The Astrophysical Journal, 804:131 (16pp), 2015 May 10 doi:10.1088/0004-637X/804/2/131

SUPERMASSIVE BLACK HOLES FROM ULTRA-STRONGLY SELF-INTERACTING DARK MATTER

Jason Pollack[1], David N. Spergel[2], and Paul J. Steinhardt[3]
[1] Walter Burke Institute for Theoretical Physics, Division of Physics, Mathematics & Astronomy, California Institute of Technology, Pasadena, CA 91125, USA; jpollack@caltech.edu
[2] Department of Astrophysical Sciences, Peyton Hall, Princeton University, Princeton, NJ 08544, USA
[3] Department of Physics and Princeton Center for Theoretical Science, Princeton University, Princeton, NJ 08544, USA
Received 2014 December 27; accepted 2015 March 5; published 2015 May 12

ABSTRACT

We consider the cosmological consequences if a small fraction ($f \lesssim 0.1$) of the dark matter is ultra-strongly self-interacting, with an elastic self-interaction cross section per unit mass $\sigma \gg 1$ cm^2 g^{-1}. This possibility evades all current constraints that assume that the self-interacting component makes up the majority of the dark matter. Nevertheless, even a small fraction of ultra-strongly self-interacting dark matter (uSIDM) can have observable consequences on astrophysical scales. In particular, the uSIDM subcomponent can undergo gravothermal collapse and form seed black holes in the center of a halo. These seed black holes, which form within several hundred halo interaction times, contain a few percent of the total uSIDM mass in the halo. For reasonable values of σf, these black holes can form at high enough redshifts to grow to $\sim 10^9\, M_\odot$ quasars by $z \gtrsim 6$, alleviating tension within the standard Λ cold dark matter cosmology. The ubiquitous formation of central black holes in halos could also create cores in dwarf galaxies by ejecting matter during binary black hole mergers, potentially resolving the "too big to fail" problem.

Key words: black hole physics – dark matter – galaxies: evolution – galaxies: halos – galaxies: structure

1. INTRODUCTION

Although Λ cold dark matter (ΛCDM) cosmology provides an excellent fit to the observational data on $\gtrsim$Mpc scales (Planck Collaboration et al. 2013), its success is less certain over the strongly nonlinear, $\lesssim$ kpc regime relevant to the substructure within galactic halos. The deviation of galactic cores from an expected cuspy density profile (Moore 1994; Moore et al. 1998) and an apparent shortfall of observed Milky Way satellites relative to expectations from simulations (Klypin et al. 1999; Moore et al. 1999) originally motivated considerations that the dark matter might have non-negligible self-interactions (Spergel & Steinhardt 2000). Although a combination of improved theoretical understanding and additional observations had appeared to alleviate these problems and remove the phenomenologically interesting parameter space for self-interacting dark matter (SIDM; Gnedin & Ostriker 2001; Yoshida et al. 2000; Markevitch et al. 2004), a recent reevaluation of the constraints (Peter et al. 2012; Rocha et al. 2013) has demonstrated that SIDM with an velocity-independent elastic self-interaction cross section per unit mass $\sigma \simeq 0.1$–1 cm^2 g^{-1} $\simeq 0.2$–2 b/GeV can simultaneously meet all constraints and alleviate the discrepancies between ΛCDM and observations. (In particular, SIDM with such a cross section appears to successfully create cores of the correct size rather than cusps; it is harder to reduce substructure within halos with a cross section of the required magnitude (Vogelsberger et al. 2012; Rocha et al. 2013) without resorting to inelastic collisions or velocity dependence.)

In this paper, we exhibit a distinct area of the SIDM parameter space which is likewise both allowed by observations and potentially interesting phenomenologically. In particular, we examine the case in which most of the dark matter remains non-self-interacting (or weakly self-interacting) as in the standard ΛCDM picture, but a small fraction $f \ll 1$ of the dark matter is made up of a subdominant component that is ultra-strongly self-interacting, abbreviated as uSIDM, with $\sigma \gg 1$ cm^2 g^{-1} (where σ denotes the cross section per unit mass). Because most of the dark matter remains inert, constraints that rely on distinguishing the overall behavior of SIDM halos from their CDM counterparts are no longer relevant.

Consider, for example, the constraints placed on the SIDM cross section from observations of the Bullet Cluster (1E 0657-6). Observations reveal an offset between the gas "bullet" and the dark matter centroid of the currently merging subcluster. Under the assumption that the subcluster has already passed through the main cluster, this offset is due to stripping and deceleration of gas in the subcluster due to interactions with the main cluster itself. The observation that the dark matter has not been slowed to the same degree allows limits to be placed on the dark matter self-interaction cross section. The strongest constraint (Randall et al. 2008) comes from the measurement of the ratio of mass-to-light ratios of the subcluster and the main cluster, which is found to be 0.84 ± 0.07. Under the assumption that the subcluster and main cluster had the same initial mass-to-light ratio before merger, this means that the subcluster cannot have lost more than 23% of its mass.

In Randall et al. (2008), this measurement plus estimates of the subcluster escape velocity and merger speed were used to constrain $\sigma \lesssim 0.6$ cm^2 g^{-1} when $f = 1$. However, it is clear that, even in the extreme example that *all* of the SIDM mass in the subcluster was lost to scattering, current observations would not be able to detect the SIDM subcomponent if $f < 0.07$, within the uncertainty on the mass-to-light ratio. So constraints from the Bullet Cluster certainly do not apply when $f \ll 10^{-1}$, regardless of the size of the self-interaction cross section per unit mass σ. Even when $f \sim 0.1$, σ may not be well-constrained, since one pass through the main cluster would not suffice to strip all of the SIDM from the bullet.

We note that observations of another cluster undergoing a major merger, A520 (Mahdavi et al. 2007; Clowe et al. 2012; Jee et al. 2012, 2014) have not provided similar constraints on

the SIDM cross section; here the dark matter centroid of the subcluster is in fact coincident with the (presumably stripped) gas. Under certain assumptions, this can be taken as evidence of a nonzero dark matter self-interaction cross section per unit mass, as strong as $0.94 \pm 0.06\,\mathrm{cm^2\,g^{-1}}$ in the latest observations (Jee et al. 2014). The limited number of ongoing major merger events in the observable universe makes it hard to give an overall estimate of the self-interaction cross section from major mergers, but future surveys could potentially combine many minor merger events to measure σ with a precision of $0.1\,\mathrm{cm^2\,g^{-1}}$ (Harvey et al. 2014).

Regardless of the situation for $f = 1$ SIDM, we have seen that there are no observational constraints on a uSDIM component of the dark matter with $f \lesssim 0.1$. At the same time, of course, a small component of uSIDM by itself is unable to produce cores or dissolve substructure to any observable degree. We point out, though, that a uSIDM component of the dark matter could instead explain another potential discrepancy with the ΛCDM picture: the existence of billion-solar-mass quasars at high redshifts $z \gtrsim 6.5$–7 (for reviews, see Dokuchaev et al. 2007; Volonteri 2010; Sesana 2012; Kelly & Merloni 2012; Treister & Urry 2012; Haiman 2013). (A number of recent papers (McCullough & Randall 2013; Fan et al. 2013a, 2013b; Randall & Reece 2014; Randall & Scholtz 2014) consider models of "Double-Disk Dark Matter," where a subdominant dark matter component self-interacts dissipatively by emitting dark radiation and can cool to form disks. The dissipative nature of their interactions means that such models do not produce the effects exhibited in our paper.) In Section 2 we review the observational situation and the difficulties with explaining it within ΛCDM. In Section 3 we suggest an alternative: gravothermal collapse of an uSIDM component. We review the mechanism of gravothermal collapse, specialize to the case of a halo containing uSIDM, and solve the problem numerically. We apply the results of Section 3 to individual observations of high-redshift quasars in Section 4, then discuss broader cosmological implications in Section 5, including a potential way for uSIDM to indirectly produce cores in dwarf halos. We finally conclude in Section 6.

2. SUPERMASSIVE BLACK HOLES

Supermassive black holes (SMBHs) that grow primarily via gas accretion are Eddington-limited: the gravitational force on the accreting gas is balanced by its own radiation pressure. Hence growth via gas accretion cannot proceed faster than exponentially, with an e-folding rate bounded by the inverse of the Salpeter time (Salpeter 1964):

$$t_{\mathrm{Sal}} = \frac{\epsilon_r\,\sigma_{\mathrm{T}}\,c}{4\pi G m_p} \approx \left(\frac{\epsilon_r}{0.1}\right) 45.1\ \mathrm{Myr}, \tag{1}$$

where σ_{T} is the Thompson cross section,

$$\sigma_{\mathrm{T}} = \frac{8\pi}{3}\left(\frac{e^2}{4\pi\epsilon_0 m_e\,c^2}\right)^2, \tag{2}$$

m_p and m_e are respectively the proton and electron masses, and ϵ_r is the radiative efficiency, which ranges from $1 - \sqrt{8/9} \approx 0.057$ to $1 - \sqrt{1/3} \approx 0.42$ as the angular momentum of the black hole increases from zero to its extremal value (Shapiro 2005); in astrophysical applications, ϵ_r is typically taken to be $\epsilon_r = 0.1$. Accretion faster than the Eddington limit, $\dot{M}_{\mathrm{Edd}} = M t_{\mathrm{Sal}}^{-1}$, onto a black hole of mass M will result in a radiation pressure exceeding the gravitational force, driving outflows which should quickly halt this excessive accretion. Yet several dozen quasars with masses a few $\times 10^9\,M_\odot$ have been detected at redshifts $z \gtrsim 6$, including a quasar, ULAS J1120+0641, with mass $2.0^{+1.5}_{-0.7} \times 10^9\,M_\odot$ at redshift $z = 7.085$ (Mortlock et al. 2011; Venemans et al. 2012). Using the Planck Collaboration's best-fit cosmological values (Planck Collaboration et al. 2013), $z = 7.085$ corresponds to 747 Myr after the Big Bang, so even continuous Eddington accretion since the Big Bang can only increase the mass of a seed black hole by a factor of 1.6×10^7. If we make the standard assumption that black hole seeds are formed from Pop III stars, the seed cannot have formed before around $z \sim 30$, so the maximum growth factor shrinks by another order of magnitude, to 1.75×10^6, requiring a seed black hole mass $\sim 10^3\,M_\odot$.

More generally, in order to explain the observed abundance of $\sim 1\,\mathrm{Gpc^{-3}}$ billion-solar-mass quasars at $z \simeq 6$ (Haiman 2013) within ΛCDM, we must form 10^2–$10^3\,M_\odot$ seed black holes soon after the beginning of baryonic structure formation and grow these black holes continuously at near-Eddington rates for ∼800 Myr. Some simulations have shown this can be achieved (Li et al. 2007), but only by making optimistic assumptions about cooling and star formation (Tegmark et al. 1997; Gao et al. 2007), fragmentation (McKee & Tan 2007; Stacy et al. 2009; Turk et al. 2009), photoevacuation (Johnson & Bromm 2007; Abel et al. 2007; Yoshida et al. 2007), black hole spin (Bardeen et al. 1972; Zhang et al. 1997; Narayan & McClintock 2012), and black hole mergers (Fitchett 1983; Haiman 2004; Merritt et al. 2004). We emphasize, in particular, that these results depend critically on the assumption of $\epsilon_r = 0.1$; because the e-folding time itself depends linearly on the radiative efficiency, the maximum mass formed by a given time is *exponentially* sensitive to its value. Because quasar masses are inferred by measuring their luminosities and assuming they are Eddington-limited, increasing the assumed radiative efficiency will decrease the inferred quasar mass by ϵ_r^{-1}. However, this reduction in required mass is made negligible by the much larger number of e-folds required to reach it. Recent work, both theoretical (Shapiro 2005) and observational (Trakhtenbrot 2014), has found $\epsilon_r \gtrsim 0.2$, which would be catastrophically incompatible with an assumption of black hole growth driven by Eddington accretion.

One alternative is to allow for extended periods of super-Eddington accretion. Super-Eddington accretion of baryons is known to be possible, for example when outflows of gas and radiation are collimated (Shakura & Sunyaev 1973; Jiang et al. 2014), and extended periods of super-Eddington growth could account for the observed supermassive high-redshift quasars (Volonteri & Silk 2014; Madau et al. 2014). However, estimates of quasar masses and luminosities at low redshifts using emission line widths indicate that, at least in the late universe, the vast majority of quasars are constrained to radiate at the Eddington limit (Kollmeier et al. 2006), or possibly well below it (Steinhardt & Elvis 2010; Steinhardt et al. 2011). (Alternatively, black holes could grow by accreting non-baryonic matter, which, provided it does not radiate, is not Eddington-limited. Ostriker (2000) pointed out that $f = 1$

SIDM could accrete efficiently onto seed black holes (assumed in Ostriker 2000 to originate in stellar collapse) and contribute 10^6–$10^9\,M_\odot$ to SMBH masses. Because we take $f \ll 1$, this mechanism will not be important for us; we will use uSIDM primarily to create seeds, not to grow them after their formation.)

In this paper, we will therefore neglect the possibility of extended super-Eddington accretion. We will assume that growth of black holes from baryonic accretion is limited to exponential growth with an e-folding time given by the Salpeter time (Equation (1)). In order to facilitate comparision of uSIDM to the standard picture, we will, however, allow for continuous accretion of baryons at this limit once a seed black hole has formed, despite the potential issues mentioned in the previous paragraph. In other words, we attempt to modify the mechanism by which black hole seeds are formed, while leaving the simplest conventional mechanism for their growth from seeds to SMBHs intact. It would be easy to combine our results with more realistic baryon accretion histories.

Finally, we note that future observations in the near-infrared, e.g., with the *James Webb Space Telescope* and Wide Field Infrared Survey Telescope, and in the radio, e.g., with the Square Kilometre Array, should be able to detect (or place limits on the density of) even intermediate-mass ($\sim 10^5\,M_\odot$) quasars out to $z \sim 10$ (Haiman & Loeb 1998, 2000; Haiman et al. 2004; Whalen et al. 2013), providing vastly more information about the formation and growth of high-redshift quasars.

3. GRAVOTHERMAL COLLAPSE

Motivated by the tensions within the standard (ΛCDM) picture discussed in the previous section, we propose an alternative mechanism for black hole seed formation: the gravothermal collapse (Lynden-Bell & Wood 1968) of the uSIDM component of a dark matter galactic halo. The simplest form of gravothermal collapse occurs in a population of gravitating point particles with elastic short-range interactions. The classic illustration of the mechanism is globular clusters, where the point particles are stars. Stellar short-range interactions are not purely elastic, so in this case collapse is eventually halted by binary formation. A gas of SIDM, however, has only elastic interactions, so core collapse continues until relativistic instability results in the formation of a black hole, which promptly Bondi accretes (Bondi 1952) the optically thick core of SIDM that surrounds it.

In this section we make this intuitive picture precise. Full expressions will be given below, but in brief we find that the uSIDM component of a galactic halo undergoes gravothermal collapse in ~460 halo relaxation times, forming a black hole which contains ~2% of the uSIDM mass of the galaxy. The halo relaxation time is a complicated expression that depends on the halo mass and time of formation as well as the uSIDM properties, but we show in the following sections that, for reasonable values of uSIDM fraction f and cross section per unit mass σ, there exist halos that can easily form seed black holes, and grow them using uSIDM and baryons, to achieve $10^9\,M_\odot$ SMBHs by redshift 6.

Before formulating the problem, we first review the gravothermal collapse mechanism itself. Intuitively, gravothermal collapse depends on the simple observation that gravitationally bound systems have negative specific heat. For a virialized system, this is immediate:

$$0 = 2T + V = T + E \rightarrow E = -T. \tag{3}$$

Now consider two systems, an inner, gravitationally bound system with negative specific heat and an outer system surrounding it with positive specific heat—the inner and outer parts of a globular cluster, for example. Evolution toward equilibrium will direct both mass and heat outward, causing both the inner and the outer system to increase in temperature. A possible physical mechanism is a two-body scattering in the inner system which sends one star closer to the core (where it gains potential energy and thus speeds up, increasing the temperature of the inner system) and kicks one star out to the periphery (where its higher speed increases the temperature of the outer system). Importantly, we see that the inner system *shrinks* as it heats up.

Now two outcomes are possible, depending on the specific heat of the two systems as a function of their masses. If the outer system always has the smaller (magnitude of) specific heat, its temperature will eventually grow to exceed that of the inner system, and the entire assemblage of masses will reach equilibrium. On the other hand, if the outer system grows in mass too quickly, its specific heat will become too large and its temperature will never catch up to the inner system. Hence the inner system will continue shrinking in mass and growing in temperature until the thermodynamic description breaks down. This is precisely the *gravothermal catastrophe* (Lynden-Bell & Wood 1968). In the case of a globular cluster (at least an idealized one with uniform-mass stars), the gravothermal collapse process is halted by binary formation, which acts as an energy sink (Heggie 1975; Hut et al. 1992). If the uSIDM interacts purely via elastic scattering, however, no bound state formation is possible, and gravothermal collapse can drive the core to relativistic velocities, where it undergoes catastrophic collapse into a black hole via the radial instability (Zel'dovich & Podurets 1966; Shapiro & Teukolsky 1985a, 1985b, 1986).

3.1. The Gravothermal Fluid Equations

We now consider the gravothermal collapse of a general two-component dark matter halo, where the self-interacting component comprises some fraction f of the mass of the halo. At this stage we do not yet specialize to the uSIDM case, with $f \ll 1$. To avoid confusion, we will therefore refer to the two different components of the halo as SIDM (making up a fraction f of the total mass of the halo) and (ordinary) CDM (making up the remainder), denoting the SIDM as uSIDM only when $f \ll 1$. To simulate the collapse, we employ the gravothermal fluid approximation (Lynden-Bell & Eggleton 1980; Balberg et al. 2002; Koda & Shapiro 2011), which reduces the problem to a set of coupled partial differential equations that can then be solved numerically. First consider the general case for an $f = 1$ fluid, i.e., a halo composed entirely of SIDM. A spherically symmetric ideal gas of point particles in hydrostatic equilibrium with arbitrary conductivity κ obeys the following equations (Lynden-Bell & Eggleton 1980):

$$\frac{\partial M}{\partial r} = 4\pi r^2 \rho \tag{4}$$

$$\frac{\partial\left(\rho\nu^2\right)}{\partial r} = -\frac{GM\rho}{r^2} \tag{5}$$

$$\frac{L}{4\pi r^2} = -\kappa \frac{\partial T}{\partial r} \tag{6}$$

$$\frac{\partial L}{\partial r} = -4\pi\rho r^2 \nu^2 \left(\frac{\partial}{\partial t}\right)_M \ln \frac{\nu^3}{\rho}, \tag{7}$$

where $\nu(r)$ is the one-dimensional velocity dispersion and $L(r)$ the total heat radiated *inward* through a sphere of radius r. Equation (4) simply defines the integrated mass distribution in terms of the density. Equation (5) is the statement of hydrostatic equilibrium: we inserted Euler's equation into the Poisson equation for a spherically symmetric potential and used the equation of state for an ideal gas, $p = \rho\nu^2$. Equation (6) states that the heat flux is proportional to the temperature gradient, with proportionality constant given by the conductivity κ. Equation (7) is the second law of thermodynamics, inserting the specific entropy of an ideal gas of point particles $u = \frac{k_B}{m} \ln(\frac{T^{3/2}}{\rho})$ and using the relation $\nu^2 = k_B T/m$. This gives a set of four differential equations with four dependent variables $\{M, \rho, \nu, L\}$ and two independent variables $\{r, t\}$. (The temperature T is directly related to ν by $\nu^2 = k_B T/m$.)

To make progress, we need an expression for the form of the thermal conductivity κ in terms of our physical parameter, the elastic scattering cross section per unit mass σ. Dimensional analysis alone will not suffice: we have one time scale, the fluid relaxation time

$$t_r \equiv 1/(a\rho\sigma\nu), \tag{8}$$

with $a = \sqrt{16/\pi} \approx 2.257$ for hard-sphere interactions, but two length scales, the mean free path $\lambda \equiv 1/(\rho\sigma)$ and the Jeans length or gravitational scale height $H \equiv \sqrt{\nu^2/(4\pi G\rho)}$. Following (Balberg et al. 2002; Koda & Shapiro 2011), we find the unique length scales in the two limiting cases, the short mean free path (smfp) regime $\lambda \ll H \rightarrow \ell_{\rm smfp} = \lambda$ and the long mean free path (lmfp) regime $\lambda \gg H \rightarrow \ell_{\rm lmfp} = H$, and combine them in reciprocal to get a final length scale, $\ell \equiv \left(\ell_{\rm smfp}^{-1} + \ell_{\rm lmfp}^{-1}\right)^{-1}$. In the smfp regime, transport theory tells us that

$$\frac{L}{4\pi r^2} \approx -\frac{3}{2} a^{-1} b \rho \frac{\lambda^2}{t_r} \frac{\partial \nu^2}{\partial r}. \tag{9}$$

The coefficient b is calculated perturbatively in Chapman–Enskog theory (Lifshitz & Pitaevskii 1981), $b = 25\sqrt{\pi}/32 \approx 1.385$. In the lmfp regime, the flux equation is well approximated as

$$\frac{L}{4\pi r^2} \approx -\frac{3}{2} C \rho \frac{H^2}{t_r} \frac{\partial \nu^2}{\partial r}, \tag{10}$$

where C is a constant setting the scale on which the two conduction mechanisms are equally effective, which depends on the shape of the initial density profile. For an initial Navarro–Frenk–White (NFW) profile, C is determined by N-body simulations Koda & Shapiro (2011) to be $C \approx 290/385 \approx 0.75$. Hence the final expression is

$$\frac{L}{4\pi r^2} = -\frac{3}{2} ab\nu\sigma \left[a\sigma^2 + \frac{b}{C}\frac{4\pi G}{\rho\nu^2}\right]^{-1} \frac{\partial \nu^2}{\partial r}. \tag{11}$$

Now consider the more general case, $f \neq 1$. Hydrostatic equilibrium is separately satisfied for each species of particle, but the gravitational potential is of course sourced by both species, giving the coupling between the two components. Because the non-SIDM component is taken to be collisionless, it has $\sigma = 0$, so $L^{\rm ni} = 0$. So the total system is governed by six partial differential equations with six dependent variables $\{M, \rho^{\rm int}, \rho^{\rm ni}, \nu^{\rm int}, \nu^{\rm ni}, L^{\rm int}\}$ and two independent variables $\{r, t\}$:

$$\frac{\partial M}{\partial r} = 4\pi r^2 \left(\rho^{\rm int} + \rho^{\rm ni}\right) \tag{12}$$

$$\frac{\partial \left(\rho^{\rm int}\left(\nu^{\rm int}\right)^2\right)}{\partial r} = -\frac{GM\rho^{\rm int}}{r^2} \tag{13}$$

$$\frac{\partial \left(\rho^{\rm ni}\left(\nu^{\rm ni}\right)^2\right)}{\partial r} = -\frac{GM\rho^{\rm ni}}{r^2} \tag{14}$$

$$\frac{L^{\rm int}}{4\pi r^2} = -\frac{3}{2} ab\nu^{\rm int}\sigma \left[a\sigma^2 + \frac{b}{C}\frac{4\pi G}{\rho^{\rm int}\left(\nu^{\rm int}\right)^2}\right]^{-1} \frac{\partial \left(\nu^{\rm int}\right)^2}{\partial r} \tag{15}$$

$$\frac{\partial L^{\rm int}}{\partial r} = -4\pi\rho^{\rm int} r^2 \left(\nu^{\rm int}\right)^2 \left(\frac{\partial}{\partial t}\right)_M \ln \frac{\left(\nu^{\rm int}\right)^3}{\rho^{\rm int}} \tag{16}$$

$$0 = \left(\frac{\partial}{\partial t}\right)_M \ln \frac{\left(\nu^{\rm ni}\right)^3}{\rho^{\rm ni}}. \tag{17}$$

As before, the first equation gives the total mass distribution, while the second and third enforce hydrostatic equilibrium. The fourth determines how the SIDM fluid conducts heat and the fifth how the flux gradient affects the fluid. Finally, the sixth equation ensures that the entropy of the collisionless component is conserved, $3\dot{\nu}/\nu = \dot{\rho}/\rho$. Notice that the fraction f does not appear in the differential equations themselves, but only in the boundary conditions: we must have

$$\frac{\int_0^\infty 4\pi r'^2 \rho^{\rm int}(r') dr'}{\int_0^\infty 4\pi r'^2 \rho^{\rm ni}(r') dr'} = \frac{f}{1-f} \tag{18}$$

at all times.

3.2. Initial Conditions

In principle, Equations (12)–(17) can be solved exactly given appropriate boundary conditions at $r = 0$ and $r = \infty$ and a set of initial radial profiles that obey the equations. In practice, this is computationally impossible: even finding the initial profiles for an arbitrary σ is infeasible. Balberg, Shapiro, and Inagaki (Balberg et al. 2002), considering the $f = 1$ case, took the $\sigma \rightarrow 0$ limit, which admits a self-similar solution where separation of variables is possible, then found the eigenvalues of the resulting system of ordinary spatial

differential equations and took the resulting profiles as their initial conditions for the more general $\sigma \neq 0$ case.

We will instead *assume* that SIDM self-interactions are unimportant during the process of halo formation, so that the the SIDM and collisionless components have the same initial profile. This allows us to use the results of (collisionless) ΛCDM simulations. We simplify further by approximating the initial halo by an NFW profile,

$$\rho_{\rm NFW}(r) = \frac{\rho_s}{(r/r_s)(1 + r/r_s)^2}, \tag{19}$$

where ρ_s and r_s are the characteristic density and scale radius, respectively. Since the NFW profile has a characteristic radius, we can state our assumption more precisely: we assume that halo formation proceeds much faster than heat conduction, which is true when the dynamical timescale of collapse is much less than the relaxation timescale due to collisions:

$$t_{\rm dyn}(r_s) \ll t_{\rm rel}(r_s) \approx \frac{1}{\tau_s} t_{\rm dyn}(r_s) \rightarrow \tau_s \ll 1; \tag{20}$$

i.e., so long as the halo is *optically thin at its characteristic radius*. Again, if the optical depth is small,

$$\tau \equiv f\rho_{\rm NFW} r\sigma \ll 1 \rightarrow \sigma f \leqslant \frac{1}{\rho_s r_s}, \tag{21}$$

typical SIDM particles have not yet undergone any self-interaction by the time of halo formation, so we are justified in assuming they follow the same initial profile as the collisionless dark matter, $\rho_0^{\rm int}(r) = f\rho_{\rm NFW}(r)$.

Before checking the validity of this assumption, we comment on the consequences of taking a different initial profile. The NFW profile is particularly simple: its form means that the optical depth at small radii, $r \ll r_s$, is independent of radius, so a small characteristic optical depth implies that the central regions are also optically thin despite the presence of a cusp. Modern ΛCDM simulations, however, have tended to find density profiles more complicated than the NFW profile. Profiles with cores or at least less cuspy behavior, e.g., Einasto profiles (Merritt et al. 2005; Graham et al. 2006), will have $\tau \ll 1$ everywhere if $\tau_s \ll 1$. Below we will see that SIDM halos with initial NFW density profiles grow cores on a scale of tens of halo relaxation times anyway, so shallower initial profiles will only result in slightly smaller times before black hole formation. Profiles with more cuspy behavior, e.g., generalized NFW or Zhao profiles (Zhao 1996) with inner slope $\alpha \gtrsim 1$, will unavoidably have regions at very small radii in the optically thick regime. Below we will see that SIDM halos with initial NFW profiles first evacuate the cusp to form cores before beginning the gravothermal collapse process, and it seems reasonable to conclude that the same thing will happen for non-pathological cuspier profiles. We conclude that imposing a different profile should not significantly change the behavior investigated below.

When is the assumption that $\tau_s \ll 1$ justified? Recall that the characteristic radius ρ_s and radius r_s for an NFW profile are given in terms of the halo virial mass M_Δ and concentration c:

$$r_\Delta \equiv c r_s, \tag{22}$$

$$M_\Delta \equiv M(r_\Delta) = \int_0^{r_\Delta} 4\pi r^2 \rho_{\rm NFW}(r) dr = 4\pi \rho_s r_s^3 \left[\ln(1 + c) - \frac{c}{1 + c}\right], \tag{23}$$

$$\rho_s \equiv \delta_c \rho_{\rm crit}(z). \tag{24}$$

The density contrast δ_c is in turn given by

$$\delta_c = \frac{\Delta}{3}\frac{c^3}{K_c}, \tag{25}$$

where $K_c \equiv \ln(1 + c) - c/(1 + c)$. The problem thus reduces to finding an expression for Δ, the virial overdensity. In the spherical collapse model, this is given by $\Delta \sim 18\pi^2 \Omega_m^{0.45}$ for a flat universe (Lahav et al. 1991; Eke et al. 1996; Bryan & Norman 1998; Neto et al. 2007); Δ hence approaches the familiar value of 178 in the matter-dominated era.

Inserting these expressions into (21) above yields an inequality for σf in terms of c and M_Δ, along with the redshift of virialization z:

$$\sigma f \leqslant \frac{1}{\rho_s r_s} = (4\pi)^{-1/3} M_\Delta^{-1/3} \left(\frac{\Delta \rho_{\rm crit}(z)}{3}\right)^{-2/3} K_c\, c^{-2} \tag{26}$$

$$= 24.56\ {\rm cm^2\ g^{-1}} \times \left(\frac{M_\Delta}{10^{12}\, M_\odot}\right)^{-1/3} \times \left(\frac{\rho_{\rm crit}(z)}{\rho_{\rm crit}(z = 15)}\right)^{-2/3} K_c\, c^{-2}. \tag{27}$$

In the second line we have inserted the typical halo parameters we will consider below: $z = 15$, $M_\Delta = 10^{12}\, M_\odot$.

It remains to insert plausible values for the concentration c. Individual halos of mass M_Δ formed at a fixed redshift z will have varying concentrations, but there should be some mass- and redshift-dependent median concentration, $c(M_\Delta, z)$. Prada et al. (2012) used the Millennium (Springel et al. 2005; Boylan-Kolchin et al. 2009), Bolshoi (Klypin et al. 2010), and MultiDark (Riebe et al. 2011) simulations to examine the shape of the $c(M_\Delta, z)$ curve with varying mass and redshifts. They found that for each redshift considered (from $z \sim 0$–6) the concentration formed a U-shaped curve: it was minimized at a certain value of the mass, but increased steeply both above and below this mass. Furthermore, they found that both the minimum value of the concentration and the mass at which this minimum was realized decreased with increasing redshift. At the large redshifts we consider, the cluster-sized halos needed to form SMBHs are far more massive than the bottom of the U-shaped curve; accordingly, the fitting formulae given in Prada et al. (2012) predict that the concentration for these halos will be extremely large, of the order of $c \sim 10^5$ for the halo parameters above. If this were true, the initial density profiles of these large, early halos would be extremely concentrated, so that their inner regions are extremely thick even for $\sigma f \gtrsim 10^{-6}$. In this case the simulations presented in this paper would not be reliable.

We emphasize, however, that the fitting formulae of Prada et al. (2012) were devised using simulated halos only out to $z \sim 6$; they should not be trusted so far away from their domain of validity. In particular, Ludlow et al. (2012, 2014) found that the upturn of concentration at large masses is due to the presence of not-yet-virialized haloes in the simulations; when only virialized halos are considered, the concentration parameter levels off at high masses. Because we are considering virialized halos here, we accordingly expect more moderate values of the concentration parameter. To confirm this expectation, we have consulted the high-redshift halo catalogs of the FIRE simulation (Hopkins et al. 2013), which attempted to resolve an overdense region at high redshift. The catalogs use the Amiga Halo Finder (Knollmann & Knebe 2009) to measure c in the same way as defined in Prada et al. (2012). We are interested in the concentration parameters of the most massive halos formed at a given redshift. Perhaps unsurprisingly, we find that, even at $z \sim 30$, halo concentrations range from 2 to 11, similar to the values found at lower redshifts in the simulations consulted in Prada et al. (2012), rather than the much higher values predicted by naively applying the fitting formulae. We do not attempt to construct the full $c(M_\Delta, z)$ curve at high redshifts on the basis of this limited data, but we do assume that realistic halos will take concentrations in this observed range.

The upper bound on σf for which $\tau_s \leqslant 1$ ranges from 0.32–2.65 $\mathrm{cm^2\,g^{-1}}$ as concentrations decreasing from 11 to 2 are inserted into Equation (27). In the remainder of this paper we will typically set $c = 9$, which gives a bound of 0.425 $\mathrm{cm^2\,g^{-1}}$. In Section 5 below we will find that this bound is of the same order of magnitude as the cross section needed to produce the desired high-redshift SMBHs using uSIDM. Accordingly, there is a surprisingly small region of parameter space where both the assumption of an initial NFW profile is valid and the desired black holes are produced. We will discuss this further in Section 5. For now, we note only that the qualitative results of this paper should still hold even when our assumption of an initial NFW profile is invalid. Outside of this range, we expect that gravothermal collapse should still occur —in fact, it should occur *faster* because core formation will have begun even before virialization—but the particular expressions given here will no longer be valid.

3.3. Integration of the Equations

Given the initial conditions, we can proceed to integrate the system of Equations (12)–(17). We first move to a dimensionless form of the problem by choosing fiducial mass and length scales $\{M_0, R_0\}$. Then the remaining dependent variables are given naturally in terms of these quantities, e.g., $\nu_0 = \sqrt{GM_0/R_0}$. Full expressions for all dependent variables in terms of M_0 and R_0 are given in Section 5 of Balberg et al. (2002). The cross section per unit mass is now expressed dimensionlessly by $\hat{\sigma} = \sigma/\sigma_0$, $\sigma_0 = 4\pi R_0^2/M_0$. It is convenient to use the two quantities already specified in the NFW profile, $\{\rho_s, r_s\}$; we therefore take

$$\left\{M_0 = 4\pi R_0^3 \rho_s,\ R_0 = r_s\right\}. \tag{28}$$

Note that we have made a different choice of $\{M_0, R_0\}$ than Balberg et al. (2002), since we consider a cuspy NFW profile rather than a cored one and thus work with characteristic rather than central quantities. Finally, the timescale is set by the initial relaxation time at the characteristic radius,

$$t_{r,c}(0) = 1/\left(f a \rho_s \nu_s \sigma\right), \tag{29}$$

so the independent variable can also be made dimensionless. Dimensionless quantities are written with tildes (e.g., $\tilde{\rho}$, $\tilde{t}$). The resulting initial profiles for an $f = 0.01$ halo are shown in Figure 1.

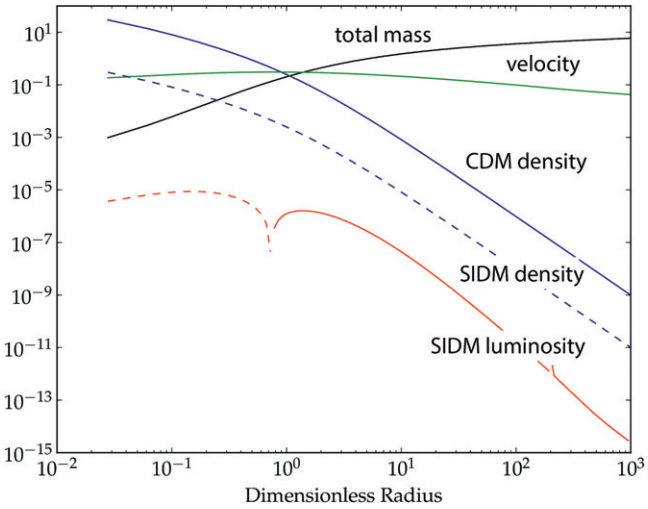

Figure 1. The dimensionless initial profiles for an NFW halo, with $f = 0.01$. The CDM and SIDM have the same velocity profile, and their density profiles have the same shape but a normalization differing by $f/(1+f)$. As expected for SIDM, the initial luminosity at small radii is negative (shown as dashed on the plot), indicating that the cusp is being forced outward as a core begins to form. The glitch in the luminosity at $\tilde{r} = 200$ is a numerical artifact.

We solve the problem by spatially discretizing into N concentric spherical shells, initially evenly logarithmically spaced in radius. At each timestep, we first apply the effects of heat conduction, which increases the energy within each shell, then adjust the profile to maintain hydrostatic equilibrium. The heat conduction step is simple: we determine the luminosity profile from the density and velocity dispersion using the dimensionless, discretized form of Equation (15), then adjust the energy of each shell accordingly (using $dU \equiv L dt$ for a finite but small timestep). Timesteps are chosen so that the change of (dimensionless) specific energy $\tilde{u}_i = 3\tilde{v}_i^2/2$ is not large: we require $\Delta\tilde{u}_i/\tilde{u}_i < \varepsilon \ll 1$ for each shell i, typically taking $\varepsilon = 0.001$. This means that as the gravothermal catastrophe approaches and core temperatures and densities become large, the size of timesteps will decrease dramatically: as expected, we cannot integrate through the collapse because the fluid approximation itself breaks down there.

To carry out the hydrostatic equilibrium step, we use the method of Lagrangian zones, in which the radius of each shell is adjusted while the mass it contains is left constant. The relaxation process, which involves long-range gravitational interactions rather than heat conduction via collisions, is entropy-preserving, so it preserves the adiabatic invariants $A_i \equiv \tilde{\rho}_i \tilde{V}_i^{5/3}$ for each shell i. After the heat conduction step, each shell is temporarily out of hydrostatic equilibrium, so that Equation (13) is violated by some amount Δ_i. The problem is to adjust the density, velocity dispersion, and radius of each shell i, such that hydrostatic equilibrium is again satisfied ($\Delta_i = 0\ \forall i$) while preserving the adiabatic invariants. The assumption of adiabaticity, along with the use of Lagrangian zones to keep the mass of each shell fixed, fixes the density and

velocity changes as functions of the set of changes of radii $\Delta\tilde{r}_i$. Hence, the requirement of hydrostatic equilibrium gives a system of differential equations for the changes of radii, which, when linearized, is tridiagonal (since the thickness of each shell depends not only on its own central radius but that of its nearest neighbors). The resulting system is solved using a standard linear algebra library.[4]

3.4. Results

Unfortunately, the above procedure is still insufficient to integrate Equations (12)–(17) in full generality. The problem is that, because the SIDM and collisionless dark matter are *separately* in hydrostatic equilibrium, the method of Lagrangian zones will result in different sets of radii for the two species. But in order to perform subsequent timesteps, we need the total mass distribution at each radius for both types of DM. For computationally feasible numbers of shells ($N \sim 400$), interpolation is not accurate enough to preserve numerical stability and the distributions cannot be integrated all the way up to the point of gravothermal collapse.

We can, however, consider the two limiting cases, which happen to be the cases we are interested in. In the pure SIDM case $f = 1$, there is only one species and the problem does not arise. In the uSIDM case, $f \ll 1$, we can ignore the gravitational backreaction of the uSIDM component on the collisionless DM and assume that it maintains an NFW profile throughout, allowing the calculation of its mass distribution analytically at every point. (We could instead use some approximation to the backreaction which does not require us to track the collisionless profile, e.g., adiabatic contraction (Blumenthal et al. 1986; Gnedin et al. 2004), but the difference this makes will be negligible for small f. Even for moderate $f \lesssim 0.1$, the effect will be less important than the backreaction of the baryonic matter also present in the halo. We are not tracking baryons either, so for the sake of simplicity we neglect backreaction entirely.) We expect that the two cases should yield similar results, because the temporary violation of Equation (13), the hydrostatic equilibrium condition, after each heat conduction timestep is overwhelmingly due to the increase on the lhs of the equation, from heat conduction, rather than from interactions with the collisionless component, on the rhs of the equation. This is just the statement that the self-interaction is much larger than gravitational strength. We indeed find that this is the case, at least qualitatively. Consider Figure 2, which shows the early evolution of $f = 0.01$ and $f = 1$ halos with the same value of $\hat{\sigma}$. We see that behavior is indeed qualitatively the same: in both cases, a core begins to form as heat conduction dissolves the initial cusp. Note that the time scales are different: in the uSIDM case the relaxation time is increased by a factor of f^{-1} since the uSIDM density is a factor of f lower. So Figure 2 suggests that uSIDM evolution is the same as the $f = 1$ case, just f^{-1} times slower.

We will focus on the $f = 1$ case in the following, and then rescale our final results by f^{-1} as just described. Figures 3 and 4 show the entire evolution of an $f = 1$ halo with $\hat{\sigma} = 0.088$ (chosen to allow comparison with Koda & Shapiro 2011) from initial NFW profile through to gravothermal collapse. First consider Figure 3, which shows the evolution of the density profile. Although the halo is initially in an NFW profile, the initial negative luminosity at small radii causes the cusp to empty out, driving evolution toward the cored, self-similar profile found by Balberg et al. (2002), as was already seen in Figure 2 above. When the self-similar profile is reached after a few tens of relaxation times, the luminosity profile becomes everywhere positive, and the core increases in density while its mass steadily shrinks. While the entire profile is in the lmfp regime, evolution is self-similar, and the central density increases steadily. Inevitably, there comes a time, about 450 relaxation times after virialization, when the inner density increases enough that the most central regions enter the smfp regime, and the core bifurcates into a very dense outer core and an inner core which transitions between the two regions.

Importantly, once the smfp regime has been reached, mass loss from the inner core is no longer efficient: the inner core has become so thick that evaporation is only possible from its boundary, not from the entire volume. This means that the mass in the inner core is essentially constant over the very short time ($\lesssim 10 t_{r,c}(0)$) between breaking of self-similarity and catastrophic collapse. As mentioned in Section 3.3 above, the size of successive timesteps decreases rapidly as the gravothermal catastrophe approaches, so this short time takes very many (increasingly small) timesteps to integrate over, and the time of collapse can be precisely given as 455.65 relaxation times after the start of integration. Because evaporation is inefficient after the loss of self-similarity, the mass in the inner core is still nonzero at the moment of collapse, unlike in the globular cluster case, and a black hole will form. Figure 4 shows that the inner core at collapse contains a mass of around $0.025 M_0$. Because the fluid approximation breaks down, we do not know that the entire inner core will collapse directly into a black hole, but, because it is optically thick, Bondi accretion (Bondi 1952) is extremely efficient. Hence, we expect that the black hole will rapidly grow to encompass the entire region regardless.

4. SUPERMASSIVE BLACK HOLES FROM uSIDM

We have found that halos with a uSIDM component (and pure SIDM halos, on much longer timescales) grow black holes of mass $M_{\rm BH} \equiv 0.025 f M_0$ in a time $455.65 t_{r,c}(0)$. Given the considerations discussed above, uSIDM can help explain the existence of massive high-redshift quasars if the resulting black holes are large enough and form early enough that baryonic accretion can grow them to $\sim 10^9\, M_\odot$ by $z \gtrsim 6$. It remains to evaluate M_0 and $t_{r,c}(0)$ in terms of the halo parameters and use this requirement to place constraints on the uSIDM parameters $\{\sigma, f\}$.

4.1. Halo Parameters

Instead of using the characteristic NFW parameters $\{\rho_s, r_s\}$, it is convenient to again parameterize a halo by its virial mass M_Δ and concentration c. In dimensionless units, the mass contained within the ith shell is

$$\tilde{M}_i = \int_0^{\tilde{r}_i} \tilde{\rho}\tilde{r}^2 d\tilde{r} = \int_0^{\tilde{r}_i} \tilde{r}^{-1}(1+\tilde{r})^{-2}\tilde{r}^2 d\tilde{r} = -\frac{\tilde{r}_i}{1+\tilde{r}_i} + \ln(1+\tilde{r}_i), \tag{30}$$

but the virial radius $r_\Delta \equiv c r_s$, so

$$\frac{M_{\rm BH}}{M_\Delta} = \frac{\tilde{M}_{\rm BH}}{\tilde{M}(c)} = \frac{0.025 f}{\ln(1+c) - c/(1+c)}. \tag{31}$$

[4] http://www.gnu.org/software/gsl

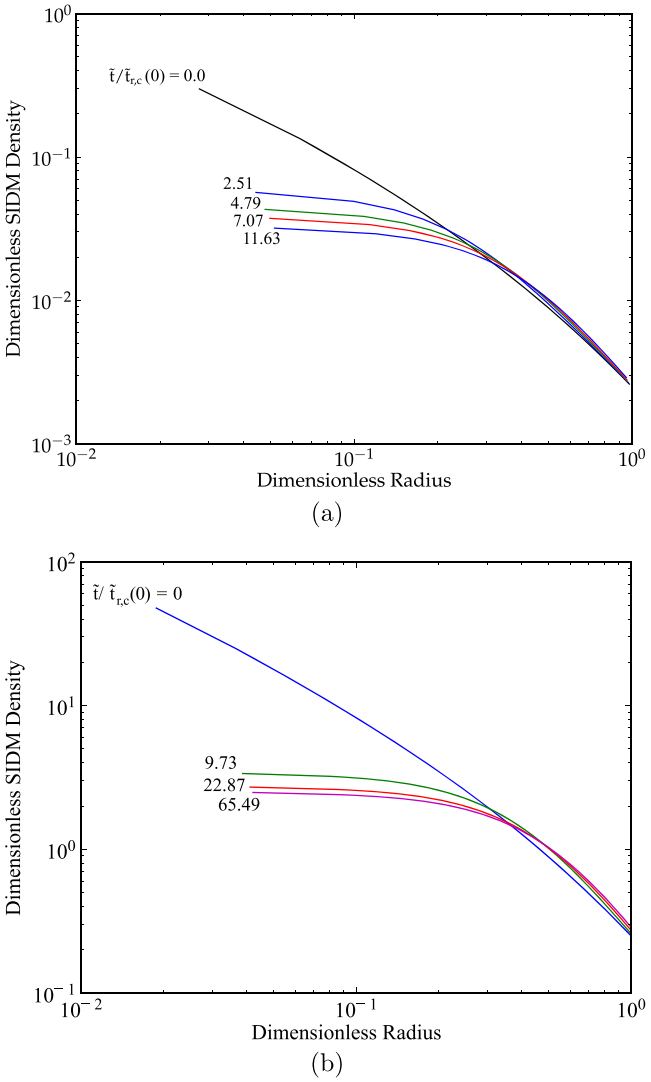

Figure 2. (a) Evolution of SIDM density profiles, starting with an $f = 0.01$ NFW halo. Only the inner part of the halo is shown; the outer part still asymptotes to r^{-3} as in Figure 1 above for all halos. From top to bottom, profiles are at 0.0, 2.51, 4.79, 7.07, and 11.63 central relaxation times. Because $t_r \propto \rho^{-1}$, this corresponds to integrating for ∼1000 relaxation times in an $f = 1$ halo. However, comparison to the $f = 1$ results below suggests that the density profile flattens in the same manner, just f^{-1} times slower: evidently the non-interacting dark matter has little influence on the central SIDM evolution. (b) Evolution of an $f = 1$ halo starting from NFW initial conditions. For clarity, only the inner portion of the density profile is shown: the outer profile has not yet changed significantly at this stage. From top to bottom, profiles are at 0.0, 9.73, 22.87, and 65.49 central relaxation times. As in the $f = 0.01$ case, the density profile is flattening as a core develops.

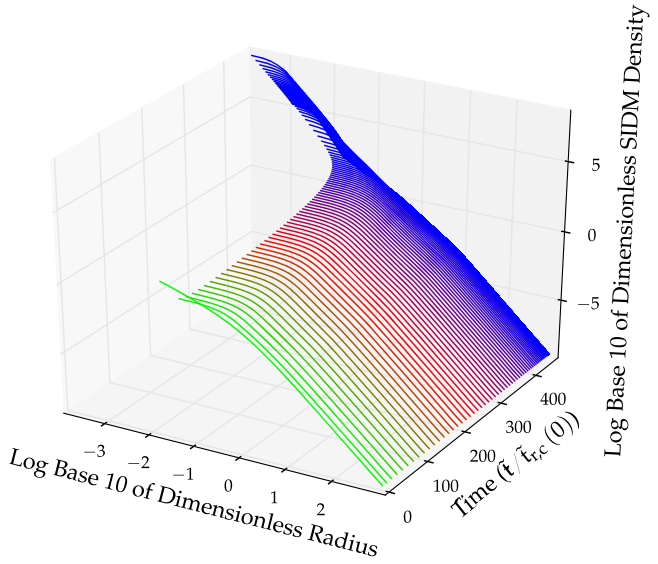

Figure 3. Runaway collapse of an $f = 1$ SIDM halo with $\hat{\sigma} = 0.088$, starting from an initial NFW profile. The inner profile starts cuspy, rapidly shrinks to a self-similar profile (as in Balberg et al. 2002 and our Figure 2) with a $\tilde{\rho} = 1$ core, then slowly increases in density in a self-similar manner. After ∼450 relaxation times, the core of the halo becomes optically thick, and self-similarity is broken: the core splits into a very dense inner core and an outer core which transitions between the two regions. Catastrophic collapse occurs as $\tilde{t}/\tilde{t}_{r,c}(0)$ approaches ∼455.65.

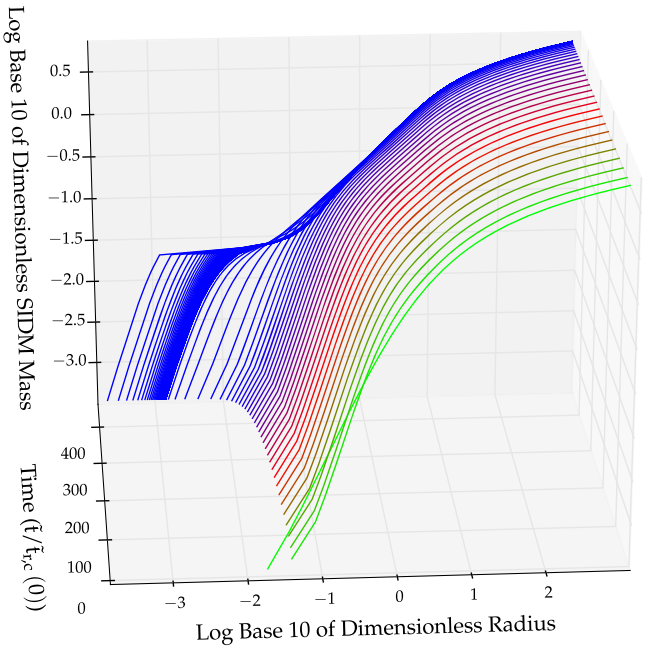

Figure 4. Mass profile history of a cored SIDM halo with $\hat{\sigma} = 0.088$, starting from an initial NFW profile. Once the core enters the optically thick regime, around $\tilde{t}/\tilde{t}_{r,c}(0) = 450$, the inner core contains a constant total mass, around 2.5%–3% of the characteristic mass M_0.

This gives the desired expression for the seed black hole mass $\tilde{M}_{\rm BH}$ in term of the halo and uSIDM parameters. The denominator ranges from ∼ 0.5 to 2 for realistic values of the halo concentration, so the BH mass is a few percent of the total uSIDM mass in the halo.

Recall that the relaxation time is $t_{r,c}(0) = 1/(a f \rho_s v_s \sigma)$, i.e., the scattering time at the characteristic radius. The lower end of the interesting range for $f = 1$ SIDM is ∼0.1 cm^2 g^{-1}, for which the relaxation time at the characteristic radius of a Milky Way scale halo is approximately a Hubble time. To grow a black hole in galactic halos by $z \sim 6$, the relaxation time needs to be ∼10^4 times smaller to ensure ∼500 relaxation times by the time the universe was a twentieth of its present age. This, however, does not mean that $\sigma \approx 1000\, f^{-1}$cm^2 g^{-1}. Recall that $\rho_s = \delta_c \rho_{\rm crit}$, where δ_c is a function of c and the cosmology given below, and the critical density goes as $(1+z)^3$ in the matter-dominated era. Also $r_s \propto r_\Delta \propto (M_\Delta/\rho_{\rm crit})^{1/3}$ implies $v_s \propto \sqrt{M_\Delta/r_\Delta} \propto M_\Delta^{1/3} \rho_{\rm crit}^{1/6}$. Hence the mass and approximate

redshift dependence of the relaxation time are

$$t_{r,c}(0) \propto (1+z)^{-7/2} M_\Delta^{-1/3} \tag{32}$$

and we expect that σf need not be that much larger than the interesting range for $f = 1$, i.e., we expect $\sigma f \gtrsim 0.1$–$1\ \mathrm{cm^2\ g^{-1}}$.

At this point, the reader might worry that this conclusion combined with the observation of non-collapsed cores in the nearby $(z \sim 0)$ universe rules out the existence of standard $(f \approx 1)$ SIDM. We emphasize, however, that large values of σf mean that core collapse in times much smaller than the age of the universe is *possible*, but not, we expect, *typical*; it occurs only in the rare halos which virialize at very high redshifts and remain uninterrupted, i.e., do not experience major mergers, for long enough to complete the gravothermal collapse process. See Section 5.1 below for further discussion of this point.

The exact expression for the halo relaxation time is

$$t_{r,c}(0) = \frac{1}{af\sigma}\left(\frac{K_c^2}{4\pi G^3}\right)^{1/6} \delta_c^{-7/6} \rho_{\mathrm{crit}}(z)^{-7/6} M_\Delta^{-1/3} \tag{33}$$

$$= 0.354\ \mathrm{Myr} \times \left(\frac{M_\Delta}{10^{12}\, M_\odot}\right)^{-1/3} \left(\frac{K_c}{K_9}\right)^{3/2} \left(\frac{c}{9}\right)^{-7/2} \times \left(\frac{\rho_{\mathrm{crit}}(z)}{\rho_{\mathrm{crit}}(z=15)}\right)^{-7/6} \left(\frac{\sigma f}{1\mathrm{cm^2\ g^{-1}}}\right)^{-1}, \tag{34}$$

where $K_c \equiv \ln(1+c) - c/(1+c)$, $\delta_c = (\Delta/3)c^3/K_c$, and Δ, the virial overdensity, is $18\pi^2\Omega_m^{0.45}$ for a flat universe, approximately 178 in the matter-dominated era. So the relaxation time, and hence the collapse time, is given in terms of the halo and uSIDM parameters. To match observations, we need some seed black holes to grow by a large enough factor via Eddington accretion to reach $M_{\mathrm{BH}} \approx 10^9\, M_\odot$ by $z \gtrsim 6$; this leads to an inequality on σ when the halo parameters and f are specified.

4.2. Explaining Observations

Let us spell out the procedure more precisely. An observation of a particular high-redshift quasar at redshift z_{obs} yields a value for the luminosity, which corresponds to a SMBH of mass M_{SMBH} once the measured luminosity is identified with the Eddington luminosity and a particular value for the radiative efficiency ϵ_r is assumed. (We have already discussed potential issues with these assumptions in Section 2 above; in the remainder of the paper, we will take the published observations at face value and assume their quoted SMBH masses, which take $\epsilon_r = 0.1$ as input, are correct.)

At the same time, the uSIDM framework developed in this paper tells us that NFW halos of viral mass M_Δ and concentration c virialized at redshift z form seed black holes of mass M_{BH} in a time $455.65 t_{r,c}(0)$, i.e., the seed black holes are formed at redshift z_{coll}, where

$$t(z_{\mathrm{coll}}) - t(z) = 455.65 t_{r,c}(0), \tag{35}$$

and the time $t(z)$ after the Big Bang corresponding to redshift z is given by the usual cosmology-dependent expression,

$$t(z) = t_0 \int_0^{1/(1+z)} \frac{da}{\dot{a}}. \tag{36}$$

Equations (31) and (33) then give expressions for these quantities in terms of the halo properties $\{M_\Delta, c, z\}$ and the uSIDM parameters $\{\sigma, f\}$.

There is still one parameter that must be specified: the fraction of SMBH mass which is due to accretion of baryons as opposed to the initial seed black hole. For simplicity, we will assume that the central black hole accretes continuously at the Eddington limit from the time of formation to the time at which it is observed. Of course more complicated growth histories are both possible and likely. Nevertheless, this simplifying assumption allows us to specify the fraction by instead giving N_e, the number of e-folds of accretion at the Eddington limit. This finally allows us to compute the observable quantities: we have

$$t(z_{\mathrm{obs}}) = t(z_{\mathrm{coll}}) + N_e\, t_{\mathrm{Sal}}, \tag{37}$$

$$M_{\mathrm{SMBH}} = M_{\mathrm{BH}} \exp(N_e). \tag{38}$$

To find acceptable values of σ and f given the SMBH observables, we must specify (or marginalize over) the halo parameters and the baryonic contribution to the SMBH mass. The latter quantity directly sets (Equation (37)) the redshift of seed black hole collapse, z_{coll}, which yields the required collapse time and thus the required value of σf via Equation (33). Knowing the growth due to accretion of baryons also tells us (Equation ((38)) the required seed black hole mass M_{BH}, which specifies f via Equation (31).

4.3. Examples

As an example, consider again ULAS J1120+0641, with mass $M_{\mathrm{SMBH}} \approx 2 \times 10^9\, M_\odot$ at $z_{\mathrm{obs}} = 7.085$. To grow four orders of magnitude $(N_e = \ln 10^4)$ by Eddington-limited baryon accretion, for example, we must form a seed black hole with mass $M_{\mathrm{BH}} = 2 \times 10^5\, M_\odot$ by $z_{\mathrm{coll}} = 12.9$. With a halo of mass $M_\Delta = 10^{12}\, M_\odot$ and concentration $c = 9$ formed at redshift $z = 15$, we find that $t_{r,c}(0) = 0.354$ Myr $\times (1\ \mathrm{cm^2\ g^{-1}})/(\sigma f)$. In order for 455.65 relaxation times to have passed in the 64.5 Myr between $z = 15$ and $z_{\mathrm{coll}} = 12.9$, we must have $\sigma f = 2.50\ \mathrm{cm^2\ g^{-1}}$. From (31), we require $f = 1.12 \times 10^{-5}$ to get the correct seed mass, so $\sigma = 2.23 \times 10^5\ \mathrm{cm^2\ g^{-1}} = 3.97 \times 10^5$ b/GeV. The large value of σ is unsurprising: we chose to start with a halo much larger than the seed black hole we wanted to form, so f had to be small and σ large in order to compensate.

Alternatively, we could start with the same halo but produce the black hole entirely from uSIDM. The relaxation time is unchanged: $t_{r,c}(0) = 0.354\ \mathrm{Myr} \times (1\ \mathrm{cm^2\ g^{-1}})/(\sigma f)$. But now the black hole need not form until $z = 7.085$, 479 Myr after halo formation, so the required value of σf is smaller, $\sigma f = 0.336\ \mathrm{cm^2\ g^{-1}}$. Again applying Equation (31) yields $f = 0.112$, $\sigma = 2.99\ \mathrm{cm^2\ g^{-1}} = 5.36$ b/GeV, coming much closer to the classic SIDM cross section.

Of course, in the absence of direct measurements of the host halo of ULAS J1120+0641 the problem is underdetermined. The point is that σf takes reasonable values of $\mathcal{O}(1)\ \mathrm{cm^2\ g^{-1}}$, well within the regime described by the gravothermal fluid approximation starting from an initial NFW profile.

5. DISCUSSION

The examples in the previous section show that *individual* observations of high-redshift quasars can successfully be explained within the uSIDM paradigm. For uSIDM to be fruitful, however, we should ideally be able to find (or rule out) a consistent choice of these parameters that successfully explains the *cosmological* abundance of high-redshift quasars. It is unsurprising that some choice of cross section per unit mass σ and fraction f can reproduce one particular observation, e.g., ULAS J1120+0641, but it is more suggestive if that choice can reproduce the entire observed number density of SMBHs as a function of mass and redshift. The minimal requirement for a viable uSIDM model is that it explain (or at least not conflict with) what has currently been observed. That means producing the correct abundance of $\sim 10^9\, M_\odot$ quasars at redshift 6–7, as has already been discussed, and ensuring that SMBHs are not overproduced in the nearby (lower-redshift) universe. Beyond that, one would like to make concrete predictions for the next generation of experiments, which should be sensitive to smaller masses and higher redshifts.

This task is difficult for a number of reasons. The essential problem is that a number of nuisance parameters must be constrained or marginalized over in order to connect the uSIDM properties to the SMBH distribution (and then further to the quasar distribution). Even in the simplified setup described above there were already the e-folds of baryonic accretion, N_e, and the halo parameters M_Δ and c. In the cosmological context, these nuisance parameters are promoted to entire unknown functions that are currently only poorly constrained by observations and simulations. Even when constraints or functional forms are available, they are often trustworthy only in regimes far separated from the ones of interest to us here (for example, in the low-redshift universe, or in a lower mass range). We have already encountered this problem in Section 3 above, when considering the concentrations of massive NFW profiles at high redshifts.

Nevertheless, in the remainder of this section we attempt to estimate the constraints that our existing knowledge places on the uSIDM parameter space. We first explain the source of our cosmological uncertainty and means by which it could be improved. Next we note a different source of tension within ΛCDM, independent of the existence of high-redshift SMBHs, that could be relieved by uSIDM. Finally, we present tentative maps of the uSIDM parameter space relevant to the resolution of these tensions.

5.1. Cosmological Caveats

Predicting the cosmological consequences of gravothermal collapse given a choice of the uSIDM parameters requires a unified picture of the SIDM profile at galaxy formation in terms of the halo mass and redshift, which will be easier given proper N-body simulations of halos containing uSIDM. There are several reasons why using the fluid approximation to simulate an isolated halo does not suffice.

First, although the process of gravothermal collapse can be quite short on cosmological timescales, which is why it allows massive quasars to form faster than in the standard ΛCDM picture, we have seen that it is long in terms of halo time scales (several hundred characteristic relaxation times). It is therefore necessary for the halo to remain essentially undisturbed for this length of time in order for core collapse to occur and seed black holes to form. The beginning and end of the collapse process—the elimination of the initial cusp and the catastrophic collapse itself after the core becomes optically thick—are driven entirely by dynamics in the innermost part of the profile, so we might expect them to be insensitive to accretion or mergers in the outer halo. Figures 2 and 3 make clear, though, that these stages are very short compared with the length of the overall process. The vast majority of the time required for collapse involves the slow increase of density in the core as mass flows inward from the outer halo, which we expect to be sensitive to accretion or mergers. In other words, the halo must be isolated for several hundred relaxation times. Strong interactions with other masses, such as major mergers, will disrupt the collapse process, essentially resetting the clock for seed black hole formation. Even accounting for more controlled accretion via minor mergers will technically necessitate the tracking of substructure within the collapsing halo, since it breaks the spherical symmetry required by the fluid approximation, although we expect it will not change our qualitative conclusions. Such tracking of substructure is only truly possible using N-body simulations.

More importantly, determining how often the collapse process is disrupted, and therefore predicting the spectrum of black hole masses as a function of redshift for particular values of σ and f, in order to compare with existing and upcoming observations, requires detailed cosmological information. We need not only the halo mass function at very high redshifts (up to redshift 15 in the above example, and ideally out to at least $z \gtrsim 30$–50) but also information on halo shape (the concentration parameter $c(M_\Delta, z)$ at the same high values of z, in the case that the halos form in NFW profiles) and, most importantly, detailed merger probabilities and histories as functions of mass and redshift. Even when analytical approximations to these quantities at $z \lesssim 1$ exist, it is unclear how confidently they can be extrapolated to $z \sim 50$. Hence dedicated N-body simulations are desirable. We will briefly note some additional interesting results, beyond the prediction of the history of the black hole mass function, which could be investigated given this cosmological information.

5.2. The Too Big to Fail Problem

This paper has noted that gravothermal collapse of uSIDM can produce seed black holes in the center of virialized halos. We have primarily been concerned with using this mechanism to explain the abundance of massive high-redshift quasars, but we now mention a few other areas where it could prove useful. We emphasize that these are logically independent of the quasar issue: we should not necessarily expect that the same choice of uSIDM parameters will be useful in both cases.

First, it is intriguing that there exists a well-known (and relatively tight) relation between the properties of a host galaxy and the massive black hole it contains, the M–σ relation (Magorrian et al. 1998; Ferrarese & Merritt 2000; Gebhardt et al. 2000), which suggests some sort of causal mechanism connecting the central portions of the galaxy containing the black hole with the more distant regions where the velocity dispersion is measured. Gravothermal collapse naturally provides one such mechanism, and it would be suggestive if it produced the correct relation for some choice of the uSIDM parameters. At a minimum, it should not spoil the observed relationship in nearby galaxies; this has been used previously to constrain the cross section of $f = 1$ SIDM (Hennawi & Ostriker 2002; Hu et al. 2006).

More speculatively, the presence of central black holes in dwarf galaxies could resolve the "too big to fail" problem (Boylan-Kolchin et al. 2011, 2012), in which the central densities of the brightest Milky Way satellites have much lower central densities than the most massive subhalos in ΛCDM simulations of Milky Way sized galaxies. (Importantly, the problem extends beyond the Milky Way, to other dwarf galaxies in the Milky Way (Garrison-Kimmel et al. 2014) as well as extreme dwarf galaxies in the field (Papastergis et al. 2015). One way to resolve the problem is to invoke physics not present in the simulations to reduce the central densities (within ∼1 kpc of the subhalo center) by a factor of order unity. If all of the dark matter is self-interacting, it naturally smooths out cusps to form cores, which could provide the needed reduction in density (Peter et al. 2012). Dark matter with a constant elastic cross section of $\sigma \simeq 0.6$ cm^2 g^{-1} (Zavala et al. 2013), or a velocity-dependent cross section (Vogelsberger et al. 2012), succeeds in reproducing the hosts of the observed Milky Way dwarfs. But the small fractions $f \ll 1$ we consider in this paper cannot solve the problem in this manner; another method of removing substantial mass from the central ∼kpc is needed.

Under some circumstances, it is possible that black holes could provide the needed reduction in mass. Merging black hole binaries emit gravitational waves anistropically and thus receive an impulsive kick, up to several hundred km s^{-1}. This energy can be distributed to the surrounding baryons and kick out a substantial portion of the central mass, forming a core (Boylan-Kolchin et al. 2004; Merritt et al. 2004; Lippai et al. 2008). In particular, a binary black hole ejects roughly its own mass in stars in the process of coalescing, so if the final black hole is formed through a series of mergers, 5–10 times its mass will be evacuated from the cusp, more if most of the mergers are of similarly-sized black holes, corresponding to major mergers of halos, rather than repeated accretion of small black hole onto a large one (Milosavljevic & Merritt 2001, 2003; Milosavljevic et al. 2002; Merritt & Milosavljevic 2005). Such a scenario is only viable if the required binary black hole mergers are sufficiently common within dwarf galaxies or their progenitors. Although the standard cosmological model predicts the presence of black holes in the center of nearly all large halos, it is not clear that ΛCDM produces enough black holes within the smaller halos, which are the progenitors of dwarf galaxies. Here we propose instead to use uSIDM to produce them.

Solving the Too Big to Fail problem using black holes formed from uSIDM requires a particular sequence of events: first, small halos must remain isolated enough to form seed black holes; second, the probability of major mergers must become large enough that essentially all of the Milky Way satellites have binary black holes (repeatedly) coalesce within them in order to reduce their central densities sufficiently. During the epoch of matter domination, we see that the black hole formation time for a halo of fixed mass goes as $(1+z)^{-7/2}$ (Equation (32)), while we expect the merger timescale to be set roughly by the Hubble time, $H^{-1}(z) \sim (1+z)^{-3/2}$. So halos of a given mass that form before some critical redshift will indeed grow black holes before they merge. In the next subsection we consider the parameter space where black hole seeds are ubiquitously formed in the progenitors of today's dwarf galaxies.

5.3. Parameter Space

5.3.1. High-redshift Quasars

In Section 4.3, we presented two possible routes to produce a SMBH matching observations. Here we move from specific examples to a discussion of the entire parameter space relevant to the production of high-redshift quasars like ULAS J1120+0641. Recall that we have six input parameters: $\{M_\Delta, c, z, N_e, \sigma, f\}$, respectively the halo mass, concentration, redshift of virialization, e-folds of Eddington-limited accretion after collapse, and uSIDM cross section per unit mass and fraction. We specify the halo properties as above: $M_\Delta = 10^{12}\,M_\odot$, $c = 9$, $z = 15$. We then use Equation (37) and the redshift at which a quasar is observed, in this case $z_{\rm obs} = 7.085$, to eliminate N_e, leaving a two-dimensional parameter space for production of black holes by this time. Finally, the requirement that the mass of the quasar match observations, $M_{\rm SMBH} \approx 2 \times 10^9\,M_\odot$, combined with the assumption of continuous Eddington-limited growth since black hole formation, reduces the parameter space to one dimension, a curve $\sigma(f)$.

We present the parameter space in Figure 5. The one-dimensional curve $\sigma(f)$, where continual Eddington-limited accretion since black hole formation results in a SMBH with $M_{\rm SMBH} = 2 \times 10^9\,M_\odot$ at $z_{\rm obs} = 7.085$, is the solid black line. Because the baryonic accretion history after seed black hole formation is uncertain, as discussed in Section 2, we also indicate with the shaded regions the entire portion of the full σ–f plane in which black holes of any size smaller than $M_{\rm SMBH}$ are produced by $z_{\rm obs}$.

There are several constraints on this reduced parameter space. First is the simple requirement that gravothermal collapse indeed occurs before $z_{\rm obs} = 7.085$. We have already seen in Section 4.3 that this constrains $\sigma f \geqslant 0.336$ cm^2 g^{-1}. Second is the requirement that the black hole produced by gravothermal collapse must not be larger than the observed mass of ULAS J1120+0641. Combined with additional assumptions about baryonic accretion, this excludes the entire region above the black curve in Figure 5. Even without the assumptions, this still constrains the black hole mass via Equation (31), and therefore the uSIDM fraction f, provided that a black hole is actually produced. Again, the resulting constraint was calculated in Section 4.3: $f \leqslant 0.112$. Larger values of f would produce black holes which contained too large a portion of the mass of the entire halo.

Finally, recall that our expressions for the collapse time and resulting black hole mass are based on simulations. As discussed in Section 3.2, the simulations assume the uSIDM is initially in NFW profile, which is only valid if uSIDM interactions were slow compared to the timescale of halo formation, i.e., when the halo is initially optically thin. This places a constraint on σf as a function of the halo parameters, given by Equation (27). For our chosen values this gives $\sigma f \leqslant 0.425$ cm^2 g^{-1}.

Because σf directly sets the collapse time via Equations (33) and (35), the upper bound on σf is also a lower bound on the time of formation of a black hole from an initially optically thin uSIDM halo: in this case, we must have $z_{\rm coll} \leqslant 7.90$. In turn, this places an upper bound on the number of e-folds of growth from baryons that can occur before $z_{\rm obs} = 7.085$, via Equation (37). We find $N_e \leqslant 2.24$, i.e., black holes formed from optically thin uSIDM halos have time to grow less than an order of magnitude from baryons. In particular, we cannot trust

the precise results of our simulations for the example we considered in Section 4.3, with $N_e = \ln 10^4$. Note, however, that the upper bound on σf, and thus on $z_{\rm coll}$, is independent of $z_{\rm obs}$: it depends only on z, the redshift of halo formation. Most high-redshift quasars are seen near $z_{\rm obs} \sim 6$: ULAS J1120 +0641 is an outlier. Black holes at redshift 6 have had time for another 180 Myr $\sim 4t_{\rm Sal}$ of baryonic growth, so they can grow by up to a factor of ∼500 from baryons.

We have seen that there is an extremely narrow range, $0.336\ {\rm cm}^2\ {\rm g}^{-1} \leqslant \sigma f \leqslant 0.425\ {\rm cm}^2\ {\rm g}^{-1}$, in which the uSIDM halo considered here is optically thin at virialization but nevertheless rapidly collapses to form a black hole. How can we explain the closeness of these two bounds? In general, they are not independent. The upper bound requires that the initial halo be optically thin, i.e., that the scattering cross section be less than the "characteristic cross section" of the halo, $1/(\rho_s r_s)$. But the lower bound requires that collapse not take too long, i.e., that the scattering time at the characteristic radius, $1/(\sigma f \rho_s v_s)$ is small compared with a Hubble time. These bounds can be simultaneously satisfied when $H^{-1} \sim r_s/v_s$. But, ignoring concentration dependence and numerical factors, $v_s \sim \sqrt{GM_\Delta/r_s} \sim \sqrt{G\rho_{\rm crit} r_s^2} \sim Hr_s$, so $r_s/v_s \sim H^{-1}$ as desired. That is, both bounds exhibit the same mass dependence, and their redshift dependence is identical when z and $z_{\rm coll}$ are similar, as can be verified from Equations (26), (33), and (35). For our particular choice of $c = 9$, the numerical factors are nearly canceled by the concentration dependence, so the bounds are especially close.

We emphasize again, however, that the upper bound on σf (the red dashed curve in Figure 5) is not a true physical exclusion of the uSIDM parameter space above it. It merely signals that the fluid approximation used in this paper is no longer valid outside this space. As discussed extensively in Section 3.2, we expect that profiles in which the uSIDM starts optically thick should in fact undergo collapse even faster. The requirement of an optically thin initial profile would only be physical if starting otherwise led to fragmentation, turbulence, or some other mechanism by which the core was destroyed or core collapse avoided.

Figure 5 presents the uSIDM parameter space for a particular choice of halo parameters $\{M_\Delta, c, z\}$. We briefly consider how constraints on the parameter space are changed when these parameters are altered. First consider the halo mass M_Δ. The collapse time (Equation (33)) scales as $M_\Delta^{-1/3}$, so smaller values of the halo mass require values of σf to form black holes in the same time. At the same time, the value of σf required for the halo to start initially optically thin (Equation ((26)) has the same scaling with halo mass. So decreasing M_Δ will shift both bounds to higher values of σf (up and to the right on the σ–f plane), but it will not *qualitatively* change the shape of the allowed parameter space. This shift accounts for the main difference between the high-redshift quasar parameter space and the dwarf satellite parameter space we will consider next. We note, however, that at relatively low redshifts there is a well-known black hole–bulge relation (Magorrian et al. 1998; Marconi & Hunt 2003; Haring & Rix 2004), $M_{\rm SMBH} \sim 10^{-3} M_{\rm bulge}$. If this relation persists at high redshifts, we should not depart too far from $M_\Delta \sim 10^{12} M_\odot$ to explain $M_{\rm SMBH} \sim 10^9 M_\odot$.

Next consider the concentration parameter c. Again consulting Equation (33), we see that the collapse time depends strongly on concentration, scaling roughly as $c^{-7/2}$. The collapse time depends more strongly on the concentration than does the optically thin condition, so for small enough values of c it will be impossible to form black holes before a given redshift starting from an optically thin halo. When the other halo parameters and $z_{\rm obs}$ are kept fixed, we find that this critical

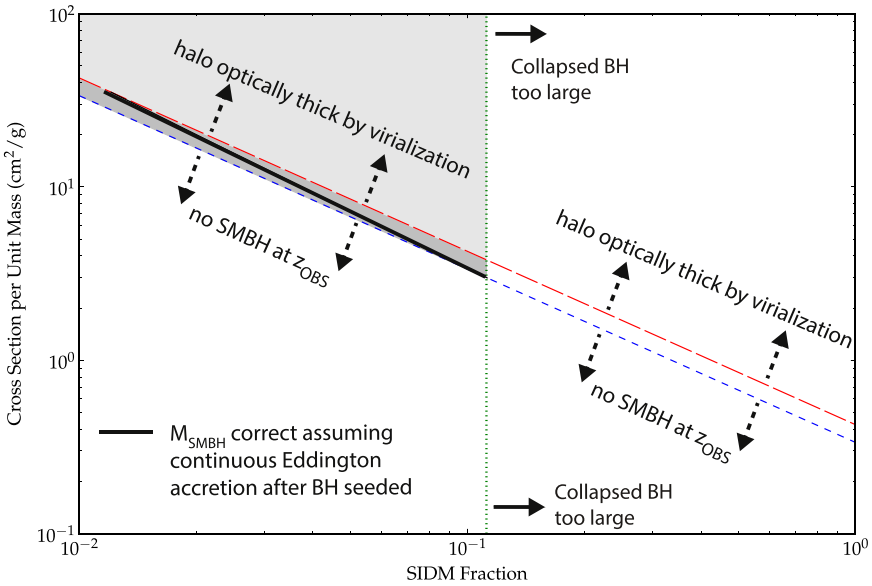

Figure 5. uSIDM parameter space for production of massive high-redshift quasars. We have used the numbers considered in the example above: $M_{\rm SMBH} \approx 2 \times 10^9 M_\odot$, $z_{\rm obs} = 7.085$, $M_\Delta = 10^{12} M_\odot$, $c = 9$, $z = 15$. The solid line plots values of σ and f that result in an SMBH of the desired size at the time of observation, assuming continuous Eddington accretion from the time the core collapses and the seed black hole is formed. The green dotted vertical line marks the largest allowed value of f. To its right, collapsed black holes are already larger than $M_{\rm SMBH}$. To its left, collapsed black holes form smaller than $M_{\rm SMBH}$, but can grow larger by accreting baryons. The points on the blue dashed line all result in collapse precisely at the redshift of observation; below this line, a black hole has not yet formed by $z_{\rm obs}$. Points on and above the red dashed line result in a halo that is already optically thick at the time of virialization, i.e., optically thick at the characteristic radius for the initial NFW profile. As discussed in the text, the methods used in this paper are not directly applicable here, but we still expect gravothermal collapse. Numerical values for all of these bounding lines are given in the text.

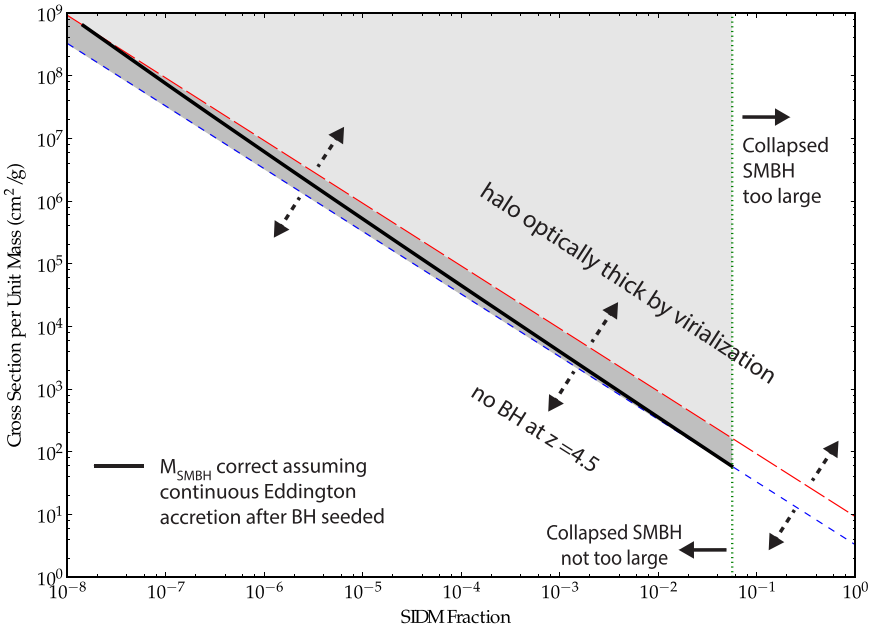

Figure 6. uSIDM parameter space for production of black holes in dwarf galaxies. The parameters used are $M_{\rm SMBH} = 10^5\,M_\odot$, $z_{\rm obs} = 4.5$, $M_\Delta = 10^8\,M_\odot$, $c = 9$, and $z = 15$. The solid line plots values of σ and f that result in an SMBH of the desired size at the time of observation, assuming continuous Eddington accretion from the time the core collapses and the seed black hole is formed. The green dotted vertical line marks the largest allowed value of f. To its right, collapsed black holes are already larger than $M_{\rm SMBH}$. To its left, collapsed black holes form smaller than $M_{\rm SMBH}$, but can grow larger by accreting baryons. The points on the blue dashed line all result in collapse precisely at the redshift of observation; below this line, a black hole has not yet formed by $z_{\rm obs}$. Points on and above the red dashed line result in a halo that is already optically thick at the time of virialization, i.e., optically thick at the characteristic radius for the initial NFW profile. As discussed in the text, the methods used in this paper are not directly applicable here, but we still expect gravothermal collapse. Numerical values for all of these bounding lines are given in the text.

value of c is 7.4. Conversely, by going to larger and larger values of c we can form black holes by any desired time at smaller and smaller values of σf. However, extremely high values of the concentration parameter correspond (unsurprisingly) to extremely concentrated halos, with $r_s \ll r_\Delta$. It is not clear that such halos are actually produced in ΛCDM.

Finally, consider the redshift of halo formation z. Once more consulting Equation (33), we see that we can take the collapse time to zero by increasing z. Heuristically, this is because the critical density, and thus characteristic density, increases with increasing redshift, so a just-virialized halo is closer to the densities needed to start the catastrophic collapse process. However, producing large virialized halos at higher and higher redshifts becomes increasingly unphysical given the bottom-up structure formation mechanism in ΛCDM. Decreasing z has the opposite effect: halos of a given size become more common, but larger values of σf are required to produce a black hole by a given $z_{\rm obs}$. Like the case of small concentration parameter, for small enough z it is impossible to form black holes starting from optically thin halos before a given time. In this case the bound on the redshift of formation for the halo considered here is $z > 13.53$.

5.3.2. Dwarf Galaxies

Recall from the previous subsection that one resolution to the Too Big to Fail problem is the formation of cores in dwarf galaxies if matter is ejected during binary black hole mergers. Our goal here is to specify the parameter space in which uSIDM produces black holes in the progenitors of dwarf galaxies before the epoch in which binary mergers are common. As in the case of high-redshift quasars above, we start by specifying a set of typical values for the halo parameters $\{M_\Delta, c, z\}$. Boylan-Kolchin et al. (2012) compared the Milky Way dwarf galaxies to subhalos around similarly-sized galaxies in the Aquarius simulations (Springel et al. 2008) to derive probable values for the virial mass M_Δ and maximum central velocity $v_{\rm max}$ of each halo at the time of its infall into the main Milky Way halo.

Recall that the maximum velocity of an NFW profile is

$$v_{\rm max} = 0.465\sqrt{\frac{c}{K_c}}\,v_\Delta, \tag{39}$$

with $K_c = \ln(1+c) - c/(1+c)$, at radius

$$r_{\rm max} = 2.163 r_s, \tag{40}$$

as can easily be verified numerically using the definition of the NFW density profile (19) and $v = \sqrt{GM(r)/r}$. Rearranging gives an expression for K_c/c in terms of M_Δ, $v_{\rm max}$, and $z_{\rm infall}$. In particular, since $r_\Delta \sim r_s \sim \rho_{\rm crit}(z)^{-1/3}$, we have $K_c/c \sim \rho_{\rm crit}(z)^{1/3}$. But K_c/c has a maximum value of 0.216, so above some value of $z_{\rm infall}$, there is no possible NFW profile with the given values of M_Δ and $v_{\rm max}$. For the derived values for the Milky Way dwarfs, we find in general that $z_{\rm infall} \lesssim 6$. As the infall redshift moves lower for each particular dwarf, the concentration increases from a minimal value $c = 2.16$. (Actually, K_c/c attains its maximum at $c = 2.16$ and approaches zero both as $c \to 0$ and $c \to \infty$, but we neglect the former branch, with $c < 2.16$, as unphysical.) In particular, we choose $z_{\rm infall} \sim 4.5$, which results in $c \sim 9$ for a typical dwarf, the same as considered above.

A typical Milky Way dwarf in (Boylan-Kolchin et al. 2012) has $M_\Delta = 2 \times 10^8\,M_\odot$ at the time of infall. If Too Big to Fail

is to be explained by means of binary black hole mergers, a typical dwarf should have undergone a major merger, so that a binary black hole merger occurs in the first place. We will therefore take $M_\Delta = 10^8\,M_\odot$, $c = 9$, $z_{\rm obs} = 4.5$ as our typical parameters.

Figure 6 presents the parameter space for our typical dwarf halo. We have taken the black hole mass to be $M_{\rm SMBH} = 10^5\,M_\odot$, in accordance with the black hole–bulge relation, and again assumed that the redshift of formation of the progenitor halo is $z = 15$. The various bounds in the Figure are attained in the same manner as they were for the quasar bounds shown in Figure 5, so we simply quote them here and refer to the discussion above for details of their calculation. The lower bound on σf, which comes from requiring collapse before $z_{\rm obs} = 4.5$, is $\sigma f \geqslant 3.26\ {\rm cm^2\ g^{-1}}$. The upper bound on f, which is calculated using Equation (31) and scales linearly with $M_{\rm SMBH}$, is $f \leqslant 0.056$.

The upper bound on σf, set by the requirement of an optically thin initial profile, is $\sigma f \leqslant 9.16\ {\rm cm^2\ g^{-1}}$. This corresponds to an upper bound on the redshift of collapse, $z_{\rm coll} \leqslant 7.90$. (As discussed above, the upper and lower bounds have the same mass dependence, and we are considering the same values of z and c as we did for high-redshift quasars, so we recover the same bound on the collapse time.) Once again, this gives an upper bound on the number of e-folds of growth from baryons, via Equation (37). Because we have a much longer time for growth after black hole formation than we did in the high-redshift quasar case, it is much looser: $N_e \leqslant 15.2$. Once again, the allowed range on σf is a factor of only a few. But the significantly looser bound on N_e means that black holes can grow by a factor of nearly 4×10^6. So f can be decreased by over six orders of magnitude from its maximal value, and σ increased by a corresponding amount, while still maintaining an optically thin initial profile and allowing reasonably large black holes to form. This explains the much larger range in σ and f seen on Figure 6 compared with Figure 5.

5.3.3. Both Simultaneously?

We have just seen that the minimum value of σf needed to produce massive high-redshift quasars is about an order of magnitude lower than those needed to produce black holes in dwarf galaxies before major mergers. We can understand this qualitatively from the expression for the halo relaxation time, (Equation (33)): it scales as $M_\Delta^{-1/3}$. The black holes in dwarf galaxies have about twice as long to form, until $z_{\rm obs} = 4.5$ instead of $z_{\rm obs} = 7.085$, so σf is scaled by a factor of $10^{4/3}/2 \approx 10$.

This scaling of σf implies that the uSIDM parameters that produce black holes in dwarfs are a strict subset of those that produce high-redshift quasars. It is then easy to choose values which solve both: one simply takes $\sigma f \geqslant 3.26\ {\rm cm^2\ g^{-1}}$ and chooses compatible values of σ and f to taste. Because σf is significantly larger than the minimum value needed to produce high-redshift quasars, the uSIDM halos which produce them will start initially optically thick, above the (red dashed) upper bound on σf shown in Figure 5. In this optically thick regime, the expressions for the gravothermal collapse time and black hole seed mass derived from the simulations of Section 3 should be taken as limits: we expect that gravothermal collapse should occur a slightly shorter time after halo formation and result in slightly larger seed black holes.

As an example, consider the $\{\sigma, f\}$ values that fall in the one-dimensional parameter space discussed at the beginning of this subsection, where continuous Eddington accretion from the time of black hole collapse until $z_{\rm obs}$ just produces a SMBH with the observed value of $M_{\rm SMBH} = 2 \times 10^9\,M_\odot$ (the solid black line in Figure 5). If $\sigma f = 3.26$, the smallest possible value needed to also produce black holes in dwarfs by redshift 4.5, we would find using Equations (33) and (37) that $z_{\rm coll} = 13.3$. (Strictly speaking, Equation (37) is not valid in the context of an optically thick initial halo, since it uses an expression for the collapse time derived from an initially optically thin NFW profile. We are simply using it here for the sake of illustration.) In this case there is time for 9.54 e-folds of baryonic accretion before $z = 7.085$, and the initial black hole has mass $7 \times 10^4\,M_\odot$. This determines the USDIM parameters for this example as $\sigma = 8.14 \times 10^5\ {\rm cm^2\ g^{-1}}$, $f = 4.01 \times 10^{-6}$.

As the beginning of this section emphasized, it is largely beyond this scope of this paper to describe a fully consistent uSIDM cosmology. Nevertheless, this subsection suggests that the uSDIM paradigm is flexible enough to resolve both of the potential tensions within ΛCDM discussed here. It is possible that a single species of uSIDM could in fact resolve both tensions simultaneously. In investigating this question further, it will be important to move beyond the simplifying assumptions employed herein, especially the stipulations of an initial optically thin profile and a cosmologically isolated profile.

6. CONCLUSION

In this paper, we considered a minimal extension of the SIDM parameter space, in which a self-interacting component comprises only a fraction of the dark matter. For $f \lesssim 0.1$, this evades all prior constraints on SIDM models. We highlighted the uSIDM regime, where the SIDM component is subdominant but ultra-strongly self-interacting, with $f \ll 1$ and $\sigma \gg 1\ {\rm cm^2\ g^{-1}}$. In the setup considered here, the presence of uSIDM leads to the production of black holes with a mass of around 2% of the total uSIDM mass in the halo at very early times. In particular, such black holes can act as seeds for baryon accretion starting soon after halo formation, alleviating potential difficulties with accommodating massive quasars at high redshifts within the standard ΛCDM cosmology. If black holes are formed ubiquitously in dwarf halos before they undergo mergers, they may also resolve the Too Big to Fail problem by ejecting matter from cores during black hole mergers. More detailed cosmological simulations are needed to confirm the conclusions of this paper and suggest other potential observational consequences of uSIDM.

Setting aside the detailed predictions, this paper has demonstrated that multi-component dark matter can have strong effects on small scales while still evading existing constraints. In the toy model discussed here, the strong effect was the result of the gravothermal catastrophe. Gravothermal collapse of a strongly-interacting dark matter component is a novel mechanism for production of seed black holes, potentially one with many implications. Given its appearance in the simple extension of ΛCDM considered here, it is plausible that gravothermal collapse and its observational consequences, such as seed black hole formation, are generic features of more detailed models. It is important to consider, and then observe or constrain, this and other observational

consequences that are qualitatively different from the predictions of the standard cosmological model.

Our discussion has been purely phenomenological, so it is reassuring to note the existence of a class of hidden-sector models (Boddy et al. 2014) that naturally produce a subdominant strongly-interacting dark matter component, with self-interaction cross-sections ranging as high as $\sigma \sim 10^{11}\,\mathrm{cm}^2\,\mathrm{g}^{-1}$. Very interestingly, some models give both a dominant component with $\sigma \simeq 0.1$–$1\,\mathrm{cm}^2\,\mathrm{g}^{-1}$, as needed to alleviate discrepancies between ΛCDM and observations, and a uSIDM component with $\sigma \simeq 10^5$–$10^7\,\mathrm{cm}^2\,\mathrm{g}^{-1}$, which could produce seed black holes via the mechanism described in this paper.

We thank Shmulik Balberg, James Bullock, Renyue Chen, Phil Hopkins, Jun Koda, Sasha Muratov, Lisa Randall, Paul Shapiro, Stu Shapiro, Charles Steinhardt, and Naoki Yoshida for helpful discussions. We thank especially Sasha Muratov for measuring concentration parameters at high redshifts in the FIRE runs and providing us with the resulting halo catalogs. This research is funded in part by DOE Grant #DE-SC0011632, and by the Gordon and Betty Moore Foundation through Grant #776 to the Caltech Moore Center for Theoretical Cosmology and Physics.

REFERENCES

Abel, T., Wise, J. H., & Bryan, G. L. 2007, ApJL, 659, L87
Balberg, S., Shapiro, S. L., & Inagaki, S. 2002, ApJ, 568, 475
Bardeen, J. M., Press, W. H., & Teukolsky, S. A. 1972, ApJ, 178, 347
Blumenthal, G. R., Faber, S., Flores, R., & Primack, J. R. 1986, ApJ, 301, 27
Boddy, K. K., Feng, J. L., Kaplinghat, M., & Tait, T. M. P. 2014, arXiv:1402.3629[hep-ph]
Bondi, H. 1952, MNRAS, 112, 195
Boylan-Kolchin, M., Bullock, J. S., & Kaplinghat, M. 2011, MNRAS, 415, L40
Boylan-Kolchin, M., Bullock, J. S., & Kaplinghat, M. 2012, MNRAS, 422, 1203
Boylan-Kolchin, M., Ma, C.-P., & Quataert, E. 2004, ApJL, 613, L37
Boylan-Kolchin, M., Springel, V., White, S. D., Jenkins, A., & Lemson, G. 2009, MNRAS, 398, 1150
Bryan, G., & Norman, M. 1998, ApJ, 495, 80
Clowe, D., Markevitch, M., Bradac, M., et al. 2012, ApJ, 758, 128
Dokuchaev, V., Eroshenko, Y., & Rubin, S. 2007, arXiv:0709.0070
Eke, V. R., Cole, S., & Frenk, C. S. 1996, MNRAS, 282, 263
Fan, J., Katz, A., Randall, L., & Reece, M. 2013a, PhRvL, 110, 211302
Fan, J., Katz, A., Randall, L., & Reece, M. 2013b, PDU, 2, 139
Ferrarese, L., & Merritt, D. 2000, ApJL, 539, L9
Fitchett, J. M. 1983, MNRAS, 203, 1049
Gao, L., Abel, T., Frenk, C., et al. 2007, MNRAS, 378, 449
Garrison-Kimmel, S., Boylan-Kolchin, M., Bullock, J. S., & Kirby, E. N. 2014, MNRAS, 444, 222
Gebhardt, K., Bender, R., Bower, G., et al. 2000, ApJL, 539, L13
Gnedin, O. Y., Kravtsov, A. V., Klypin, A. A., & Nagai, D. 2004, ApJ, 616, 16
Gnedin, O. Y., & Ostriker, J. P. 2001, ApJ, 561, 61
Graham, A. W., Merritt, D., Moore, B., Diemand, J., & Terzic, B. 2006, AJ, 132, 2701
Haiman, Z. 2004, ApJ, 613, 36
Haiman, Z. 2013, in Astrophysics and Space Science Library, Vol. 396, The First Galaxies, ed. T. Wiklind, B. Mobasher, & V. Bromm (Berlin: Springer-Verlag), 293
Haiman, Z., & Loeb, A. 1998, ApJ, 503, 505
Haiman, Z., & Loeb, A. 2000, arXiv:astro-ph/0011529[astro-ph]
Haiman, Z., Quataert, E., & Bower, G. C. 2004, ApJ, 612, 698
Haring, N., & Rix, H.-W. 2004, ApJ, 604, 89
Harvey, D., Tittley, E., Massey, R., et al. 2014, MNRAS, 441, 404
Heggie, D. 1975, MNRAS, 173, 729
Hennawi, J. F., & Ostriker, J. P. 2002, ApJ, 572, 41
Hopkins, P. F., Keres, D., Onorbe, J., et al. 2013, arXiv:1311.2073[astro-ph.CO]
Hu, J., Shen, Y., Lou, Y.-Q., & Zhang, S. 2006, MNRAS, 365, 345
Hut, P., McMillan, S., Goodman, J., et al. 1992, PASP, 104, 981
Jee, M. J., Hoekstra, H., Mahdavi, A., & Babul, A. 2014, ApJ, 783, 78
Jee, M., Mahdavi, A., Hoekstra, H., et al. 2012, ApJ, 747, 96
Jiang, Y.-F., Stone, J. M., & Davis, S. W. 2014, arXiv:1410.0678
Johnson, J. L., & Bromm, V. 2007, MNRAS, 374, 1557
Kelly, B. C., & Merloni, A. 2012, AdAst, 2012, 970858
Klypin, A. A., Kravtsov, A. V., Valenzuela, O., & Prada, F. 1999, ApJ, 522, 82
Klypin, A., Trujillo-Gomez, S., & Primack, J. 2010, arXiv:1002.3660
Knollmann, S. R., & Knebe, A. 2009, ApJS, 182, 608
Koda, J., & Shapiro, P. R. 2011, MNRAS, 415, 1125
Kollmeier, J. A., Onken, C. A., Kochanek, C. S., et al. 2006, ApJ, 648, 128
Lahav, O., Lilje, P. B., Primack, J. R., & Rees, M. J. 1991, MNRAS, 251, 128
Li, Y.-X., Hernquist, L., Robertson, B., et al. 2007, ApJ, 665, 187
Lifshitz, E. M., & Pitaevskii, L. P. 1981, Physical Kinetics
Lippai, Z., Frei, Z., & Haiman, Z. 2008, arXiv:0801.0739 [astro-ph]
Ludlow, A. D., Navarro, J. F., Angulo, R. E., et al. 2014, MNRAS, 441, 378
Ludlow, A. D., Navarro, J. F., Li, M., et al. 2012, MNRAS, 427, 1322
Lynden-Bell, D., & Eggleton, P. P. 1980, MNRAS, 191, 483
Lynden-Bell, D., & Wood, R. 1968, MNRAS, 138, 495
Madau, P., Haardt, F., & Dotti, M. 2014, ApJL, 784, L38
Magorrian, J., Tremaine, S., Richstone, D., et al. 1998, AJ, 115, 2285
Mahdavi, A., Hoekstra, H., Babul, A., Balam, D., & Capak, P. 2007, ApJ, 668, 806
Marconi, A., & Hunt, L. K. 2003, ApJL, 589, L21
Markevitch, M., Gonzalez, A., Clowe, D., et al. 2004, ApJ, 606, 819
McCullough, M., & Randall, L. 2013, JCAP, 1310, 058
McKee, C. F., & Tan, J. C. 2007, arXiv:0711.1377 [astro-ph]
Merritt, D., & Milosavljevic, M. 2005, LRR, 8, 8
Merritt, D., Milosavljevic, M., Favata, M., Hughes, S. A., & Holz, D. E. 2004, ApJL, 607, L9
Merritt, D., Navarro, J. F., Ludlow, A., & Jenkins, A. 2005, ApJL, 624, L85
Milosavljevic, M., & Merritt, D. 2001, ApJ, 563, 34
Milosavljevic, M., & Merritt, D. 2003, ApJ, 596, 860
Milosavljevic, M., Merritt, D., Rest, A., & van den Bosch, F. C. 2002, MNRAS, 331, L51
Moore, B. 1994, Natur, 370, 629
Moore, B., Ghigna, S., Governato, F., et al. 1999, ApJL, 524, L19
Moore, B., Governato, F., Quinn, T. R., Stadel, J., & Lake, G. 1998, ApJL, 499, L5
Mortlock, D. J., Warren, S. J., Venemans, B. P., et al. 2011, Natur, 474, 616
Narayan, R., & McClintock, J. E. 2012, MNRAS, 419, L69
Neto, A. F., Gao, L., Bett, P., et al. 2007, MNRAS, 381, 1450
Ostriker, J. P. 2000, PhRvL, 84, 5258
Papastergis, E., Giovanelli, R., Haynes, M. P., & Shankar, F. 2015, A&A, 574, A113
Peter, A. H., Rocha, M., Bullock, J. S., & Kaplinghat, M. 2012, arXiv:1208.3026
Planck Collaboration, et al. 2013, arXiv:1303.5076
Prada, F., Klypin, A. A., Cuesta, A. J., Betancort-Rijo, J. E., & Primack, J. 2012, MNRAS, 428, 3018
Randall, L., & Reece, M. 2014, PhRvL, 112, 161301
Randall, L., & Scholtz, J. 2014, arXiv:1412.1839
Randall, S. W., Markevitch, M., Clowe, D., Gonzalez, A. H., & Bradac, M. 2008, ApJ, 679, 1173
Riebe, K., Partl, A. M., Enke, H., et al. 2011, arXiv:1109.0003
Rocha, M., Peter, A. H., Bullock, J. S., et al. 2013, MNRAS, 430, 81
Salpeter, E. 1964, ApJ, 140, 796
Sesana, A. 2012, AdAst, 2012, 805402
Shakura, N., & Sunyaev, R. 1973, A&A, 24, 337
Shapiro, S. L., & Teukolsky, S. A. 1985a, ApJ, 298, 58
Shapiro, S. L., & Teukolsky, S. A. 1985b, ApJ, 298, 34
Shapiro, S. L., & Teukolsky, S. A. 1986, ApJ, 307, 575
Shapiro, S. L. 2005, ApJ, 620, 59
Spergel, D. N., & Steinhardt, P. J. 2000, PhRvL, 84, 3760
Springel, V., Wang, J., Vogelsberger, M., et al. 2008, MNRAS, 391, 1685
Springel, V., White, S. D., Jenkins, A., et al. 2005, Natur, 435, 629
Stacy, A., Greif, T., & Bromm, V. 2009, arXiv:0908.0712
Steinhardt, C. L., & Elvis, M. 2010, MNRAS, 402, 2637
Steinhardt, C. L., Elvis, M., & Amarie, M. 2011, arXiv:1103.4608
Tegmark, M., Silk, J., Rees, M. J., et al. 1997, ApJ, 474, 1
Trakhtenbrot, B. 2014, ApJL, 789, L9
Treister, E., & Urry, C. M. 2012, AdAst, 2012, 516193
Turk, M. J., Abel, T., & O'Shea, B. W. 2009, Sci, 325, 601
Venemans, B., McMahon, R., Walter, F., et al. 2012, ApJL, 751, L25
Vogelsberger, M., Zavala, J., & Loeb, A. 2012, MNRAS, 423, 3740

Volonteri, M. 2010, A&ARv, 18, 279
Volonteri, M., & Silk, J. 2014, arXiv:1401.3513
Whalen, D. J., Heger, A., Chen, K.-J., et al. 2013, ApJ, 778, 17
Yoshida, N., Oh, S. P., Kitayama, T., & Hernquist, L. 2007, ApJ, 663, 687
Yoshida, N., Springel, V., White, S. D., & Tormen, G. 2000, ApJL, 544, L87
Zavala, J., Vogelsberger, M., & Walker, M. G. 2013, MNRAS Letters, 431, L20
Zel'dovich, Y. B., & Podurets, M. A. 1966, SvA, 9, 742
Zhang, S., Cui, W., & Chen, W. 1997, ApJL, 482, L155
Zhao, H. 1996, MNRAS, 278, 488