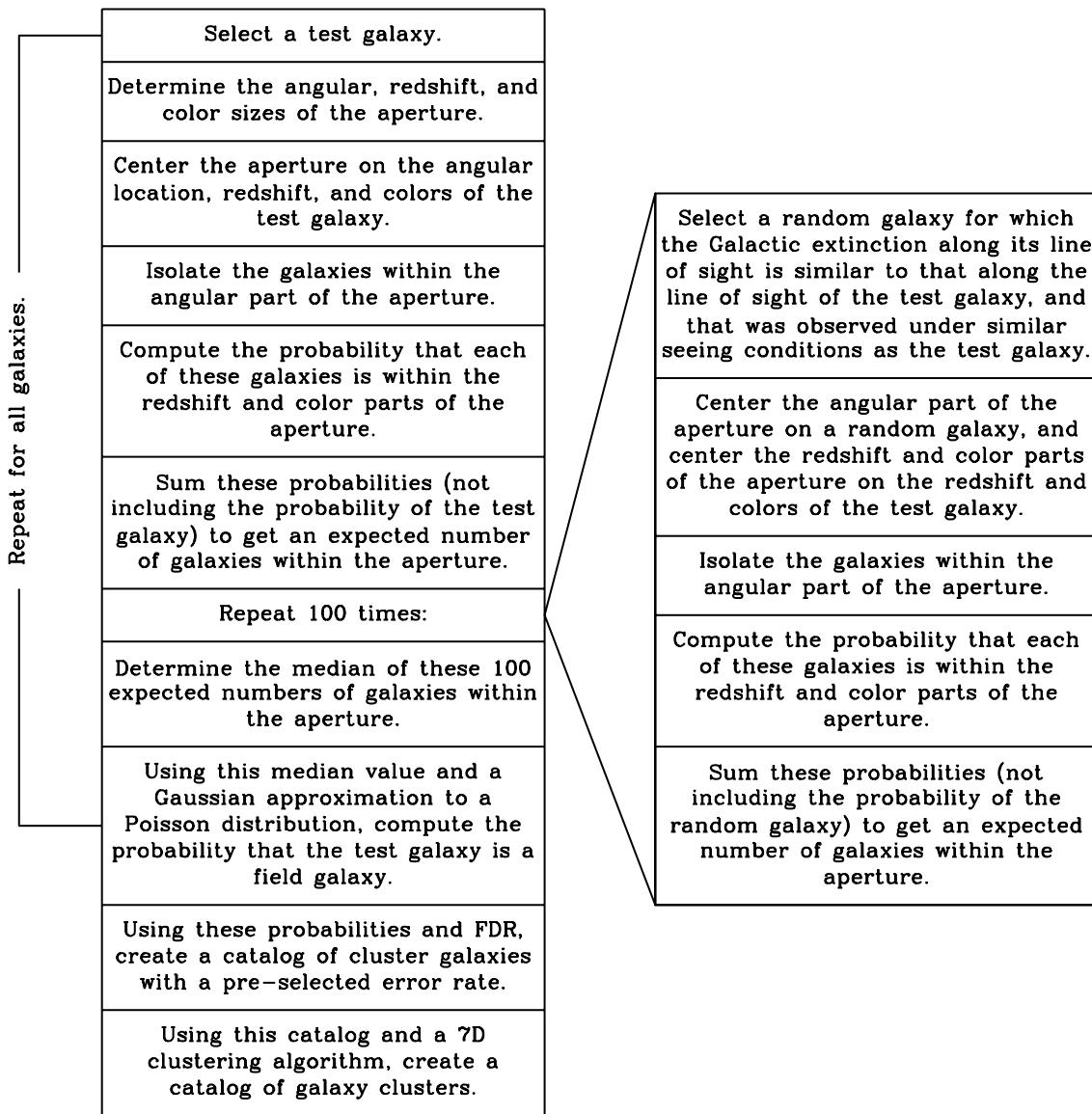


FIG. 2.—Flowchart describing the C4 cluster-finding algorithm.

types, e.g., our algorithm would detect a cluster dominated by a “blue” population of galaxies (see Fig. 21).

2.1. Defining the Seven-Dimensional Search Aperture

Every target galaxy in the data set has a uniquely defined location in a seven-dimensional data space. For example, the position of the i th galaxy is defined as

$$\mathbf{r}^i = [\text{R.A.}^i, \text{decl.}^i, z^i, m_u^i - m_g^i, m_g^i - m_r^i, m_r^i - m_i^i, m_i^i - m_z^i], \quad (1)$$

where m_x^i are the five passband Petrosian magnitudes from the SDSS PHOTO version 5.4 data reductions (typically abbreviated u , g , r , i , and z). No k -corrections are used herein.

To look for clusters in this seven-dimensional data space, we need to define a search aperture. Clearly, the size of this aperture will have an effect on the types of clusters we find in this data space. We begin by using a projected radius that is fixed in comoving coordinates and specifies the right ascension and declination aperture surrounding the target galaxy. The exact cos-

ological model used makes little difference over the redshift range we examine here ($z \sim 0.1$). This aperture can be tuned to find the size that optimizes completeness and purity in the mock galaxy catalogs.

We next define the redshift (or line of sight) dimension of the C4 search aperture. For the spectroscopic SDSS sample, all galaxies have known redshifts and we simply place a z -constraint around the target galaxy. For the SDSS photometric sample, one would need estimated redshifts or else this constraint must be dropped entirely. We have chosen to convert redshift to comoving distance under an assumed model, but one could also simply let the length of the redshift dimension vary with redshift.

Finally, we must define the color part of our search aperture. The sizes of the four color dimensions will be driven by the well-established intrinsic color-magnitude relation (CMR) seen in clusters (see Fig. 1) and the expected errors on the SDSS magnitudes. The CMR is known to have a linear relationship with a small negative slope (with increasing magnitude) and small scatter (Bower et al. 1992). Therefore, the size of the “color box” should be set to capture the full range of colors in the CMR, from the brightest to the faintest cluster galaxies in any given

cluster in our data. We also include the known statistical (1σ) uncertainties in the individual galaxy magnitudes. For the SDSS main galaxy spectroscopic sample, these errors are minimal (less than 0.1% at $m_r = 17.7$). We sum in quadrature these statistical errors and also a systematic uncertainty via

$$\delta C_{xy} = \sqrt{\sigma_{xy}^2(\text{stat}) + \sigma_{xy}^2(\text{sys})}, \quad (2)$$

where $\sigma_{xy}^2(\text{stat})$ is the observed error for the two magnitudes (x, y), summed in quadrature. Here $\sigma_{xy}^2(\text{sys})$ is a measure of the inherent scatter in the CMR (see below). Therefore, for each i galaxy the size of the color box is given by

$$\delta C^i = [\delta C_{ug}^i, \delta C_{gr}^i, \delta C_{ri}^i, \delta C_{iz}^i]. \quad (3)$$

We have used the Petrosian magnitudes reported by the SDSS throughout, as it is better suited for the analyses of galaxies in the SDSS spectroscopic sample (see Stoughton et al. 2002). However, our final cluster catalog is robust against the use of Petrosian versus model magnitudes. We do not apply evolutionary or k -corrections to our data, as we are looking for galaxies clustered in both position and color around another galaxy: for a given redshift and color of a galaxy, any excess of neighboring galaxies with similar colors should occur independently of any evolutionary effects and k -corrections.

Once we have defined the search aperture around a target galaxy, we then “count” the number of neighboring galaxies within that aperture. To do this, we demand that any neighboring galaxy fit exactly within the spatial part of the aperture (R.A., decl., and redshift), as these dimensions are accurately known.

In the color dimensions, we allow for uncertainties in both the color box of the target galaxy and the individual colors of surrounding galaxies. Specifically, we replace the color boxes with Gaussians having widths specified by equation (4), which “softens” the sides of the four-dimensional color box. We also treat the errors on the individual galaxies as Gaussians. We then measure the joint probability that any galaxy falls within the color box of the target galaxy. We then sum these probabilities for all neighboring galaxies and report this as the “number count” of neighboring galaxies.

2.2. Building the Count Distributions

The next step in the C4 algorithm is to build a distribution of expected number counts for each target galaxy, given that it was in a random position. We place the seven-dimensional aperture of the target galaxy around 100 randomly chosen galaxies and “count” the neighbors as described above. We allow for the fact that our algorithm can be run on the SDSS photometric data, in which case the seeing conditions and Galactic extinction can have a large effect on the selection function of the SDSS photometric sample. The random galaxies can be selected such that they have the same seeing and reddening as the target galaxy. However, on the complete SDSS spectroscopic sample we ignore this constraint. From these 100 randomly chosen locations in the data, we construct a distribution of counts for the seven-dimensional aperture of the target galaxy. So long as we expect no more than half of the galaxies to be in clustered environments (i.e., have higher counts with respect to the mean), the medians of these distributions are robust descriptors of the distributions.

2.3. Determining Probabilities

By this stage, we have defined a unique aperture for the target galaxy. We have measured the number of neighboring galaxies

within its aperture and built a count distribution from 100 random locations at the same redshift of the target galaxy. We then ask the question, how likely is the observed neighboring galaxy count given the distribution of neighboring galaxy counts for a specified seven-dimensional aperture? The exact form of the distribution of neighboring galaxy counts depends on the number of counts measured. For example, in the photometric SDSS data, where there are millions of galaxies, the distributions of neighboring galaxy counts is Gaussian. However, in the spectroscopic data, the count distributions can sometimes be small and Poissonian. As a compromise, we adopt the Gaussian approximation to the Poisson distribution. In order to justify the basis of Poissonian statistics, we need to meet the following requirements in the count distribution: (1) the count within an aperture of zero volume is zero, (2) each of the 100 randomly chosen counts must be independent, (3) the count values depend only on the size of the aperture, (4) the aperture size does not change when building the count distributions, and (5) no two counts come from the same location. Requirements (1), (2), (4), and (5) are already met in our algorithm, while (3) requires that the randomly selected points have an underlying count distribution that is also random. This, of course, is not true for all galaxies, as galaxies are known to cluster and galaxies within clusters have a higher neighbor count than those in the field. However, if a majority of the randomly selected galaxies are “fieldlike,” i.e., less clustered than the older elliptical population in clusters and groups, then we can expect requirement (3) to hold. At worst, this assumption produces a small bias by slightly raising our probabilities, resulting in a loss of statistical power (which would affect the C4 completeness), and so our Poisson assumption is a conservative one.

The Gaussian approximation to the Poisson distribution has the convenient feature that the width of the Gaussian is equal to the square root of the mean of the Poisson distribution. Thus, when we build the count distributions based on the 100 random locations, we only need to calculate the median, which then fully describes the Gaussian approximation. Thus, the probability that a target galaxy looks like a field galaxy is determined solely from the count around the target and the median of the counts around the 100 random locations.

2.4. Repeat

Once the above steps are performed on the first target galaxy, we then repeat for all galaxies in sample. This is conducted in no specific order. Once we have looped over the entire sample, every galaxy has a probability that it is a “field galaxy.” These probabilities are ranked so that a threshold can be applied to separate the cluster galaxies from the field galaxies.

2.5. Choosing a Threshold

Miller et al. (2001c) present a new thresholding technique known as the false discovery rate, originally devised by Benjamini & Hochberg (1995). This technique allows one to choose a statistically meaningful threshold, in the sense that the fraction of false positive detections over the total number of detections is controlled. We apply the techniques discussed in Miller et al. here. Briefly, this involves choosing a priori the maximum fraction of acceptable false discoveries (α) one is willing to tolerate. The p -values (probabilities) are rank listed (from lowest to highest), and a line of slope α is drawn. Where the two lines intersect for the first time defines the threshold one must use to guarantee the fraction of false discoveries (see Miller et al. [2001c] and Hopkins et al. [2002] for some examples). After

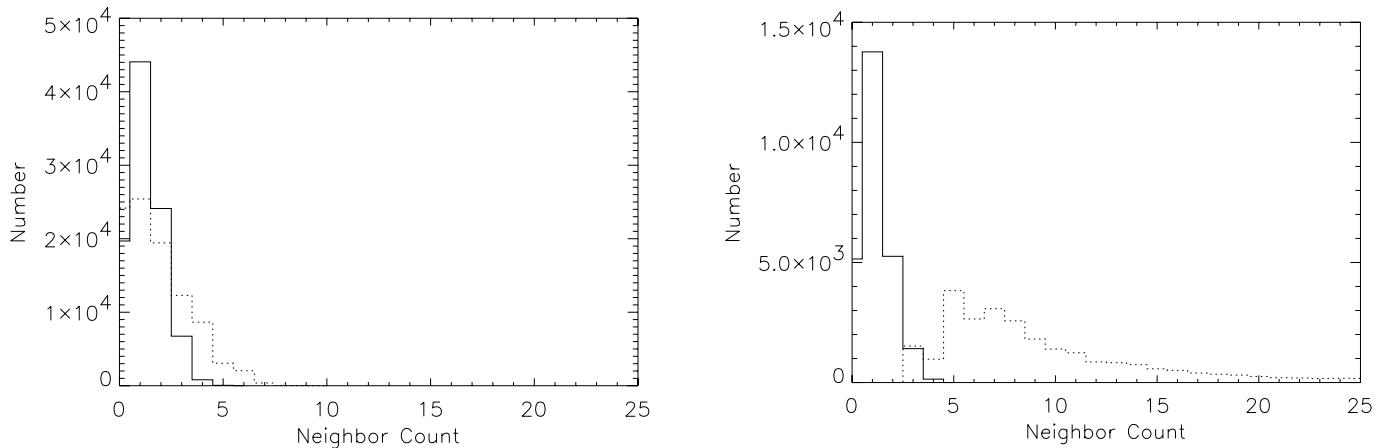


FIG. 3.—Counts from the field based on the median of 100 random locations for each galaxy (solid lines) and counts around each specific galaxy (dotted lines). On the left we show fieldlike galaxies, while on the right we show cluster-like galaxies. Note that cluster-like galaxies have more neighbors than the median of the field.

applying the FDR technique, all galaxies above the threshold are called “cluster-like” or C4 galaxies and are then used to identify C4 cluster centers.

We test our probability model and whether our FDR threshold can separate fieldlike and cluster-like galaxies on a galaxy-by-galaxy level. In Figure 3, we show the distribution of median counts for the count distributions around all galaxies in the mock SDSS catalog (solid line) compared to the individual counts around galaxies (dashed line). We have split our sample into fieldlike and cluster-like galaxies as described above. Note that the counts around cluster-like galaxies are significantly higher than around fieldlike galaxies. Also note that the median counts of the field distributions, and also the counts around fieldlike galaxies, are similar and Poisson. This shows that the medians of the random distributions are representative of fieldlike galaxies. This justifies our use of the median to represent the random distributions and indicates that the probability threshold (discussed below) cleanly separates fieldlike and cluster-like galaxies.

2.6. Identifying the Clusters

The C4 algorithm works by identifying galaxies clustered in a positional and color space. Once the seven-dimensional aperture

and threshold are defined, galaxies with low probabilities of being fieldlike (as defined by random positions on the sky) are identified as clustered C4 galaxies (10% of all galaxies using our fiducial parameters). We then find centers of the clumps of C4 galaxies and call these C4 clusters. The algorithm (to this point) does not define the galaxy membership of the clusters (see § 5).

We were motivated by the spherical overdensity method applied by Evrard et al. (2002) to identify halos in the Hubble volume (HV) simulation. This adds a level of consistency when comparing our observed C4 catalog to the mock C4 cluster samples. We begin by measuring the distance to the sixth-nearest projected neighbor for each C4 galaxy (using only C4 galaxies). We do this in redshift shells of $\Delta z = 0.02$. The nearest neighbor distances are ordered from the smallest to largest, and the C4 galaxy with the smallest sixth-nearest neighbor distance is assigned as the center of the first C4 cluster. We then exclude all C4 galaxies from this list out to a projected radius corresponding to 15 times the background density of C4 galaxies centered on this first cluster. This choice of enhancement is arbitrary. However, this same overdensity is used when examining the real data or the mock catalogs. We have checked to make sure that the distribution of neighbor distances is the same in both the real SDSS

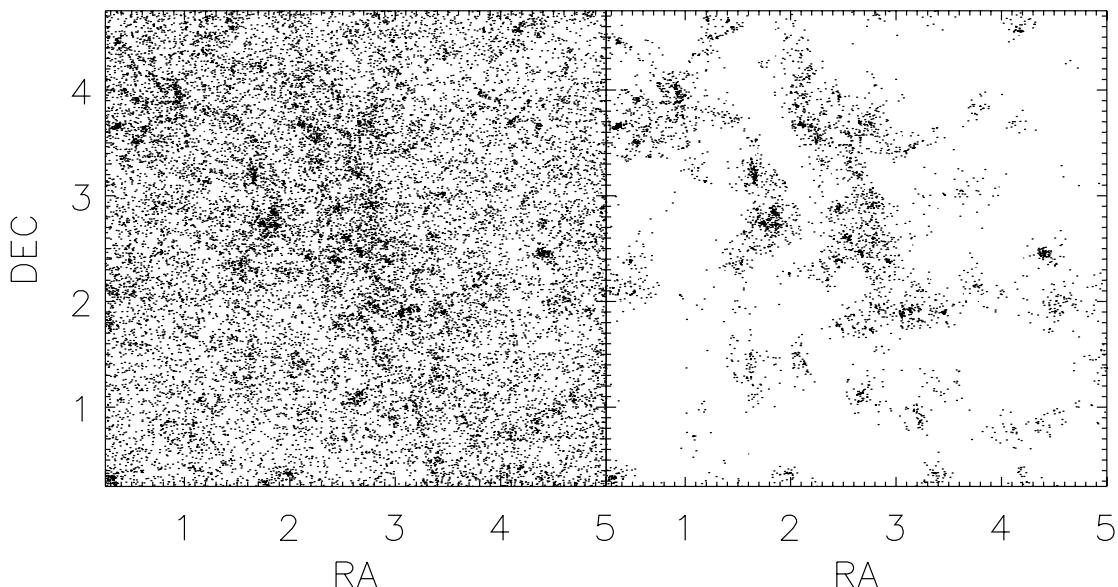


FIG. 4.—Projected galaxy distribution of the simulations before (left) and after (right) the C4 algorithm is run and a threshold is applied to eliminate fieldlike galaxies.

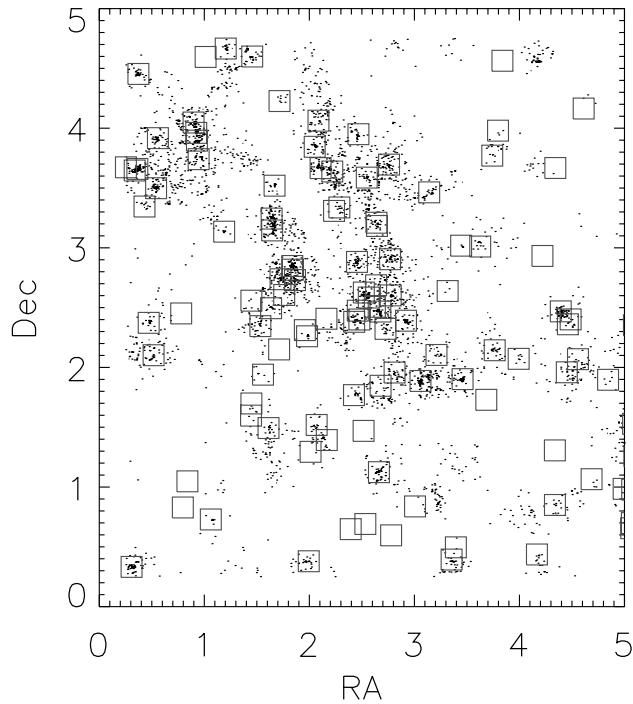


FIG. 5.—C4 galaxies, as in Fig. 4 (right), with the dark matter halo positions overplotted (squares). The halos have masses greater than $4.5 \times 10^{13} h^{-1} M_{\odot}$. [See the electronic edition of the Journal for a color version of this figure.]

data and in the mock galaxy catalog. We then move to the next highest density C4 galaxy that is not within any other C4 cluster and repeat, i.e., the C4 galaxy now with the smallest sixteenth-nearest neighbor distance becomes the center of the second cluster, and so on. The iterations are terminated when all C4 galaxies are assigned to clusters or the local densities fall below the threshold. These initial centers are then peaks in the C4 galaxy surface overdensities.

This process is shown visually in Figures 4 and 5. We investigated other methods for finding the C4 cluster centers (e.g., “friends-of-friends” algorithm) and find this method to be the best in terms of accuracy, completeness, and purity when compared to the actual halo catalog.

During this process, if there are fewer than three C4 neighbors around any cluster center, we exclude it as a possible cluster. Likewise, if we determine that less than 10% of all galaxies in a $1 h^{-1}$ Mpc aperture around the cluster are classified as C4 galaxies, we exclude it. (The number of clusters excluded because of this criterion is less than 2% of the total found.) Finally, we exclude any cluster that has fewer than eight members within $1 h^{-1}$ Mpc so that we may measure a reliable velocity dispersion. We note that these exclusions imply that C4 completeness (see § 4) is lower than it should be, since these excluded systems are often real systems.

2.7. Other Algorithm Considerations

2.7.1. Survey Edges

Edge effects are taken into account during this procedure: C4 galaxies are required to be farther than 7.5' from any edge of the data sample. This allows use of all galaxies within a circle of 15' diameter when making our counts. Edges have their greatest affect on the catalog at the lowest redshifts. To help minimize the problems of survey edges, we only include cluster candidates whose initial redshift (estimated from the C4 galaxies) is above

$z = 0.03$. Of course, we do find clusters below this redshift (see Fig. 25, which is a known X-ray cluster at $z = 0.027$), but the C4 algorithm cannot be tuned to work statistically at such a low redshift.

2.7.2. The Magnitude Limit

The current incarnation of the catalog is run on apparent magnitude-limited surveys. The SDSS spectroscopic main sample was targeted using galaxies brighter than $m_r = 17.77$, while the luminous red galaxy (LRG) sample includes additional elliptical-type galaxies to $m_r = 19.5$. While technically this is an input parameter that could be tuned, we instead consider the effect on completeness, purity, and total number of clusters after a single limit is chosen. When measuring the cluster properties (see § 5), we apply an absolute magnitude limit as well.

3. A MOCK SDSS GALAXY CATALOG

One advantage of the C4 catalog over many previous cluster catalogs is our use of realistic cosmological N -body simulations to refine the algorithm and to determine the completeness and purity of the catalog. Since the C4 algorithm is highly dependent on clustering in *both* spatial and color space, any mock galaxy catalog must have relations between galaxy color and density that mimic those found in real data. Similarly, the luminosity functions of the mock catalogs and the data must be similar. Simple bias schemes that produce a population of galaxies above some luminosity cut (e.g., Cole et al. 1998) will not produce the information we need to find clusters. Mock catalogs created from more detailed semianalytic models of galaxy formation (e.g., Kauffmann et al. 1999; Eke et al. 2004) do produce colors and luminosities; unfortunately, at this stage none of these models are yet producing galaxies with properties that reproduce those seen in the SDSS data closely enough. The mock catalog we use here thus takes a very empirical approach, in which we aim to populate a dark matter simulation with galaxies whose properties (especially the luminosity and color distributions as a function of environment) closely match those seen in SDSS data. The method for creating these catalogs is described briefly in Wechsler (2004) and in detail in R. H. Wechsler (2005, in preparation, hereafter W05); here we just give a rough outline. A complementary approach, using conditional luminosity functions constrained by galaxy clustering, has been developed by Yang et al. (2004).

The W05 catalogs are constructed using the distribution of dark matter in the Λ CDM HV simulation. The simulation follows 10^9 particles of mass $2.25 \times 10^{12} h^{-1} M_{\odot}$ in a periodic cubical volume with side length $3 h^{-1}$ Gpc, in a flat Λ CDM universe with $\Omega_m = 0.3$, $\sigma_8 = 0.9$, and $h = 0.7$. We use a subset of the “MS” sky survey output described by Evrard et al. (2002), which mimics the collection of data on the past light cone of an observer located at the center of the volume. We only look at halos more massive than $4.5 \times 10^{13} h^{-1} M_{\odot}$ (see below). The large size of the simulation allows the creation of a full-sky survey out to a depth of $z_{\max} = 0.57$ and thus can also be used to test cluster-finding algorithms that use only the SDSS photometric data and extend to higher redshift.

Briefly, the algorithm for creating the W05 catalogs consists of constraining the relation between local dark matter density and luminosity in the r -band such that the mock galaxies match the luminosity-dependent two-point correlation function measured in the SDSS data (Zehavi et al. 2002). We then measure the local galaxy density in the r -band, for both the data (see Gómez et al. 2003) and the simulation, and assign the colors of real SDSS

galaxies to mock galaxies that have similar luminosities and local galaxy densities. All relevant details of these new simulations are presented in W05, including the prescriptions used to assign galaxy properties to each dark matter particle and the extensive tests performed on these simulations to ensure they closely mimic the real data.

The mass resolution of the simulation allows us to include galaxies brighter than about $0.4 L^*$, and to resolve halos more massive than $4.5 \times 10^{13} h^{-1} M_\odot$. Therefore, these simulations can only resolve bright galaxies in intermediate-to-massive clusters of galaxies, e.g., a cluster like the Coma Cluster would contain about 500 galaxies in these mock SDSS catalogs. This makes these simulations well suited to the brighter (and more massive) galaxies in the SDSS spectroscopic sample (which is what we use in this work), which only includes dimmer galaxies at the lowest redshifts $z < 0.05$. As discussed above, we place a constraint on any C4 cluster that there be at least eight galaxies within $1 h^{-1}$ Mpc of the cluster centers. For halos at the minimum well-resolved mass in the simulation, only 1% have fewer than eight galaxy members, so nearly all are intrinsically detectable by the C4 algorithm. This indicates that these simulations allow us to reliably estimate the completeness of our catalog over the full mass range of interest, but higher resolution simulations will be required to fully characterize the purity of the smallest C4 systems.

Another advantage of using the HV simulation is that we can directly relate any detected clusters in these mock SDSS catalogs to the measured halo masses discussed by Jenkins et al. (2001) and Evrard et al. (2002). In other words, we can directly relate the observables (e.g., velocity dispersions and summed total cluster optical luminosities) to the dark matter halos used by Jenkins et al. (2001) for constructing the cosmological mass function. There is no ambiguity in the definition of mass between the theoretical models and the observables.

Since the real data will contain photometric errors, we have added errors to each mock galaxy magnitude. However, we note that for the spectroscopic sample, the SDSS Petrosian (1976) magnitude errors are tiny (the largest fractional errors are $<0.1\%$ in the r -band). As described in § 2.1, the size of the color box is at least an order of magnitude larger than the errors on the colors. Thus, this step is not necessary for the spectroscopic data. However, when the full photometric SDSS sample is used (to $m_r = 22$), the photometric errors can in fact dominate the color box. Therefore, we have built in this mechanism at this stage. We note that we use the true SDSS errors when the algorithm is run on the real SDSS data. In the mock catalogs, we use the median observed error for galaxies in the SDSS main galaxy sample (see Strauss et al. 2002). Our results are insensitive to changes ($<50\%$) in the value of this constant error.

The mock galaxy catalog is complete down to an absolute magnitude of $M_r = -19.6$, which corresponds to $m_r = 19.2$ at $z = 0.17$ (which is the highest redshift cluster we find in the real data). We apply an apparent magnitude limit of $m_r \leq 17.7$ to the SDSS mock catalog to mimic the real magnitude-limited SDSS main galaxy sample (see Strauss et al. 2002). Once the magnitude limit is applied and the errors are added, we are able to run the exact same C4 algorithm on the mock galaxy catalogs and identify clusters as outlined in § 2. For this work, we have used a volume that is larger and more contiguous than the SDSS Second Data Release (DR2). However, edge effects are handled identically in the data as they are in the mock catalogs (see § 2.7.1). While we have not applied the SDSS targeting algorithm, in § 7.1 we study this issue in detail and find that the effect on completeness and purity is small.

4. COMPLETENESS, PURITY, AND TUNING THE C4 ALGORITHM

We use the SDSS mock galaxy catalogs to test the C4 algorithm, fine-tune the choice of parameters, and measure the completeness and purity of the catalog (i.e., the selection function). We do this by running the C4 algorithm on the mock galaxy catalog and comparing the found C4 clusters to the known halos from Evrard et al. (2002). To make the comparison, we apply a matching algorithm to associate C4 clusters with halos. We have investigated several prescriptions for matching these two data sets and have found that our matches are robust against the details of the matching algorithm. Here we present results based on matching a dark matter halo with any C4 cluster within a projected distance corresponding to one virial radius and within $\Delta z = 0.005$. We discuss this matching in more detail in § 4.2. To estimate purity, we match clusters to any simulated halo within the estimated r_{200} of the “observed” C4 cluster, while for the completeness measurements we match each “observed” C4 cluster to the nearest dark matter halo within $\Delta z = 0.005$ and the projected r_{200} of the halo.

This method for matching allows for multiple matches. In other words, when measuring completeness, multiple C4 clusters can be matched to one HV halo, and similarly, when measuring purity, multiple halos can be matched to a single C4 cluster. There are many ways to deal with this problem. For instance, when multiple halos match to one C4 cluster, we could take the most massive halo as the fiducial match; or, we could take the halo that has the most similar luminosity to the C4 cluster, or use any other method. We have chosen to simply take the match that is closest in separation on the sky (and within $\Delta z = 0.005$). We have investigated a few of the other methods we mentioned and find no clear winner. The C4 algorithm finds fewer clusters in the mock catalog than there are real HV halos (i.e., the C4 algorithm is never 100% complete). As seen and discussed in the following sections, this completeness drops with halo mass such that the C4 algorithm can miss up to 50% of the halos at masses $\sim 5 \times 10^{13} h^{-1} M_\odot$. This means that there will always be more multiple halo matches to the C4 clusters than vice versa. On average, 50% of the C4 clusters have multiple halos within $\Delta z = 0.005$ and r_{200} , while only 5% of halos have multiple C4 clusters within those same constraints.

After the matching is done, we plot the cumulative quantity:

$$\text{Purity}(L_r) = \frac{\text{Number}(>L_r) \text{ C4 Matched to Halos}}{\text{Number}(>L_r) \text{ C4 Clusters Found}}, \quad (4)$$

$$\text{Completeness}(M_{200}) = \frac{\text{Number}(>M_{200}) \text{ Halos Matched to C4}}{\text{Number}(>M_{200}) \text{ Total Halos}}, \quad (5)$$

where M_{200} is the mass within a radius that is 200 times the critical density and L_r is the summed luminosity of the cluster member galaxies as defined in detail in § 5. Since completeness is defined against the “true” halos from the mocks, we plot completeness versus halo mass. On the other hand, purity is measured from the point of view of the measured clustered catalog, and so purity is plotted against the observable optical luminosity. It is important to keep in mind that the high-mass (or high luminosity) systems are rare and so the purity and completeness measurements can be noisy in these regimes.

In Figure 6 (left), we present the completeness and purity of the mock C4 catalog as a function of different radii for search

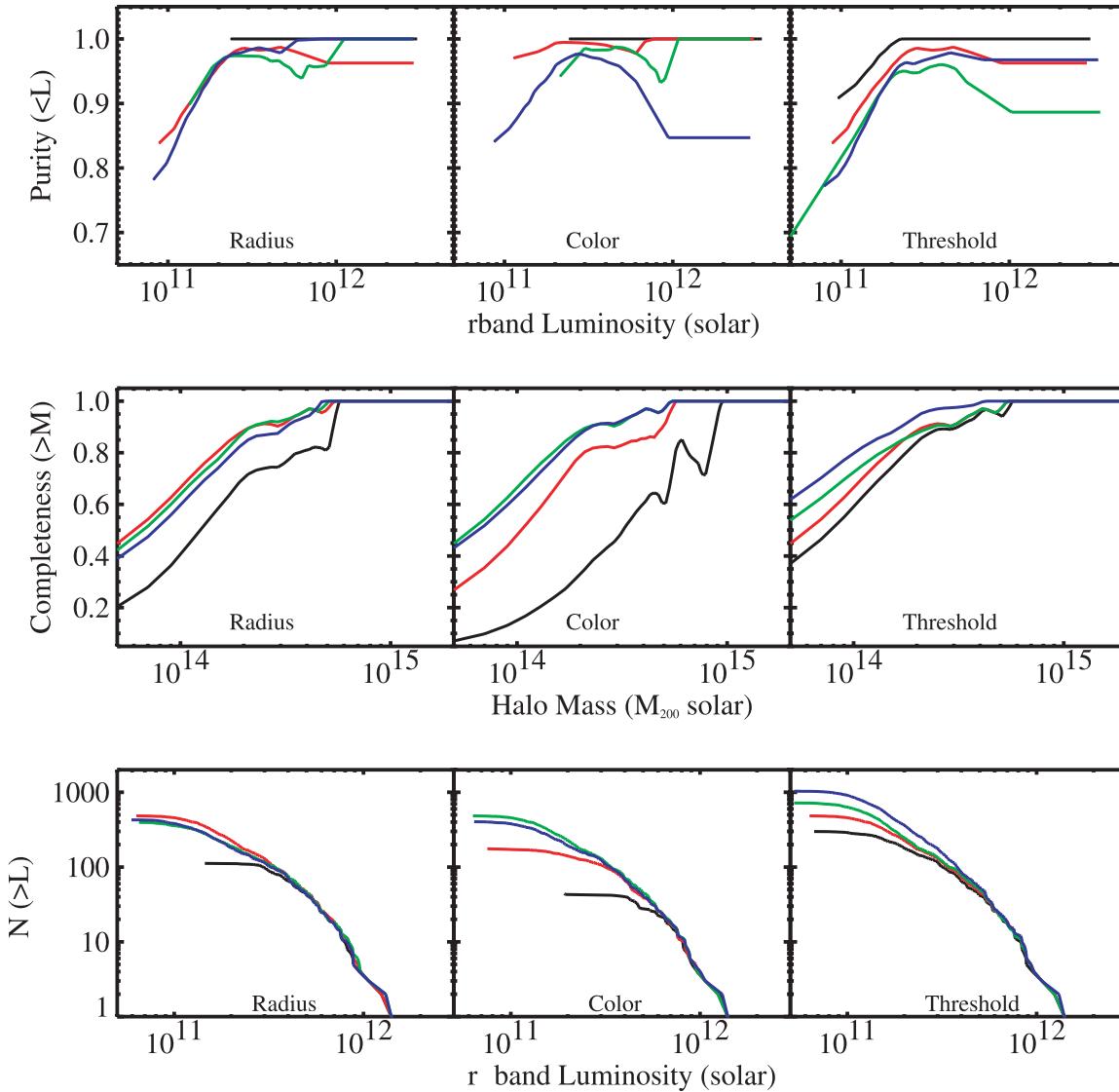


FIG. 6.—How the choice of box size, threshold, and redshift bin affects completeness and purity for clusters found using the simulated mock galaxy data. Also shown is the cumulative number plot [$N(>L)$] for the mock catalog clusters. The variation of the box size is described in § 2.1; the variation of the FDR threshold is described in § 2.5. Generally, the size of the box increases from black to red to green to blue. The purity and cumulative number counts are measured against the summed r -band luminosity of the discovered clusters, while the completeness is measured against the halo mass.

apertures of 500, 1000, 2000, and 6000 h^{-1} kpc. In this figure, the other dimensions of the search aperture are fixed at the final values as discussed below. A radius of 500 h^{-1} kpc (black line) appears to be too small, as it significantly lowers the completeness of our sample for all but the most massive systems (although it does produce the purest sample). However, larger search radii make little difference to the completeness or purity of the algorithm. The highest completeness and purity occur when a comoving radius of 1 h^{-1} Mpc is used (red line).

We varied the redshift dimension of the seven-dimensional box to be a comoving length of 25, 50, 100, and 200 h^{-1} Mpc. The size of the aperture in the redshift direction must be large enough to allow for significant (and unknown) peculiar velocities of galaxies within massive clusters of galaxies, and therefore, our three-dimensional positional aperture is shaped like a narrow cylinder. Using these tests, we find that our final cluster catalog is *independent* of the length of the line-of-sight aperture. We attribute this to the fact that there are not many clusters or groups that lie directly along the line of sight that also have simi-

lar global colors. Alternatively, one could argue that by not using k -corrections for our SDSS colors, we have already accounted for the redshift dimension in the “color box.” We set the redshift dimension of the search aperture to 50 h^{-1} Mpc.

In the middle panel of Figure 6, we show the completeness and purity for the mock SDSS catalog as a function of the “color box” size, holding constant the spatial part of the search aperture. We examine only the effect of changing $\sigma_{xy}(\text{sys})$, using $\sigma_{ug}(\text{sys}) = \gamma 0.15$, $\sigma_{gr}(\text{sys}) = \gamma 0.12$, $\sigma_{ri}(\text{sys}) = \gamma 0.1$, and $\sigma_{iz}(\text{sys}) = \gamma 0.1$. These values represent reasonable widths for the CMR, decrease with increasing wavelength (as indicated in Fig. 1), and are motivated by the results of Goto et al. (2002). However, we note that our algorithm is not attempting to model the CMR. Thus, we allow the color box size to be a free parameter in our algorithm by varying γ as 1, 2, 4, and 6. We then use the mock galaxy catalogs to fine-tune this variable. We note that the median of $\sigma_{xy}(\text{stat}) = 0.02$ for our data changes very little over our magnitude range (recall that these are the bright galaxies in the spectroscopic SDSS data). Thus, $\sigma_{xy}(\text{sys})$ is the dominant term in equation (4).

As seen in Figure 6 (*middle*), the smallest “color box” dimension produces a very pure but highly incomplete (*black line*) sample, as was the case for the smaller radial aperture. As we increase the size of the color box, we increase the completeness while decreasing the purity. For the final algorithm, we choose $\gamma = 4$, which has the highest completeness (for $M \geq 10^{14} M_{\odot}$ systems), while still maintaining a high level of purity.

In Figure 6 (*right*), we show the completeness and purity of the C4 sample as a function of the FDR threshold. We vary α from 0.05, 0.10, 0.15, 0.20, and 0.50. We note that our least conservative threshold ($\alpha = 0.5$) produces the highest completeness, but at the expense of purity. By lowering the FDR threshold, one simply increases the number of C4 galaxies being selected, but these extra galaxies either increase the detection likelihood of clusters already detected at higher FDR thresholds or form a background that decreases the purity. Changing the threshold by a factor of 4 only improves the C4 completeness for $M_{200} \leq 1 \times 10^{14} M_{\odot}$ systems by $\sim 10\%$, but at the price of decreasing the purity for such systems by 10%. Based on this, we choose $\alpha = 0.1$ (*red line*), which provides both a high purity and high completeness. This is preferred over maintaining a higher completeness (e.g., $\alpha = 0.5$), as it gives users of the C4 catalog the confidence to pick and choose real clusters for any scientific analyses. Also, as we discuss in § 4.2, gains in the measured completeness are as much a result of random matches as they are of a more efficient algorithm. When this final threshold is applied, approximately 90% of all galaxies are excluded as not being in color and spatially clustered environments.

In Figure 7, we show purity, completeness, and number for clusters in shells of equal volume, increasing in redshift from black to red to green to blue to violet, covering redshift ranges of $[0.03, 0.075]$, $[0.075, 0.093]$, $[0.093, 0.107]$, $[0.107, 0.118]$, and $[0.118, 0.128]$, respectively. As with Figure 6, these panels use the mock galaxy catalogs. However, unlike Figure 6, whose clusters numbered in the many hundreds, these smaller volume bins contain ~ 100 –200 clusters and so the results are noisier. As expected, completeness decreases with increasing redshift but varies little out to $z = 0.107$ for all masses. Beyond that, completeness drops steeply. Purity is fairly constant (to within 10%) over all redshift ranges; however, the lowest redshift bin is the purest. The C4 catalog is $>90\%$ complete and $>95\%$ pure for systems more massive than $\sim 2 \times 10^{14} M_{\odot}$ (or brighter than $\sim 3 \times 10^{11} L_{\odot}$) and out to a redshift of $z \sim 0.12$. In Figure 7 (*bottom*), we see that the number function is mostly dependent on the completeness in each redshift shell. The most complete bin (lowest redshift) shows the highest number of low-luminosity (or mass) clusters. As completeness dwindles with redshift (and mass), the number of found halos decreases similarly. The excess of $N(>L)$ in Figure 7 (*bottom*) for the highest redshift bin (violet) is due to a single bright, massive halo in the simulations.

4.1. The Strength of Color Clustering

We have run a series of tests to determine whether our choice of all four colors is necessary for our stated goals (high purity and known completeness) compared to using just a subset of these colors. Specifically, we ran the C4 algorithm using each of the four colors separately, as well as using subsets of the colors, e.g., $g - r$ and $r - i$ but not $u - g$ or $i - z$.

In Figure 8 (*left*), we show how the purity of the C4 catalog changes as we add more color information. The highest purity comes from using all four colors. As expected, the reddest color selection, $i - z$, does well (even though the z -band magnitudes have greater photometric uncertainties). Combining two colors does

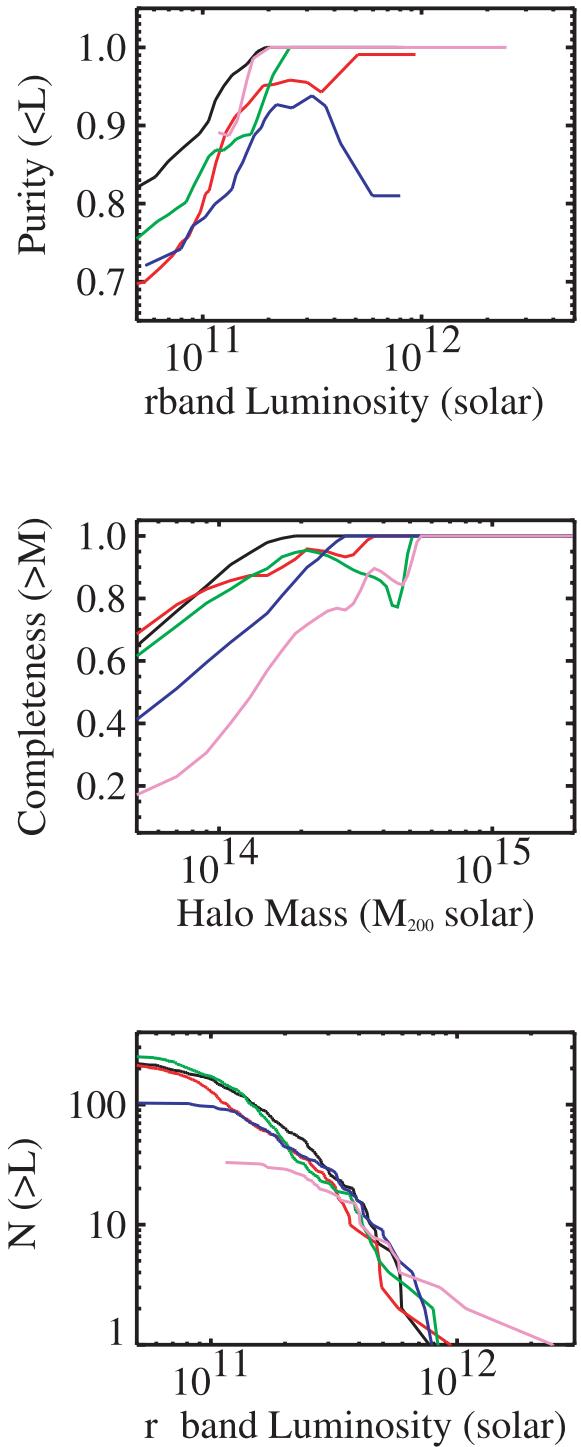


FIG. 7.—Completeness, purity, and number of halos brighter than L_r for equal-volume shells, each at increasing redshift. The shells are described in the text. They go from the lowest to the highest redshift as black, red, green, blue, and violet.

reasonably well and is better than only using a single color. We conclude that the choice of four colors gives us our highest purity.

We next examine the completeness as a function of our color choices. In Figure 8 (*right*), we present completeness as a function of halo mass for the same color selections as used above in the purity case (*solid lines*). The highest purity case (using all four colors) results in the lowest completeness. At first glance, it

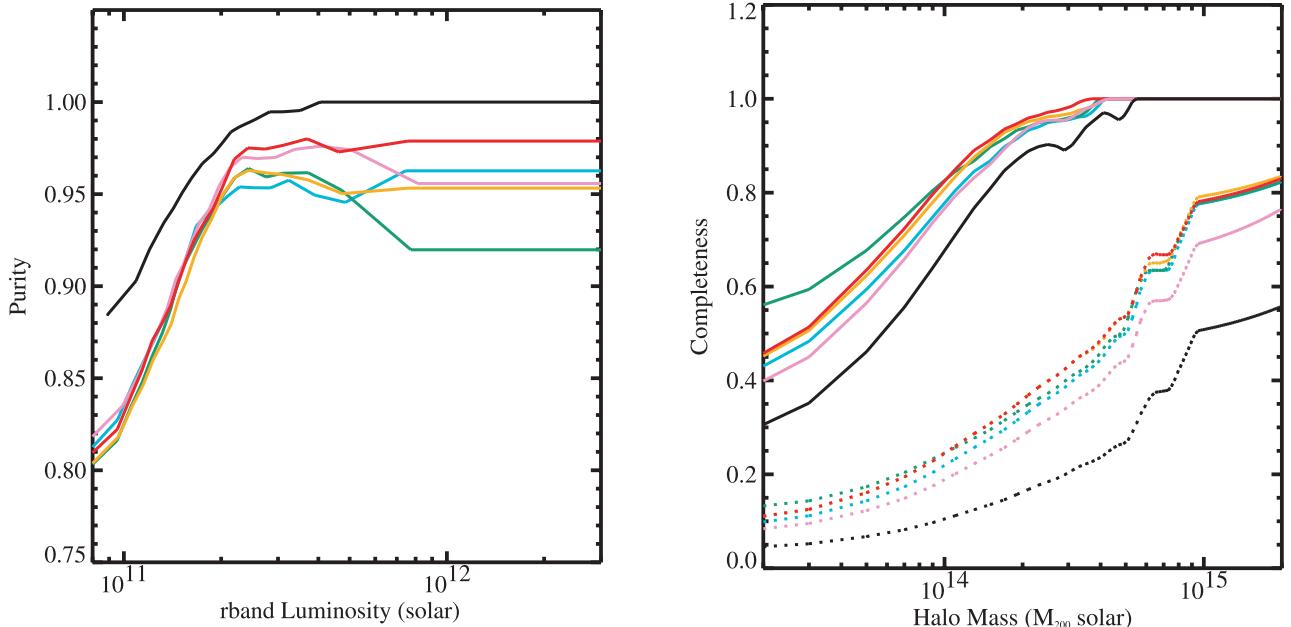


FIG. 8.—Left: Purity of the C4 catalog as a function of different color selections. The black line (highest purity) comes from using all four colors. In both panels, the green, turquoise, orange, and red lines correspond to using only $u' - g'$, $g' - r'$, $r' - i'$, and $i' - z'$, respectively. The purple line corresponds to using both $g' - r'$ and $r' - i'$ only. Right: Measured completeness of clusters found using the C4 algorithm (solid lines). The dotted lines indicate the completeness measured for a random selection of clusters of the same sample size as found by the algorithm. The black line (the smallest sample with the lowest true and random completeness) comes from using all four colors.

appears that the use of single- or double-color criteria produces better results, with 10% higher completeness than the four-color case. This is, however, misleading: the higher formal completeness is in fact entirely due to random matches (Fig. 8, *dotted lines*). For example, at $M = 1 \times 10^{14} M_\odot$, the completeness increases from $\sim 70\%$ for the four-color criteria to 80% for the single $i - z$ color selection. At the same time, the random matches (described below) increase by 10% and the purity decreases by $\sim 10\%$. In other words, the single- and double-color criteria have approximately twice as many detected “clusters” as the four-color criteria, producing a much greater chance for random matches. Of course, as seen in Figure 8, a larger fraction of these detected “clusters” are spurious. We discuss the random matches in more detail in § 4.2.

As a final test, we should mention here that we experimented with shuffling the colors of the galaxies in the mock catalogs, while keeping their positions fixed, and reran the C4 algorithm. We found only a few of the closest richest clusters, which again demonstrates the power of color clustering in four dimensions.

4.2. Random Matches

These first results raise the issue of the number of random matches one would expect given *any* sample. We quantify random matches by selecting the same number of clusters as found by the C4 algorithm but centered on random galaxies in the mock catalog. For example, if we find 934 clusters in the mock catalog using the algorithm, we select 934 galaxies at random from the same mock catalog and use them as our cluster centers. We then match these with the dark matter halo catalog using the same criteria as before. We show the “completeness” from a sample of randomly placed cluster centers as the dotted line in Figure 8 (*right*). As expected, the “completeness” of the random matches monotonically increases with cluster mass because the number of clusters as a function of mass monotonically decreases, while the number of matches remains fixed. In other words, for a

fixed number of random positions, a greater fraction of rare rich clusters is recovered compared to the numerous poor clusters.

This conclusion is as much a statement of our matching criteria as it is of one’s ability to randomly find clusters. What does not appear in this analysis is the scatter in the cluster observables at fixed halo mass due to accidental (i.e., random) matches. In Figure 9, we show how well we recover the halo observables after we find C4 clusters in the mock catalogs and match to the halos. We show the difference between the recovered and the “true” summed optical luminosities and richnesses (see § 5). These figures show that we recover the true observables to typically within 20%.

Inherent in these analyses is our ability to match halos to the C4 clusters. Keep in mind that halos from simulations are themselves messy, nonspherical systems whose boundaries are dependent on the exact identification algorithm (Lacey & Cole 1994; White 2002). The halo sample we employ is based on a spherical overdensity approach. Details of the finding algorithm and the resultant halo samples are published in Evrard et al. (2002). A more detailed exploration of matching clusters to dark matter halos is presented in W05.

4.3. The Final C4 Algorithm Parameters

The parameters of our final algorithm are (1) an aperture on the sky corresponding to $1 h^{-1}$ Mpc projected at the redshift of the target galaxy; (2) a redshift box corresponding to a fixed comoving $\pm 50 h^{-1}$ Mpc around the target galaxy; and (3) a four-dimensional color box of width specified by equation (5) and

$$[\sigma_{ug}^{\text{sys}}, \sigma_{gr}^{\text{sys}}, \sigma_{ri}^{\text{sys}}, \sigma_{iz}^{\text{sys}}] = [0.6, 0.48, 0.4, 0.4]. \quad (6)$$

We apply a probability threshold that results in no more than 10% contamination.

These parameters have been tuned to produce a cluster catalog with the highest possible purity and similarly high completeness.

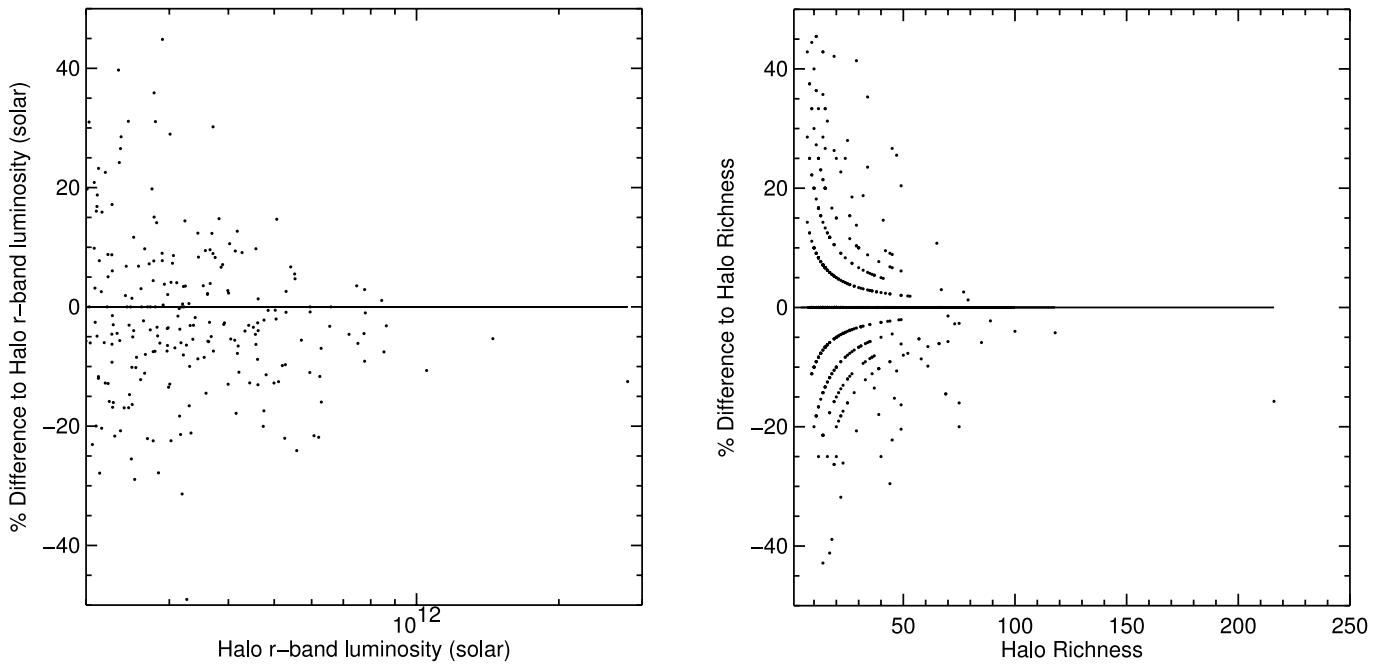


FIG. 9.—*Left:* Percentage difference between the measured cluster optical luminosity and the “true” halo luminosity. *Right:* Same as the left panel, but for richness. The method of measuring luminosities and richness is described in § 5. For the C4 clusters, we use the centroid of the found cluster, while for the halos we use the centroid as reported in the halo catalog.

We note that Figure 6 shows that the measured purity and completeness are very robust to modest changes in the tunable parameters. The algorithm is demonstrably robust.

4.4. Summary of C4 Catalog Purity and Completeness

In Figure 10 we present the final purity and completeness of the C4 catalog based on our optimal parameter choices as discussed above (over all redshifts). Recall, purity is defined as the percentage of systems detected in the mock SDSS catalog, using the C4 algorithm, and matched to any dark matter halo (more massive than 4.5×10^{13}) in the HV simulation. We also measure the purity as a function of velocity dispersion, using only those clusters that contain 10 or more galaxies. We find that our C4 catalog is 100% pure for such systems and thus do not present this result in a figure.

As one can see in Figure 10 (*top, solid line*), the purity of the C4 sample remains at 100% for the most massive systems, dropping to $\sim 90\%$ for the remainder. The C4 catalog is more than 99% pure for luminosities larger than $3 \times 10^{11} L_\odot$. The high purity of the C4 catalog is a direct product of our search for clusters in a high-dimensional space.

In Figure 10 (*bottom, solid line*), we also show the completeness of the C4 algorithm as a function of halo mass (M_{200}), as selected from the mock SDSS catalog. This figure demonstrates that the C4 catalog remains more than 90% complete for systems with $M_{200} \gtrsim 2 \times 10^{14} M_\odot$. Below this mass the catalog becomes progressively more incomplete and is only 55% complete at $M_{200} \simeq 1 \times 10^{14} M_\odot$. The completeness is only 30% for the lowest mass systems probed here ($M_{200} \simeq 2 \times 10^{13} M_\odot$).

4.5. Questions about the C4 Methodology

We address here three common questions raised about the C4 approach. These are:

1. Why focus on the photometric data in the C4 algorithm, when the redshifts (i.e., the three-dimensional positions in redshift space) are known?

2. Why use all four SDSS colors ($u - g$, $g - r$, $r - i$, and $i - z$)? Why not use the spectra of the galaxies instead of the broadband filters?

3. Does the algorithm miss clusters with younger stellar populations?

In Figure 11, we address the first question and demonstrate the power of using the color information in addition to the spatial coordinates. Here we show the projection of the SDSS seven-dimensional search aperture (four colors and one spatial coordinate) onto the different color-color planes for both a cluster and a field region. In Figure 11, we have placed the same size of physical aperture over two galaxies: one in a clustered environment (Fig. 11, *left panels*) and the other in a field like environment (Fig. 11, *right panels*). The galaxy on which the color-color plots are centered is the target galaxy and is identified by the open circle.

The blue circles are all galaxies within the spatial part of the search aperture. Visually, one might be able to detect the spatial clustering by noticing that there are more galaxies in the clustered versus the field environment (388 vs. 327). The red circles are galaxies that lie within the color box in all three figures. Note that this “color box” (*green box*) is the same, in location and size, for both the cluster and the field environments. The overdensity of galaxies in the clustered environment now becomes much more apparent—there are 19 cluster galaxies (*red circles*) that have both similar positions and similar colors to the target galaxy, while there remains only one galaxy (with similar colors and position) in the fieldlike environment. This process increases the signal-to-noise ratio of the cluster overdensity (compared to the field overdensity) from 388/327 to 19/1, so that the slight overdensity in the three-dimensional position space becomes an extreme overdensity in the much sparser seven-dimensional data space used here. Figure 11 demonstrates the elimination of projection effects and the strength of color in galaxy clustering. It also demonstrates the enhancement of the overdensities one can achieve by using positions *and* colors in our clustering algorithm.

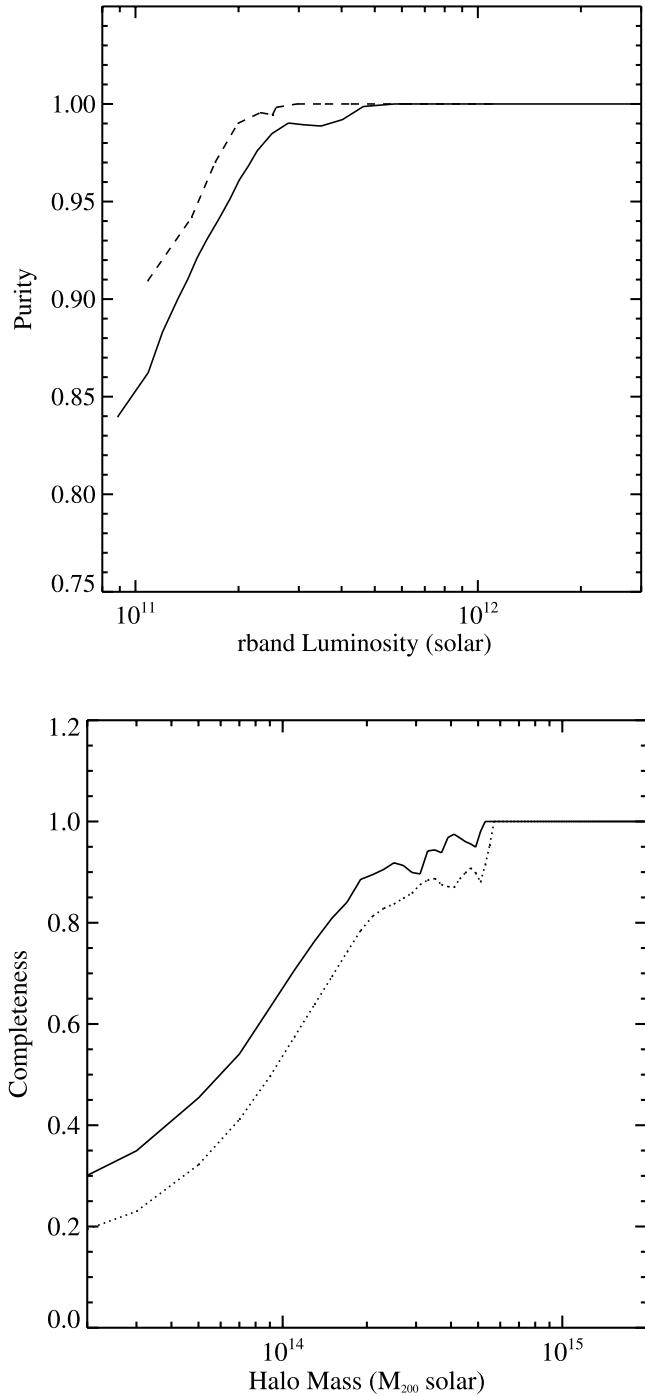


FIG. 10.—Final measured purity and completeness of our C4 catalog using the mock SDSS catalog. The solid line is measured before the fiber collision algorithm is applied (see text). The dotted line shows the effect of omitting galaxies because of fiber collisions.

To address the second question, we remind the reader of our analysis in § 4.1, where we show that four colors do better than just one or two. We also stress that there are physical reasons for wanting to use all four SDSS colors. For example, the u -band is an excellent measurement of recent star formation in galaxies (see Hopkins et al. 2003) and is below the D4000 feature. Therefore, the $u - g$ color allows us to discriminate between star-forming and passive galaxies; this is demonstrated in Figure 1, in which we see a large color difference ($\simeq 1$ mag) between galaxies with and without strong H α emission in the $u - g$ color-magnitude plot (*top left*). The g , r , and i passbands are the most sensitive

photometric passbands available and therefore have the smallest photometric errors, which results in a tight “red sequence” in the $g - r$, $r - i$ color-color plane (see Fig. 1). Finally, the z -passband is useful, as it provides the best measurement of the old stellar population for these low-redshift galaxies and is the least affected by Galactic reddening. The larger errors on the z -band photometry do not compromise the C4 algorithm, as we take the observed errors into account when constructing the seven-dimensional search aperture.

With regard to using the spectra instead of the SDSS colors, we note that the five SDSS passbands (u , g , r , i , and z) cover a larger wavelength range than the spectra. The central wavelengths of the SDSS photometric filters are 3550, 4770, 6230, 7620, and 9130 Å, respectively, covering a wavelength range from $\simeq 3300$ Å to 1 μm. In comparison, the spectra only cover a wavelength range of 3900–9100 Å. From a computational standpoint, using the spectral data instead of the photometric colors would require working in a many thousand-dimensional data space. Even with a million galaxy spectra, such a high-dimensional data space would be severely underpopulated, leading to statistical problems in finding any clustering in the data. In summary, the seven-dimensional data space discussed herein is very effective for our task of finding clusters and groups of galaxies, as the dimensionality is sufficient to eradicate projection effects while remaining manageable in size.

Perhaps more importantly, the ability to create mock catalogs with color clustering is currently a challenge (which we think has been met by the catalogs used herein). We are still a long way from having mock catalogs that have galaxies with synthesized spectra that match the environmental trends seen in the data. Thus, the colors allow us to achieve our goals of maximizing completeness and eliminating projection effects, as tested against the mock galaxy catalogs.

To address the final question, in Figure 12 we present two groups of galaxies found by the C4 algorithm that possess very different galaxy properties. The first group of galaxies in Figure 12 (*top*) contains galaxies that appear redder and more elliptical-like—as expected in a typical group of galaxies. The second group in Figure 12 (*bottom*) contains much bluer, more disklike galaxies. These groups were both detected as overdensities of galaxies with similar positions and colors (in $u - g$, $g - r$, $r - i$, and $i - z$), thus demonstrating that the C4 algorithm does not exclude systems dominated with younger stellar populations. The details of the galaxies in these two groups are presented in Table 1. Likewise, in Figure 21 we show a cluster comprising mostly blue star-forming galaxies at redshift $z = 0.11$. Note the lack of any prominent E/S0 ridgeline in this system. As discussed in § 2, our algorithm is only insensitive to systems that would contain spatially clustered galaxies that cover a broad range of spectral types. However, galaxy types are not broadly classified but bimodal (spirals or elliptical, star forming or passive). Thus, to first order, every cluster will contain at least 50% of one of the two major types of galaxies, and the algorithm will find such color clustering.

5. MEASURED CLUSTER PROPERTIES

For each C4 cluster we measure a set of quantities that includes the cluster centroid, the velocity dispersion, and the summed r -band luminosity. In addition, we characterize the substructure and local large-scale structure of each cluster.

5.1. Cluster Centroids

We measure three different cluster centroids: (1) the peak in the C4 density field, (2) the luminosity-weighted mean centroid,

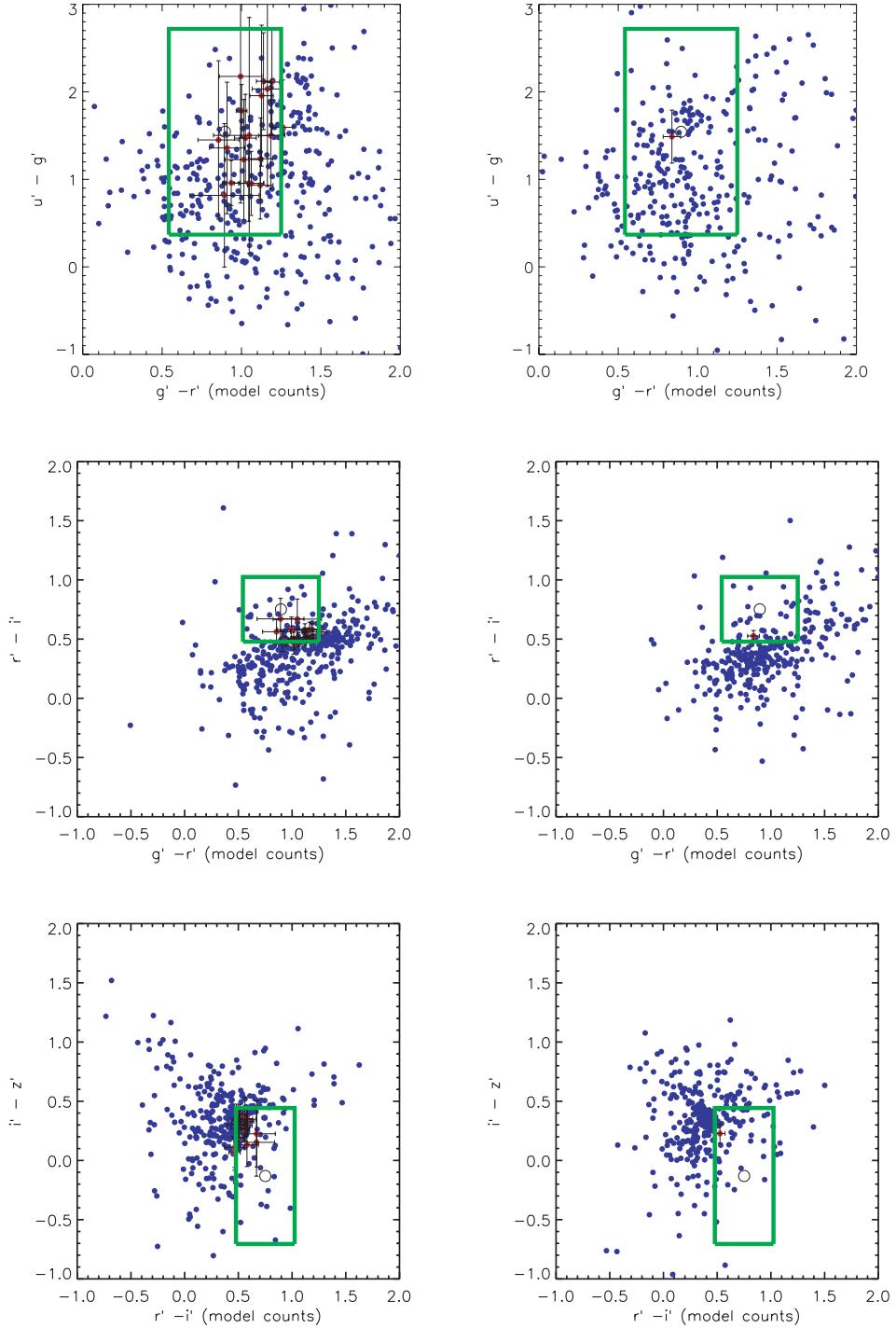


FIG. 11.—Different SDSS color-color planes for an example cluster galaxy (*left*) and a randomly chosen field galaxy (*right*). The blue circles show all galaxies within the spatial part of our search box (right ascension, declination, redshift), while the red circles show those galaxies that are also within the color part of our seven-dimensional search aperture (green rectangles).

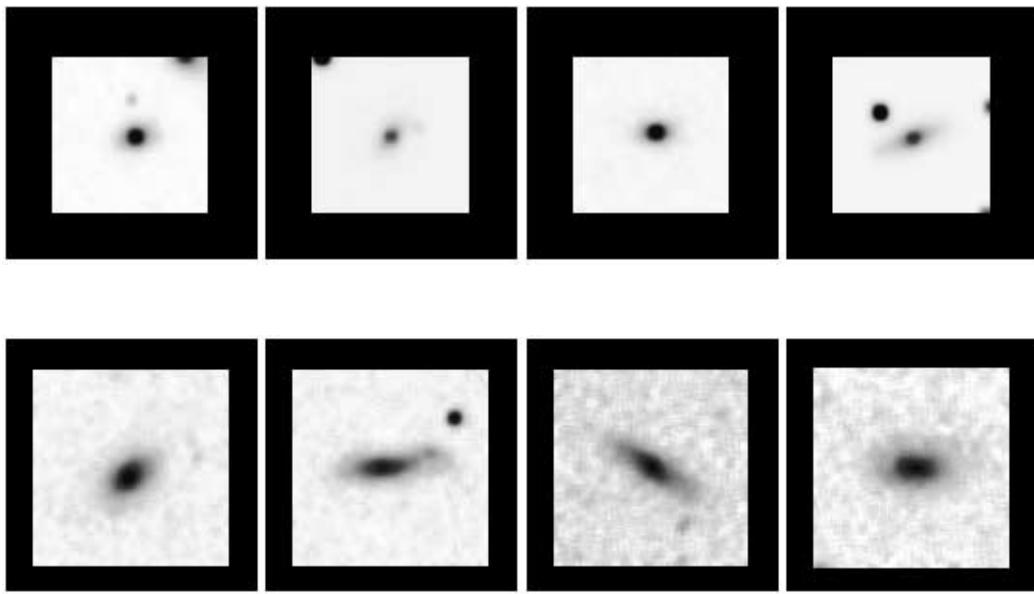


FIG. 12.—Images of two sets of four galaxies that are clustered in both position and color. The top four are elliptical galaxies at a redshift of ~ 0.027 , and all lie within a projected distance of $350 h^{-1}$ kpc. The bottom four are galaxies that are more disklike and have a younger stellar population (i.e., they are bluer). The bottom four galaxies are at a redshift of ~ 0.04 , and all lie within a circle that is $450 h^{-1}$ kpc in radius. The positions and colors of the galaxies in each of the sets are all very similar and are listed in Table 1. [See the electronic edition of the Journal for a color version of this figure.]

and (3) the position of the brightest cluster galaxy (BCG). Method 1 was discussed in § 2.6. Method 2 uses all galaxies within $1 h^{-1}$ Mpc of the initial centroid (method 1) and within four velocity dispersions (see § 5.2). We then calculate an r -band luminosity weighted center. Method 3 attempts to identify the BCG. The BCG is taken as the brightest galaxy within $500 h^{-1}$ kpc of the initial centroid (method 1), within four velocity dispersions (see § 5.2), that has no strong $H\alpha$ emission ($< 4 \text{ \AA}$). We then report the position of the BCG.

The cluster redshift we report is determined via the biweighted statistic of Beers et al. (1990). We use only those galaxies within $1 h^{-1}$ Mpc of the initial centroid and within $\Delta z = 0.02$ of the peak in the redshift histogram as described in § 5.2.

5.2. Velocity Dispersions

To calculate the velocity dispersion of galaxies in our clusters, we perform an iterative technique based on the robust biweighted statistic of Beers et al. (1990). Having defined the sky-

positional centroid of each cluster, we construct a redshift histogram of all galaxies (regardless of their C4 classification) within a projected radius of $1 h^{-1}$ Mpc. We then search this histogram for a peak and tentatively identify this peak as the velocity center of the cluster. (As a lower limit, this peak must contain at least three galaxies within 1000 km s^{-1} of each other.) We iterate by only keeping galaxies within a projected $1.5 h^{-1}$ Mpc of the cluster centroid and within $\Delta z = 0.02$ of the velocity center defined above. We then compute the biweighted mean recessional velocity for these galaxies and measure their biweighted velocity dispersion, σ_v^{est} .

We stress that the above procedure is performed on all available galaxies in the SDSS spectroscopic sample. For consistency, we check that each cluster contains a certain number of C4 galaxies and reject those clusters for which this fraction is below 10% of all galaxies within $1 h^{-1}$ Mpc and $\Delta z = 0.02$. Only a small fraction (2%) of clusters originally identified using the C4 galaxies fail this test and are not included in the final cluster sample. We also reject a small number of clusters that do not contain enough galaxies to measure an accurate velocity dispersion. At least eight galaxies are required to define the velocity dispersion, consistent with previous limits to get a reliable value (see Collins et al. 1995). Note that most (80%) of our clusters in the real SDSS data contain ≥ 20 galaxies with which to measure the velocity dispersion. Clusters that do not meet this criterion are removed from the main sample. To get our final velocity dispersion measurements (σ_v), we recalculate it for each cluster using only galaxies within 4 times the estimated velocity dispersion (σ_v^{est}) discussed above. This is similar to the standard sigma-clipping method used in the literature. The accuracy of these final measurements, which are in the observed reference frame, are discussed below.

5.3. Summed Optical Cluster Luminosity

To calculate the total summed r -band optical luminosity for each cluster, we convert the apparent magnitudes of all cluster members into optical luminosities, using the conversions

TABLE 1
GALAXY POSITION AND COLORS IN FIGURE 12

R.A.	Decl.	Redshift	$u' - g'$	$g' - r'$	$r' - i'$	$i' - z'$
Red Galaxies (Fig. 12, top)						
219.409	3.946	0.025	5.01	0.67	0.31	0.20
219.481	3.984	0.029	5.38	0.75	0.37	0.22
219.778	3.999	0.027	4.00	0.72	0.37	0.22
219.887	3.925	0.029	4.86	0.56	0.22	0.14
Blue Galaxies (Fig. 12, bottom)						
228.312	4.513	0.036	0.89	0.34	0.43	0.29
228.445	4.251	0.041	1.19	0.32	0.42	0.00
228.574	4.064	0.042	1.13	0.29	0.45	0.40
228.277	4.195	0.037	1.09	0.45	0.40	0.12

NOTE.—Units of right ascension and declination are degrees.

in Fukugita et al. (1996), and sum them. All magnitudes are k -corrected according to Blanton et al. (2003b) and extinction-corrected according to Schlegel et al. (1998). Cluster membership is confined to galaxies within $4 \sigma_v$ in redshift space and a projected radius of $1.5 h^{-1}$ Mpc on the sky. The SDSS main galaxy sample is designed to be complete ($>95\%$) to $m_r = 17.77$ and the C4 cluster sample is complete ($>90\%$) to $z = 0.11$. Thus, to minimize effects from the SDSS selection function, we use an absolute magnitude limit of $M_r < -19.9$, which is an apparent magnitude of $m_r = 17.8$ at $z \simeq 0.11$, when measuring the optical luminosities. Clusters beyond $z = 0.11$ will need to have their optical luminosities corrected for this incompleteness.

5.4. Structure Contamination Flag

We define a “structure contamination flag” (SCF) to measure the degree of isolation in redshift space for each cluster. Specifically, we examine radial variations of the velocity dispersion for each cluster, noting large radial variations from the mean velocity dispersion. Clusters that are embedded in surrounding large-scale structure can have significant velocity contamination. SCF increases with increasing standard deviations in the velocity dispersion profiles. We calculate the biweight velocity dispersion within $500, 1000, 1500, 2000$, and $2500 h^{-1}$ kpc radial bins, as described above, and determine the standard deviation. We then assign an SCF based on the ratio of the standard deviation of the dispersions over the mean of the velocity dispersions. We use three bins, $SCF = [0, 2]$, in approximating the bottom, middle, and top thirds of the distribution of the ratio. A cluster with $SCF = 0$ has a ratio of less than 15%, whereas $SCF = 2$ has a ratio $>30\%$. The sky plot and velocity profile are shown for a real SDSS cluster with a high $SCF = 2$ cluster in Figures 20, 22, and 23. Note that the velocity dispersion as a function of radius is highly erratic, producing a large standard deviation about the mean. Figures 22 and 23 show two clusters separated by less than 0.2 and $\Delta z = 0.01$. Note that the velocity dispersion profile increases systematically as the galaxies from the neighboring system are picked up with increasing aperture. These clusters both have $SCF = 2$. A cluster with $SCF = 1$ is shown in Figure 21 and one with $SCF = 0$ is shown in Figure 24. Note here that the velocity dispersion profiles are nearly constant. This SCF does not necessarily quantify the substructure of the main cluster but rather identifies clusters whose velocity dispersions may be inaccurate because of nearby large-scale structure.

5.5. Dressler-Shectman Statistic

In addition to quantifying the local large-scale structure for each cluster, we can use the Dressler-Shectman (DS) substructure statistic to search for local differences in a cluster’s mean recession velocity and velocity dispersion. For each cluster member (within $1.5 h^{-1}$ Mpc and $4 \sigma_v$), we compute a local recession velocity and local velocity dispersion using the 10 nearest neighbors to the galaxy that are also within $4 \sigma_v$ of the cluster redshift. We then compute the difference between these local quantities and the mean recession velocity and velocity dispersion for the whole cluster. The cumulative differences are then used as a measure of the cluster substructure. To compute the significance of this measurement, we shuffle the galaxy velocities within the cluster and repeat the exercise 1000 times. Using these Monte Carlo realizations, we calculate the probability that the observed cumulative differences would be obtained at random, given the spatial positions of the galaxies. A low probability indicates that the substructure is significant. For

more details, see Dressler & Shectman (1988) and Oegerle & Hill (2001).

6. SCALING RELATIONS AND THEIR SCATTER

Any galaxy cluster catalog will have observables that can be related to the underlying halo dark matter mass. Typically, researchers have used some sort of galaxy number count, i.e., richness (Abell 1958; Yee & Ellingson 2003). While the C4 clusters certainly have galaxy number counts, we also measure the summed r -band luminosity of the galaxies within each cluster. We also measure a velocity dispersion for all clusters within our spectroscopic data (using eight or more galaxy members within a projected radius of $1 h^{-1}$ Mpc and within four velocity dispersions). In this section, we determine which cluster observables scale best with the halo masses in the mock galaxy catalogs. We stress that this section does not *quantify* any absolute scaling laws (or their scatter), which requires a detailed analysis of the role of cosmology and of the sensitivity of this scaling to the halo occupation. This will be presented in an upcoming paper. Here we simply study how the scatter changes as we vary parameters of the cluster-finding algorithm and measures of the local foreground/background contamination. In short, the scaling relations presented in this section are used solely to guide our choice of the best cluster observable when relating to mass, as qualified by the scatter, and we caution the reader not to use them to draw cosmologically dependent conclusions.

6.1. Structure Contamination and Velocity Dispersion

For virialized systems, the velocity dispersion is an obvious choice when attempting to measure the mass of a cluster. So first, we examine the validity of our method to recover the velocity dispersions. In Figure 13, we compare our measured velocity dispersions for C4 clusters in the mock catalog against the known particle line-of-sight velocity dispersions for the halos in the simulations. We use only clusters with 10 or more galaxies within $1.0 h^{-1}$ Mpc and within $4 \sigma_v$. We find an excellent one-to-one agreement with $\sim 15\%$ scatter (using only those mock clusters with $SCF = 0$). This scatter doubles if we use all clusters. Therefore, the SCF is essential in recovering the true velocity dispersions of the systems. We also show the comparison when we use only the C4 galaxies to measure the velocity dispersion. Note that the C4 galaxies are not strongly affected by local structure contamination. However, there are fewer systems with enough (>10) galaxies to measure an accurate dispersion.

In Figure 14 we present the correlation between halo mass and the measured velocity dispersions of C4 clusters in the mock SDSS catalog. We present the best (robust fit) linear relationship between these two physical quantities. There is also significant asymmetrical scatter about this relationship, with many low-mass systems possessing an apparently high velocity dispersion. Part of this asymmetry and scatter is due to the number of galaxies used to measure the velocity dispersion. In Figure 14 (*middle*) we split the sample into clusters with fewer than 25 members and more than 25 members within $1.5 h^{-1}$ Mpc. As expected, clusters with a larger number of members have higher mass halos, but we also observe that the scatter decreases by $\sim 50\%$ for the regression when only the high-number systems are used. Unfortunately, if we place a cut on the number of cluster members to reduce the scatter in this scaling law, we also constrain ourselves to only the higher mass systems.

As an alternative to using only clusters with the most galaxies for accurate velocity dispersions, we plot in Figure 14 (*top row*) the relationship between M_{200} and velocity dispersion, but now

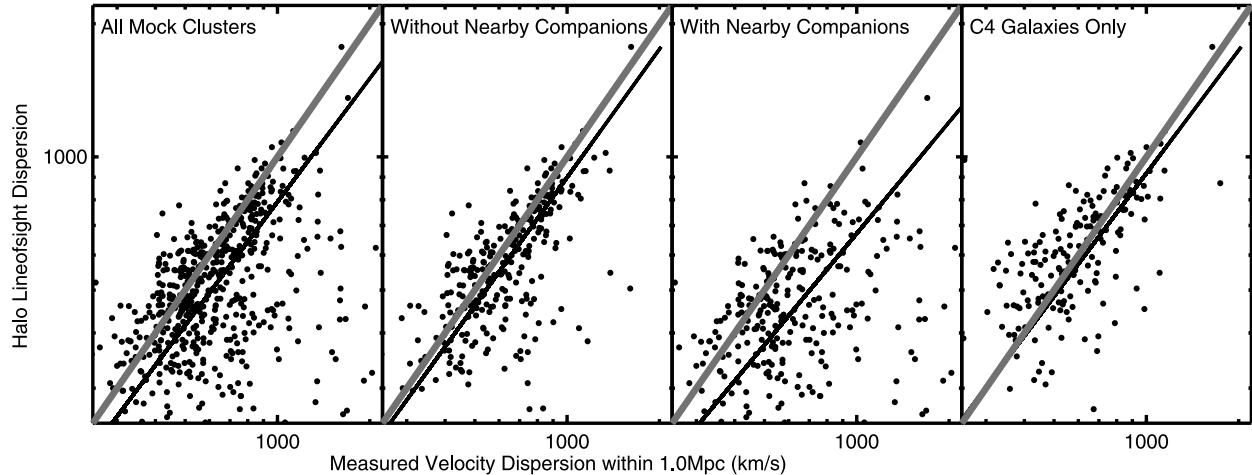


FIG. 13.—Comparison using the mock clusters of our measured velocity dispersions to the line-of-sight velocity dispersions measured directly on the halo particles in the HV simulations. The thick gray line is the one-to-one line, while the black line is the best fit when SCF = 0. We use only clusters with 10 or more members within a projected radius of $1 h^{-1}$ Mpc and within $4 \sigma_v$. The best-fit relation occurs when SCF = 0, where the rms scatter is $\sim 15\%$.

separated on the SCF. The scatter in the relation is reduced by a factor of 2 after we include only those clusters with SCF = 0. In addition, the scatter is reduced over the whole mass range of the clusters. This demonstrates that a majority of the scatter in Figure 14 is due to SCF > 0 systems, which by definition have a known companion within $1.5 h^{-1}$ Mpc $^{-1}$ of the main cluster. The end result is that velocity dispersions are typically overestimated when there is nearby large-scale structures. It may further indicate that these clusters are not virialized systems due to a recent, or ongoing, merging event. Figure 14 (top) also shows that for the SCF = 0 systems, one can obtain a tight, linear relationship between M_{200} and velocity dispersion.

Finally, we also show in Figure 14 (bottom row) how the DS statistic alters the scatter in the relation between halo mass and velocity dispersion. Here we only use those clusters with 25 or more member galaxies to ensure that the DS statistic is not dominated by Poisson noise. The DS statistic does not seem to help reduce the scatter in the relation. Since the DS statistic is really trying to measure internal velocity substructure in clusters, this result may imply that the substructure is not strongly correlated with mass. Note that the SCF tends to find structure *outside* the cluster, since it looks for variations in the velocity dispersion in increasing radial bins (all the way out to 2.5 Mpc beyond a cluster's center). Thus, the SCF and the DS statistic are mapping contamination in different ways.

6.2. Structure Contamination and Richness

Richness (or galaxy member counts) is an often used measure of cluster mass (see Yee & Ellingson 2003 and references therein). We examine multiple richness measures, including counts within metric radii (500, 1000, 1500, 2000, and $2500 h^{-1}$ kpc) and counts of only C4 galaxies (all to an absolute magnitude limit of $M_r = -19.8$ or L^*). We find the richness measured within 1 or $1.5 h^{-1}$ Mpc to have the lowest scatter against halo mass. Since we are working with spectroscopic data (or mock spectroscopic data), we also apply a constraint that members be within ± 4 times the cluster velocity dispersion. Thus, our richnesses are not affected by foreground/background projection, which is typically the case in optical cluster catalogs. In Figure 15, we show how halo mass scales with richness in the mock catalogs. Note that the scatter in mass versus richness is comparable to that seen in mass versus velocity dispersion (Fig. 14, top) when SCF = 0. It is also

worth noting that when the SCF is ignored, the scatter increases by a third.

6.3. Structure Contamination and Summed Optical Luminosity

While richness (as defined here) is a fixed, integer quantity, the summed *r*-band optical luminosity of a cluster takes into account any environmental dependence in luminosity, even by galaxy type (Hogg et al. 2003; Baldry et al. 2004). We examine multiple luminosity measures, including those within metric radii (500, 1000, 1500, 2000, and $2500 h^{-1}$ kpc), using galaxies brighter than $M_r = -19.8$ (L^*). We find the cluster luminosity measured within 1 or $1.5 h^{-1}$ Mpc to have the lowest scatter against halo mass. Since we are working with spectroscopic data (or mock spectroscopic data), we also apply a constraint that members be within ± 4 times the cluster velocity dispersion.

In Figure 16 (top), we show the halo mass as a function of the summed cluster *r*-band luminosity (within $1 h^{-1}$ Mpc to $M_r = -19.8$), again separated by the value of the SCF. We note here that the scatter on the relation only drops by $\sim 10\%$ after applying the SCF cut, demonstrating that the summed luminosity as a proxy to cluster mass is less contaminated by nearby large-scale structure than velocity dispersion. We also note that the scatter is smaller for mass versus luminosity than it is for either mass richness or mass-velocity dispersion. Thus, we conclude that, in the mock catalogs, the summed *r*-band luminosity is the best measure of the cluster dark matter halo mass.

Figure 17 shows the halo mass as a function of the summed *r*-band cluster luminosity with a variety of best-fit relations and their scatter. We observe a linear correlation between mass and optical cluster luminosity and the best robust-fit linear relationships (mass vs. luminosity and luminosity vs. mass) are shown (light-gray solid and dashed lines, respectively). One might note that the relation is not precisely linear, and so we have also used a nonparametric technique to more accurately trace this scaling relation (black solid line). We plot a 1σ confidence band around our best nonparametric fit (black dashed lines). The simultaneous bands were determined using the method of Sun & Loader (1994). The gray-shaded band is the 1σ scatter in this relation and includes effects from matching (§ 4.2), measurement error, and any intrinsic scatter in the relationship.

As a consistency check, we present in Figures 16 and 15 (bottom) the correlation between the velocity dispersion and

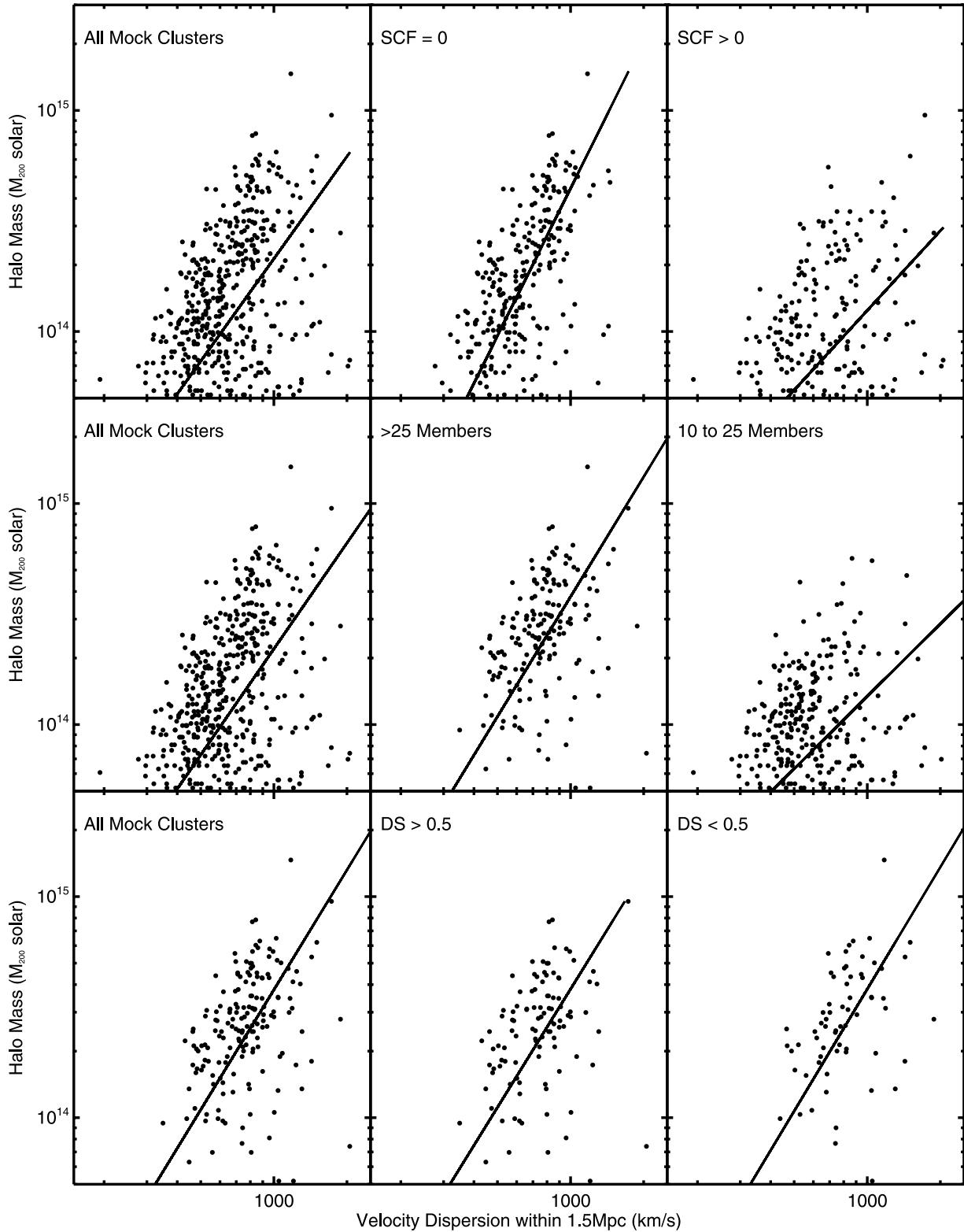


FIG. 14.—Relationship between M_{200} and the velocity dispersion as a function of different methods of contamination. *Top:* SCF. *Middle:* Number of galaxies used to measure the dispersion. *Bottom:* Dressler-Shectman statistic.

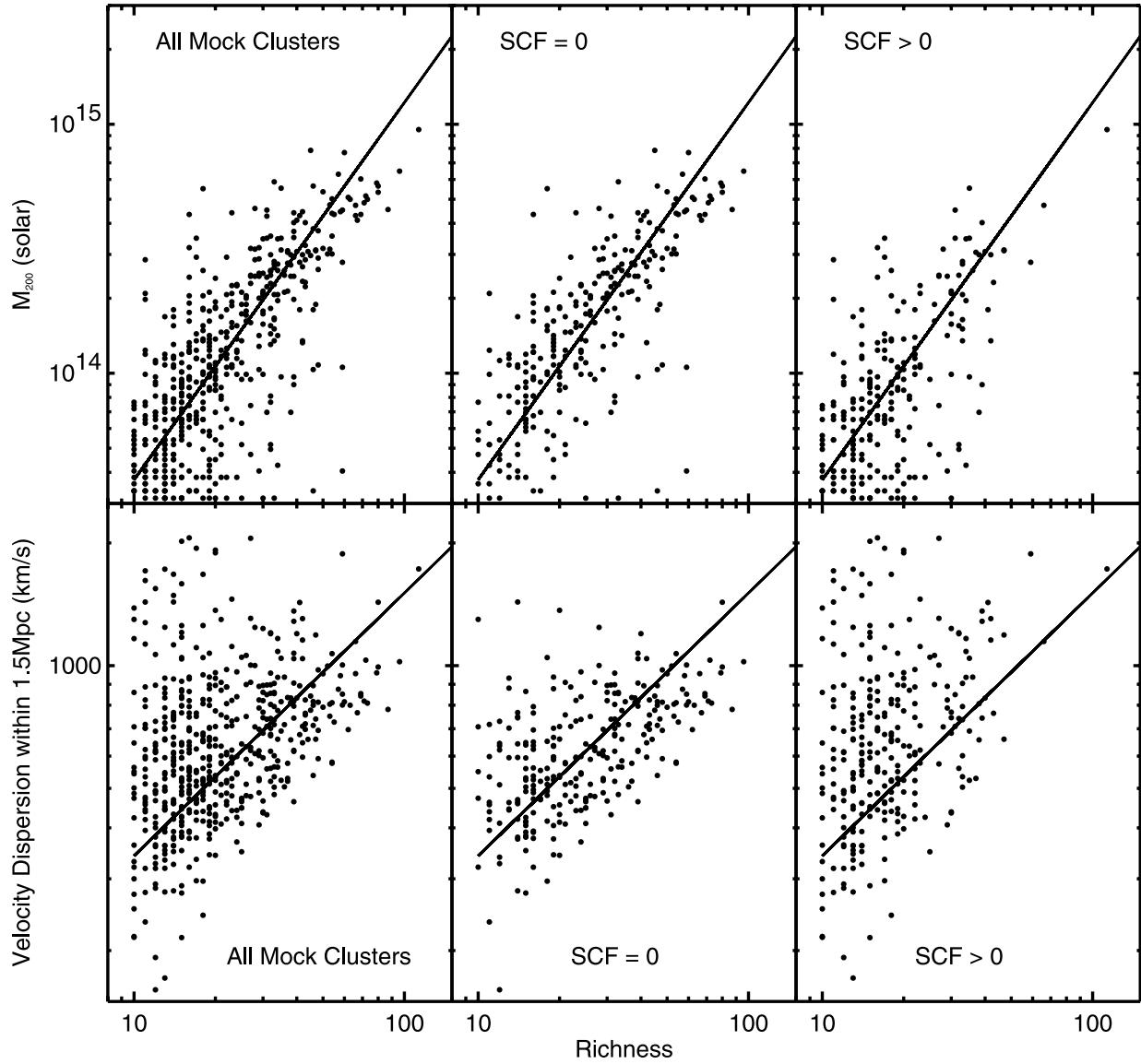


FIG. 15.—Relationship between the richness as measured with $1 h^{-1} \text{ Mpc}$ for members within ± 4 velocity dispersions of the cluster, using the mock clusters, against halo mass (top) and velocity dispersion (bottom). We show the effect of the SCF. The best-fit relation is for all clusters without a nearby companion.

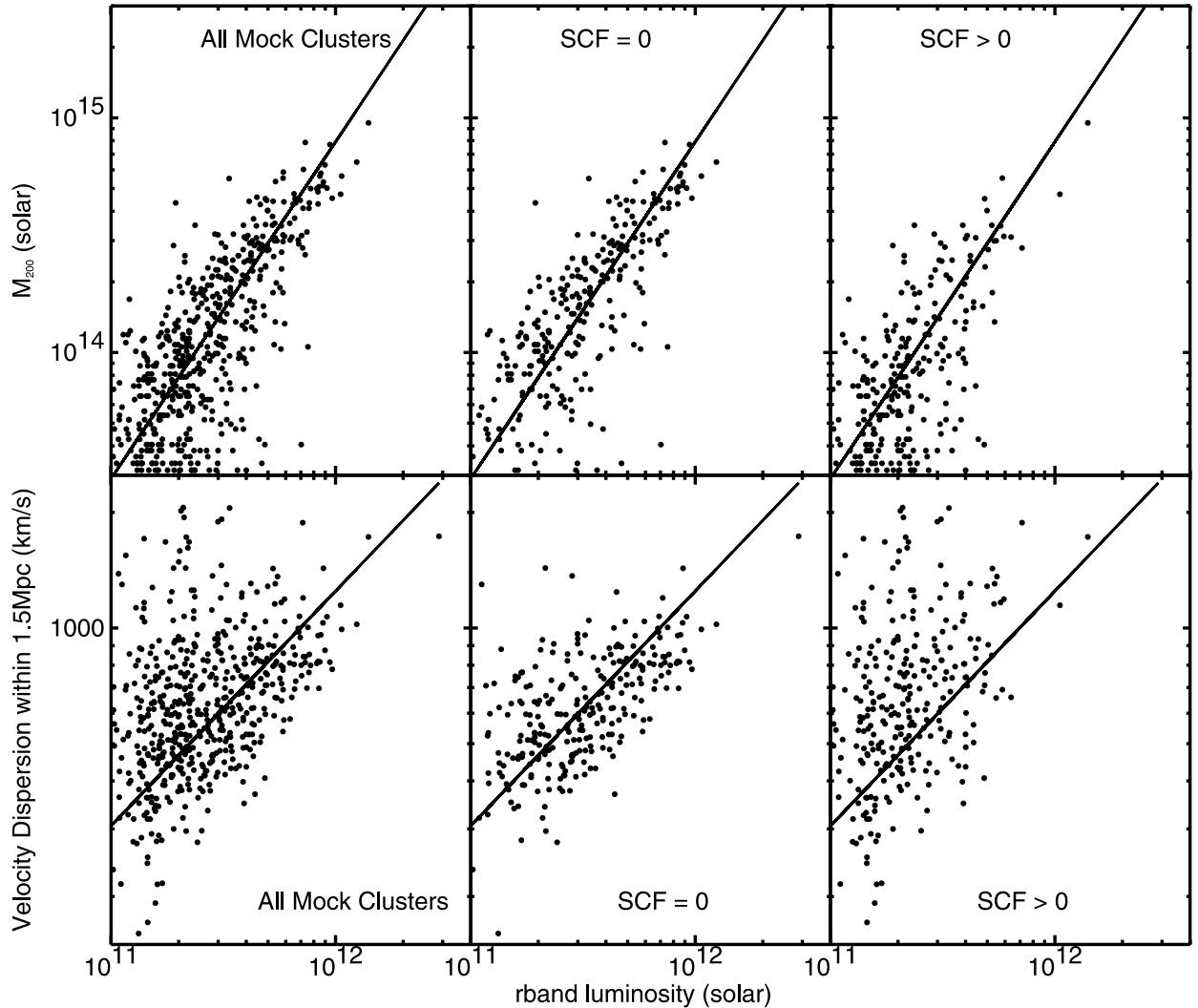


FIG. 16.—Relationship between total r -band cluster luminosity and the halo mass (top) and velocity dispersion (bottom), using the mock clusters, as a function of the SCF. The best-fit relation is for all clusters without a nearby companion. Note that unlike velocity dispersion, the scatter in mass vs. luminosity is not significantly reduced when only clusters with $\text{SCF} = 0$ are used. In addition, the scatter in the mass-luminosity relation without cuts is already as small as in Fig. 14 (top middle) with $\text{SCF} = 0$.

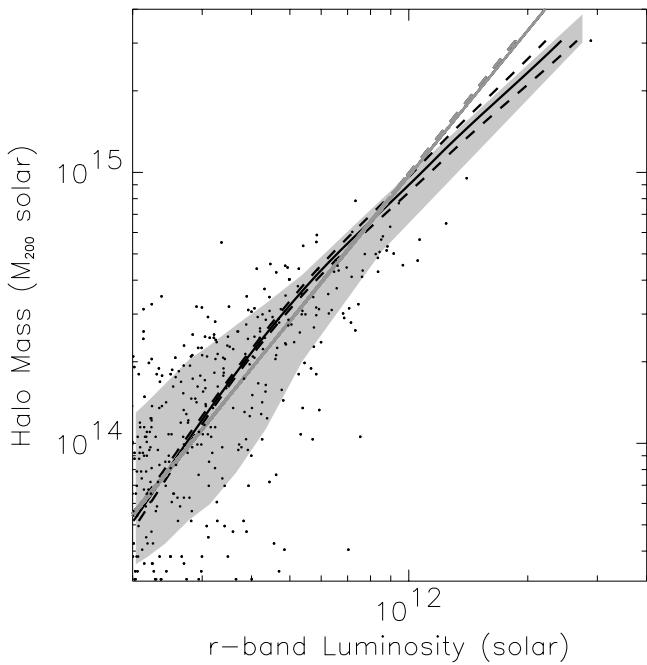


FIG. 17.—Summed r -band cluster luminosities vs. the halo mass for the simulations. The gray solid and dashed lines are fits to mass vs. luminosity and luminosity vs. mass, respectively. The black solid line is a nonparametric fit to the data, and the black dashed lines encompass the 1σ confidence band around this fit. The shaded gray region encompasses the 1σ irreducible scatter of the data about the best nonparametric fit.

richness or summed total cluster r -band luminosity for clusters in our mock SDSS data set. As seen in Figure 14, the scatter is similarly reduced by a factor of 2 when we use only $SCF = 0$ clusters. Our conclusion is the same as above: our velocity dispersions can be overestimated when there is intervening large-scale structure to interfere with the measurement of the velocity dispersion.

In the above sections (§§ 5 and 6), we have examined how foreground/background contamination from large-scale structure (or other clusters) affects the observables of velocity dispersion, richness, and optical luminosity. We pointed out a few example clusters for which the SCF varies from 0 to 2, thus destroying any clean measure of the velocity dispersion. We showed that the scaling laws that use the summed r -band luminosity and/or richness are less affected by local large-scale structure than velocity dispersion. We conclude that the optical luminosity is the cleanest observable measure of halo mass (using the mock catalogs).

7. APPLICATION TO SDSS DR2 DATA

In this section, we discuss the application of the C4 cluster-finding algorithm to the SDSS spectroscopic galaxy data (Strauss et al. 2002). Technical details of the SDSS instrumentation and operations can be found in Gunn et al. (1998), York et al. (2000), Hogg et al. (2001), Stoughton et al. (2002), Smith et al. (2002), and Pier et al. (2003), as well as at the SDSS Web site.²⁰ In this paper, we focus on the photometric and spectroscopic data of the DR2 of the SDSS (Abazajian et al. 2004), i.e., we apply the C4 algorithm to $\simeq 250,000$ spectroscopically observed galaxies, which cover $\sim 2600 \text{ deg}^2$ of the sky. We apply some quality

constraints to the sample. Specifically, we take all objects spectroscopically identified as galaxies (including the deeper LRG sample; Eisenstein et al. 2001), excluding those objects with SDSS warning flags set for low-confidence redshift, no currently available spectrum, no red end, or no blue end. Also, the $z\text{Conf}$ variable must be greater than 0.7. This results in 249,725 unique galaxies.

Using the methodology outlined above, we find 748 clusters and groups of galaxies in the SDSS DR2 data. This is a relatively small number of systems compared to other published SDSS photometric cluster catalogs (Goto et al. 2002; Kim et al. 2002; Bahcall et al. 2003), but we stress that our sample is confined to systems with at least eight *spectroscopic* redshifts within $1 h^{-1} \text{ Mpc}$ of the cluster center. This limits our sample to nearby clusters, in the redshift range $0.02 < z < 0.17$. Other SDSS cluster surveys are based only on the SDSS photometry and therefore extend to higher redshift and contain many more clusters, at the expense of not having measured velocity dispersions or precisely determined membership for all of the clusters.

7.1. Fiber Collisions

To this point, we have made no correction in our analysis for the problem of “fiber collisions.” As discussed in Strauss et al. (2002), no two spectroscopic fibers can be placed within $55''$ of each other in any given SDSS spectroscopic plate. These fiber collisions have been minimized by overlapping adjacent plates (i.e., tiling) and result in at most 20% of all SDSS main sample galaxies not being targeted over all surface densities of the galaxies (see Blanton et al. 2003a). We study fiber collisions in two ways: (1) for the mock catalogs, we run the C4 algorithm before and after fiber collisions are applied; and (2) for the real data, we run the algorithm on the spectroscopy (i.e., after fiber collisions) and add SDSS photometric cluster galaxies that were missed by the tiling algorithm. Since we do not know the redshifts of these missed SDSS targets, it is impossible to run the algorithm on the real data prior to fiber collisions unless estimated redshifts are used. Unfortunately, even very good photo- z ’s (i.e., with $\Delta z \sim 0.05$) have errors larger than the redshift box used in our algorithm.

The mock catalogs allow us to study the effect of fiber collisions on completeness and purity, while both the mock and the real data provide an independent analysis of the effect fiber collisions have on cluster luminosity. In the mock catalogs, we find clusters both before and after fiber collisions, so the measured properties can change in many ways. For instance, the cluster centers and velocity dispersions can change, both of which would add to or subtract from cluster membership. However, in the real data the cluster centers and velocity dispersions are determined solely through the spectroscopic sample. We can only correct the total cluster luminosity by adding in the missed photometric galaxies that are within the projected radius. To minimize foreground/background contamination, we also constrain these to have colors and magnitudes near the E/S0 ridgeline of the clusters. Thus, unlike the mock catalogs, in which the fiber-corrected luminosities can be higher or lower, in the real SDSS data the fiber-corrected summed r -band luminosities will always be the same or higher.

7.1.1. Fiber Collisions in the Mock Catalogs

To create our mock catalogs with fiber collisions, we find all mock galaxy pairs within $55''$ and randomly choose the brightest galaxy in each pair. We do this such that 70% of all galaxies with nearest neighbor separations of less than $55''$ are retained. This mimics (albeit conservatively) the targeting algorithm described

²⁰ See <http://www.sdss.org>.

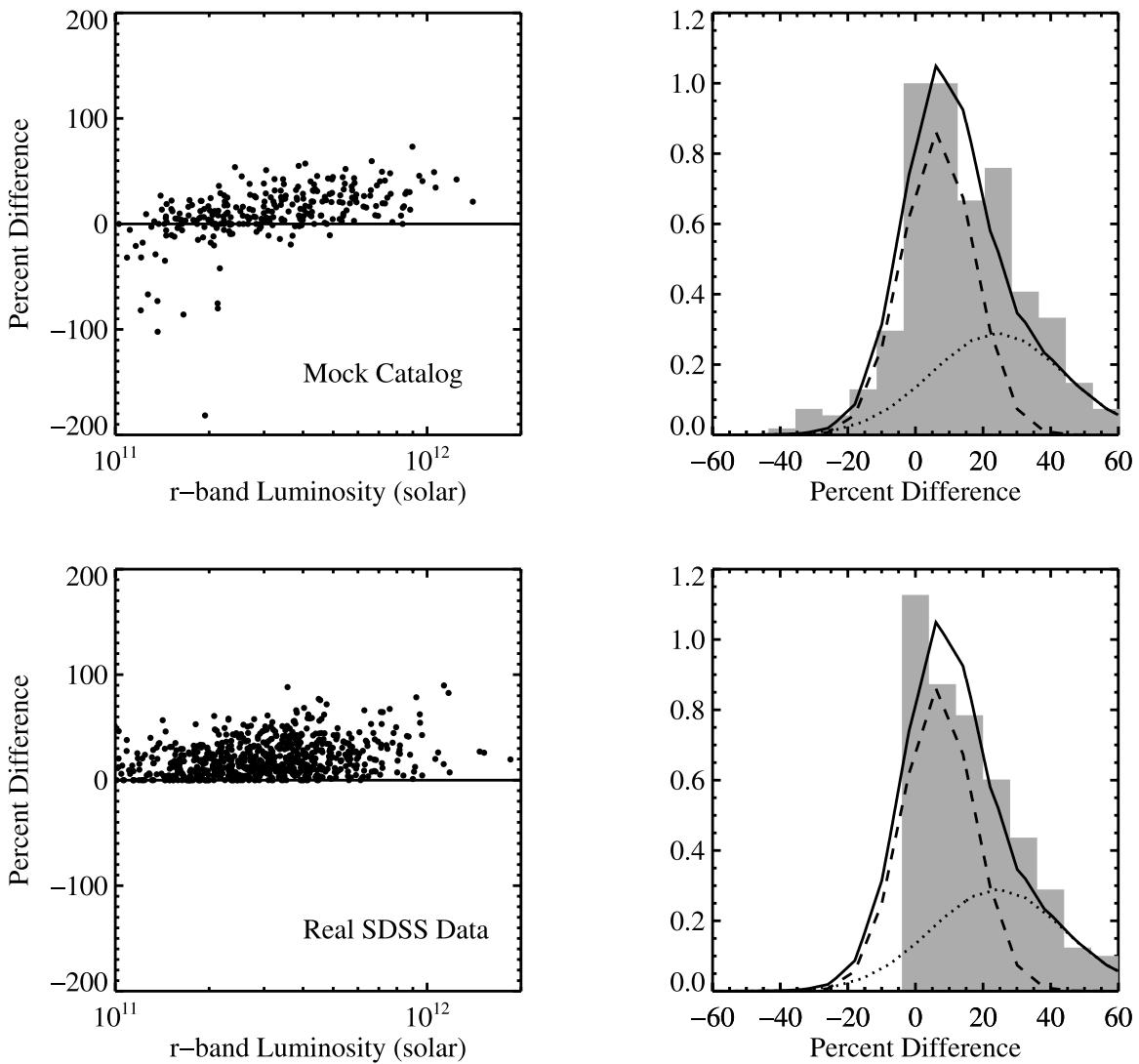


FIG. 18.—Comparison using the mock clusters of our summed r -band luminosities before and after we remove mock galaxies due to fiber collisions. The clusters are systematically brighter (by 5%) before we make the fiber collision corrections. The Gaussian is a fit to those clusters with revised luminosities after the fiber collisions were put in and has a width of $\sim 10\%$.

in Blanton et al. (2003a). We then rerun our cluster finder on these revised mock catalogs.

We find that, regardless of fiber collisions, the completeness and purity remain unchanged for the more massive systems. As the mass (or luminosity) of the clusters decreases, the completeness after applying fiber collisions drops slightly ($\sim 5\%$). As a consequence of finding fewer systems, the purity increases slightly after fiber collisions are applied (again by 5%). This can be seen in Figure 10, in which the solid and dashed lines show measurements taken before and after fiber collisions, respectively. Thus, fiber collisions play a minimal role in finding clusters.

However, the summed optical r -band luminosities are systematically underestimated, as seen in Figure 18 (top). In this figure we show the difference in the measured mock cluster luminosities before and after the application of (i.e., minus) the fiber collision algorithm. Bright massive clusters are more affected than small dim clusters. The histogram has been normalized to the average number of clusters, with differences between 0% and 10%. As indicated by the peak of the histogram, many of the cluster luminosities remain unchanged. Most are systematically brighter before we remove mock galaxies due to fiber collisions (as expected). A few are dimmer *before* galaxies are

removed from fiber collisions. As described above, the clusters found after the fiber collision algorithm is applied have different membership criteria than those found before (because of a different centroid or velocity dispersion). Thus, we expect a few clusters to be brighter after galaxies are removed from the mock catalogs.

The distribution of the histogram in Figure 18 is not a simple Gaussian. We have investigated this and find that the sum of two Gaussians is statistically a good fit. The two Gaussians are explained by two different populations of missed targets: typical cluster galaxies and BCGs. In the mock catalogs, we find that the BCG would be missed by fiber collisions $\sim 38\%$ of the time. These BCGs are significantly brighter than typical cluster galaxies and they account for the tail in the histogram of Figure 18. To confirm this, we have fitted the histograms of the BCGs and non-BCGs in Figure 18 separately (as shown by the dotted [BCG] and dashed [non-BCG] curves). The sum of these two Gaussians (*solid line*) is then a good fit to the overall distribution. The centers of these Gaussians tell us (on average) how much fiber collisions lower the cluster luminosities. When the BCG is missed, clusters are dimmer by 24%, while when the BCG is spectroscopically observed, clusters are dimmer by only 6%.

7.1.2. Fiber Collisions in the SDSS Data

Since we cannot find clusters in the real SDSS spectroscopic data before fiber collisions, we attempt to correct the SDSS C4 clusters by adding galaxies that were targeted for observation but missed because of the tiling algorithm. We extract all galaxies within a projected $1 h^{-1}$ Mpc of the cluster center and brighter than $m_r = 19.6$ and $M_r = -19.8$, where we assume the galaxies to be at the redshift of the cluster. We then add this additional galaxy light to the cluster luminosities and look for a new BCG. To avoid adding too much noise from foreground/background galaxies, we require these additional galaxies to have colors that lie within the cluster's E/S0 ridgeline and also be no more than 2 mag dimmer than the originally identified BCG.

In Figure 18 (*bottom*), we show the differences in the cluster luminosities before and after the fiber corrections. Recall that, by design, the corrected cluster luminosities will always be as bright as or brighter than they were before the correction (since we only add in galaxies from the photometry). As in the mock catalogs described above, the BCGs are missed $\sim 39\%$ of the time. Also similar to the mock catalogs, the peak of the histogram is at zero difference. For comparison, we overplot the Gaussians from the mock catalog histograms (Fig. 18, *top*). The close match between the mock catalogs and the real SDSS data in terms of the tails of these distributions as well as of the fraction of missed BCGs indicates that (1) the fiber collision algorithm we use on the mock catalogs is close to the tiling of the SDSS and (2) that the cluster luminosity differences have two components (BCG and non-BCG) and can be corrected.

7.2. Galaxy Deblending and Targeting

To be targeted for spectroscopy by the SDSS (Blanton et al. 2003a), a galaxy must first be identified in the photometric data. Since we are only looking at the brightest galaxies in the SDSS, we assume that star-galaxy separation is a nonissue (see Scranton et al. 2002). However, for various reasons, the photometric deblender occasionally fails to correctly identify a galaxy. An extreme example is for Abell 1539 (R.A. = $186^{\circ}5813$, decl. = $62^{\circ}5563$), for which the multiple identification of a satellite trail caused the deblender to reach its maximum number of objects within a given field. Thus, many galaxies within the cluster were not identified.

While a paper presenting the SDSS deblender algorithm and tests is currently under preparation (R. H. Lupton et al. 2005, in preparation), we made a minimal effort to quantify any SDSS deblender issues, specifically with regard to galaxy clusters. To do so, we identified 110 Abell clusters within the sky coverage of the SDSS DR2 spectroscopic survey and visually examined each using the SDSS SkyServer cutout tool. This Web-based tool allows the user to look at any SDSS photometric field and overlay photometric or spectroscopic SDSS galaxies (those both targeted and observed). One can also click on the galaxies to determine their redshifts, magnitudes, etc.

The purpose of this exercise was to determine whether SDSS deblending or targeting problems could affect the C4 algorithm. To perform this analysis, we take the following steps:

1. Using the SDSS Finding Chart cutout tool, look at a $7' \times 7'$ field around each Abell cluster center.
2. Identify the BCG, its coordinates, and its *r*-band magnitude.
3. Determine the fraction of spectroscopic targets that actually have spectroscopy.
4. Determine whether the BCG was targeted and observed spectroscopically.

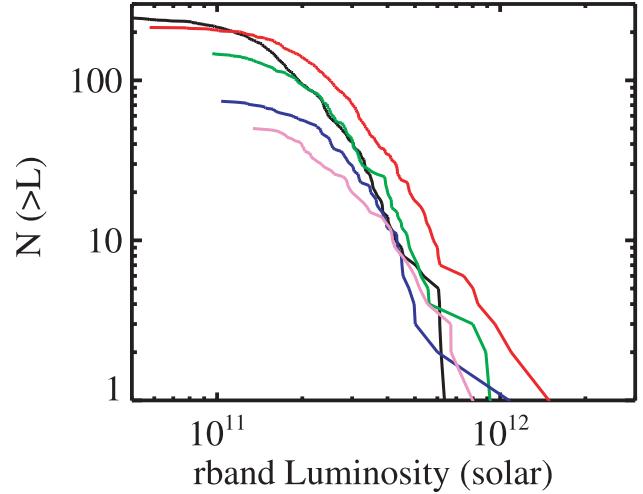


FIG. 19.—Number of SDSS C4 clusters brighter than L_r for equal-volume shells, each at increasing redshift. The shells are described in the text. They go from the lowest to the highest redshift as black, red, green, blue, and violet. The red line covers a volume that includes the SDSS Great Wall (Gott et al. 2003), where we find additional massive clusters compared to the other four equal volumes.

5. Count the number of galaxies that were missed by the deblender.

6. Count the number of galaxies that are brighter than the SDSS spectroscopic limit ($m_r = 17.77$) but not targeted.

In a few cases, no cluster was obvious to the eye and so the BCG parameters and other measures were skipped (Abell 130, 630, 682, 685, 796, 2195, 2433, and 2703).

We find that 32% of the Abell clusters are not affected by fiber collisions at all and that 73% of cluster cores have more than 70% of their targeted galaxies actually observed spectroscopically by the SDSS tiling algorithm. The worst case is Abell 2644, for which only 4 of the 10 targeted galaxies were observed. The cluster BCGs were all targeted and 70% were observed, which is slightly better than the global C4 average of 61%, as discussed above. The SDSS deblender seemed to miss objects in less than 10% of the clusters. In most of these few cases (i.e., except Abell 1539 mentioned above), only one obvious galaxy object was missed in the entire field (which usually contains hundreds of objects).

Our conclusions from this visual examination of over 100 Abell clusters are that the SDSS deblender works extremely well for these low-redshift systems. Likewise, the targeting and tiling algorithms perform as expected.

7.3. The SDSS-C4 Clusters

In Figure 19, we present the number of clusters as a function of *r*-band luminosity for the same equal-volume shells described in § 2.7.2. In a comparison with Figure 7 (*bottom*), we see the same trend of finding fewer low-mass systems at higher redshifts. A major difference between Figure 19 and Figure 7 (*bottom*) is the redshift shell $z = [0.075, 0.093]$, which contains the Sloan Great Wall (Gott et al. 2003) at redshift $z = 0.08$ (Figs. 7 and 19, *red line*). In this one bin of volume, we seem to find many more systems compared to the other equal volumes. In addition, there appears to be more massive clusters in this bin than in the other equal-volume bins, while the number of low-mass systems is little changed.

In Table 2, we present the 748 clusters detected using the C4 algorithm on the DR2 data set. For each cluster, we present an SDSS C4 ID (col. [1]); the (J2000.0) right ascension and

TABLE 2
THE SDSS-C4 CLUSTER CATALOG (DR2)

ID (1)	M1 R.A. (deg) (2)	M1 Decl. (deg) (3)	M2 R.A. (deg) (4)	M2 Decl. (deg) (5)	BCG R.A. (deg) (6)	BCG Decl. (deg) (7)	Redshift (8)	σ_v (km s $^{-1}$) (9)	Richness (10)	Lum $_r$ (M_\odot) (11)	SCF (12)	DS (13)	Other Names (14)
3002.....	258.1272	64.0166	258.1362	64.0183	258.1200	64.0608	0.08078	1355	150	1.933993E12	0	0.02	RBS 1630, ABELL 2255
1031.....	213.7458	-0.3874	213.7301	-0.4146	213.7405	-0.3496	0.14037	931	48	1.584852E12	1	0.99	ABELL 1882
3153.....	133.9645	42.1373	133.9973	42.1182	133.9380	42.0673	0.13867	607	21	1.558037E12	1	0.91	No matches
1047.....	229.2174	-0.9029	229.1859	-0.9629	229.2174	-0.9029	0.11856	870	84	1.511027E12	0	0.00	ABELL 2051
3238.....	257.6715	56.7729	257.7072	56.7694	257.5823	56.7573	0.12508	274	13	1.448392E12	2	0.18	No matches
3077.....	240.2678	53.9485	240.3145	53.9002	240.3670	53.9474	0.10725	611	30	1.282911E12	2	0.77	RXC J1601.3+5354, ABELL 2149
3001.....	255.6148	33.4949	255.6578	33.5058	255.6381	33.5167	0.08847	1599	104	1.232508E12	1	0.95	RX J1702.5+3330, ABELL 2245
1039.....	228.8332	4.3846	228.8090	4.3502	228.8088	4.3862	0.09817	971	56	1.159921E12	0	0.94	ABELL 2048
1055.....	202.7347	-1.7053	202.7637	-1.7449	202.7960	-1.7273	0.08470	1112	82	1.149299E12	0	0.00	ABELL 1750
2002.....	358.5570	-10.4192	358.5397	-10.3877	358.5570	-10.4190	0.07612	976	85	1.147930E12	0	1.00	ABELL 2670
1043.....	234.1718	-2.0237	234.1550	-2.0297	234.1242	-1.9634	0.14477	832	41	1.123676E12	0	0.95	ABELL 2094
1135.....	180.2013	3.4281	180.1875	3.3906	180.2303	3.4488	0.13586	1595	35	1.069101E12	0	0.18	ABELL 1437
3130.....	186.8255	63.3870	186.8714	63.4071	186.9634	63.3848	0.14477	556	16	1.055957E12	1	0.12	ABELL 1544, RX J1227.8+6323
1140.....	176.3583	-2.3925	176.3784	-2.4031	176.3790	-2.4531	0.12399	2182	36	1.033189E12	1	0.94	ABELL 1373, ABELL 1373E
2005.....	19.8822	14.8913	19.8902	14.8660	19.9095	14.8980	0.12850	773	40	9.780194E11	0	0.08	ABELL 0175

NOTES.—“M1” and “M2” refer to method 1 and method 2, respectively. Table 2 is published in its entirety in the electronic edition of the *Astronomical Journal*. A portion is shown here for guidance regarding its form and content. The names in col. (14) are ABELL, Abell (1958); RBS, *ROSAT* Bright Source Catalog (Voges et al. 1999); RX, the Hamburg *ROSAT* All Sky Survey (RASS) catalog of optical identifications (Bade et al. 1998); ZwCl, Zwicky compact groups and clusters (Zwicky & Herzog 1963, 1966, 1968; Zwicky & Kowal 1968; Zwicky et al. 1961, 1965); WX, Wei et al. (1999); RXC, Oppenheimer et al. (1997); WBL, Catalog of Nearby Poor Clusters (White et al. 1999); CAN, Wegner et al. (1996); SCC, Söchting et al. (2002); MACS/VMF, Vikhlinin et al. (1998); and KDCS, Deep Range Survey (Postman et al. 2002).

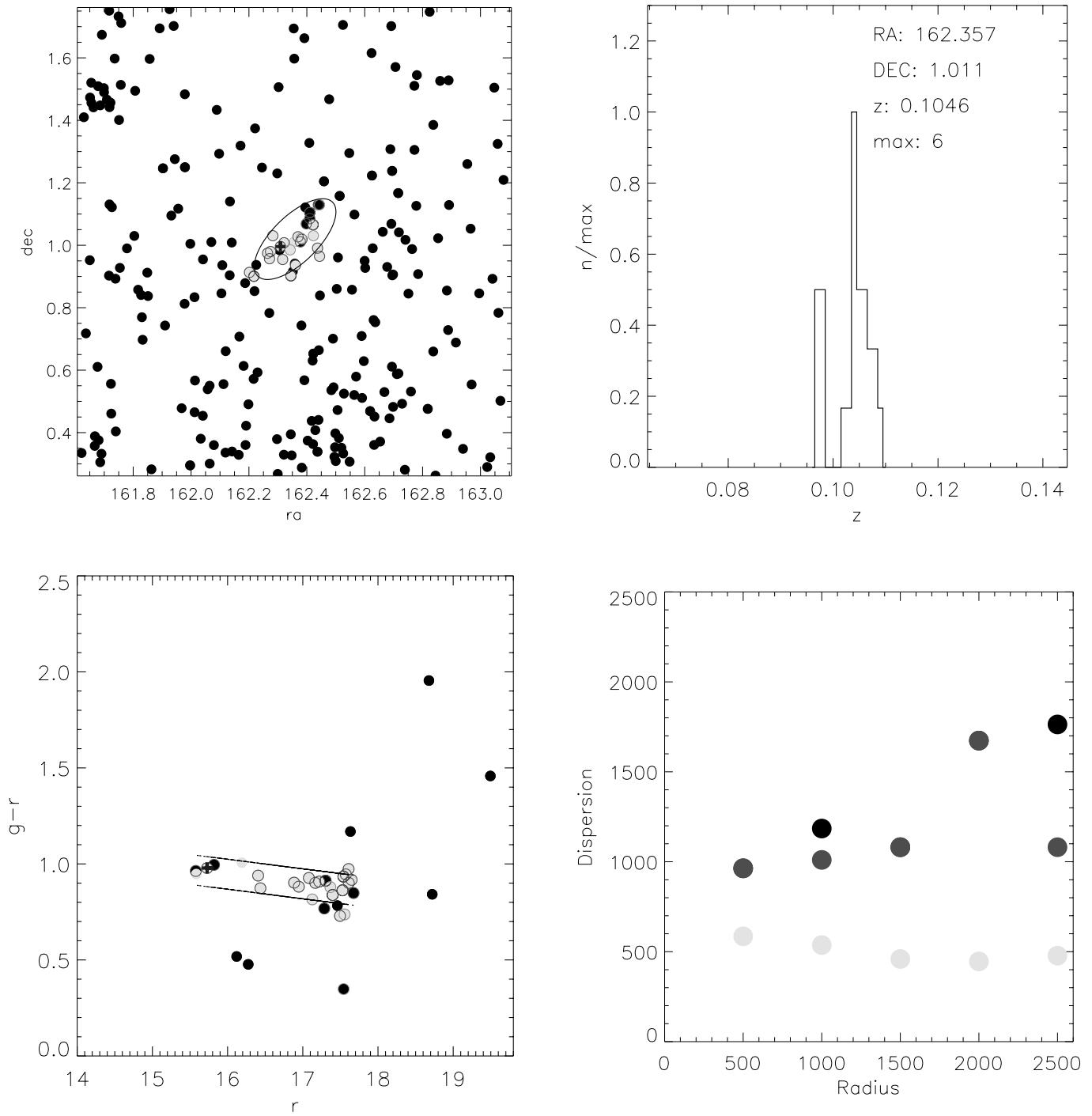


FIG. 20.—SDSS DR2–C4–1028 sky plots (top left), velocity histograms (top right), $g - r$ vs. r CMR (bottom left), and the velocity dispersion profile (bottom right). Red and green circles are for cluster members within $1.5 h^{-1}$ Mpc and within $\pm 4 \sigma_v$, with red having weak H α emission and green having strong H α emission. Blue circles are for the BCG. Yellow filled circles are for the C4 galaxies. Black filled circles are for all galaxies in the sky plot; all galaxies within $1.5 h^{-1}$ Mpc in the CMR; and all galaxies within $1.5 h^{-1}$ Mpc and within $\pm 4 \sigma_v$ in the velocity dispersion profile. This specific cluster is highly elliptical, and the galaxy velocity dispersion profile has a high level of scatter in the outer radii, causing the SCF to be set to 2. Note that the C4 galaxy velocity dispersion profile has much less scatter. [See the electronic edition of the Journal for a color version of this figure.]

declination of the centroid via method (1) in § 5.1 (cols. [2] and [3]); the right ascension and declination via method (2) in § 5.1 (cols. [4] and [5]); the right ascension and declination of the BCG (cols. [6] and [7]); the mean redshift of the cluster (col. [8]); the velocity dispersion (to convert to the rest frame of the cluster, divide by $1 + z$) (col. [9]); the richness (col. [10]); the summed r -band cluster luminosity (corrected for missed targets) (col. [11]); the SCF (col. [12]); the Dressler-Shectman substructure proba-

bility (col. [13]); and the names of other clusters matched to within $10'$ of the cluster centroid (col. [14]). Note that the match is not performed against redshift, and thus some of these matches are in projection only. More information on these names and additional matches can be found on the C4 Web site.

In Figures 20–25, we show the sky plots, CMRs, velocity histograms, and velocity dispersion profiles for six C4 clusters detected in the DR2. We have made available these plots for all

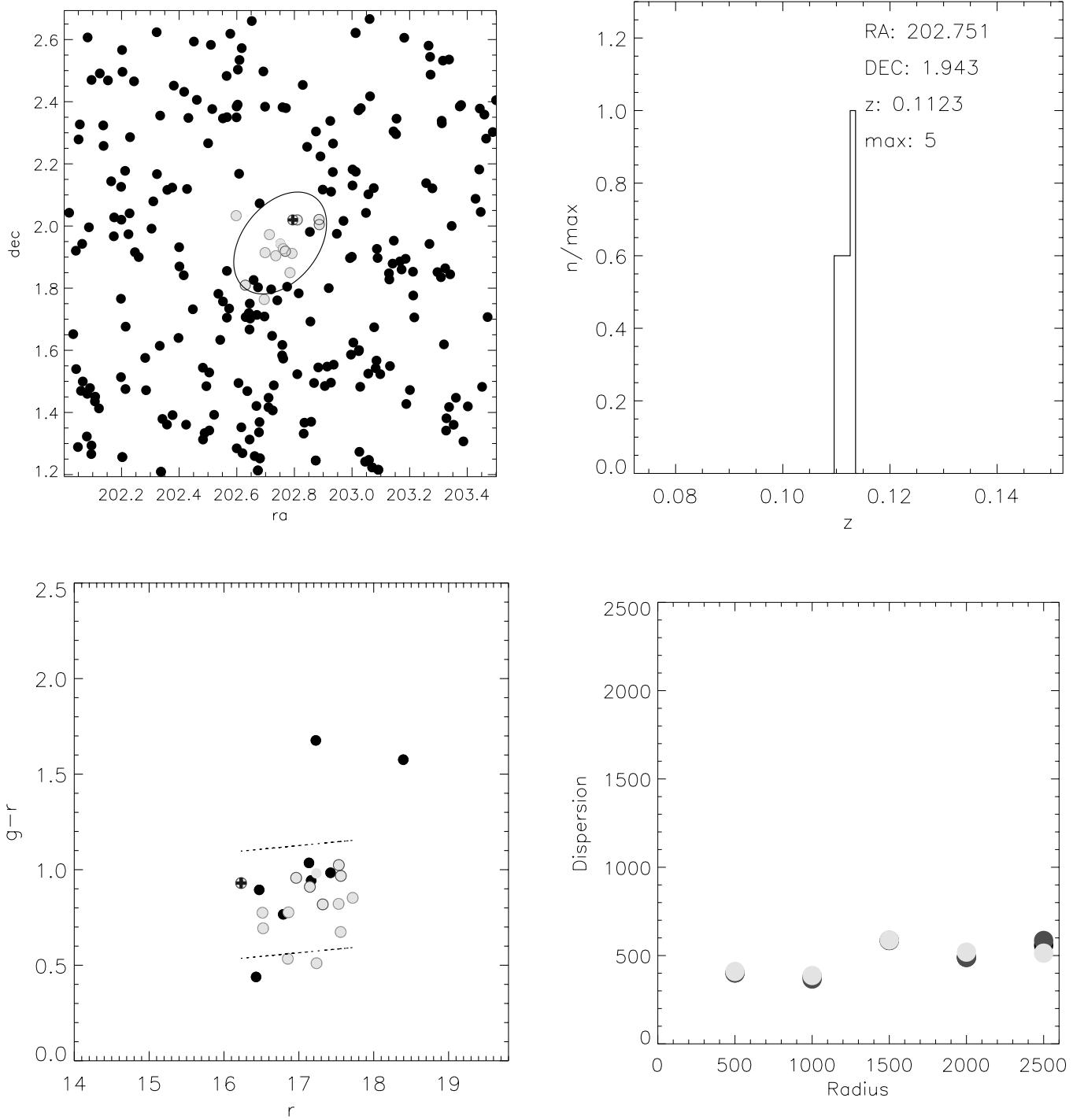


FIG. 21.—Same as Fig. 20, but for SDSS DR2-C4-1088. The core of this cluster is dominated by a galaxy population with strong H α emission and blue colors, indicative of star formation. This cluster has SCF = 0. [See the electronic edition of the Journal for a color version of this figure.]

748 C4 DR2 clusters via the World Wide Web. These plots provide for quality assurance, allowing users of this catalog to inspect the individual systems and determine whether they suit their scientific requirements.

An electronic version of the catalog given in Table 2 can be downloaded from the C4 Web site.²¹ We intend to update this C4 catalog as more SDSS data become available and we therefore encourage the reader to check this Web site periodically.

7.4. Comparison to Other Cluster Catalogs

In this subsection we take a closer look at the C4 cluster sample by comparing it to two different samples from the literature: the Abell (1958) catalog and the RASS-SDSS catalog (Popesso et al. 2004). These two have remarkably different selection algorithms (which are fully described in their respective references). In short, the Abell catalog was created by eyeball examination of the *optical* Palomar All Sky Survey plates. On the other hand, the RASS-SDSS sample identifies the optical counterparts to *X-ray-discovered* clusters taken from the

²¹ See <http://www.ctio.noao.edu/~chrism/C4>.

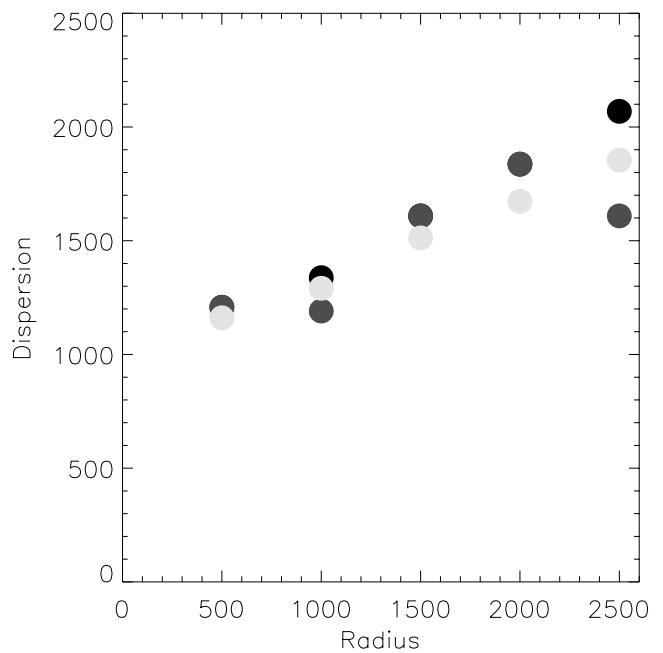
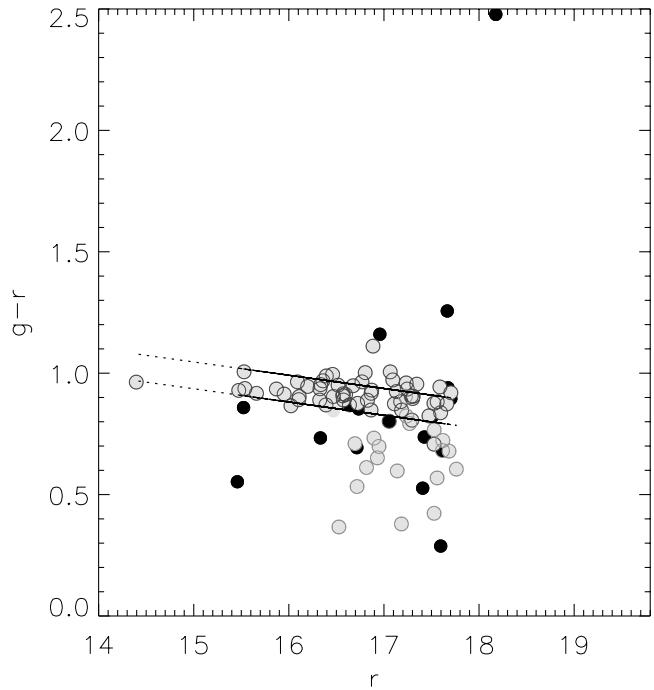
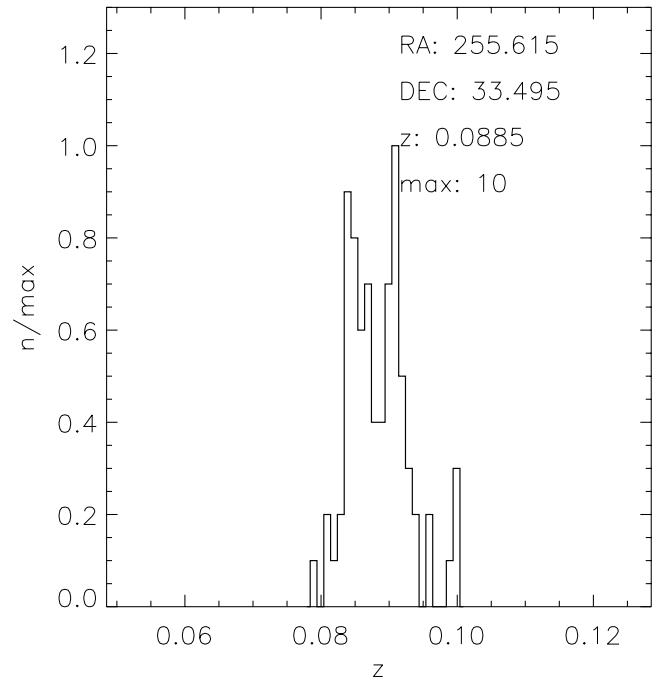
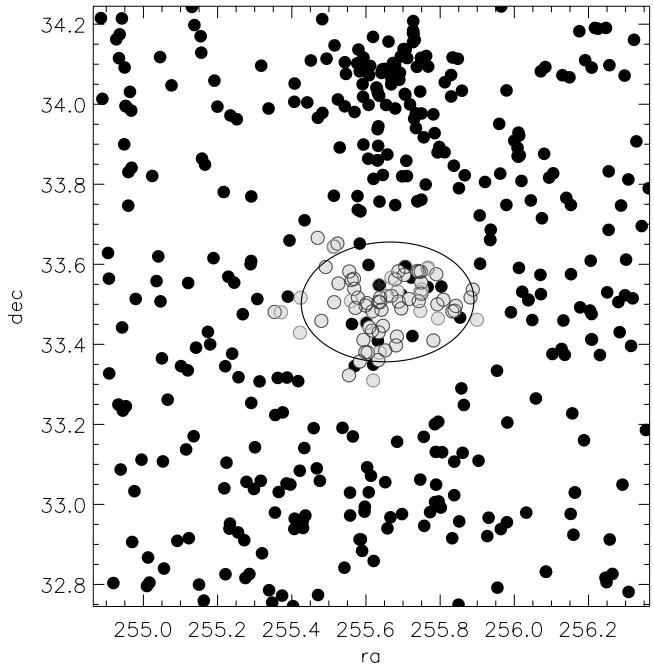


FIG. 22.—Same as Fig. 20, but for SDSS DR2-C4-3001. This cluster is the southern component to a binary system (note the concentration of galaxies at decl. = 34°1). Note how the velocity dispersion profile increases (by a factor of 2) as one probes to larger and larger radii, thus bringing in galaxies from the northern cluster. This cluster has SCF = 2. [See the electronic edition of the Journal for a color version of this figure.]

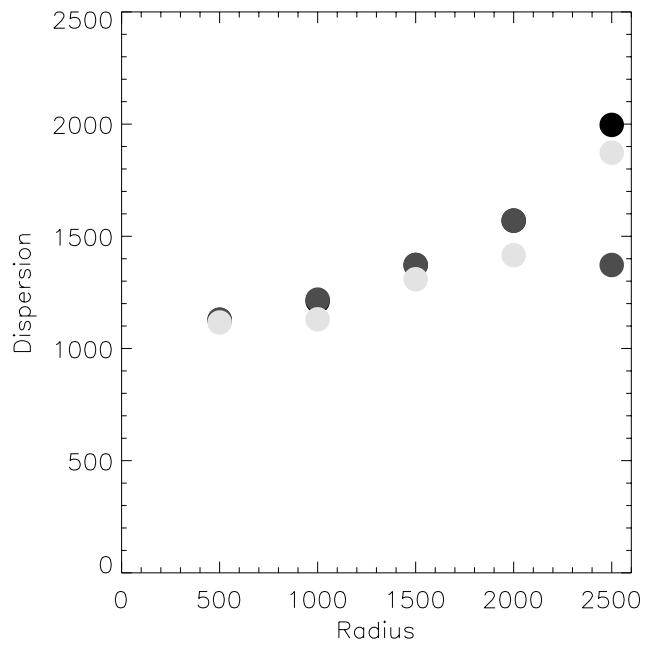
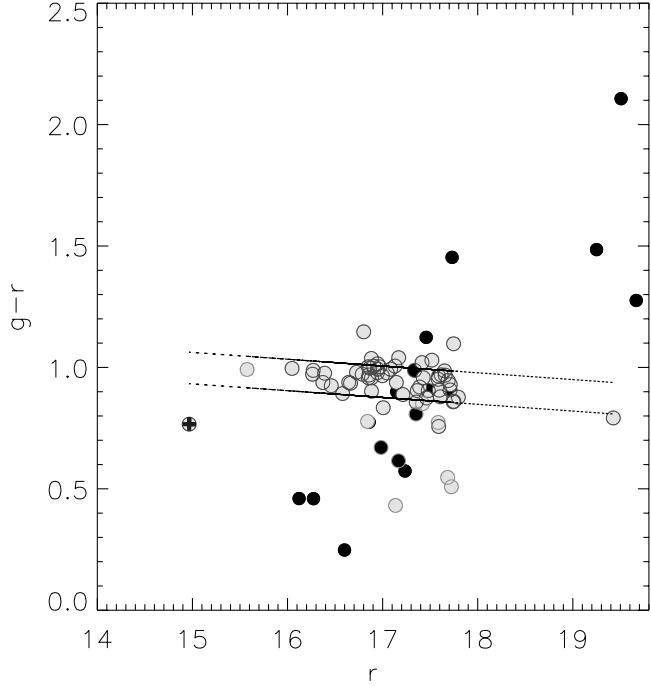
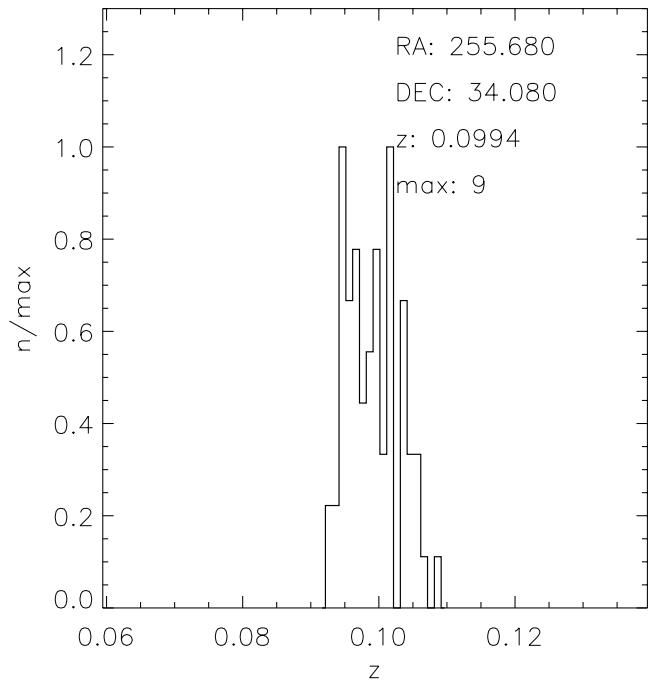
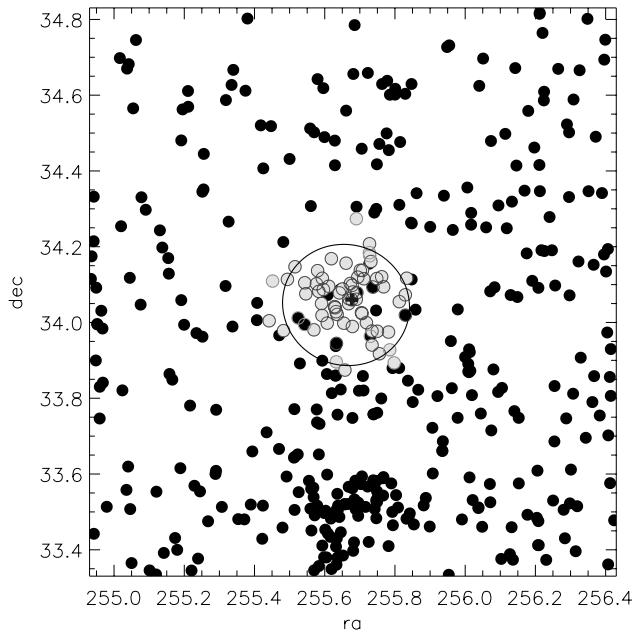


FIG. 23.—Same as Fig. 20, but for SDSS DR2–C4–3004. This is the northern counterpart to Fig. 22. [See the electronic edition of the Journal for a color version of this figure.]

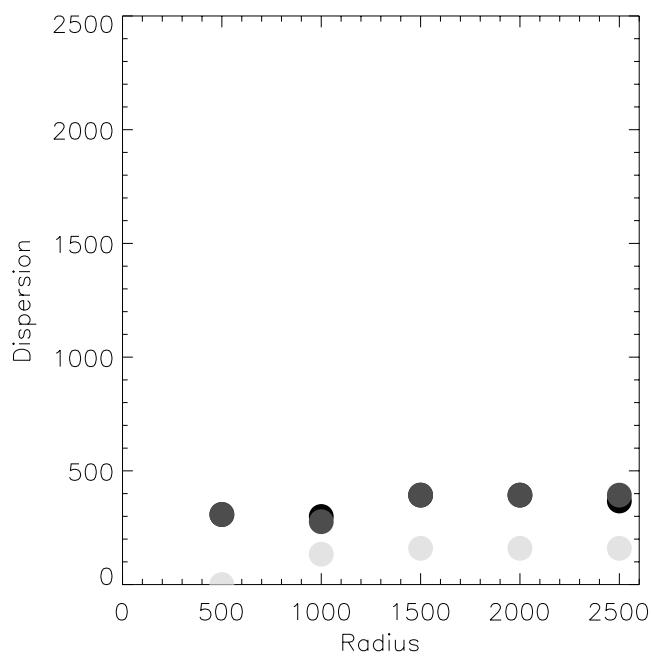
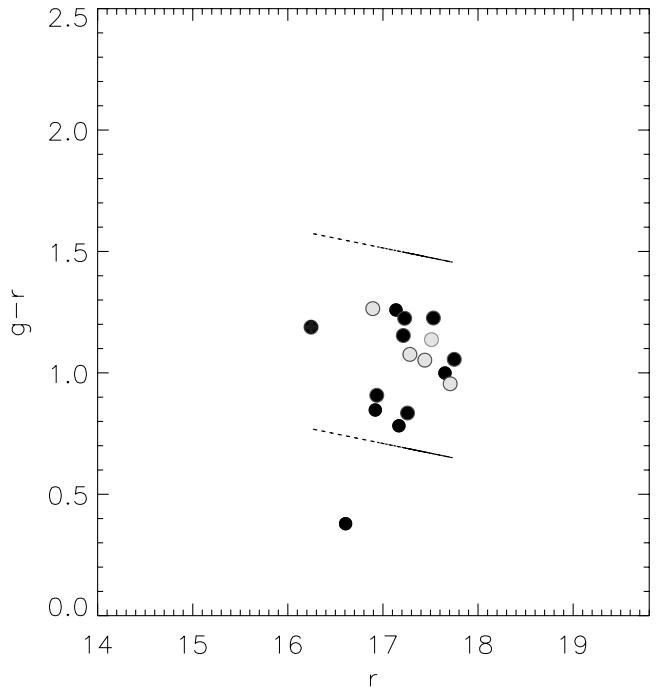
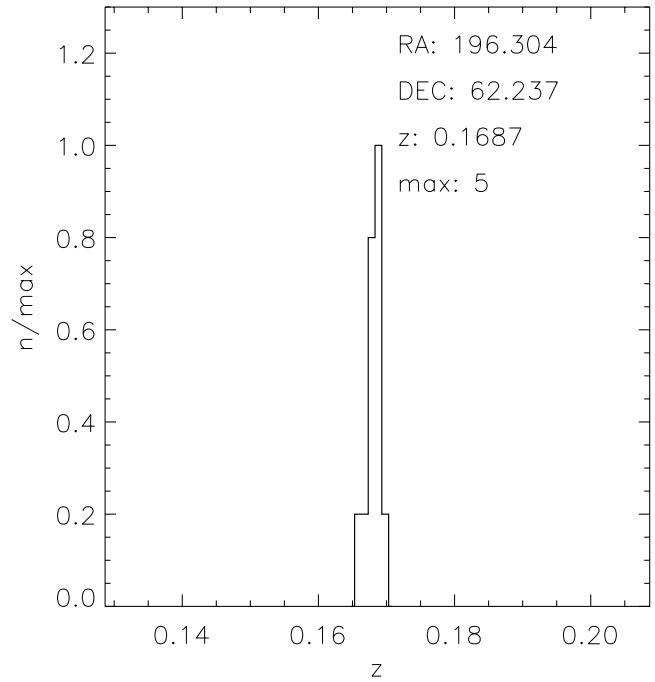
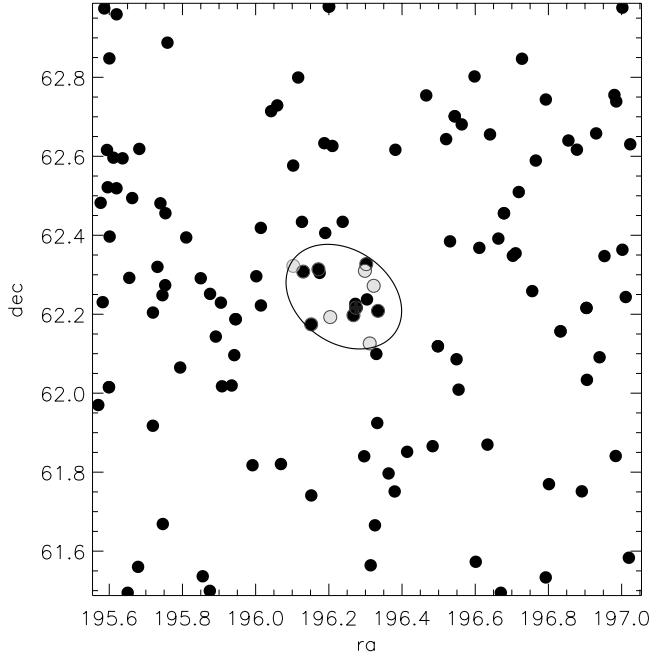


FIG. 24.—Same as Fig. 20, but for SDSS DR2–C4–3269. This is one of our higher redshift clusters ($z \sim 0.168$). This is a relatively low-mass system (400 km s^{-1}). Note how the C4 galaxies (yellow) have a smaller dispersion than the red galaxies (red). This cluster has SCF = 0. [See the electronic edition of the Journal for a color version of this figure.]

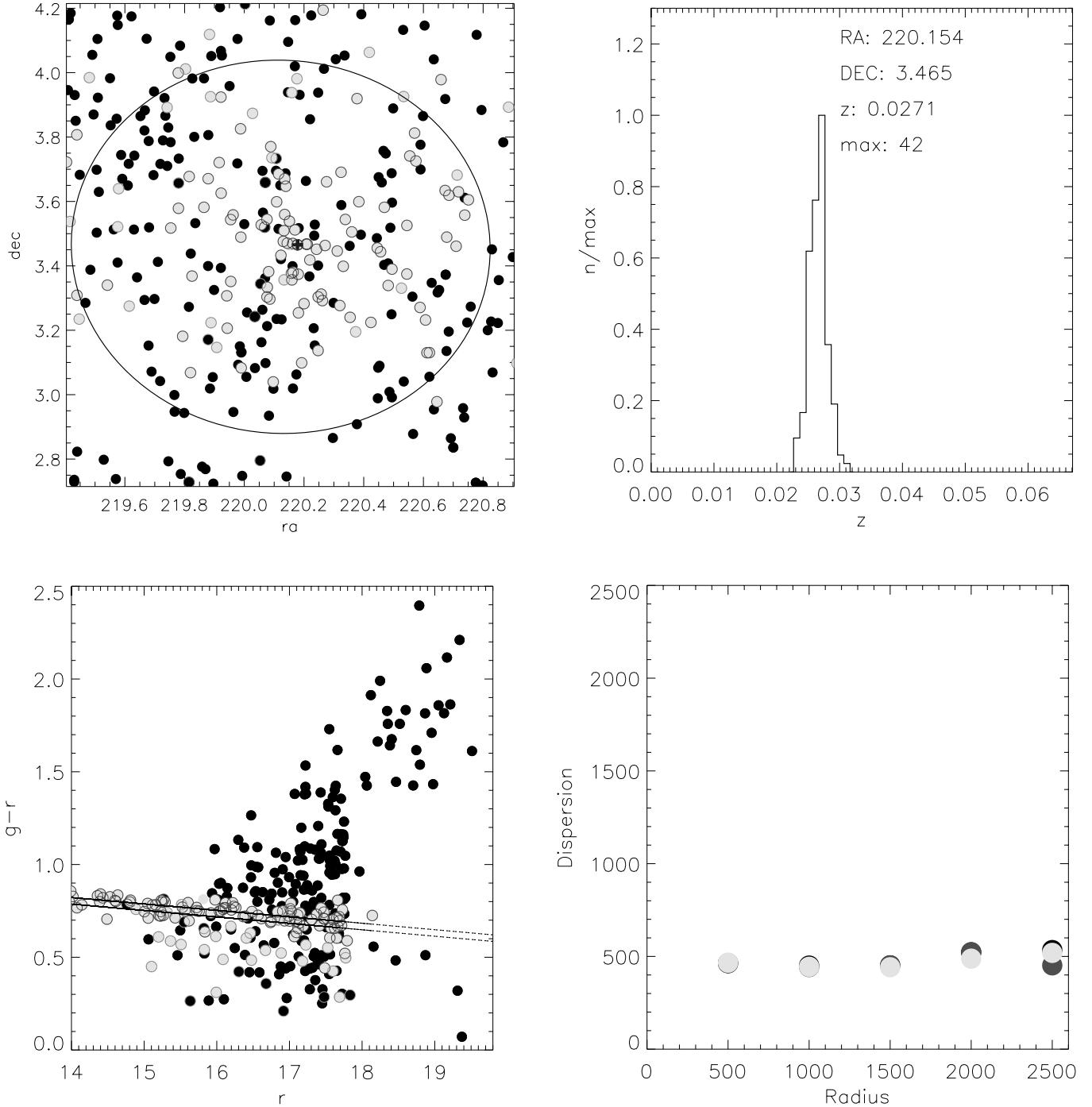


FIG. 25.—Panels and symbols are the same as in Fig. 20. This is one of the lowest redshift systems ($z \sim 0.027$) found by the C4 algorithm. This cluster has SCF = 0. Note in the sky plot how the C4 galaxies are always cluster members (i.e., there are no yellow filled circles without a green or red outline). This figure shows how well the C4 algorithm does in separating cluster members from the background (black circles). Because of its low redshift, this cluster is not included in the final catalog. [See the electronic edition of the Journal for a color version of this figure.]

TABLE 3
ABELL CLUSTERS NOT RECOVERED BY THE C4 ALGORITHM

Abell ID	R.A.	Decl.	Redshift	Abell Richness	Notes
87.....	10.75653	-9.79327	0.05500	1	Blended with Abell 85
116.....	13.96094	0.63749	0.06650	0	Too few spectra
190.....	20.91811	-9.85647	0.10150	0	
237.....	25.24066	0.26942	0.10320	0	Not a cluster
605.....	119.07229	27.39939	0.10960	0	Not a cluster
733.....	135.32976	55.62065	0.11590	1	
756.....	138.10278	48.47721	0.11760	1	Too few spectra
869.....	146.55203	2.35163	0.11740	0	Too few spectra
912.....	150.29539	-0.10799	0.04460	0	
919.....	151.23901	-0.69325	0.09540	1	Too deep ($z = 0.2$)
975.....	155.65785	64.63030	0.11860	1	
1032.....	157.57814	4.01031	0.07940	0	Blended with Abell 1024
1215.....	169.87431	4.34328	0.04940	1	Too deep ($z = 0.156$)
1289.....	172.90517	60.75701	0.11180	0	Not a cluster
1322.....	174.24931	63.22324	0.11140	0	Too deep ($z = 0.245$)
1389.....	177.34004	-1.37796	0.10320	0	Not a cluster
1402.....	178.13921	60.42187	0.10620	0	Not a cluster
1404.....	178.08986	-2.81113	0.10320	0	Too few spectra
1406.....	178.30972	67.88883	0.11780	1	Too few spectra
1432.....	179.92520	68.10468	0.11350	0	Too deep ($z = 0.265$)
1459.....	181.06020	2.50469	0.08070	1	
1477.....	182.22612	64.07182	0.11090	1	Too deep ($z = 0.24$)
1630.....	192.93535	4.56138	0.06480	1	
1729.....	201.00092	-3.36041	0.11440	1	Too few spectra
2023.....	226.45502	2.85714	0.05470	1	Too few spectra
2053.....	229.31754	-0.68262	0.11270	1	Blended with A2051
2356.....	323.93530	0.12408	0.11610	1	
2433.....	334.61554	14.01783	0.08800	0	Too deep ($z = 0.12$)
2644.....	355.28537	0.09426	0.06930	1	Too few spectra
2705.....	1.49696	15.79528	0.11470	1	No cluster

NOTE.—Units of right ascension and declination are degrees.

literature. Only 36% of the RASS-SDSS clusters are also Abell clusters.

As opposed to the C4 catalog, neither the Abell nor the RASS-SDSS cluster catalogs have selection functions, and so we can make no quantitative claims about their completeness or purity. However, since the C4 catalog is >90% pure regardless of redshift and/or mass and >90% complete for masses above $2 \times 10^{14} M_{\odot}$ and redshifts below $z = 0.12$ (see Fig. 7), we expect to be able to recover the more massive clusters from other samples.

7.4.1. The Abell C4 Clusters

We identified those Abell clusters that lie in the area of sky coverage for the SDSS DR2. We then match the two cluster samples and find any C4 clusters that are within 10' of an Abell cluster. We find 346 Abell clusters that meet these criteria and 123 are uniquely matched to C4 clusters (only five Abell clusters

are matched to two different C4 clusters). However, to more fairly compare the two samples we must place a redshift constraint on the two samples so that some minimum level of completeness is met by the algorithms. The Abell cluster catalog is thought to be fairly complete within $0.03 \leq z \leq 0.12$ (see Miller et al. 2002). As shown in Figure 7, the C4 sample is also >90% complete in this redshift range for masses greater than $2 \times 10^{14} M_{\odot}$. When we apply this redshift constraint, there are 104 Abell clusters, and 71% of them are also found by the C4 algorithm. We took a closer examination of the 30 missing Abell clusters (see Table 3) and found that six of them are actually too deep for the C4 algorithm, eight have too few spectra, and six are not clusters at all (as determined through a visual examination of the photometry and spectra around their Abell centers). So, if we exclude these 20 from the list of Abell clusters that the C4 algorithm should be able to find, then the recovery rate goes up to 88%. If

TABLE 4
RASS-SDSS CLUSTERS NOT RECOVERED BY THE C4 ALGORITHM

RASS-SDSS ID	R.A.	Decl.	Redshift	X-Ray Luminosity (W)	Notes
RXC J0736.4+3925	114.10400	39.43290	0.11700	4.81×10^{37}	
RXC J0953.6+0142	148.42310	1.71180	0.09800	0.97×10^{37}	Blended with SDSS DR2-C4-1304
RXC J1303.9+6731	195.98540	67.51770	0.10600	0.35×10^{37}	Too deep ($z = 0.22$)
RXC J1511.5+0145	227.88969	1.76420	0.03700	0.07×10^{37}	Too close to survey edge
RXC J2324.3+1439	351.08771	14.66450	0.04200	0.97×10^{37}	SDSS DR2-C4-2001 is 14' away

NOTE.—Units of right ascension and declination are degrees.

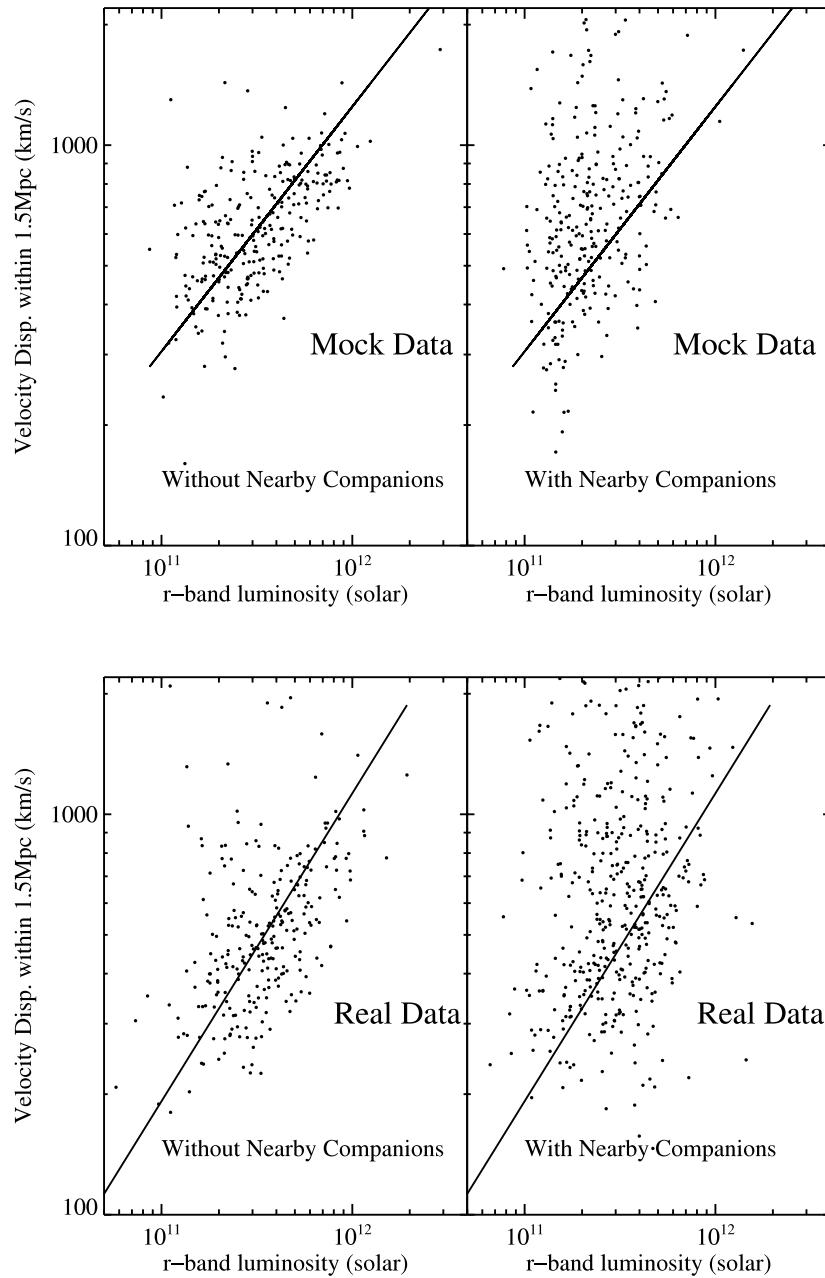


FIG. 26.—Relationship between our main cluster observables, the velocity dispersion, r -band cluster luminosity, and the SCF. The top panels are for the mock data, while the bottom panels are for the real SDSS data. The scatter in the scaling relation between velocity dispersion and r -band cluster luminosity is reduced by a factor of ~ 2 once nearby large-scale structure is accounted for. We show the best-fit linear relation to the data without nearby companions (*left*) for both the real data and the mock catalog.

we include another three Abell clusters that were found but were blended into nearby clusters, the recovery rate is 92%. Thus, the C4 algorithm can recover $\sim 90\%$ of Abell clusters, and those that are missed are typically found at redshifts at which the C4 algorithm is less than 90% complete ($z > 0.12$).

7.4.2. The RASS-SDSS C4 Clusters

We applied the same area constraints to the Popesso et al. (2004) sample as we did to the Abell clusters. Of the 97 RASS-SDSS clusters that lie within the area of the SDSS DR2 spectroscopic survey, we find 53 using the C4 algorithm. However, when we apply the same volume constraints as for the Abell sample, the C4 algorithm finds 39 of the 43 RASS-SDSS clusters within $0.03 \leq z \leq 0.12$. Five missing RASS-SDSS clusters (one

has a mistaken redshift) are listed in Table 4, along with notes made after additional examination. The RASS-SDSS recovery rate goes up to 98% if we account for a slightly larger matching radius at very low redshifts, edge effects, and the occasional cluster deblending issue.

8. DISCUSSION

As mentioned in § 1, our goal with the C4 catalog was to construct an optical cluster catalog with a well-determined selection function, i.e., one in which we know both the completeness and purity of the catalog as a function of the underlying dark halo mass and the cluster observables. To achieve this, we have used a new breed of mock galaxy catalogs that are designed to match the distribution of galaxy colors and luminosities and

their environmental dependence in the real SDSS data. We have then run the exact same cluster-finding algorithm on the mock catalog as we run on the data; this allows us to calculate the completeness and purity of the cluster catalog and examine how these quantities change as we adjust different parameters in the cluster-finding algorithm.

We have thus used the mock galaxy catalogs to fine-tune our algorithm such that it maximizes both our completeness and purity. Using the W05 mock SDSS catalogs, we find that the C4 catalog has less than 10% contamination over all cluster masses and luminosities and redshifts examined. The catalog has less than 5% contamination for clusters with masses greater than $2 \times 10^{14} M_{\odot}$ and is more than 90% complete for clusters with masses larger than $2 \times 10^{14} M_{\odot}$ and within $z = 0.12$.

We have also used the W05 mock catalog to study key cluster scaling relations and the scatter about these relations to determine which cluster observable-to-mass scaling has the smallest scatter. For the first time, we are able to directly relate cluster observables (velocity dispersion, summed cluster optical luminosities, and richness) to the halo masses used to derive theoretical mass functions (Jenkins et al. 2001; Evrard et al. 2002). The scaling relation with the smallest scatter (measured perpendicular to the best fit) is halo mass versus velocity dispersion when only systems with SCF = 0 are considered (i.e., only those with no surrounding large-scale structure; see Fig. 14, middle). The relation between halo mass and summed *r*-band luminosity has only slightly more scatter than the best case for velocity dispersion. However, this scatter is not significantly affected by limiting to systems without close companions; i.e., the sample is less affected by intervening large-scale structure than the velocity dispersion is. Thus, while the velocity dispersion (which has a simple physical relationship to mass via the virial theorem) correlates well with halo mass, the optical luminosity may be more appropriate when the velocity dispersion is contaminated or cannot be accurately measured.

A much more thorough analysis of the scaling relations and how well they can be calibrated against simulations will be presented in W05; a detailed comparison of various observational scaling relations for C4 and maximum likelihood BCG (maxBCG) clusters (Annis et al. 1999) will be presented in T. McKay et al. (2005, in preparation).

We present this C4 catalog to demonstrate the power of the SDSS spectroscopic data in finding nearby clusters of galaxies and to advance our understanding of cluster scaling relations and galaxy evolution in dense environments (for which the sample has been used in the studies of Gómez et al. 2003, Miller et al. 2003, and Balogh et al. 2004). In future work, we will use the C4 catalog to constrain cosmological parameters and to statistically

understand cluster properties such as cluster profiles and luminosity functions. By the end of the SDSS, we estimate that the C4 catalog will contain $\simeq 2500$ clusters, making it one of the largest, most homogeneous catalogs of nearby clusters in existence.

As an illustration of the quality of the C4 catalog, we present in Figure 26 the scaling relation between the total *r*-band cluster luminosity and the velocity dispersion, separated by the value of the companion flag, for both the real SDSS data (bottom) and the mock SDSS catalog (top). These cluster properties are our main observables, and in the real data their interpretation does not rely on the details of the generation of the mock catalog. This plot again demonstrates the power of the SCF in reducing the scatter on these scaling relations. Note the striking similarity between the scaling relations in the mock data and the real data. Within the uncertainties, the only difference between the two is in the number of clusters available, which is higher in the mock catalog. The best fits to these scaling relations are almost identical in the simulations and the real data. In summary, the clusters in our mock universe are statistically identical to those in the universe, suggesting that the underlying cosmological model of our mock catalogs— $\Omega_{\text{matter}} = 0.3$, $\Omega_{\Lambda} = 0.7$, and $\sigma_8 = 0.9$ —is a good representation of reality. Detailed comparison of these scaling relations will be presented in a future work.

R. H. W. acknowledges support provided by the National Aeronautics and Space Administration (NASA) through Hubble Fellowship grant HST-HF-01168.01-A awarded by the Space Telescope Science Institute. The authors acknowledge the hospitality of the Aspen Center for Physics, where some of this work was completed. Funding for the creation and distribution of the SDSS Archive has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, NASA, the National Science Foundation, the US Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society. The SDSS Web site is at <http://www.sdss.org>. The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the University of Chicago, Fermilab, the Institute for Advanced Study, the Japan Participation Group, The Johns Hopkins University, Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy, the Max-Planck-Institute for Astrophysics, New Mexico State University, University of Pittsburgh, Princeton University, the US Naval Observatory, and the University of Washington. This research has made use of the NASA/IPAC Extragalactic Database, which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA.

REFERENCES

- Abazajian, K., et al. 2004, AJ, 128, 502
- Abell, G. O. 1958, ApJS, 3, 211
- Abell, G. O., Corwin, H. G., & Olowin, R. P. 1989, ApJS, 70, 1
- Adami, C., Ulmer, M. P., Romer, A. K., Nichol, R. C., Holden, B. P., & Pildis, R. A. 2000, ApJS, 131, 391
- Annis, J., et al. 1999, BAAS, 31, 1391
- Bade, N., et al. 1998, A&AS, 127, 145
- Bahcall, N. A., Fan, X., & Cen, R. 1997, ApJ, 485, L53
- Bahcall, N. A., et al. 2003, ApJS, 148, 243
- Baldry, I. K., Glazebrook, K., Brinkmann, J., Ivezić, Ž., Lupton, R. H., Nichol, R. C., & Szalay, A. S. 2004, ApJ, 600, 681
- Balogh, M., et al. 2004, MNRAS, 348, 1355
- Baum, W. A. 1959, in IAU Symp. 10, The Hertzsprung-Russell Diagram, ed. J. L. Greenstein (Paris: IAU), 23
- Beers, T. C., Flynn, K., & Gebhardt, K. 1990, AJ, 100, 32
- Benjamini, Y., & Hochberg, Y. 1995, J. R. Stat. Soc. B, 57, 289
- Blakeslee, J. P., et al. 2003, ApJ, 596, L143
- Blanton, M. R., Lin, H., Lupton, R. H., Maley, F. M., Young, N., Zehavi, I., & Loveday, J. 2003a, AJ, 125, 2276
- Blanton, M. R., et al. 2003b, AJ, 125, 2348
- Böhringer, H., et al. 2000, ApJS, 129, 435
- Bower, R. G., Lucey, J. R., & Ellis, R. S. 1992, MNRAS, 254, 589
- Carlstrom, J. E., Holder, G. P., & Reese, E. D. 2002, ARA&A, 40, 643
- Chester, C., & Roberts, M. S. 1964, AJ, 69, 635
- Cole, S., Hatton, S., Weinberg, D. H., & Frenk, C. S. 1998, MNRAS, 300, 945
- Collins, C. A., Guzzo, L., Nichol, R. C., & Lumsden, S. L. 1995, MNRAS, 274, 1071
- Cooray, A., & Sheth, R. 2002, Phys. Rep., 372, 1
- Dalton, G. B., Efstathiou, G., Maddox, S. J., & Sutherland, W. J. 1992, ApJ, 390, L1
- Diaferio, A., Kauffmann, G., Colberg, J. M., & White, S. D. M. 1999, MNRAS, 307, 537
- Dressler, A., & Shectman, S. A. 1988, AJ, 95, 985

- Ebeling, H., Edge, A. C., Fabian, A. C., Allen, S. W., Crawford, C. S., & Boehringer, H. 1997, *ApJ*, 479, L101
- Edge, A. C., & Stewart, G. C. 1991, *MNRAS*, 252, 414
- Eisenstein, D. J., et al. 2001, *AJ*, 122, 2267
- Eke, V. R., et al. 2004, *MNRAS*, 348, 866
- Evrad, A. E., et al. 2002, *ApJ*, 573, 7
- Frenk, C. S., White, S. D. M., Efstathiou, G., & Davis, M. 1990, *ApJ*, 351, 10
- Fukugita, M., Ichikawa, T., Gunn, J. E., Doi, M., Shimasaku, K., & Schneider, D. P. 1996, *AJ*, 111, 1748
- Gerke, B. F., et al. 2005, *ApJ*, 625, 6
- Gioia, I. M., Henry, J. P., Maccacaro, T., Morris, S. L., Stocke, J. T., & Wolter, A. 1990, *ApJ*, 356, L35
- Gladders, M. D. 2002, Ph.D. thesis, Univ. Toronto
- Gladders, M. D., & Yee, H. K. C. 2000, *AJ*, 120, 2148
- Gómez, P. L., et al. 2003, *ApJ*, 584, 210
- Goto, T., et al. 2002, *AJ*, 123, 1807
- Gott, J. R., III, Jurić, M., Schlegel, D., Hoyle, F., Vogeley, M., Tegmark, M., Bahcall, N., & Brinkmann, J. 2003, preprint (astro-ph/0310571)
- Gunn, J. E., et al. 1998, *AJ*, 116, 3040
- Henry, J. P., & Arnaud, K. A. 1991, *ApJ*, 372, 410
- Henry, J. P., et al. 1995, *ApJ*, 449, 422
- Hogg, D. W., Finkbeiner, D. P., Schlegel, D. J., & Gunn, J. E. 2001, *AJ*, 122, 2129
- Hogg, D. W., et al. 2003, *ApJ*, 585, L5
- Hopkins, A. M., Miller, C. J., Connolly, A. J., Genovese, C., Nichol, R. C., & Wasserman, L. 2002, *AJ*, 123, 1086
- Hopkins, A. M., et al. 2003, *ApJ*, 599, 971
- Jenkins, A., Frenk, C. S., White, S. D. M., Colberg, J. M., Cole, S., Evrard, A. E., Couchman, H. M. P., & Yoshida, N. 2001, *MNRAS*, 321, 372
- Kauffmann, G., Colberg, J. M., Diaferio, A., & White, S. D. M. 1999, *MNRAS*, 303, 188
- Kepner, J., Fan, X., Bahcall, N., Gunn, J., Lupton, R., & Xu, G. 1999, *ApJ*, 517, 78
- Kim, R. S. J., et al. 2002, *AJ*, 123, 20
- Kochanek, C. S., White, M., Huchra, J., Macri, L., Jarrett, T. H., Schneider, S. E., & Mader, J. 2003, *ApJ*, 585, 161
- Lacey, C., & Cole, S. 1994, *MNRAS*, 271, 676
- Lasker, B. M. 1970, *AJ*, 75, 21
- Lee, B. C., et al. 2004, *AJ*, 127, 1811
- Lucey, J. R. 1983, *MNRAS*, 204, 33
- Lumsden, S. L., Nichol, R. C., Collins, C. A., & Guzzo, L. 1992, *MNRAS*, 258, 1
- Mandolesi, N., Villa, F., & Valenziano, L. 2002, *Adv. Space Res.*, 30, 2123
- McClure, R. D., & van den Bergh, S. 1968, *AJ*, 73, 313
- Miller, C. J., Batuski, D. J., Slinglend, K. A., & Hill, J. M. 1999, *ApJ*, 523, 492
- Miller, C. J., Krughoff, K. S., Batuski, D. J., & Hill, J. M. 2002, *AJ*, 124, 1918
- Miller, C. J., Nichol, R. C., & Batuski, D. J. 2001a, *ApJ*, 555, 68
- . 2001b, *Science*, 292, 2302
- Miller, C. J., Nichol, R. C., Gómez, P. L., Hopkins, A. M., & Bernardi, M. 2003, *ApJ*, 597, 142
- Miller, C. J., et al. 2001c, *AJ*, 122, 3492
- Oegerle, W. R., & Hill, J. M. 2001, *AJ*, 122, 2858
- Oppenheimer, B. R., Helfand, D. J., & Gaidos, E. J. 1997, *AJ*, 113, 2134
- Ostrander, E. J., Nichol, R. C., Ratnatunga, K. U., & Griffiths, R. E. 1998, *AJ*, 116, 2644
- Oukbir, J., & Blanchard, A. 1992, *A&A*, 262, L21
- Perlmutter, S., & Schmidt, B. P. 2003, *Supernovae and Gamma-Ray Bursters* (Berlin: Springer), 195
- Petrosian, V. 1976, *ApJ*, 209, L1
- Pier, J. R., Munn, J. A., Hindsley, R. B., Hennessy, G. S., Kent, S. M., Lupton, R. H., & Ivezić, Ž. 2003, *AJ*, 125, 1559
- Popesso, P., Böhringer, H., Brinkmann, J., Voges, W., & York, D. G. 2004, *A&A*, 423, 449
- Postman, M., Lauer, T. R., Oegerle, W., & Donahue, M. 2002, *ApJ*, 579, 93
- Postman, M., Lubin, L. M., Gunn, J. E., Oke, J. B., Hoessel, J. G., Schneider, D. P., & Christensen, J. A. 1996, *AJ*, 111, 615
- Reichart, D. E., Nichol, R. C., Castander, F. J., Burke, D. J., Romer, A. K., Holden, B. P., Collins, C. A., & Ulmer, M. P. 1999, *ApJ*, 518, 521
- Romer, A. K., et al. 2000, *ApJS*, 126, 209
- . 2004, in *Clusters of Galaxies: Probes of Cosmological Structure and Galaxy Evolution*, ed. J. S. Mulchaey, A. M. Dressler, & A. Oemler (Cambridge: Cambridge Univ. Press), 47
- Sand, D. J., Treu, T., & Ellis, R. S. 2002, *ApJ*, 574, L129
- Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, *ApJ*, 500, 525
- Scranton, R., et al. 2002, *ApJ*, 579, 48
- Smith, J. A., et al. 2002, *AJ*, 123, 2121
- Söchtig, I. K., Clowes, R. G., & Campusano, L. E. 2002, *MNRAS*, 331, 569
- Stoughton, C., et al. 2002, *AJ*, 123, 485
- Strauss, M. A., et al. 2002, *AJ*, 124, 1810
- Sun, J., & Loader, K. 1994, *Ann. Stat.*, 22, 1328
- Sutherland, W. 1988, *MNRAS*, 234, 159
- Tully, R. B., Mould, J. R., & Aaronson, M. 1982, *ApJ*, 257, 527
- Viana, P. T. P., & Liddle, A. R. 1996, *MNRAS*, 281, 323
- Vikhlinin, A., McNamara, B. R., Forman, W., Jones, C., Quintana, H., & Hornstrup, A. 1998, *ApJ*, 502, 558
- Visvanathan, N., & Sandage, A. 1977, *ApJ*, 216, 214
- Voges, W., et al. 1999, *A&A*, 349, 389
- Wechsler, R. H. 2004, in *Clusters of Galaxies: Probes of Cosmological Structure and Galaxy Evolution*, ed. J. S. Mulchaey, A. M. Dressler, & A. Oemler (Cambridge: Cambridge Univ. Press), 53
- Wegner, G., Colless, M., Bagley, G., Davies, R. L., Bertschinger, E., Burstein, D., McMahan, R. K., & Saglia, R. P. 1996, *ApJS*, 106, 1
- Wei, J. Y., Xu, D. W., Dong, X. Y., & Hu, J. Y. 1999, *A&AS*, 139, 575
- White, M. 2002, *ApJS*, 143, 241
- White, R. A., Bliton, M., Bhavsar, S. P., Bornmann, P., Burns, J. O., Ledlow, M. J., & Loken, C. 1999, *AJ*, 118, 2014
- Wittman, D. 2002, *Gravitational Lensing: An Astrophysical Tool* (Berlin: Springer), 55
- Yang, X., Mo, H. J., Jing, Y. P., van den Bosch, F. C., & Chu, Y. 2004, *MNRAS*, 350, 1153
- Yang, X., Mo, H. J., van den Bosch, F. C., & Jing, Y. P. 2005, *MNRAS*, 356, 1293
- Yee, H., & Ellingson, E. 2003, *ApJ*, 585, 215
- York, D. G., et al. 2000, *AJ*, 120, 1579
- Zehavi, I., et al. 2002, *ApJ*, 571, 172
- Zwicky, F. 1952, *PASP*, 64, 247
- Zwicky, F., & Herzog, E. 1963, *Catalogue of Galaxies and of Clusters of Galaxies*, Vol. II (Pasadena: Caltech)
- . 1966, *Catalogue of Galaxies and of Clusters of Galaxies*, Vol. III (Pasadena: Caltech)
- . 1968, *Catalogue of Galaxies and of Clusters of Galaxies*, Vol. IV (Pasadena: Caltech)
- Zwicky, F., Herzog, E., & Wild, P. 1961, *Catalogue of Galaxies and of Clusters of Galaxies*, Vol. I (Pasadena: Caltech)
- Zwicky, F., Karpowicz, M., & Kowal, C. T. 1965, *Catalogue of Galaxies and of Clusters of Galaxies*, Vol. V (Pasadena: Caltech)
- Zwicky, F., & Kowal, C. T. 1968, *Catalogue of Galaxies and of Clusters of Galaxies*, Vol. VI (Pasadena: Caltech)