Gravitational lensing and radio interferometers probe sub-galactic dark matter structure

(story of my life)

Saghar Asadi

2012–2014
Department of Astronomy
Stockholm University

Contents

1 Structure formation

This section needs a HARD polishing! If for no other reason, then simply because it's prone to plagiarism as I'm adopting the framework from this reference of mine and have not changed it much yet!!
Things I need to be able to make the conversion back and forth (by heart!):

- $v_{circ} - M_{vir}$

- $M_{vir} - M_{Ein}$ for my redshift combinations and density profile (at least relatively/roughly)

- $r(pc) - r(arcsec)$ in both source and lens plane

1.1 Concordance model

The concordance model of cosmology is a special case of the FLRW world model currently accepted by most astronomers as the theoretical framework of cosmology, where gravity is the dominant interaction in our Universe which is homogeneous and isotropic on "large enough", i.e., $> \sim 100$ Mpc, scales (cosmological principle). These models have been on the heart of modern cosmology from the beginning of the 20$^{\text{th}}$ century, when the cosmological principle was mainly backed up with philosophical arguments such as the *Copernican principle* stating that we do not occupy a special place in the Universe, rather than detailed observational evidence. As opposed to a few decades later when cosmic isotropy was confirmed by the cosmic microwave background, first by satellite COBE (ref?) and then WMAP (ref?), and very recently PLANCK(ref?). Besides, the homogeneity of the Universe is now backed up by large surveys such as the 2–degree field galaxy redshift survey (2dfGRS), and the Sloan digital sky survey (SDSS).
Here I can go through the RW metric and its terms, but this is not important in the case of my project, so I'll skip it altogether and just talk about its physical interpretation instead!
Dark matter as the currently second dominant energy contribution of the Universe is thought to be non–baryonic, cold (i.e. non–relativistic), very (?!) weakly interacting massive particle (WIMP). Some of these characteristics are inferred from cosmological evidence. For instance, observation of small–scale cosmic structure implies the DM to be non–relativistic, while the theory of cosmic nucleosynthesis results in the DM being non–baryonic. (elaborate and cite!!)

1.2 Newtonian theory of structure formation

Our Universe is neither homogeneous nor isotropic on scales smaller than 100 Mpc. Galaxies and galaxy clusters are huge overdens regions with respect to the mean matter density. They make up the cosmic web consisting of sheets and filaments. Even though the detailed account of structure formation is only possible in the framework of general relativity, the Newtonian approach is sufficient to give a simple analytical description of most of the processes. This section describes the Newtonian theory of structure formation (within our horizon) in the framework of the Concordance model, i.e. a flat geometry.
The current large–scale structure in the Universe grew (by gravitational instability) from small perturbations in the homogeneous FLRW model. In order to solve perturbed hydrodynamic equations, we keep the first perturbed terms only (linear perturbation), and treat it in the framework of Newtonian theory implying (explain how exactly!) the following assumptions:

- The scale of perturbations is small compared to the horizon ($\ell \ll ct_0$)

- The flow is non–relativistic ($|v| \ll c$) and ($p \ll \rho c^2$)

Moreover, we assume that the Universe is filled with an inhomogeneous, dissipationless (p is very small!), and ideal fluid. (me: Therefore, we are basically dealing with DM particles as they are currently understood, and not baryons which are clearly dissipative!). Moreover, we are assuming what arises is adiabatic perturbations, meaning that the relative number density of flows remain constant. It also implies that different flows do not transfer energy and the conservation of energy is satisfied by each flow independently, $\rho_I \propto a_I^{-3}$

Friedmann equations can be solved for various flows as long as they do not interact beyond gravity. This is the case for different fluids in the structure formation (i.e. post–inflation) era. The matter density, velocity field, pressure, gravitational potential, and entropy per unit mass of the fluid are shown as $\rho(t,\mathbf{r})$, $\mathbf{v}(t,\mathbf{r})$, $p(t,\mathbf{r})$, $\Phi(t,\mathbf{r})$, and $S(t,\mathbf{r})$, respectively. where t is the cosmic time and $\mathbf{r}$ denotes the physical (not comoving!) coordinates. Therefore, we can write the equation of mass conservation (*continuity equation*):

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho \mathbf{v}) = 0, \tag{1.1}$$

the *Euler equation*

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} = -\frac{\nabla P}{\rho} - \nabla\Phi, \tag{1.2}$$

the *Poisson equation*

$$\nabla^2 \Phi = 4\pi G\rho, \tag{1.3}$$

and the conservation of entropy

$$\frac{\partial S}{\partial t} + (\mathbf{v}.\nabla)S = 0, \tag{1.4}$$

and an equation of state $P = P(\rho, S)$. Now, we add the perturbation term to each solution. Therefore we have:

$$\rho(t,\mathbf{r}) = \bar{\rho}(t) + \delta\rho(t,\mathbf{r})$$

$$\mathbf{v}(t,\mathbf{r}) = \bar{\mathbf{v}}(t,\mathbf{r}) + \delta\mathbf{v}(t,\mathbf{r})$$

$$p(t,\mathbf{r}) = \bar{p}(t) + \delta p(t,\mathbf{r})$$

$$\Phi(t,\mathbf{r}) = \bar{\Phi}(t,\mathbf{r}) + \delta\Phi(t,\mathbf{r})$$

$$S(t,\mathbf{r}) = \bar{S}(t) + \delta S(t,\mathbf{r})$$

The velocity field is homogeneous in the comoving coordinate system (and not the physical system we are working on)(are we assuming that? is it the direct consequence of homogeneity and isotropy?, else where does it come from?). The the comoving distance $D(t)$ is related to the proper distance as $D(t) = \frac{D_{prop}(t)}{a(t)}$. Moreover, $H^2 = (\frac{\dot{a}}{a})^2$ by definition(?). Therefore, the homogeneous velocity field is given by the Hubble Law as

$$\bar{\mathbf{v}}(t,\mathbf{r}) = H(t)\mathbf{r} \tag{1.5}$$

The continuity equation for the homogeneous background becomes then

$$\dot{\bar{\rho}} + 3\bar{\rho}H(t) = 0$$

and the divergence of Euler equation, combined with the Poisson equation gives

$$\dot{H} + H^2 = -\frac{4\pi G\bar{\rho}}{3}$$

Now, we substitute perturbed fluid equations to hydrodynamic equations 1.1, 1.2, 1.3, and 1.4 and keep only the linear terms.

$$\frac{\partial \delta\rho}{\partial t} + \bar{\rho}\nabla.\delta\mathbf{v} + \nabla.(\delta\rho\bar{\mathbf{v}}) = 0 \tag{1.6}$$

$$\frac{\partial \delta\mathbf{v}}{\partial t} + (\delta\mathbf{v}.\nabla)\bar{\mathbf{v}} + (\bar{\mathbf{v}}.\nabla)\delta\mathbf{v} = -\frac{\nabla}{\rho}(c_s^2\,\delta\rho + \sigma\,\delta S) - \nabla\delta\Phi \tag{1.7}$$

$$\nabla^2\delta\Phi = 4\pi G\delta\rho \tag{1.8}$$

$$\frac{\partial \delta S}{\partial t} + (\bar{\mathbf{v}}.\nabla)\delta S = 0 \tag{1.9}$$

In the case of the Euler equation we use

$$\frac{1}{(\bar{\rho} + \delta\rho)} \simeq \frac{1}{\bar{\rho}} + \frac{\delta\rho}{\bar{\rho}^2} + \cdots$$

Moreover, we substituted the perturbed EoS at the first order with $\delta P = c_s^2 \, \delta\rho + \sigma \, \delta S$, where $c_s^2 = (\partial \bar{P}/\partial \bar{\rho})_{\bar{S}}$ is the square of the sound speed and $\sigma = (\partial \bar{P}/\partial \bar{S})_{\bar{\rho}}$. Note that the EoS of an ideal gas is given by $P = \frac{\Re_*}{\mu}\rho T$, where T is related to S with the speed of sound, c. (how?). In other words, EoS of an ideal gas is given by P = Now, by going from physical coordinate system $(t, \mathbf{r})$, to comoving $(t, \mathbf{x})$ defined as $x = r/a(t)$ and the proper time $d\tau = dt/a(t)$, we can simplify our hydrodynamic equation set by using transformations below.

$$\begin{aligned}
\mathbf{r} &\rightarrow a(t)\mathbf{x} \\
\nabla_{\mathbf{r}} &\rightarrow \frac{1}{a(t)}\nabla_{\mathbf{x}} \\
\left.\frac{\partial}{\partial t}\right|_{\mathbf{r}} &\rightarrow \left.\frac{\partial}{\partial \tau}\right|_{\mathbf{x}} - \frac{1}{a(t)}\bar{\mathbf{v}}.\nabla_{\mathbf{x}}
\end{aligned}$$

According to 1.5, one can substitute $\bar{\mathbf{v}}$ with $H\mathbf{r}$ in the physical coordinate system and therefore $\dot{a}\mathbf{x}$ in the comoving system of coordinates. Therefore,

$$\nabla.\bar{\mathbf{v}} = 3H \tag{1.10}$$

and

$$(\delta\mathbf{v}.\nabla)\bar{\mathbf{v}}/a = \delta\mathbf{v}H \tag{1.11}$$

Then the linearized hydrodynamic equations become:

$$\delta\dot{\rho} - \frac{1}{a}v.\nabla\delta\rho + \frac{\bar{\rho}}{a}\nabla.\delta v + \frac{1}{a}3\delta\rho H = 0 \tag{1.12}$$

$$\delta\dot{v} - \frac{1}{a}v.\nabla\delta v + \delta vH + \frac{1}{a}(v.\nabla)\delta v = -\frac{\nabla}{a\bar{\rho}}(c_s^2\delta\rho + \sigma\delta S) - \frac{1}{a}\nabla\delta\Phi \tag{1.13}$$

$$\frac{1}{a^2}\nabla^2\delta\Phi = 4\pi G\delta\rho \tag{1.14}$$

$$\delta\dot{S} - \frac{1}{a}v.\nabla\delta S + \frac{1}{a}(v.\nabla)\delta S = 0 \tag{1.15}$$

Now, this guy takes the FT of the above equation set with respect to the comoving coordinates, to obtain them for a given mode $\mathbf{k}$, but I do not understand why!(beginning of page 32 in his manuscript!), then discusses three individual perturbation modes (one entropy mode, two vortical modes and two adiabatic modes) individually and concludes that the only interesting mode for structure formation in an expanding Universe would be the adiabatic modes, while the most general solution is a superposition of all these five! **I leave the structure formation here now and go through halo discussions first, then come back to this later!**

2 Dark matter (9/17)

The "missing matter" in the Universe has represented itself both locally, i.e. in the galactic disk and around the Sun, and globally in the past century and even though the problem has been approached from the particle–physics point of view in addition to the cosmological one, it is still considered as one of the biggest challenges of both! On the one hand, modern data suggests an insignificant amount of dark matter in the Solar vicinity which is made of baryonic but not luminous matter such as faint stars or Jupiter–like objects. In short, according to our current understanding of the Universe, dark matter is the dominant matter component of the Universe.

2.1 Cosmological evidence (9/17)

2.1.1 Masses of galaxy clusters

In the beginning of 1930s, Zwicky (**cite!**) was first to notice the order–of–magnitude difference in dynamical mass of the Coma cluster with the mass derived from luminosities of individual galaxies. The discrepancy was not recognized as an issue until after its persistence in various cases (**cite!**). The previously–unknown population of dark (or high mass–to–light–ratio) matter was proven to dominate the mass budget of the Universe. The mean density of the matter was shown to be ~ 0.2 of the critical density of the Universe and different models for the dark matter started appearing in the literature. Firstly, the missing matter was thought to be in faint stars or hot gas but both failed to explain the situation.
Since (non–baryonic) dark matter is needed for the gravitational clustering and structure formation in the Universe as well as resolving the controversy (what controversy?**) about Big Bang nucleosynthesis, essential information about the nature of dark matter comes from studying the cosmic web and CMB.

2.1.2 Galactic rotation curves

In the end of the 20th century, both optical(Rubin et al. 1978, 1980) and radio(Bosma 1978) data confirmed the flat rotational curves of galaxies at large radii providing important evidence for the presence of massive halos around galaxies. This was followed by new X–ray observations of hot gas in galaxy clusters confirming that the hot gas reservoir cannot explain the missing mass either. The X–ray data, additionally, provided new measurements of the velocity dispersions of galaxies in clusters, confirming the previous dynamical mass estimates(Forman et al. 1972, Gursky et al. 1972, Kellogg et al. 1973). Another independent probe for mass of galaxy clusters was gravitational lensing, which served as yet another support for the presence of a dark massive component.(**cite!**)

2.1.3 Gravitational lensing

The mathematical framework of gravitational lensing is a commonly–accepted framework in astronomy for a long time. However, only after the general relativistic corrections to the Newtonian formalism of gravitational lensing, the dark matter problem arose in this context. Gravitational lensing, as an accurate and reliable gravitational mass measurement of the lens, confirmed the discrepancy between the gravitational and luminous mass both on galactic and galaxy cluster scales.

2.1.4 CMB temperature fluctuations

As discussed in section 1.2, structures in the Universe are believed to have grown from small fluctuations in the density field of the Universe after the era of recombination. Density fluctuations are of the same order as temperature fluctuations (**cite!**), therefore measuring temperature fluctuations of the CMB reveals the order of magnitude of density fluctuations in the early Universe. However, taking only the baryonic matter into account, these fluctuations are two small to give rise to any structure formation in our expanding Universe. While the required amount of matter in the Universe is $\sim 0.2 - 0.3$ in units of the critical cosmological density (**cite?**), the big bang nucleosynthesis puts an upper limit of ~ 0.04 on the amount of baryonic matter (**cite!**) supporting the non–baryonic nature of dark matter. Moreover, if dark matter is non–baryonic,

its density fluctuations can start growing already at the radiation–dominated era while the growth of baryonic matter is damped by radiation, therefore this could explain small density fluctuations of the CMB. The first non–baryonic candidate for dark matter particles was neutrino. However, simulations based on this model (Doroshkevich & Shandarin 1978) indicated that neutrino–dominated dark matter cannot give rise to small–scale structure in the distribution of galaxies, due to the cut–off at the power–spectrum of these rapidly–moving particles. This also suggested that suitable candidates for dark matter particles not only need to be dissipationless, but also required to be much more massive than neutrinos (Blumenthal et al. 1982, Bond et al. 1982, Peebles, 1982). Therefore, the first numerical cosmological simulations based on cold dark matter (CDM) was performed by (Melott et al. 1983) which could represent the small structures of the Universe much more accurately.

The discrepancy between the expected amplitude of temperature fluctuations in the CMB data based on structure formation (density/temperature fluctuations of the order of 10^{-3}) cannot be explained without a large amount of non–baryonic matter in the radiation–dominated era. Another strong argument supporting the presence of some matter in addition to the baryonic matter in the Universe is also coming from the CMB. (**the first peak in baryonic acoustic oscillations**)

2.2 Detection methods

Indirect detection of DM particles

Actually, there is a category of methods, this author calls "very indirect" methods for detecting dark matter and all cosmological/astrophysical effects induced by dark matter is included in this category. On the other hand, gamma–rays and neutrinos are "not–so–indirect" probes of dark matter in his book.

3 main processes to study, (a) *pair–annihilation*, (b) *decay of DM particles into SM ones*, (c) *elastic scattering between DM and SM particles*, where SM means standard model particles, i.e. baryons! Observing the rate of SM particle production places constraints on DM particles via processes (a) and (b). Momentum (mass) of DM particles is set by observing energy scale of SM "messengers" via (c). The type of SM particles produced in (a) is strongly dependent on details of DM particle model.

In the very early Universe, when matter and radiation was coupled, at those high temperatures, the particle interaction (Γ) rate was larger than the Hubble expansion rate (H), we call the point in time/temperature where $\Gamma \sim H$, "freeze–out". After this point, particle species just "redshift" whatever momentum and number density they had at this point. This point differs for various particle species, with different number densities and interaction cross sections. If the particle species freeze out while they are relativistic, they are called *hot relics*, otherwise they are *cold relics* because the freeze–out temperature is much lower than the mass of the particle. The details of DM particle model puts an upper limit as well as a lower limit to the mass of *cold* dark matter particles. This lower limit corresponds to the mass of first structures that gravitationally collapsed in the Universe, and for a typical WIMP model is $\sim M_{\oplus} \simeq 10^{-6} M_{\odot}$ (**cite!**). However, this cut–off limit changes significantly in various models with substantial consequences for the dark matter predictions on small scales!

self–annihilating dark matter and gamma–ray emission

Despite all the effort for direct detection of dark matter particles, the nature of these particles is still unknown. However, there are recent claims of a double–peak gamma–ray spectrum detections with the Fermi satellite (LAT), both in the center of our galaxy and nearby galaxy clusters (Weniger 2012, Tempel et al. 2012a, Hektor et al, 2013) which is interpreted as a signal of dark matter annihilation. (**consider elaborating more on this and self–annihilating dark matter in general!**)

Apart from gamma–ray hints of DM from the galactic center, the diffuse isotropic gamma–ray background (IGRB) at the new Fermi IGRB measurements could also place constraints on $M_{\min}$ of dark matter halos. "A more detailed analysis of the IGRB, with new Fermi IGRB measurements and modeling of astrophysical backgrounds, may be able to probe values of $M_{\min}$ up to $\sim 1 M_{\odot}$; for the 130 GeV candidate and $\sim 10^{-6} M_{\odot}$; for the light DM candidates. Increasing the substructure content of halos by a reasonable amount would further improve these constraints."

Modified theories of gravity

The relativistic modification of gravity came around after the Newtonian theory of gravity failed to predict correct position of Mercury in its orbit. A few decades later, and not even general relativity could estimate the correct dynamics for the galaxies without including a large amount of a mysterious invisible matter, soon called the "dark matter". Even today, around 80 years after Zwicky first mentioned the discrepancy, we have no clue what dark matter is and if this is the right approach at all, in comparison to further modification in our theory of gravity. Therefor, another appealing idea, suggested by (Milgrom & Bekenstein 1987) was questioning the validity of the Newtonian law of gravity in large distances. This theory which is now know as modified Newtonian dynamics (MOND) has been able to explain a number of observational data without any dark matter. However, there are still a number of observational evidence in favor of non–baryonic dark matter which is inexplicable with MOND.

3 Small–scale ΛCDM vs. observational evidence

Dark matter seems to be dominating the matter content of our Universe. Our understanding of DM is particles interacting only through gravity. Therefore, the simplest way to perform cosmological N–body simulations is to run them with pure dark matter. While DM contribution in the mass budget of the Universe is so much more than the baryonic matter that a pure DM simulations hardly counts as a real simplification, this approach simplifies the problem by far! As much as this happened to be a good solution for getting a handle on how the Universe works on large–scale, we keep facing systematic inconsistencies when it comes to small scales structure. The CDM–based cosmological simulations robustly predict the cosmic structure formation which does not seem to need any baryonic matter inclusion. We tend to assume that "mass follows light" therefore galaxies are considered as tracers of dark matter halos. Besides, since dark matter has –yet– remained unobserved, this is a practical assumption to test our hypotheses. Turns out that the observed cosmic web and filaments are in good agreement with predictions from CDM–only simulations. However, to test the small–scale predictions, one needs to make detailed comparisons between maps of galaxies with certain characteristics and corresponding halos. There are certain aspects to this comparison, from which I will mention those that seem to be fundamentally troublesome.

1. Abundance matching (The missing–satellites problem)

2. Central slope (The core–cusp problem)

3. Normalization (The too–big–to–fail problem)

4. Spatial distribution (The disk of satellites)

3.1 Abundance matching

The first and simplest test for small–scale CDM–based simulations is comparing low–mass end of predicted dark matter mass function with the faint–end of luminosity function of observed galaxies. Connecting the dark halo mass function and dwarf galaxy luminosity function needs a linking assumption which lies at the heart of abundance–matching techniques; *Galactic luminosity is a monotonic function of halo mass.* (Klypin et al. 1999, ?) show that the CDM predictions already fail at this test. Of course, this is an area where observations suffer from different kinds of biases. Therefore, the first potential solution to the mismatch is that the source of discrepancy is practical rather than physical. However, despite the corrections based on detection thresholds and incompleteness, beside further detections of 12 (?) ultrafaint dSphs of the MW in (2005 ?), the mismatch is still persisting within more than an order of magnitude. The more physical explanation could be due to baryonic physics. This is a much more complicated solution to investigate and several mechanisms has been suggested in relation to this solution all of which suppress star fomarion in low–mass CDM subhalos. Some of these mechanisms are following: "inefficient cooling of atomic and molecular gas in halos of virial temperature $T_{vir} \lesssim 10^4$ K ($M_{vir} \lesssim 10^8 M_\odot$), loss of baryons to tidal and ram– pressure stripping and/or to photo–evaporation induced by cosmic reionization (Me: Doeen't Bullock et al. 2010 argue that there is more DM loss in tidal stripping processes than there is baryon loss?), and inefficient accretion of baryons due to supersonic relative velocity between baryons and dark matter shortly after recombination." Whatever the particular mechanism is, the star formation efficiency needs to be less than "normal" in low–mass CDM halos.

3.2 Central slope

One of the long–standing systematic issues of CDM halos is their density profiles. Until recently (?), the main consensus was that the "universal" profile describing simulated dark matter halos was a two–parameter density profile, suggested by NFW 1997, where the central logarithmic density slope changes as $1/r$. The logarithmic slope of this profile gets less steep (?) with increasing r and can be described as $\rho(r) \propto r^{-3}$ around the virial radius. Beside the halo mass, M_{200}, the second parameter is the *concentration parameter* of the halo, $c \equiv r_{200}/r_s$, where r_{200} is the virial radius and r_s indicates the *scale radius,* i.e. where the

logarithmic slope is −2. The concentration parameter describes how the logarithmic slope changes. However, soon the dynamical measurements of stellar and gas content in the central kpc of dwarf galaxies (?) showed that these measurements favor a constant–density for these regions. While further, high–resolution measurements of this kind from various low–surface brightness dwarf galaxies and dark matter–dominated galactic rotation curves support the cored profiles, the higher resolution CDM simulations also tend to indicate more consistency with a three–parameter density profile rather than the traditional NFW ones. The extra parameter describing dark halo density profiles is the shape(?) index, α, which gives the density profile more flexibility in shape (chetori tosif konam ino??). However, still fails to describe observed dark matter halos inhabiting baryonic matter.

On the other hand, even though the density profiles of simulated halos are universal, i.e. independent of the mass scale, the high–mass end of the halo mass function and luminous part of galaxy luminosity function do not indicate any discrepancies regarding the central DM slope. The reason is that there exist several independent methods to derive the luminous and dark matter content of luminous/massive galaxies in their central kpcs. Resolved stellar and gas dynamical measurements probe the baryonic component, and the X–ray temperature maps plus weak and strong gravitational lensing measurements can trace the mass in these galaxies. Therefore, it is easier to decide whether it is the dark matter or the stellar mass dominating the central kpcs of the galaxy. However, the challenging part is subtracting the dynamical mass contribution from stars and deriving the dark matter–only density profile, which needs precise modeling of stellar populations and mass function in the galaxy. "Recent estimates of such combined measurements suggest that the *total* density profiles of relaxed clusters are consistent with cusped CDM halos, while the DM–only contribution is significantly shallower near the center, in cores of ∼ 10 kpc–sized."

3.3 Normalization

Recently, (Boylan–Kolchin, Bullock and Kaplinghat, 2011) pointed out another discrepancy in the structural properties of DM subhalos and dSphs in the local group. While one would naively expect that the most luminous dwarf galaxies correspond to the most massive subhalos, high–resolution cosmological simulations persistently produce massive subhalos ($M_{vir} \geq 10^{10} M_{\odot}$) which are too concentrated in their central kpc to host any of the satellite galaxies around the Milky Way or Andromeda. ([?]). The combination of measured effective (half–light) radius R_e and velocity dispersion σ of a dSph with simple dynamical arguments (chi daghighan? + ref.) can constrain the total mass of the dwarf within R_e. The Upper mass limits derived for the 8 most massive satellites of the MW, (excluding the Magellanic clouds), are systematically smaller than those of the 10 most massive subhalos in high–resolution cosmological simulations, within the same radii (Aquarius/Millennium/LV2 in this case?). Therefore, the most massive subhalos cannot host the most luminous dSphs of the MW. Consequently, unless MW halo is a very peculiar system, the most massive subhalos of the galaxy remain dark while the most luminous dSphs reside in less–massive dark matter halos. One naive explanation of this discrepancy could be that the halo of the MW is less massive than previously estimated and therefore, the local massive dSphs should be compared to simulated subhalos of a less–massive halo. However, "the simplest formulation of the TBTF problem is as following: the mean densities inferred within the central few–hundred pc of the most luminous dSphs are systematically smaller than predicted in CDM simulation." Therefore, the problem is pointing to a more fundamental difference in simulated CDM halos and dSphs than the uncertainty in the MW halo mass.

Another proposed solution to the problem is that one should be cautious about the infall redshift of the halos when make an analogous between the luminosity function of dSphs and the mass function of their inhabit dark halos. READ MORE!

3.4 Spatial distribution

One major concern is the fact all these problems with the CDM paradigm arise on the same scale. (See (Penarrubia et al. 2012))

The "missing satellite" and "core/cusp" problems

Collision–less dark matter simulation, do not contain baryons and therefore the most massive subhalos in these simulations are associated to the most luminous satellites. DM subhalo mass function is divergent on the low–mass scale, meaning for a consistent abundance match between satellite galaxies and DM subhalos, a strong suppression of star formation in low–mass (i.e. $M < 10^{10} M_\odot$) subhalos is required. The abundance of collapsed CDM subhalos increase exponentially toward small scales as $dn/dM_{vir} \propto M_{vir}^{-1.9}$. N–body simulations including pure ΛCDM, assume the gravitational interaction between WIMPs to be dominating the structure formation. As a result, one would expect galaxies to form within dark matter halos distributed in a mass function of the form $dn/dM_{vir} \propto M_{vir}^{-1.9}$, diverging in the low–mass end (Springel et al. 2008). Moreover, these halos form in a universal density profile centrally diverging as $\rho \propto r^{-1}$, while the logarithmic slope decreases in larger radii, i.e. $\rho \propto r^{-3}$ for $r > r_s$ (Navarro et al. 1996). On the other hand, dwarf and low–surface–brightness galaxies show flat mass–density profiles in the center (Oh et al. 2011) as well as a much flatter mass function than the one predicted by the simulations (Klypin et al. 1999, Moore et al. 1999). Either non–CDM models or non–gravitational interactions (like DM annihilation and such?) suggest solutions to these problems. However, there is not yet a single prescription resolving both discrepancies simultaneously. One proposed solution to the core/cusp problem lies in supernova gas flows, transferring energy to the dark matter component and flattening central density profiles. There are a number of proposed solutions to the suppression of galaxy formation in low–mass halos, e.g. "SN feedback, inefficient cooling of ISM, cosmic reionization, UV radiation, and acoustic oscillations" (e.g. Tassis et al. 2008, Bovill & Riccotti 2009, Sawala et al. 2010, Bovy & Dvorkin 2012). However, non has been successful to mark the discrepancy solved, moreover, the provided explanations do not work in coupling the missing satellite problem to the core–cusp issue. According to their model, "the (minimum) star formation efficiency required to form DM cores in dSphs leads to an over–abundance of luminous satellites!". Finally they conclude that the high–z core–formation cannot actually resolves the tension of opposite demands on star–formation efficiency *on the same mass scales*, but only shifts it to lower luminosities.

The problem arises when no M(total halo mass)–L relation can be found in the faint–end of galaxy luminosity function from observations of ultrafaint dSphs. Two possible explanations: The physical one is that the scatter in L at fixed V_{max} is very large and is due to "photoionization or suppression below the atomic cooling limit". In other words (my interpretation), the total baryonic content is so low that once a small population of stars form within the halo they photoionize the rest of the baryons and there is not enough matter for efficient cooling (there only exists atomic cooling processes which doesn't offer sufficient cooling for the star formation cycle to continue), therefore, even though the halo is quite massive, the luminosity remain really low. The other possible explanation, on the other hand, is the product of a selection bias, as "most ultrafaint galaxies inhabit halos with $M_{300} \leq 10^7 M_\odot$, but they are too diffused to have been discovered." CONCLUSIONS AND DISCUSSION ro bekhun! (Bullock et al. 2010)

DISCUSSION ro bekhoon!(Penarrubia et al. 2012) (Penarrubia et al. 2012) make an estimate of the amount of energy required for cusp removal in the case of two massive MW dwarf spheroidal with relatively large observed cores (the dark halo core extends well–beyond the luminous radius of the dwarf!). They rule out the NFW profile for the halos of these galaxies with high confidence levels. They show that to reconstruct the observed large cores in thesw dwarf spheriodals from the cuspy DM profiles of the same M_{vir}, $\sim 10^{53} - - 10^{55}$ erg energy is required. Assuming the SNeII as the unique sorce of exerting such amount of kinematic energy to the gas in the central kpc of the dwarf, the model requires either (i) very efficient (close to 100%) SNeII energy transfer (into kinetic energy of the gas). This value, however, would be in disagreement with the relation between metallicity and luminosity of dSphs which suggests a value of $\sim 0.5\%$ instead (Governato et al. 2010) (ii) star–formation history peaking at very high redshifts (z≥6) (why??) (iii) a very top–heavy IMF, therefore large number of SNe with less efficiency in energy transfer, (iv) "substantial satellite disruption or other stochastic effects to ease the substructure abundance constraints" (what??), otherwise keeps the low–mass end of galaxy mass functions cuspy. These small galaxies lack enough star formation, therefore SN–driven gas flow is insignificant. However, this requires high star formation efficiency in dwarf spheroidals with $M_{vir} \leq 10^{10} M_\odot$ (for them to show "cored" density profiles), while the missing satellite problem (flatter observed galaxy mass function than the predicted halo mass function diverging with $dn/dM_{vir} \propto M_{vir}^{-1.9}$

at the low–mass end), strong suppression of star formation in order for these low–mass halos (on the same mass–scale) to remain dark (Penarrubia et al. 2012). They also argue that cusps must be removed in very high redshifts ($z > 6$) which is inconsistent with stellar ages of the two studied MW dSphs.

In 2010, Bullock et al. wrote a paper on a potentially large population of low–luminosity dwarf galaxies within the halo of the MW. They argue that in case we impose a halo low–mass limit to the galaxy formation, only a handful of such dwarfs are expected in the halo of the galaxy. On the other hand, in the absence of such threshold in galaxy formation, the presence of these dwarfs, which are only different from the known ultrafaint dwarf galaxies is their slightly less–massive DM halos, implies that our common mass scale for MW dwarfs is merely an artifact of selection bias.

Halo density profiles

2–parameter NFW or 3–parameter Einasto ?

3.5 Classical CDM subhalos (NFW)

At $z = 0$, the CDM scenario predicts the existence of dark matter halos with masses ranging from $\sim 10^{15}\ M_\odot$ down to the cutoff in the density fluctuation spectrum, which is set by the detailed properties of the CDM particles. For many types of WIMPs, this cut–off lies somewhere in the range $\sim 10^{-11}$–$10^{-3}\,M_\odot$ [?]. However, alternative models involving superweakly interacting particles (superWIMPS) and MeV mass dark matter may place the cutoff as high as $\sim 10^3$–$10^8\ M_\odot$ [?, ?].

As these low–mass halos merge to form more massive ones, some temporarily survive in the form of subhalos within the larger halos. N–body simulations indicate that the subhalos within a galaxy–sized CDM halo should follow a mass function of the type:

$$\frac{\mathrm{d}N}{\mathrm{d}M_{\mathrm{sub}}} \propto M_{\mathrm{sub}}^{-\alpha}, \tag{3.1}$$

with $\alpha \approx 1.9$ [?, ?]. The relative contribution from subhalos with mass $M \gtrsim 10^5\ M_\odot$ to the dark matter surface mass density at the typical location of macroimages in a galaxy–mass lens is $f_{\mathrm{sub}} \approx 0.002$ [?], albeit with a large scatter [?].

The density profiles of isolated field halos in CDM simulations can be reasonably well described by Navarro, Frenk & White ([?]; hereafter NFW) profiles:

$$\rho(R) = \frac{\rho_{\mathrm{i}}}{(R/R_{\mathrm{S}})(1 + R/R_{\mathrm{S}})^2}, \tag{3.2}$$

where R_{S} is the characteristic scale radius of the halo. The slope of the inner density cusp ($\beta = \mathrm{d}\ln\rho/\mathrm{d}\ln r$) in this profile is $\beta = -1$, and this makes it difficult for NFW halos in the dwarf–galaxy mass range to produce millilensing effects of the type we are considering in this paper. Typically, cusp slopes obeying $\beta \lesssim -1.5$ would be required [?]. Later work has shown that models with inner cusp slopes that become progressively shallower towards the center provide even better fits to isolated halos in CDM simulations, eventually reaching inner slopes of $\beta \geq -1$ [?, e.g.]]Navarro et al. 10. In the context of detecting millilensing effects from low–mass halos, this just makes matters worse, since the central density is reduced.

Traditional universal NFW profile for dark matter halos, is a 2–parameter model which describes the mass–density of the halo by the its mass, and concentration parameter. The halo concentration parameter is defined as $c \equiv r_{vir}/r_s$, where r_s is the scale radius (the radius at which $\rho \propto 1/r^2$). The mass and concentration parameter are correlated, through a shallow slope. Besides, concentration parameter also is correlated with redshift. The is, however, not a single recipe for the detailed dependencies (see Bullock et al. 2011, others? for different opinions!).

Once a halo falls into the potential well of a larger halo and becomes a subhalo, it is stripped of material – primarily from its outskirts – due to interactions with its host halo and with other subhalos. This alters the density profile of the subhalo compared to an isolated halo of the same mass [?, e.g.]]Hayashi et al.,Kazantzidis et al., but these modifications tend to diminish rather than enhance the ability of a CDM subhalo to produce detectable millilensing effects [?].

To demonstrate that standard CDM subhalos do not provide a significant "background" of millilensing events in the observational situation that we consider, we have therefore adopted NFW profiles for these objects, since this results in an overoptimistic estimate on the millilensing effects that standard CDM subhalos are likely to produce. Even then, the chances of detecting millilensing effects from these objects turn out to be negligible in the observational situations we are considering.

To derive the R_S values of our NFW subhalo profiles, we adopt the mass–dependent concentration parameters $c = R_{vir}/R_S$ from either [?] or [?], where R_{vir} is the virial radius of the halo. Since both of these recipes predict higher concentration parameters for low–mass halos, and since more centrally concentrated profiles (i.e. profiles with higher c) are more efficient in producing millilensing effects, we calculate the subhalo concentration parameters based on their current masses rather than the masses they had prior to becoming subhalos. Since nearly all subhalos have lost considerable amounts of material [?, e.g.]]Vale & Ostriker, this also results in overly optimistic millilensing properties.

3.6 Three–parameter alternative CDM halo profile (Einasto)

The more general variation of the NFW profile, gNFW relaxes the central logarithmic slope γ

$$\rho(r) = \frac{\rho_s}{(r/r_s)^\gamma (a + r/r_s)^{3-\gamma}}, \tag{3.3}$$

where $\gamma = 1$ gives the traditional NFW profile and $\gamma = 2$ is equivalent to a singular isothermal sphere (SIS). Another 3–parameter profile is the Einasto profile, the tree dimensional version of the Sersic profile for surface brightness of elliptical galaxies. Both high-resolution measurements of central stellar and gas content of low surface brightness dwarf galaxies, and high–resolution CDM simulations, tend to indicate more consistency with a three–parameter density profile rather than the traditional NFW ones. There are various studies confirming that CDM halos are better described by the Einasto profile, than the NFW one (e.g Dutton & Maccio 2014, others!). The extra parameter describing dark halo density profiles is the shape(?) index, α, which gives the density profile more flexibility in shape, i.e. $\gamma(r) = -dln\rho/dlnr \simeq 1/r$.

Navarro et al. 2004 found the Einasto index for halos in the mass range between dwarves and clusters is between 0.12 and 0.22, with an average value of 0.17. According to Hayashi & White 2008 and Gao et al. 2008, α_E tends to increase with mass and redshift in halos of the Millennium simulation (Springer et al. 2005). From gravitational lensing point of view, Einasto profiles are more demanding to work with as for general α_E values, there is no analytical form for the surface mass density of these halos. Therefore, the lens equation needs to be solved numerically each time.

However, Einasto profile still fails to describe observed dark matter halos inhabiting baryonic matter.

Halo concentration parameter c

Halo concentration is another parameter, in addition to the inner density slope that takes part in telling different halo models apart. The halo concentration parameter is defined as $c \equiv r_{vir}/r_s$, where r_s is the scale radius (the radius at which $\rho \propto 1/r^2$). Halo virial radius as a function of redshift is calculated by (Bullock et al. 2001), while the concentration parameter follows a relation proposed by (Maccio et al. 2007).

A cosmologically–motivated halo density profile is

$$\rho(r) = \frac{\rho_0 r_s^3}{(r_c + r)(r_s + r)^2}, \tag{3.4}$$

where ρ_0 is a characteristic halo density, r_s is a scale radius (the radius at which $\rho \propto 1/r^2$), and r_c is the core radius.

Maximum circular velocity V_{max} and that of the infall epoch V_{infall}

Mass accretion history (MAH)

Issues:

1) Inner density profiles; NFW or the bulk of mass – Central cusps (nature of DM)

SPH simulations by Zolotov et al. (2012) suggest dynamics associated with star formation may flatten cores in more massive dwarfs

2) Shapes as predicted: shapes from lensing (individual clusters or stacked galaxies) – Orbits of the streams in the MW or M31 halos

3) Substructure as predicted: Effects on disks? GCs? Streams? – **Effects on strongly lensed BG object** – Satellite counts; abundances, M_* vs. V_{max} relations

Gaps in the Pal5 star stream (at > 99% confidence) may be induced by > 1000 substructures within 30 kpc with V_{max} > 1 km/s, consistent with ΛCDM predictions, discussed in Carlberg, Grillmair & Hetherington 2013

4) Correspondence of halo/subhalos to galaxies

Galaxies form by cooling and condensation of gas in halo cores, they are associated with subhalos, not halos. Not all subhalos may have a galaxy. Not all galaxies may have a subhalo in a DM–only simulation – Springel et al. (2001)

Subhalo abundance matching is sensitive to numerical resolution (Guo & White 2013)

CDM halos are formed inside–out.

Substructure contributes less than ~ 10% of the total mass.

CDM halo substructure appear to be self–similar. i.e. halos of different mass, properly scales, look alike.

The –scaled– subhalo mass function is independent of the host halo mass (Zheng et al. 2005 – Lravtsov et al. 2004)

Typically, halos have only one subhalo more massive than ~ 3% of the host halo Virial mass. (Wang et al. 2012)

Only 3 MW satellites appear to inhabit halos more massive than V_{max} ~ 30 km/s, while –on average– 10 subhalos more massive than this are present in Aquarius halos (Boylan–Kolchin et al. 2012)

The spatial distribution of subhalos is independent of halo mass. Most subhalos populate the outer regions os the halo.

The shape of the mass profile of the DM halos is roughly independent of halo mass and cosmological parameters.

Density profiles are **cuspy** and clearly differ from power laws.

Usually parameterized by a mass, M_{200}, and concentration, $c \equiv r_{200}/r_s$

Models to predict the mass and redshift dependence of the concentration (eg. Bullokc et al 2001, Eke et al 2001, Neto et al 2007, Gao et al. 2008)

The mass profile of CDM halos deviates slightly but systematically from universality in very high–resolution simulations. Therefore, a third parameter seems needed to describe the mass profile in detail, as in the Einasto profile ($\alpha = 0.17$ resembles the NFW profile). (Navarro et al. 2004)

Based on the Millennium simulations: The concentration depends systematically but **weakly** on halo mass. This relation, and its dependence on redshift have proved very difficult to reproduce. Models that use the halo accretion history rather than a formation redshift seem to work best (Wechsler 2002, Zhao et al. 2003). At a given mass, average halos are NFS–like, outliers are not. (Ludlow et al. 2013)

The origin of the concentration and shape parameters: The concentration measures the characteristic density of a halo ans is assumed to trace the formation redshift of the halo, although the definition of the latter has been unclear! Mass profile and accretion history are closely related (Lu et al. 2006, Wu et al. 2013)

The shape of the mass profile resembles that of the accretion history. Both are well approximated by the NFW shape. Similarity of mass profiles is due to similarity of mass accretion histories.

Generally, there is good agreement between the mass profiles of galaxy clusters and those of CDM simulations. Only some tension remains regarding the concentration of strong–lensing clusters AND the slope of the inner density cusp. The slightly higher concentration of strong–lensing clusters is thought to reflect projection effects of highly aspherical clusters.

cusp vs. core problems highlights for dwarf galaxies remain inconclusive. Do we really need baryons to modify substantially the dark matter halo??

Dark matter on scales of < 1 kpc: Is there CDM substructure? Methods: flux ratio and surface brightness anomalies In a galaxy–galaxy lensing system, the Einstein radius of the system is much larger than can be

explained by stars alone for any reasonable IMF, lens or dynamical model.

Dark matter on scales of 1–10 kpc: What is the DM fraction inside R_{eff}, R_{Ein}? How is DM distributed (slope/concentration)? Can DM explained by dark stars? IMF??

DM on scales of 10–100 kpc: Extended DM halos through combined strong and weak lensing

Proposed methods to detect such substructures around th eMW rely on gamma–ray emission from the annihilation of dark matter, or on the gravitational scattering of stars in the tidal streams of satellite galaxies.

There is this new method proposed by Feldmann & Spoylar on 2013, based on the substructures gravitational pull on stars of the MW disk.

The number density of dark matter halos encodes invaluable information about the primordial power spectrum, the physics of early universe and the nature of dark matter. (Moore et al. 1999, Bullock, Kravtsov & Weinberg 2000, Kuhlen, Vogelsberger & Angulo 2012)

in WDM models, a competitor of the CDM paradigm, structure formation is suppressed below the free–streaming scale of the dark matter particle, resulting in a deficit in substructure with masses below $\sim 10^9 M_\odot$ (Zentner & Bullock 2003)

A spatially unresolved measurement of the flux density from a lensed source is a factor of order μ brighter than for an unlensed source, while spatially resolved measurements can provide a factor of $\sqrt(\mu)$ higher resolution. (Schneider et al. 1992)

4 Gravitational Lensing

"Dedicated optical and radio surveys to find and study strong galaxy–galaxy lens systems: Sloan Lens ACS survey (Bolton et al. 2006), SQLS (Oguri et al. 2006), Strong Lens Legacy Survey (More et al. 2012; Ruff et al. 2011), and the BOSS Emission Line Lens Survey (Brownstein et al. 2012) in the optical and the Cosmic Lens All Sky Survey (CLASS; Browne et al. 2003)." (From Wardlow et al. 2013)

5 Radio interferometry

In this report, I briefly explain the main principles based upon which radio interferometers work. I, also, try to put radio interferometry into the context of radio astronomy as well as observational astronomy, in general, by comparing radio interferometry with optical interferometry and the advantages and disadvantages of observations in either frequency regime with respect to the other.

5.1 Radio telescopes

A radio telescope is usually made of two main parts:

- Reflector: The special shape of this reflector, i.e. a parabola, is advantageous compared to convex mirrors used in reflecting optical telescopes as the parabolic shape keeps all reflected radiation in phase and as I will repeatedly mention in the report, coherence is of great importance in radio astronomy. Keeping the same principle in mind, current radio instruments use aperture arrays keeping the electromagnetic wave coherent at the receiver by exerting appropriate delays to the signal.

- Receiver: Unlike optical telescopes which scan different points of the focal plain with numerous receivers (pixels of a CCD), radio telescopes generally work with a single receiver at the focal point of their reflector. In other words, radio telescopes see the entire sky within a single pixel only!

  Heterodyne receivers are the most common kind of receivers used in radio antennae today, and although in principle they could be arranged in 2-D arrays on the focal plane of the antenna to perform as a CCD in an optical system, the cross-talk among closely placed receivers make the calibration of each receiver cumbersome and usually inefficient. A more practical approach would be to use several parabolic reflectors (aperture arrays) simultaneously to observe several pointing on the sky.

Heterodyne receivers are the most common type of receivers at sub–mm wavelengths (and higher) and they are crucial to use if the signal modulation (both amplitude and phase of the signal) is needed for observations (the case in radio interferometers)

In order to digitally record the electromagnetic wavefront reaching the receiver, the signal needs to be sampled with a frequency at least twice the frequency of the wave (Nyquist theorem). This is an electronically challenging task, and particularly impossible for frequencies in the order of 100 GHz and higher. Therefore, the arriving signal is first mixed with the signal from a stable local oscillator (LO) with a well–known frequency slightly lower than that of the original signal. The mixed signal then has two frequencies $f_{sky} + f_{LO}$ and $f_{sky} - f_{LO}$. Therefore, we can sample the low–frequency signal with as fine resolution as desired, ignoring the higher frequency. The signal is then demixed by exerting a phase difference and is recorded in two frequency windows below and above the LO frequency, namely the *lower sideband* and the *upper sideband*, by the receiver.

5.2 Observational limitations

Observational limitations of the recorded data may come from the source or, more commonly, from our instruments and can be classified into limitation in

- resolution

- sensitivity

- dynamic range

As will be discussed further into the report, radio interferometry has brought significant improvements in the former two, although mostly at the cost of the latter.

The convolution theorem

As previously mentioned, what a radio telescope records is not the direct image of the source at different points at the same time, a radio telescope rather records signals from the entire sky into a single "pixel" and so does an array of radio antennae. Radio interferometry is based on *the convolution theorem*. According to this theorem, it can be shown that convolving two functions in Fourier space (i.e. Fourier transform of their convolution) is equivalent to multiplying the Fourier transform of the two in the ordinary space.

$$\mathscr{F}(f * g) = \mathscr{F}(f) \times \mathscr{F}(g) \tag{5.1}$$

Where convolution of one function $f(t)$ with another function $g(t)$ means shifting the function g consecutively from $t = -\infty$ to $t = +\infty$, while at each step the multiplication of the two functions (i.e. $f(t)g(t+\tau)$) is added to the convolved function $f * g$:

$$(f * g)(\tau) = \int_{-\infty}^{+\infty} f(t)g(t+\tau)dt \tag{5.2}$$

No device can achieve infinite resolution due to its diffraction limit, which results in spreading the light from a point source into its surrounding pixels, aka. point spread function (PSF) of the instrument. Therefore, anything observed with that instrument will be seen as the convolution of the true source structure with the PSF. Angular resolution of an instrument (i.e. FWHM of its PSF) scales with the size of the aperture and wavelength as $\Delta\theta \propto \lambda/D$. Therefore, the PSF itself scales with the square of the Fourier transform of the aperture which, according to the convolution theorem, equals consequently the Fourier transform of the autocorrelation of the aperture. Therefore, in order to increase the resolving power of an instrument, one needs to either increase the aperture size or decrease the observing wavelength. While the former solution might not be possible to arbitrary large extent due to different physical emission mechanisms, the former is in the heart of using the interferometry technique.

5.3 Two-element interferometer

When we talk about interferometry, clarification is needed to distinguish between the common picture of Young's optical double–slit interferometer and the concepts and techniques used in radio interferometers. In the optical case, the fring pattern made on the detector plane is a distribution of points with various intensities $I(x) = |E_1(x) + E_2(x)|^2$, and visibility is a real quantity between 0 and 1 defined as below:

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \tag{5.3}$$

Visibility in the case of radio interferometry, however, is a complex quantity describing the time correlation of the two signals reaching the two elements of the interferometer and can be expressed as below:

$$V(\tau) = < E_1(t)E_2^*(t+\tau) > \tag{5.4}$$

with τ being the geometrical time delay between the two signals due to the different paths the wave needs to reach each antenna.

$$\tau = \frac{\vec{B}.\vec{S}}{c} \tag{5.5}$$

Where $\vec{B}$ is the three-dimensional distance vector between the two elements, aka. *baseline*, $\vec{S}$ is the source vector and c is the speed of light.

Practically, the time delay τ of one signal with respect to the other is added to the signal electronically to correct for this effect before sending the two signals to the correlator. Applying the geometrical time delay τ is equivalent to projecting the baseline connecting the two telescopes on the plane perpendicular to the direction of the source. Therefore, what matters in all math hereafter is the *projected* baseline between the two antenna, rather than the actual baseline.

The correlated signal is a complex visibility describing the response of the interferometer to the two signals, both amplitude and phase, received from a point source which is a fringe pattern (on the sky) perpendicular to the projected baseline (BL_{proj}) as seen from that point on the sky. What happens in radio interferometry is that we sample the visibility values on this plane depending on the position of the source, BL_{proj}, and duration of observations. The interferometer response to an extended source is simply the superposition of the response patterns of various point sources making up the extended source. In other words, the measured visibility is the integral of source structure multiplied by the response. Therefore,

$$V_{\mathrm{observed}} = \int V_{\mathrm{response}}^{\mathrm{point\ source}} I(x,y)dxdy. \tag{5.6}$$

Where x and y are celestial coordinates and $I(x,y)$ is the intensity distribution of the extended source.

The van Cittert-Zernike theorem and the interferometer equation

As long as the source we are observing is far enough that the wavefronts from the source can be approximated as plane waves (i.e. Fraunhofer approximation applies), the complex visibility of the source only depends on the relative position of the two observed points, and the wavelength the observations are made at The visibility function then is a measure of the spatial coherence of the wavefront of the source encoding the source structure in it.

$$V^{\mathrm{AB}} = <E_A \times E_B^*> = \mathscr{F}(\frac{\vec{R}_A - \vec{R}_B}{\lambda}) \tag{5.7}$$

The left hand side is, by definition, the complex visibility of the source, $\vec{R}_A - \vec{R}_B$ is the observing baseline, and $\mathscr{F}$ is the Fourier transform of the source structure $I(x,y)$. Therefore, rewriting **(from what?)** the interferometer equation, one finds:

$$V(x,y,z) = \int_{x,y} I(x,y)e^{-\frac{2\pi j}{\lambda}(ux+vy+wz)} \frac{dxdy}{z} \tag{5.8}$$

where $z = \sqrt{1-x^2-y^2}$. If x and y are small enough so that the observing field can be approximated as flat, the wz term becomes zero and the mere geometrical delay correction means that both receivers are probing the same coherent wavefront. This interferometry equation transforms the intensities on the source plane to visibilities on the image plane and their values depend on the coordinates of the observing baseline, projected on the source plane, as well as the structure of the source and the observing frequency.

5.4 Aperture synthesis

Back to the definition of visibility in radio interferometry, one can directly conclude that a multi-element interferometer is only a combination of several two–element interferometers, sampling the (u,v) space in pairs, based on their projected baseline and the observing frequency, measuring the visibility at those sample points only.

Snapshot observations

Fourier transform is blind to the absolute baseline position, but as baseline is a vector, it does carry information about the direction of the baseline. This implies that in a multi–element interferometer, a (arbitrary) frame of reference is needed for combining the antenna pairs and calculating the transformation from the ordinary plane to the (u,v) plane. The (u,v) coverage in a snapshot is equal to the autocorrelation of the array configuration as seen by the source –except for the central point which corresponds to the measurement with the hypothetical $BL = 0$, impossible to implement in practice.

Since Fourier transform is a Hermitian transformation, a single baseline corresponds to sampling not one, but two points on the visibility plane; one at (u,v), and the other at (-u,-v). Moreover, if I(x,y) is a real number, $V(u,v) = V^*(-u,-v)$. This means that each baseline gives two separate visibility measurements.

This means that adding even one more antenna to the array has a significant effect on better sampling the (u,v) plane. One through the increase in the number of baselines ($N_{\mathrm{BL}} = \frac{N_{\mathrm{ant}}(N_{\mathrm{ant}}-1)}{2}$), and one through the double–sampling of each baseline as mentioned above. Therefore, a monofrequency snapshot made by an interferometer with N_{ant} antennae makes $2N_{\mathrm{BL}}$ discrete visibility measurements where longer baselines in the (u,v) plane correspond to smaller scales on the image plane and vice versa. Besides, as seen in equation 5.8, the conversion between the (x,y) plane and (u,v) plane is scaled by the observing frequency. Therefore, a wider observing bandwidth would result in a wider radial coverage of the (u,v) plane. (As (u,v) coordinates are measured in wavelength, one way of filling the Fourier space is to observe at multiple wavelengths simultaneously)

The other way to upgrade the discrete sampling of (u,v) space to –closer– continuous sampling (in the tangential direction), is longer exposure time. In this mode, the Earth rotation would lead to (continuous) change in projected baselines of the array as seen by the source and hence more (u,v) coverage.

Two extreme regions of the (u,v) plane are usually the most challenging to fill. Long baselines (corresponding to small scales on the image plane) are difficult to achieve for obvious reasons and therefore the longest baseline in the array sets the angular resolution of the reconstructed image and obviously the impossibility of infinite baselines make the infinite resolution also impossible! On the other hand, short spacings on the (u,v) plane could also be challenging to achieve because of the minimum spacing forced by antenna sizes. This leads to a fundamental feature of any (u,v) sampling known as "zero-spacing hole". The lack of visibility measurements on this scale corresponds to missing out on any possible (smooth) large-scale structure on the source such as diffused emission. In cases where an unrecovered diffused component does exist in the original source, the integrated flux of the source measured using the interferometer would be smaller than the integrated flux of the target measured by a single-dish telescope. This is a tell–tale signature of the presence of a large unrecovered structure in the source and is known as "the missing flux" problem. The problem arises due to the poor (or lack of) sampling the short spacings on the (u,v) plane. Therefore, the intensity profile of an internally Gaussian source, lacks data points in the center thus visibility interpolation tends to flatten out the peak. In many cases, sampling the short spacings with a more compact array or a single–dish telescope helps constrain a better fit to the light profile of the source.

Aperture synthesis and image reconstruction

Aperture synthesis is the technique of reconstructing an image from incomplete sampling of the Fourier space. The technique is based on modeling the visibility plane based on the existing samplings. Therefore, the more complete the (u,v) coverage, the closer the model image to the real one seen by the antennae. However, there will always be holes in the sampled (u,v) plane limiting the fidelity and quality (i.e. dynamic range $= \frac{\mathrm{peak}}{\mathrm{RMS}}$ of the reconstructed image.)

Fourier transform of the (u,v) coverage gives the so-called "dirty beam" (equivalent to the point spread function, PSF), while the Fourier transform of the measured visibilities on the (u,v) plane results in the "dirty image" of the source, which is, as expected, the actual source structure convolved with the dirty beam. Now that we have images of both the beam and the source, we can deconvolve the beam from the image. Remember that deconvolution is non-linear and non-unique! (See section 5.4 CLEAN algorithm)

Despite what it looks like, modeling is actually applied while working in the image plane. The process is such that the dirty beam and dirty image are calculated based on the (u,v) coverage and direct visibility measurements, respectively. For this purpose, the (u,v) plane is required to be gridded[1]. At this step, Nyquist sampling of the (u,v) plane introduces a limitation in the image space, i.e. grid size/cell size needs to be chosen such that the final PSF includes at least 3 pixels in the image plane. The next step in image reconstruction is deconvolving the beam from the dirty image. The actual interpolation happens here! Deconvolution is not a unique process, hence requires not only griding the (u,v) plane but also making assumptions about the deconvolving beam. One assumption need to be made is how to weight the measured visibilities. This question, in turn, has two parts: How are the measured visibilities in one (u,v) pixel weighted? and what

---

[1] In fact, this is the issue one faces when using *Fast* Fourier transform (FFT), while *direct* Fourier transform does not require griding the (u,v) plane. One would expect the resulting image to be nearly the same as the model obtained by natural weighting. However, direct FT is so much slower that it is not worth practicing as far as the visibilities are sampled wisely!

is the relative weight of various pixels on the (u,v) plane? Different choices at this step result in different angular resolution and dynamic range limitations in the model image.

Natural weighting as it sounds is simply weighting visibilities inversely proportional to the variance of their distribution. Therefore, short–spacings in the (u,v) plane , with closer samplings, are weighted more than longer baselines which results in higher dynamic range in the modeled image at the cost of a coarser angular resolution. In other words, the dirty beam has larger FWHM, but less prominent side lobes. Therefore, it is best for imaging faint and smooth sources.

Uniform weighting is the same as natural weighting with an additional factor inversely proportional to the number of visibilities in each pixel. Therefore, it boosts the contribution from long–spacings by giving the same weight to visibilities at all (u,v) distances, even though those at long distances are sampled more coarsely. This property makes this approach more suitable for sources with rich and complicated structure. Although, the source needs to be bright enough for this approach to work best. This is because, the described weighting scheme results in a narrow(er than naturally–weighted) PSF, with smaller side lobes and therefore, better angular resolution. However, the RMS of the residuals of this model tends to be large(r than that of natural by at least an order of ~ 2), i.e. poorer sensitivity.

Briggs weighting is a tunable combination of both approaches above. The tunability of Briggs' recipe is applied through the *robust* parameter. This parameter changes in the range of -2 to +2 from uniform to natural weighting. Usually, robust=0.5 gives a good compromise between the two extreme approaches. Although it still depends on the preferences based on the scientific target and come on, this is part of the charm of the method!

(u,v) tapering is useful when the observed source is extended and most of the information lies in short baselines. Therefore, in order to improve the SNR, we can overweight the measurements made using these baselines even further and decrease the weight of measurements where we have a lower signal, i.e. long baselines. This can be done via Gaussian circular tapers in the (u,v) plane. If the source has a very extended structure, which is remained unrecovered due to the short–spacing hole in the (u,v) coverage, even with natural weighting, then (u,v) tapering is the last resort to recover that diffused emission.
As a matter of fact, one might want to get rid of an extended emission they are not interested in, and focus more on the small–scale clumpy structure of their source. In which case, one only needs to apply an inverse Gaussian taper to the (u,v) plane.

Deconvolving the dirty beam and CLEAN algorithm

While interferometry data are incomplete due to lack of information on –most of– spatial frequencies, we can interpolate the (u,v) plane by applying our prior information about the source structure as the criteria required to proceed with deconvolution. Clearly, the more our prior information about the source, the more reliable the model image is.
The problem of deconvolving dirty beam from the dirty image has non–unique solutions. An intuitive way to show it is assuming a source with non–zero visibilities everywhere but where we are making our measurements, i.e. our (u,v) coverage. This class of sources remain invisible to the interferometer, as long as the (u,v) coverage is the same. Now, one can add *any* combination of these "invisible" sources to our reconstructed image without the measurement set changing. Improving the (u,v) coverage will decrease the number of these combinations and therefore, increase the image fidelity, and yet *high* fidelity is not the same as *infinite* fidelity. Therefore, the reconstructed image always remains a model of our real data. I know that I have brought this up several times by now but this is in the heart of the subject and I cannot emphasize it too much!
CLEAN is the oldest, fastest and best deconvolution algorithm practiced to day for image reconstruction in radio interferometry :)) Talking very briefly (and naively), what this algorithm does is to start at the maximum flux point in the image, subtract the PSF at that point and add a 2-D Gaussian component (with the

same FWHM of the PSF and scaled flux) to the CLEAN image instead then repeat these three steps for the peak in the residual map until either of the criteria below is met:

- Noise limit: If the peak in the residual map has a smaller value than $\alpha\sigma_{\mathrm{RMS}}$

- Dynamic range limit: If the peak in the residual map has a smaller value than $\frac{1}{\alpha} \times f_{\max}$. of the dirty beam
  (α is an arbitrary coefficient set by the scientific purpose of the image.)

- Maximum number of CLEAN cycles: This is an arbitrary criterion!

Finally, what the algorithm produces are three matrices below:

CLEAN model: A set of point sources subtracted from the dirty image, where flux peaks located.

CLEAN beam: The corresponding Gaussian to the dirty beam. In other words, the PSF model. If the dirty beam is elliptical, then the CLEAN beam is also a similarly–elliptical Gaussian.

CLEAN image: It is simply the CLEAN model convolved with the CLEAN beam and is the final product of the image reconstruction procedure.
While different deconvolution algorithms apply different criteria to the source structure, the CLEAN algorithm's basic assumption is that the source is made of discrete Dirac delta functions, i.e. point sources. Therefore, it tends to convert smooth and diffused components into clumpy ones. While one does not loose real emission by not CLEANing it, it is possible to get artificial features by CLEANing noise!
The current commonly–used implementation of CLEAN is the Cotton–Schwab CLEAN, consisting two same–level, linked cycles. Start with a **Minor cycle** as below:

1. Locate the peak intensity in the dirty image.

2. a) Subtract a dirty beam centering at the point of the peak, from the dirty image.

   b) Compute the residual image = image used in step (1) – product of (2.a).

   c) Scale down the flux density in residual image to **g** times the image peak, where **g** is the arbitrary CLEAN gain value.

3. a) Add a point–like component to the CLEAN model, at the position of the peak.

   b) Set the flux density of the point in CLEAN model to the peak of the PSF.

   c) Convolve the newly–added point of the CLEAN model with the CLEAN beam and add it to the CLEAN image

4. Use the product of step (2.b) in step (1).

A **Major cycle**, then, follows the minor cycle as below:

1. Compute model visibilities corresponding to the CLEAN model made in the minor cycle.

2. Compute the residual visibilities = observed visibilities – model visibilities made in step(1).

3. FFT the residual visibilities and feed it to the next minor cycle

While the CLEANing process seems complete within the minor cycle, the major cycle is needed because the algorithm (minor cycle only) cannot subtract anything from marginal parts of the image, where we have no information about possible sources in those regions whose side lobes (only!) affect our dirty image.

6 Sources: quasars vs. SMGs

6.1 Quasars

6.2 sub–mm galaxies (8/28)

SMGs are a population of high–redshift (redshift distribution peaking at $z = 2.2$ and $\sigma_z = 1.3$ (see e.g. Chapman et al. 2003, & 2005 as well as Figure 3 Swinbank et al. 2006)), massive, star–forming and gas–rich galaxies (Ivison et al. 2000, Smail et al. 2002). They show properties expected from progenitors of luminous Ellipticals, although other ideas such as SMGs being dusty QSOs or even more mundane galaxies are also out there. Besides, Swinbank et al. 2006 show that SMGs could explain the formation of the brightest local ellipticals although semi–analytical galaxy formation models (e.g. Durham GALFORM+GRASIL dust modeling code) did not predict SMGs. Swinbank et al. 2006's model of SMG produce the radio emission via two mechanisms: (i) thermal bremsstrahlung from H_{II} regions and (ii) synchrotron radiation powered by acceleration of relativistic electrons in SN remnants. Figure 2 in Swinbank et al. 2006 indicates the cumulative number counts/deg^2 of observed SMGs with various surveys with SCUBA on the JCMT at 850 μm as a function of sub–mm flux density, overlayed with the model predictions. As confirmed by the observational data, their model suggests that the surface density of SMGs with radio flux densities above 30 μJy is 75%.

Swinbank et al. 2008 show that "the predicted far–IR properties of model galaxies with $S_{850} > 5$ mJy and $S_{1.4} > 30$ μJy are in good agreement with observations." They also show that "the predicted redshift distribution of sub–mm galaxies provides a reasonable fit to the observational data with a median redshift $z = 2.0$. The radio selected subset of sub–mm galaxies are predicted to make up approximately 75% of the population and peak at $z = 1.7$, in a reasonable agreement with the observed radio detected fraction and redshift distribution."

Typical masses of SMGs are $\sim 4 \times 10^{11} M_\odot$ within $\sim 20 kpc$. Gas fractions are $> 25\%$ within central ~ 4 kpc, stars contribute another 25–50% (depending upon M/L)

Space density of SMGs is $5 \times 10^{-5} Mpc^{-3}$ (Swinbank et al. 2008)

The SMGs used in this work follow Swinbank et al. 2008 estimation of $S_{850\mu m=352.7GHz} = 5 mJy$ (I adopt the unlensed source to have $S_{336.5GHz} = 5 mJy$) at $z = 2$. As Swinbank et al. 2008 (and probably other, but cite!) show, sizes of these galaxies are "poorly constrained and usually taken to be 4-8 kpc (which is approximately the size seen in HST observations and in resolved spectroscopic imaging)"

Looking at the rest frame spectral energy distribution of galaxies in a broad range, one expects a peak in flux density at ~3000 GHz, lying within ALMA band 10 receiver's coverage for a galaxy at $z = 2$. The source then is expected to appear ~ 4 times brighter in band 9, than in band 7. This, however, is not taken into account. (add the plot with the effect of galaxy flux density!)

The "unusual (negative) sub–mm K–correction": Beyond a redshift of $\simeq 1$, as the redshift of a dusty galaxy increases, its characteristic spectral energy distribution ensures that the effect of the increasing luminosity–distance on its observed flux density is balanced by the increasing rest–frame luminosity of the galaxy. The consequence is that the galaxy's flux density is approximately independent of redshift.