The universe at \( z > 5 \): When and how did the “dark age” end?

MARTIN J. REES

Institute of Astronomy, Madingley Road, Cambridge, CB3 OHA, United Kingdom

ABSTRACT  This paper considers how the first subgalactic structures produced the UV radiation that ionized the intergalactic medium before \( z = 5 \) and the “feedback” effects of the UV radiation on structure formation. The first “pre-galaxies” may eventually be detectable by their direct UV emission, with characteristic spectral features at Lyman \( \alpha \); high-\( z \) supernovae may also be detectable. Other probes of the intergalactic medium beyond \( z = 5 \), and of the epochs of reheating and reionization, are discussed, along with possible links between the diffusion of pregalactic metals and the origin of magnetic fields.

The structures in our present universe are the outcome of more than 10 billion years of evolution. Slight irregularities imprinted at very early eras led to increasing contrasts in the density from place to place, until overdense regions evolved into bound structures. Quantitative and detailed (albeit speculative) theories for the origin of the fluctuations (in an ultra-early inflationary phase, or else as the outcome of topological defects, etc.) can now be tested against increasingly detailed and precise data.

The data are of two kinds. (i) CMB fluctuations probe irregularities on scales relevant to observed cosmic structure and on larger scales that are still in the linear regime. The detected photons have (according to most assumptions) propagated almost uninterruptedly since the recombination era, when the universe had been expanding for less than a million years. (ii) The other line of evidence on cosmic structure comes from traditional astronomy. Any acceptable theory must account for the present clustering properties of galaxies and dark matter; it must also match the actual universe at all past eras that can be probed observationally. We have for some years known about quasars with redshifts up to 5. Quasars themselves may be associated with atypical (even exceptional) galaxies, so their intrinsic properties are hard to relate to the general trend of galaxy formation. What has been especially exciting about recent developments is that the morphology and clustering of ordinary galaxies can now be probed out to similar redshifts.

Clustering in Hierarchical Models

Characteristic Scales in Cold Dark Matter (CDM) Models. I will focus on the CDM model. But this is just a “template” for some more general deductions, which essentially apply to any “bottom up” model for structure formation. There is no minimum scale for the aggregation, under gravity, of cold nonbaryonic matter. However, the baryons constitute a gas whose pressure opposes condensation on very small scales. Therefore, the gas does not “feel” the very smallest condensations. In the context relevant here, where the dark matter’s gravity dominates, the baryonic Jeans mass is

\[
M_J = 5 \times 10^6 \left( \frac{\Omega_b}{0.025h^2} \right) \left( \frac{T_g}{T_{rad}} \right)^3 M_\odot
\]

Only on scales larger than this can baryons promptly condense into bound systems, along with the dark matter. [The proviso “promptly” is nontrivial because, as emphasized by Loeb (private communication), the baryons may be able to condense later into smaller systems.] During the “dark age” the gas became even cooler than the microwave background (i.e., \( T_g < T_{rad} \)): if it had cooled adiabatically, with no heat input since recombination, \( T_g \) would, at \( z = 10 \), have been below 5 K. The smallest bound structures, with mass \( \sim M_J \), would have virialized at a temperature a few times larger than \( T_g(z) \). Larger masses would virialize at temperatures higher by a further factor \((M/M_J)^{1/2}\). This virial temperature would be reached not solely by adiabatic compression, but also because of a shock, which would typically occur before the radius had decreased by a factor of 2.

These virialized systems would, however, have a null existence as stable clouds unless they could lose energy and deflate.
due to atomic or molecular radiative processes—clouds that couldn’t cool would simply remain in equilibrium, being later incorporated in a larger scale of structure as the hierarchy built up. On the other hand, clouds that can cool radiatively will deflate. Most cooling mechanisms are more efficient at higher temperatures, as well as at higher densities. Once collapse starts, it proceeds almost isothermally, so that the internal Jeans mass falls as the density rises. A virialized, self-gravitating cloud that can cool radiatively would eventually go into free-fall collapse and (perhaps after a disc phase) fragment into smaller pieces.

Three “cooling regimes” are relevant during successive phases of the cosmogenic process, each being associated with a characteristic temperature.

(i) For a H–He plasma the only effective cooling at low temperatures (<10^4 K) comes from molecular hydrogen. Even this process cuts off below a few hundred degrees; but above that temperature it allows contraction within the cosmic expansion timescale. The H2 fraction is never high, and it is, in any case, not a very efficient coolant—indeed, systems that collapse at z < 10 fail to form enough molecules for effective cooling [e.g., see figure 1 of Tegmark, et al. (1)]—but molecular cooling almost certainly played a role in forming the very first objects that lit up the universe

(ii) If H2 is prevented from forming, so that molecular cooling is ineffective, then a H–He mixture behaves adiabatically unless T is as high as 8,000–10,000°, when excitation of Lyman α by the Maxwellian tail of the electrons provides efficient cooling whose rate rises steeply with temperature. Because of this steep temperature dependence, gas in this regime contracts almost isothermally, so that its Jeans mass decreases as the density rises.

(iii) The UV from early stars will photoionize some (and eventually almost all) of the diffuse gas. When this happens, the HI fraction is suppressed to a very low level, so there is no cooling by collisional excitation of Lyman lines; moreover, the energy radiated whenever a recombination occurs is quickly canceled by the energy input from a photoionization, so the only net cooling is via bremsstrahlung. The cooling is, in effect, then reduced by a factor of 100 (see, for instance, ref. 2). The minimum temperature (below which there is a net heating from the UV) depends on the UV spectrum and on whether the He is doubly ionized: it is in the range of 20,000–40,000°. (These three regimes refer to a H–He plasma. When heavy elements are present they can dominate the low-T cooling; ionization is still important in suppressing the most efficient channels for cooling.)

The Role of Molecular Hydrogen and the UV Feedback. The role of molecular cooling at early cosmic epochs has been considered by many authors, dating back to the 1960s; recent discussions are given, for instance, by Tegmark et al. (1) and Haiman, Rees, and Loeb (3). The exact efficiency depends on the density and, therefore, on the redshift when the first collapse occurs.

However, even at high redshifts, H2 cooling would be quenched if there were a UV background able to dissociate the molecules as fast as they formed. Photons of hν > 11.18 eV can photodissociate H2, as first calculated by Stecher and Williams (4). These photons can penetrate a high column density of HI and destroy molecules in virialized and collapsing clouds. (If the incident spectrum has a nonthermal component extending up to KeV energies, there is a counterbalancing positive feedback because the number of photoelectrons is increased, and this enhances molecule formation.)

Only a small fraction of the UV that ionized the IGM could have been produced in systems where star formation was triggered by molecular cooling. Most must have formed in systems large enough to have been able to cool by atomic line effects. There is then a further transition when the medium breaks through and becomes completely ionized: the UV background intensity gets a boost, because the contributions from remote regions (which dominate in Olbers-type integrals) are less severely attenuated. The UV is then enough to maintain the very high mean ionization implied by transparency of the IGM beyond the Lyman limit. This means that it can maintain high ionization of a cloud until it has either collapsed to an overdensity exceeding the IGM ratio of ions to neutrals or until it becomes self-shielding (which happens at more modest overdensities for large clouds). Until that happens the cooling rate will be reduced by the lack of bound electrons and consequent elimination of the (otherwise dominant) “line” contribution to the cooling.

When this third phase is reached, the thermal properties of the uncollapsed gas will resemble those of the structures responsible for the observed Lyman–forest lines in high-z quasars spectra—these are mainly filaments, draining into virialized systems. Such systems have velocity dispersions of 50 km/sec and are destined to turn into galaxies of the kind whose descendants are still recognizable. I shall return (section 4) to discuss the detectability of these early galaxies.

Evidence for Diffuse Gas at High z

CMB Fluctuations as a Probe of the Ionization Epoch. If the intergalactic medium were suddenly reionized at a redshift z, then the optical depth to Thomson scattering back to z(≫1) would be

$$\tau_\theta = 0.05h^{-1} \left( \frac{1 + z_f}{10} \right) \left( \frac{\Omega_b}{0.025h^2} \right)$$

(the generalization to a more realistic scenario of gradual reionization is straightforward). Even when this optical depth is far below unity, the ionized gas constitutes a “fog” that attenuates the fluctuations imprinted at the recombination era; the photons that are scattered at <z_i manifest a different pattern of fluctuations, characteristically on larger angular scales. This optical depth is consequently one of the parameters that can, in principle, be determined from CMB anisotropy measurements. It is feasible to detect a value as small as 0.1—polarization measurements may allow even greater precision, because the scattered component would imprint polarization on angular scales of a few degrees, which would be absent from the Sachs–Wolfe fluctuations on that angular scale originating at \( h\Omega_i \).

Twenty-one-centimeter Emission, Absorption, and Tomography. The 21-cm line of HI at redshift z would contribute to the background spectrum at a wavelength of 21(1 + z) cm. This contribution amounts to a brightness temperature of order 0.05 (1 + z)\(^{1/2}\). This is very small compared with the 2.7 K of the CMB and smaller still compared with the nonthermal background, which swamps the CMB, even at high galactic latitudes, at the long wavelengths where high-z HI should show up. Nonetheless, inhomogeneities in the HI may be detectable, because they would give rise not only to angular fluctuations but also to spectral structure (5, 6). If the same strip of sky were scanned at two radio frequencies differing by (for example) 1 MHz, the temperature fluctuations due to the CMB itself, to galactic thermal and synchrotron backgrounds, and to discrete sources would track each other closely. Contrariwise, there would be no correlation between the 21-cm contributions, because the two frequencies would be probing “shells” in redshift space whose radial separation would exceed the correlation length. Consequently, it is not necessarily unfeasible to distinguish the 21-cm background, utilizing a radio telescope with a large collecting area. That line radiation allows three-dimensional tomography of the high-z HI renders this a specially interesting technique.
For the 21-cm contribution to be observable, the spin temperature $T_S$ must differ from that of the black-body cosmic background. The gas would be detected in absorption or in emission depending on whether $T_S$ is lower or higher than $T_{rad}$. The hyperfine levels of HI are affected by the microwave background itself, by collisional processes, and by Lyman $\alpha$ (whose profile is itself controlled by the kinetic temperature). $T_S$ will therefore be a weighted mean of the CMB and gas temperatures.

Before there had been any heat input due to the development of nonlinear structures, the kinetic temperature would be lower than that of the radiation, and the 21 cm would be an absorption feature. In principle, one might be able to detect incipient large-scale structure, even when still in the linear regime, because it leads to variations in the column density of HI, per unit redshift interval, along different lines of sight (5).

When reheating occurs, the situation becomes more complicated (6). The kinetic temperature can rise due to the weak shocking and adiabatic compression that accompanies the emergence of the first (very small scale) nonlinear structure (compare section 2). When photoionization starts, there will also, around each HII domain, be a zone of predominantly neutral hydrogen that has been heated by hard UV or x-ray photons. This latter effect would be more important if the first UV sources emitted radiation with a power-law (rather than just exponential) component.

Because the signal is so weak, there is little prospect of detecting high-$z$, 21-cm emission unless the signal displays structure on (comoving) scales of several million parsec (Mpc) (corresponding to angular scales of several arc minutes). According to CDM-type models, the gas is likely to have been already ionized, predominantly by numerous ionizing sources, each of subgalactic scale, before such large structures become conspicuous. On the other hand, if the primordial gas was heated by widely spaced, quasar-level sources, each of these regions of mass $10^8$ solar masses and internal velocity dispersion greater than about 15 km/sec. What is the chance of detecting these ancient stellar systems?

Estimating the cumulative amount of activity at high redshifts is straightforward. The total integrated UV production of detecting these ancient stellar systems?

A separate estimate of the number of O and B stars at these early eras comes from the high-brightness models, the UV that reheated and photoionized the IGM before its first absorption feature. In principle, one might be able to detect Lyman $\alpha$ forest heavy elements that have been heated by hard UV or x-ray photons. These will then be scattered many times by dust. The mean surface brightness of the Lyman $\alpha$ emission line from the galaxies is simply obtained by multiplying the photon comoving number density by $c/4\pi$, which yields $10^{-17} f_i f_d Z_{-4} (\Omega_b h^2/0.023)$ photons cm$^{-2}$sec$^{-1}$arcsec$^{-2}$. The fiducial value of $\Omega_b h^2$ used here is close to that implied by a primordial deuterium abundance $D/3^2 = 2 \times 10^{-5}$, as reported by Tytler, Fan, and Burles (13).

The Likely Luminosity of Individual Galaxies. The above arguments determine the average surface brightness of the high-redshift sky in the Lyman $\alpha$ emission line or, in other words, the product of the number of galaxies per solid angle and their individual, mean Lyman $\alpha$ flux. But the ease of detection obviously depends crucially on whether the UV background comes from huge numbers of individually ultra-faint systems or (more optimistically) at least in part from systems of higher luminosity. The uncertainty here is much greater—it is essentially tied to the poorly understood efficiency of forming massive stars in low-mass dark matter halos with shallow potential wells.

If star formation were highly efficient in all collapsed halos with velocity dispersion $\sigma \gtrsim 15$ km s$^{-1}$, corresponding to the virial temperature $T = \sigma^2/k = 10^4 K$, then reionization would take place as soon as the highest-density peaks on the scales of these low-velocity dispersions collapse. For a top-hat spherical region of mass $M$ that turns around when the age of the universe is $t_{\phi}/2$, the relation between mass and velocity dispersion is $M = 2^{3/2}/(2\pi G)^{1/2} \sim 10^9 M_\odot (\sigma/20 \text{ km s}^{-1})(t_{\phi}/10^9 \text{ yr})$. If all the baryons turned into stars over a timescale of order of the free-fall time of the halo, the star formation rate would then be $0.3M_\odot \text{ yr}^{-1} (\sigma/20 \text{ km s}^{-1})^3$. For a normal IMF, this can yield a luminosity close to the peak luminosity of a supernova. The Lyman $\alpha$ luminosity from this efficient, rapid starburst
would be \(10^{52} f_{\alpha} (\sigma/20 \text{ km s}^{-1})^3 (f_b/0.1) \) photons per sec\(^{-1}\). The corresponding Lyman \(\alpha\) flux at present (for \(\Omega = 1\) and \(H_0 = 70 \text{ km s}^{-1}\text{Mpc}^{-1}\)) is \(10^{-6} f_{\alpha} (\sigma/20 \text{ km s}^{-1})^3 (f_b/0.1) [1 - (1 + z)^{-1/2}]^{-2} (1 + z)^{-1} \) photons per cm\(^2\) sec\(^{-1}\). If the sources that reionized the universe have these characteristics, then we can use the mean Lyman \(\alpha\) surface brightness obtained earlier to conclude that the number density of these sources in the sky should be \(\sim 1/\text{arcsec}^2\) for the fiducial numbers we have chosen, and a formation redshift \(z_f \sim 10\). The redshift \(z_f\) should of course depend on the detailed model for the amplitude of the primordial fluctuations on small scales.

The main problem with this scenario is that supernova explosions in galaxies of such low velocity dispersions may well expel the gas before more than a small fraction has turned to stars. This leaves us with two other options regarding the sources of the reionization photons. (i) If just a small fraction of the gas in each object turns to stars before the remainder is expelled in a wind, then each galaxy emits fewer photons. Therefore, more have to form from lower amplitude peaks. The luminosity of each galaxy would be reduced. (ii) Star formation may be so inefficient in the low-escape-velocity halos that reionization has to await the formation of more massive galaxies with \(\sigma \gtrsim 100 \text{ km s}^{-1}\) (photoionization has little effect on the cooling rate in halos with velocity dispersion \(\sigma \gtrsim 100 \text{ km s}^{-1}\); see refs. 14 and 15).

Some moderately massive galaxies are in any case expected to form before reionization, because the power spectrum at small scales in cold dark matter models flattens to a slope of about \(n = -2.7\) (so the amplitude of fluctuations decreases very slowly with scale), implying that at the same epoch, halos with \(\sigma = 70 \text{ km s}^{-1}\) (with a mass \(\sim 40\) times larger than halos with \(\sigma = 20 \text{ km s}^{-1}\)) would be collapsing from \(3 - \sigma\) peaks on this larger scale. In a Gaussian theory, the \(3 - \sigma\) peaks should contain \(\sim 10\%\) as much mass as the \(2 - \sigma\) peaks at a fixed epoch; so we see that the mass distribution of these first-generation galaxies should probably extend well above the minimum mass for efficient atomic cooling.

**Detectability of Galaxies Beyond the Reionization Redshift \(z_f\).** What is the chance of detecting these galaxies? Given the mean surface brightness of Lyman \(\alpha\) photons derived previously, the surface brightness in the UV continuum follows with little additional uncertainty, because the stars emitting most of the light in the range \(1,216 \text{ Å} < \lambda < 2,500 \text{ Å}\) are also mostly very young stars. For every Lyman \(\alpha\) photon produced from ionizing photons there should be \(\sim 10\) UV continuum photons. The mean surface brightness from all galaxies in the rest-frame UV continuum is \(10^{-5}\) photons cm\(^{-2}\) sec\(^{-1}\) arcsec\(^{-2}\). The main spectral feature that should identify any such galaxies at \(z \gtrsim 5\) is a sharp break in the UV continuum at the Lyman \(\alpha\) wavelength, due to the Gunn–Peterson trough (16). In addition, the Lyman \(\alpha\) emission line may be present, depending on dust absorption and scattering of the Lyman \(\alpha\) photons. Notice that the redshift at which the IGM was reionized is only marginally relevant to the presence of the Gunn–Peterson edge, because even if the medium was reionized at \(z \gg 5\), we know that the flux decrement caused by the Lyman \(\alpha\) forest reaches a factor of 2 at \(z = 4\) and grows rapidly with redshift. Thus, the technique of identifying galaxies at \(z = 3\) from the Lyman continuum break (17, 18) should be replaced by the Gunn–Peterson trough at \(z \approx 5\) (see refs. 19 and 20 for a careful analysis of the effects of the Lyman \(\alpha\) forest on galaxy colors).

The mean surface brightness in the rest-frame UV continuum from these galaxies can also be expressed as \(S_v = 6 \times 10^{-33} \text{ erg cm}^{-2} \text{ sec}^{-1} \text{ Hz}^{-1} \text{ arcsec}^{-2} = 32 \text{ AB arcsec}^{-2}\). The number density of these sources in the sky is \(\sim 1/\text{arcsec}^2\). If there is one galaxy per arcsec\(^2\) at redshifts \(z_f \sim 10\), each galaxy would have an AB magnitude of 32, i.e., about 3 magnitudes fainter than the detection limit in the Hubble Deep Field for the I-band (21). However, as mentioned above, galaxies with a mass 40 times larger would be \(\sim 400\) times rarer, which still implies a number density of a few galaxies per square arc minute with AB magnitudes of \(\sim 28\) (with the optimistic assumption of a rapid and efficient starburst in these more massive galaxies).

Detection with ground-based telescopes should be possible at \(z \lesssim 6\). The Keck telescope can detect point sources to AB magnitudes \(R \gtrsim 28\), and \(I \lesssim 27\), in a night of observing time [M. Rauch, personal communication; Cohen (23, 24)]. Galaxies up to \(z = 6\) can be detected in the I band; for higher redshift, the magnitude limit from a ground-based telescope degrades rapidly due to the high atmospheric background at \(\lambda > 8,500\) Å. The magnitude limit could be improved significantly with adaptive optics. The HST detects galaxies in the Hubble Deep Field down to \(F = 28.5\) (ref. 21). The New Generation Space Telescope (NGST) could image galaxies to AB magnitudes \(\sim 31\) in the near-infrared and should be able to detect galaxies to a much higher redshift (24) with a much higher number density than has been seen so far. The prospect for detecting the high-redshift galaxies responsible for the enrichment of the Lyman \(\alpha\) forest has also been analyzed by Haiman and colleagues (25) and by Cen (R. Cen, unpublished work).

In general, the lensing magnification in rich lensing clusters may be used here to stretch the magnitude limit (see also R. Cen, unpublished work). As an example, a lensing cluster with an Einstein ring radius \(b = 30'\) should magnify to \(|A| > 10\) an area of \(\sim 30\) arcsec\(^2\) in the source plane. In the example used above, about 30 galaxies with \(AB = 32\) could be in this area, which would be magnified to \(AB = 29.5\). Magnified images of high-redshift galaxies should characteristically appear in pairs around the critical lines, in a region that can be predicted from lensing models [see Miralda-Escudé and Fort (26), Kneib et al. (27), and references therein], so this should help in their identification. These numbers indicate that a new deep field (similar to the Hubble Deep Field) imaged with HST in a rich cluster, adding also the H and J filters in the near-infrared, might well identify several galaxies at \(z > 5\). In fact, the largest redshift object known at present (at \(z = 4.92\) is already a gravitationally lensed galaxy (M. Franx, G. D. Illingworth, D. D. Kelson, P. G. van Dokkum, and K.-V. Tran, unpublished work).

The foregoing discussion has assumed that these high-redshift galaxies would be sufficiently small to remain unresolved. Resolved objects would need to have higher fluxes to be detected, because the detection is limited by the sky background. The likely dimensions of the star-forming regions are \(\sim 100\) parsec (pc), corresponding to angular sizes \(\sim 0.01\) arcsec—corresponding to sources with NGST.

**The Lyman \(\alpha\) Emission Line.** Another possible way to detect the faint galaxies is to search directly for the Lyman \(\alpha\) emission line. As we discussed before, for a normal starburst spectrum we expect \(\sim 10\%\) of the UV photons to be in the Lyman \(\alpha\) emission line if dust absorption is not important. Therefore, if the sensitivity for detecting galaxies is still limited by the sky background for emission-line searches, the width of the line should be \(\Delta \lambda/\lambda < 0.01\) to allow detection on a shorter time than for the UV continuum. The width of the Lyman \(\alpha\) line in emission from a region of neutral gas with column density \(N_H = 10^{22} \text{ cm}^{-2}\) and velocity dispersion \(\sigma = 10 \text{ km s}^{-1}\) is \(\Delta \lambda/\lambda = 2 \times 10^{-3}\). Thus, the sensitivity of emission-line searches might be, at best, similar to searches for the UV continuum, and of course a much smaller range of redshift is searched because a very narrow band must be used.

The Lyman \(\alpha\) emission line may be suppressed because of internal dust extinction. In addition, any photon emitted in the blue side of the Lyman \(\alpha\) line will be scattered in the surrounding IGM, reducing the strength of the line by a factor of 2. For a galaxy observed when the surrounding IGM is still mostly neutral, even photons on the red side can be scattered because of the damping wing of the Gunn–Peterson trough (34). This width of the damped absorption from the IGM is broader than the emission lines from starburst galaxies, so any
Lyman $\alpha$ emission line should be suppressed by the IGM in any galaxy observed before the reionization. The edge of the Gunn–Peterson trough should be “rounded off” into a characteristic shape (34), except that the inhomogeneity of the IGM may alter the profile (in particular, the presence of a halo of gas accreting on the galaxy may substantially increase the column density contributing to the damped profile from gas at a redshift close to that of the galaxy).

Thus, even though the Lyman $\alpha$ emission line might not be as good a spectral signature as the Gunn–Peterson trough as a technique to search for high redshift galaxies, its presence or absence should be an important diagnostic of the ionization state of the IGM in the neighborhood of any detected galaxy.

A Note on Active Galactic Nuclei (AGNs) and Nonthermal UV Background. AGN formation requires virialized systems with larger masses and deeper potential wells (cf. ref. 29). AGNs may “take over” as the dominant UV source at redshifts below 5 (and the second ionization of He may be delayed until AGNs can provide a power-law contribution to the spectrum), but at higher $z$, when the $H$ itself was reheated and ionized, OB stars would dominate.

It is worth noting also that the diffuse IGM at redshifts of 3–5 has such a long thermal timescale that it retains a “memory” of its thermal history at even higher redshifts. The adiabatic this gas is on at $z = 5$ depends on the value of $z$ and also on the slope of the UV spectrum. This adiabatic can be constrained by observations of weak and narrow lines in the Lyman forest (30, 31).

Very High Redshift Supernovae (and Gamma-Ray Bursts?)

The intergalactic gas was already highly photoionized by $z = 5$; also, the mean abundance of heavy elements had attained a level about 0.01 solar by that time, this degree of contamination being about what would be expected if the reheating and ionization were due to OB stars (25, 32). It is straightforward to calculate how many supernovae would have gone off, in each comoving volume, as a direct consequence of this output of UV and heavy elements; there would be one, or maybe several, per year in each square arc minute of sky (33). These would be primarily of type 2. The typical observed light curve has a flat maximum lasting 80 days. One would therefore (taking the time dilation into account) expect each supernova to be near its maximum for nearly a year. It is possible that the explosions proceed differently when the stellar envelope is essentially metal-free, yielding different light curves, so any estimates of detectability are tentative. However, taking a standard type 2 light curve (which may, of course, be pessimistic), one calculates (33) that these objects should be approximately 27th magnitude in $J$ and $K$ bands even out beyond $z = 5$. The detection of such objects would be an easy task with the NGST.

With existing facilities it is marginal. The best hope would be that observations of clusters of galaxies might serendipitously detect a magnified, gravitationally lensed image from far behind the cluster.

As a speculative addendum, note that a few percent of observed gamma-ray bursts may come from redshifts as large as 5; this would be expected if the burst rate, as a function of cosmic epoch, tracks the star-formation rate. At the time of writing, data on optical and x-ray afterglows are still very sparse, but it is at least an exciting possibility that there may be occasional flashes, far brighter than supernovae, from very large redshifts.

The First Stars: Some Uncertainties

The gravitational aspects of clustering can all be modeled convincingly by computer simulations. Now the dynamics of the baryonic (gaseous) component—including shocks and radiative cooling—can be modeled this way as well. But the nature of the simulation changes as soon as the first stars (or other compact objects) form. The first stars exert crucial feedback—the remaining gas is heated by ionizing radiation and perhaps also by an injection of kinetic energy via winds and even supernova explosions. The three uncertainties here are:

(i) What is the IMF of the first stellar population? The high-mass stars are the ones that provide efficient (and relatively prompt) feedback. It plainly makes a big difference whether these are the dominant type of stars or whether the initial IMF rises steeply toward low masses, so that very many faint stars form before there is a significant feedback.

(ii) The influence of the early stars depends on whether their energy is deposited locally or penetrates into the all-pervasive medium that is not yet in contracting systems. The UV radiation could, for instance, be mainly absorbed in the gas immediately surrounding the first stars, so that it exerts no feedback on the condensation of further clumps—the total number of massive stars needed to build up the UV background, and the concomitant contamination by heavy elements, would then be greater.

(iii) Quite apart from the uncertainty in the IMF, it is also unclear what fraction of the baryons that falls into a clump would actually be incorporated into stars before being reejected. The retained fraction would almost certainly depend on the virial velocity; gas more readily escapes from the shallow potential wells relevant to dwarf galaxies. Ejection is even easier in potential wells so shallow that they cannot confine gas at the photoionization temperature. This effect may lead to a reduced efficiency during the second, as compared with the third, of the three cosmic stages discussed in section 2.

All of these three uncertainties would, for a given fluctuation spectrum, affect the redshift at which molecules were destroyed and at which full ionization occurred.

Heavy Elements and Magnetic Fields

If the main UV source is stars, there is inevitably an associated build-up of heavy elements. (In more radical pictures where black holes are involved in the early energy input, this inference doesn’t hold, because the energy supply could be gravitational rather than nuclear.) The question then arises of how this processed gas would be distributed. Would it be confined in the virialized systems, or could it spread through the entire IGM?

The ubiquity of carbon features in intermediate- and high-(N $\geq 3.10^{14}$ cm$^{-2}$) column-density systems implies that heavy elements are dispersed broadly enough to have a large covering factor. These absorption systems may be associated with the subgalactic sites of star formation that produce the first heavy elements. The nucleosynthesis sites therefore cannot be too sparse if these elements are, within the time available, to diffuse enough so that they are encountered somewhere along every line of sight through a typical high-column-density cloud. (Moreover, efficient mixing would then certainly imply bulk motions so high that their neglect would be a serious inadequacy in simulations of Lyman $\alpha$ forest formation, etc.) The absorption line data tell us the mean abundance through the relevant cloud. They are compatible with 99% of the material being entirely unprocessed, the heavy elements being restricted to 1% of the material—the early heavy elements need not be thoroughly mixed, but they must have spread sufficiently to have a large “covering factor” in the intermediate- and high-N clouds.

The first stars are important for another reason: they may generate the first cosmic magnetic fields. Moreover, mass loss (via winds or supernovae permeated by magnetic flux) would disperse magnetic flux along with the heavy elements. This flux, stretched and sheared by bulk motions, can be the “seed” for the later amplification processes that generate the largescale fields pervading disc galaxies.
Where Are the Oldest Stars?

The efficiency of early mixing is important for the interpretation of stars in our own galaxy that have ultra-low metallicity (33)—lower than the mean metallicity that would have been generated in association with the UV background at $z > 5$. If the heavy elements were efficiently mixed, then these stars would themselves need to have formed before galaxies were assembled. To a first approximation they would cluster nondissipatively; they would therefore be distributed in halos (including the halo of our own galaxy) like the dark matter itself. More careful estimates slightly weaken this inference. This is because the subgalaxies would tend, during the subsequent mergers, to sink via dynamical friction toward the centers of the merged systems. There would nevertheless be a tendency for the most extreme metal-poor stars to have a more extended distribution in our galactic halo and to have a bigger spread of motions.

The number of such stars depends on the IMF. If this were flatter, there would be fewer low-mass stars formed concurrently with those that produced the UV background. If, on the other hand, the IMF were initially steeper, there could, in principle, be a lot of very low mass (macho) objects produced at high redshift. These could provide a few percent of the halo if omega were 1 (and a higher proportion of the $\approx 30$-kpc inner halo probed by lensing searches); a larger proportion could be provided in a low-density universe.

Most of the small, first-generation galaxies by now should have merged into more massive systems (forming part of the halo population of stars in normal galaxies like the Milky Way), but some could survive until today in galactic halos or even as isolated objects. This may be the explanation of present-day dwarf spheroidal galaxies (34).

Summary

Our general conclusions are relevant to any model in which the initial fluctuations have amplitudes decreasing with scale, so that cosmic structures form “bottom-up.” Such models differ, of course, in the epoch at which “first light” would have occurred. In models with primordial baryon fluctuations of course, in the epoch at which “first light” would have occurred, the universe becomes, in effect, an HII region. This must have happened before $z = 5$. The only net cooling of a fully photoionized gas comes from bremsstrahlung, which is less effective than the collisionally excited line emission from gas that is only partly ionized. The completion of photoionization may therefore signal another pause in the cosmogonic process, associated with a further increase in the minimum scale that can collapse, and in the efficiency of cooling.

By the epoch $z = 5$, some structures (albeit perhaps only exceptional ones) must have attained galactic scales. But huge numbers of lower-mass systems should already exist at higher redshifts, and we can make quite firm estimates of their integrated UV output. This paper has addressed how we can probe their indirect effects and perhaps even detect them directly.

I am especially grateful to Jordi-Miralda Escude and would also like to thank Zoltan Haiman, Avi Loeb, and Max Tegmark for discussion and collaboration on some of the topics described here.