Populating the Black Hole Mass Gaps in Stellar Clusters: General Relations and Upper Limits

Johan Samsing1© and Kenta Hotokezaka2©

1 Niels Bohr International Academy, The Niels Bohr Institute, Blegdamsvej 17, DK-2100 Copenhagen, Denmark; jsamsing@gmail.com
2 Research Center for the Early Universe, Graduate School of Science, University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan

Abstract

Theory and observations suggest that single-star evolution is not able to produce black holes with masses in the range 3–5$M_\odot$, and above $\sim 45M_\odot$, referred to as the lower mass gap and the upper mass gap, respectively. However, it is possible to form black holes in these gaps through mergers of compact objects in, e.g., dense clusters. This implies that if binary mergers are observed in gravitational waves with at least one mass-gap object, then either clusters are effective in assembling binary mergers, or our single-star models have to be revised. Understanding how effective clusters are at populating both mass gaps have therefore major implications for both stellar and gravitational wave astrophysics. In this paper we present a systematic study of how efficient stellar clusters are at populating both mass gaps through in-cluster mergers. For this, we derive a set of closed form relations for describing the evolution of compact object binaries undergoing dynamical interactions and mergers inside their cluster. By considering both static and time-evolving populations, we find in particular that globular clusters are clearly inefficient at populating the lower mass gap in contrast to the upper mass gap. We further describe how these results relate to the characteristic mass, time, and length scales associated with the problem.

Unified Astronomy Thesaurus concepts: Black holes (162); Gravitational wave astronomy (675); Stellar dynamics (1596); Neutron stars (1108)

1. Introduction

Several gravitational wave (GW) sources have now been observed by the LIGO and Virgo GW observatories, including both binary black holes (BBHs; Abbott et al. 2016a, 2016b, 2016c, 2017a, 2017b; Zackay et al. 2019; Venemadhav et al. 2020), and binary neutron stars (BNSs; Abbott et al. 2017c). Their astrophysical origin is still unknown, but several formation channels have been suggested. Some of the recently proposed ones include: field binaries (Dominik et al. 2012, 2013, 2015; Kinugawa et al. 2014; Belczynski et al. 2016b, 2016a; Silsbee & Tremaine 2017; Murguia-Berthier et al. 2017; Rodriguez & Antonini 2018; Schröder et al. 2018), dense stellar clusters (Portegies Zwart & McMillan 2000; Lee et al. 2010; Banerjee et al. 2010; Tanikawa 2013; Bae et al. 2014; Rodriguez et al. 2015; Ramirez-Ruiz et al. 2015; Rodriguez et al. 2016a, 2016b; Askar et al. 2017; Park et al. 2017; Samsing 2018; Samsing & D’Orazio 2018; Samsing et al. 2020; Kremer et al. 2020b), active galactic nucleus (AGN) disks (Bartos et al. 2017; Stone et al. 2017; McKernan et al. 2018; Tagawa et al. 2020), galactic nuclei (GNs; O’Leary et al. 2009; Hong & Lee 2015; VanLandingham et al. 2016; Antonini & Rasio 2016; Stephan et al. 2016; Hoang et al. 2018; Hamers et al. 2018), very massive stellar mergers (Loeb 2016; Woosley 2016; Janiuk et al. 2017; D’Orazio & Loeb 2018), and single-single GW captures of primordial black holes (Bird et al. 2016; Cholis et al. 2016; Sasaki et al. 2016; Carr et al. 2016).

The question is, which of these proposed merger channels dominate the merger rate? Are several channels operating with a possible dependence on redshift? Or are the majority of GW sources formed through a still unknown mechanism? Several studies show that one can distinguish at least classes of channels, such as isolated binaries and dynamically induced mergers, by considering the observed distribution of merger masses (Zevin et al. 2017) or the relative spin orientation of the merging objects (Rodriguez et al. 2016c), as well as the orbital eccentricity at some reference GW frequency (Gültekin et al. 2006; Samsing et al. 2014; Samsing & Ramirez-Ruiz 2017; Samsing & Ilin 2018; Samsing et al. 2018b; Samsing 2018; Samsing et al. 2018a; Samsing & D’Orazio 2018; Rodriguez et al. 2018; Zevin et al. 2019; Samsing et al. 2019a, 2020). Other “indirect” probes have also been suggested, such as stellar tidal disruptions (e.g., Samsing et al. 2019b; Lopez et al. 2019; Kremer et al. 2019b). In this picture, it is now largely believed that dynamically assembled mergers are likely to have mass ratios near one (e.g., Rodriguez et al. 2018), random relative spin orientations (e.g., Rodriguez et al. 2016c), and a nonnegligible fraction of mergers with measurable eccentricity in LISA (Samsing & D’Orazio 2018; Kremer et al. 2019c), DECIGO/Tian-Qin (e.g., Chen & Amaro-Seoane 2017; Samsing et al. 2020), and LIGO (Samsing 2018). This is in contrast to isolated binary mergers, which likely have correlated spins (e.g., Kalogera 2000), a bimodal distribution for the effective spin parameter (Zaldarriaga et al. 2018; Hotokezaka & Piran 2017; Piran & Piran 2020), larger mass ratios, and which merge on orbits with eccentricities indistinguishable from ~0 near LISA and LIGO. This picture is rather clean when comparing mergers forming in highly dynamical systems, such as globular clusters (GCs) and GNs, to completely isolated field binary mergers; however, it becomes less clean when considering, e.g., the proposed subpopulation of field binaries that undergo secular interactions with nearby single or binary objects (e.g., Naoz et al. 2013; Naoz 2016; Toonen et al. 2016; Antonini et al. 2017; Silsbee & Tremaine 2017; Liu & Lai 2018; Rodriguez & Antonini 2018; Randall & Xianyu 2018a; Antonini et al. 2018; Liu & Lai 2019; Fragione & Loeb 2019; Fragione & Kocsis 2019; Hamers & Thompson 2019; Safarzadeh et al. 2020). In this case, secular exchanges of especially angular momentum can drive the
binary to merge with random spin orientations (e.g., Liu & Lai 2017) and notable eccentricities (e.g., Randall & Xianyu 2018b; Liu et al. 2019; Fragnione & Kocsis 2020), which makes it more challenging to disentangle cluster mergers from field binary mergers.

An additional outcome that is somewhat unique to dynamical environments is the formation of so-called hierarchical mergers (e.g., O’Leary et al. 2016; Fishbach et al. 2017; Gerosa & Berti 2017; Yang et al. 2019; Antonini et al. 2019; Gerosa & Berti 2019; Samsing & Ilan 2019; Rodriguez et al. 2019; Gerosa et al. 2020; Safarzadeh et al. 2020; Gayathri et al. 2020; Kimball et al. 2020; Doctor et al. 2020; Baibhav et al. 2020). Here, the picture is that compact objects (COs) that merge inside their cluster through, e.g., single-single GW captures (e.g., Samsing et al. 2020) or through chaotic few-body interactions (e.g., Samsing et al. 2014; Zevin et al. 2019) will form a new population of “second-generation” (2G) objects that are characterized by having a higher mass than the original “first-generation” (1G) population and a dimensional spin parameter around 0.7 (e.g., Berti et al. 2007). This 2G population can undergo further interactions leading to mergers with other 1G or 2G objects, which then naturally will lead to an observable modified BH mass spectrum and spin distribution. This process can in principle also lead to 3G-, 4G-, ..., NG populations, which naturally gives rise to unique observables. Looking for such hierarchical merger configurations has been proposed to be one way of probing the origin of GW mergers in very dense systems, such as GCs (e.g., Samsing et al. 2019), GNs (Antonini & Rasio 2016), and AGN disks (Yang et al. 2019). However, fine-tuned few-body configurations in the binary field population can in principle also create hierarchical mergers (e.g., Safarzadeh et al. 2020), but in this case it is highly unlikely to go beyond 2G. In any case, an observation of a hierarchical merger would strongly indicate that at least some GW sources are assembled as a result of few-body interactions.

Another interesting consequence of the hierarchical merger scenario is the possibility of populating the so-called lower mass gap (LMG) and upper mass gap (UMG), where the LMG is ~3–5M⊙ (Bailyn et al. 1998; Özel et al. 2010; Farr et al. 2011) and the UMG is marked by a lower limit of ~45M⊙ (Woosley 2017; Leung et al. 2019; Farmer et al. 2019). For example, the LMG might be populated through BNS collisions, while the UMG can be populated by BH mergers. This makes it possible for dense clusters to produce GW sources with objects in either the LMG or the UMG. If “nature” is not able to form BHs through single-star evolution in these mass gaps, then an observation of GW sources with a mass-gap object will give us insight into the fraction of mergers assembled in clusters, or at least dynamically. These mass gaps not only play a key role in stellar astrophysics, but introduce also a characteristic mass scale that can be used to, e.g., constrain the cosmological parameters (Farr et al. 2019).

Several recent studies have discussed the possibility of populating the mass gap in clusters (e.g., Rodriguez et al. 2019; Di Carlo et al. 2020; Doctor et al. 2020; Baibhav et al. 2020; Rodriguez et al. 2020; Banerjee et al. 2021). Currently, numerical studies suggest that BNS mergers are not likely to form in systems such as GCs (e.g., Ye et al. 2020). By contrast, recent observations of the orbital parameters of galactic BNSs interestingly indicate that BNSs might actually form in clusters at rates several orders of magnitude higher than suggested by the numerical studies (e.g., Andrews & Mandel 2019), which of course poses some interesting tension. Regarding BBHs, several studies have found that if the initial BH spins are low, then up to ~10% of BBH mergers from GCs could be in the form of 1G–2G binaries, with a subtraction of these being in the UMG (Rodriguez et al. 2019). To find the observable contribution from such hierarchical mergers in upcoming future GW data, several numerical techniques and models are now under development (e.g., Doctor et al. 2020); however, what is common for the majority of these models is that they are not linked to any real physical model, they are instead just generic functional forms with a few fitting parameters. This kind of model-independent approach might be useful to condense a huge stream of data into just a few fitting parameters, but gives a priori no astrophysical insight into which systems are likely and able to undergo hierarchical mergers and populate the mass gaps.

In this paper we derive a set of fundamental relations describing how effective a dense cluster can grow a 2G population from a series of in-cluster GW mergers of 1G–1G binaries, as a function of characteristic mass, length, and timescales of the 1G objects and their cluster. The core of our calculations are based on the post-Newtonian (PN) binary-single hardening model presented in Samsing (2018) and Samsing & D’Oraio (2018), where binaries are able to merge in between or during their hard binary-single interactions.

We use our derived expressions to make general statements about which clusters are able to populate the LMG through BNS mergers and the UMG through BBH mergers. Our model is fully analytical and our results are given in closed form expressions, and as a result, we are therefore only able to describe idealized clusters with a constant density and velocity dispersion (for an extension of our model, see, e.g., Antonini & Gieles 2020); however, our work serves as an important first step in connecting physical parameters with more general statements related to hierarchical mergers (see also recent work by Baibhav et al. 2020).

The paper is organized as follows. In Section 2 we introduce our dynamical cluster- and 3-body interaction model, and use it to derive results on how efficient a simple unevolving binary and single cluster population is at producing in-cluster GW mergers. In Section 3 we extend our model to include a time-dependent distribution of both singles and binaries, from which we derive a closed form solution to the upper limit on the number of 2G objects relative to 1G objects a given cluster can reach in a Hubble time. We further discuss these results in relation to populating the LMG and the UMG. We conclude our study in Section 4.

2. Formation of In-cluster Mergers

In this work we study clusters with a core populated by COs (NSs or BHs), each with the same mass m. These COs interact and can through different dynamical pathways merge through the emission of GWs either inside or outside of their cluster (e.g., Rodriguez et al. 2018; Samsing & D’Oraio 2018). The COs that merge inside the cluster give rise to a growing in-cluster population of BHs with a mass \( \sim 2m \), given that the kick velocity associated with asymmetric GW emission at merger is smaller than the cluster escape velocity (e.g., Gerosa & Berti 2019). In this work we refer to the initial population of COs by “1G” or 1. generation (1G) objects, and the population of BHs that is formed through the collision of 1G–1G binaries by
Figure 1. Illustration of our $\delta$-model described in Section 2.1. In this model we assume that all binaries dynamically form at the HB-limit inside the cluster, after which they undergo scatterings with the surrounding singles. Each scattering leads to a fixed decrease in the SMA from $a$ to $a_\delta$, as shown in the left column (Interactions). This series of hardening interactions terminates at a characteristic SMA $a_m$, which is where the binary merger inside the cluster is being ejected, or when the time passes $t_{\text{in}}$, as further described in Section 2.1.1. The middle and the right columns show the two scenarios where the binary is either ejected (Dynamical Ejection) or merges inside the cluster (In-cluster Merger), respectively. The outcome from each of these scenarios is shown in the bottom panel (Outcomes). As seen, the outcome from a dynamical ejection is 1 binary and $\sim2$ singles, where for 1G–1G in-cluster mergers the outcome is by definition a 2G object. When following the in-cluster population of 1G and 2G objects over time, the Dynamical Ejection outcome always acts as a “sink term,” where the 1G–1G In-cluster Merger outcome is the “source term” for the 2G population, as further described in Section 3.1. Note that the gray and black circles refer to 1G- and 2G objects, respectively, while the diamond symbol denotes either of these two objects.

“2” or 2. generation (2G). As described in the Introduction, the 2G population is able to populate both the lower ($3–5M_\odot$) and upper ($>45M_\odot$) BH mass gaps that are believed to be associated with the initial 1G population. For example, it might be possible to populate the $3–5M_\odot$ BH mass gap through the collision of BNSs.

Below we start by deriving and presenting a set of basic relations for describing the growth of 2G populations through in-cluster 1G–1G GW mergers. Throughout the paper we mainly illustrate results for our two fiducial mass cases: $m = 1.4M_\odot$ and $m = 30M_\odot$, which are in the relevant range for populating the LMG and the UMG, respectively.

### 2.1. Cluster Model and 3-body Dynamics

We consider a cluster described by two distinctive populations: a population of lighter stars and one of heavier COs with equal mass $m$. Due to the effect of mass segregation (e.g., Spitzer 1987), the heavier COs constantly seek to occupy the central regions to form their own subcluster, in which most of their strong interactions take place (e.g., Arca Sedda et al. 2018). We will therefore in this work mainly refer to the dynamical properties of the central CO subcluster. This subcluster we assume has a constant number density of singles $n$ and velocity dispersion $v_\text{c}$. We further denote the escape velocity of the entire cluster by $v_\text{esc} = f_{\text{esc}} \times v_\text{c}$, where we take $f_{\text{esc}} = \sqrt{12}$ in all our examples, as found in the case of a Plummer’s sphere (Binney & Tremaine 2008). Besides singles, the cluster also harbors a population of CO binaries that at early times consists of 1G–1G pairs, but at later times, through dynamical exchange interactions, can evolve to have pairs also including 2G objects. Although primordial binaries might be present in the cluster at early times, we only include dynamically formed CO binaries in our analysis as only these can provide a constant predictable source leading to in-cluster GW mergers. As further described below, this happens mainly by the CO binaries undergoing strong scatterings with the single population, also referred to as binary-single interactions (Samsing et al. 2014). In-cluster GW mergers can also form in other ways, such as through single-single GW captures (Samsing et al. 2020), secular Kozai triples (Antonini et al. 2016), and binary-binary interactions (Zevin et al. 2019); however, these pathways are generally subdominant compared to the binary-single channel. Our main discussions will therefore mostly involve mergers from the interacting binary-single population. Finally, we do not consider strong interactions between the COs and the (lighter) stars, but the stars do indirectly play a role as a heat sink that enables COs on long excursions to sink back into the core regions. In the sections below we continue by describing the basics of our cluster model.

#### 2.1.1. Binary Hardening and Outcomes

We assume that a given CO binary inside the cluster forms dynamically with a semimajor axis (SMA), $a$, equal to the hard binary (HB) limit value (e.g., Heggie 1975; Aarseth & Heggie 1976; Hut & Bahcall 1983),

$$a_{\text{HB}} = \frac{3G}{2} \times \frac{m}{\sqrt{a}},$$

which is where the binary binding energy $(Gm^2/(2a))$ equals the kinetic energy of the surrounding singles with respect to the binary $(mv^2_g/3)$.

After this, the binary undergoes scatterings with the surrounding singles, each of which leads to a decrease in the SMA of the binary from $a$ to $a_\delta$. This corresponds to a change in $a$ of $-a(1-\delta)=-a\Delta$, where we have introduced $\Delta \equiv 1 - \delta$ to shorten notations. In reality, the change per interaction in the binary binding energy $E_b$ follows approximately a power-law distribution $P(E_b) \propto E_b^{-\gamma}$ with $\gamma \sim 9/2$, depending on exactly how a strong binary-single interaction is defined (e.g., Heggie 1975; Stone & Leigh 2019). In this work we do not use the full distribution, instead we assume that each interaction leads to a fixed fractional decrease $\delta$ in the SMA, which is equal to the average value found from the distribution $P(E_b) \propto E_b^{-\gamma}$. Given that $E_b = -Gm^2/(2a)$ and $\delta \equiv \langle a \rangle / a_0$, where $a_0$ is the initial SMA and $\langle a \rangle$ is the average value of the
resulting SMA, $\delta$ is given by (see also Samsung 2018)
\[
\delta = (\gamma - 1) \int_0^1 \delta^{(\gamma-1)} d\delta = 1 - \gamma^{-1},
\]
\[
= \frac{7}{9} (\gamma = 9/2).
\tag{2}
\]

The binary keeps undergoing these so-called “hardening” interactions with the surrounding single population, until its SMA reaches one of the following three characteristic values. The first, denoted by $a_{ej}$, is the maximum SMA value for which the binary will get ejected out of the cluster if it undergoes a binary-single interaction. Note that this is a fixed value in our simple “$\delta$-model.” The second, denoted by $a_{GW}$, is the SMA for which the total integrated probability for the binary to merge at any given state from $a_{HB}$ to $a_{GW}$ equals one. The merging binary will of course have a decreasing SMA as it inspirals, but will during this time not interact with other objects. The third, denoted by $a_{HI}$, is the value it takes a Hubble time to reach through binary-single interactions alone from the initial value $a_{HB}$. The hierarchy of these three characteristic scales is set by $(v_m, n, m)$ and plays a key role in how to grow a 2G BH population inside the cluster through in-cluster mergers (see also Antonini & Rasio 2016; Antonini & Gieles 2020; Baibhav et al. 2020). For example, if $a_{ej} > a_{GW}$, then most binaries will get ejected and merge outside of the cluster, compared to if $a_{GW} > a_{ej}$, in which case all binaries will merge inside. If, however, $a_{HI} > \{a_{ej}, a_{GW}\}$ then the system will not be able to conclude even a single interaction sequence, and an effective accumulation of 2G mergers is therefore near impossible. As a result, the “relevant” value for a given system is
\[
a_m = \max (\{a_{ej}, a_{GW}, a_{HI}\}),
\tag{3}
\]
where the subscript “m” here refers to “minimum,” as this is the smallest value the SMA of the interacting binary can take. This $\delta$-model is further illustrated and described in Figure 1.

Throughout the paper we refer to the process in which the system brings a binary from $a_{HB}$ to $a_{m}$ through binary-single interactions alone as one “Interaction Cycle” (IC). Generally, after a binary has completed its IC, the cluster core will contract as a result of the lost dynamical heat source (e.g., Breen & Heggie 2013), which will trigger the formation of a new binary through a three-body interaction (e.g., Morscher et al. 2015; Banerjee 2018; Antonini & Gieles 2020; dynamical binary formation), after which the binary-single interaction process repeats. This cycle of binary formation and hardening interactions is also often referred to as “binary burning” (e.g., Kremer et al. 2020a). This balance between dynamical heat created in the core regions and the overall structure of the cluster further leads to the fact that most clusters only harbor a few binaries at any given time (see Section 3.2.1). We continue below by deriving $a_{ej}, a_{GW}$, and $a_{HI}$. We also refer the reader to Antonini & Rasio (2016), Antonini & Gieles (2020), and Baibhav et al. (2020) for complementary discussions on this.

### 2.1.2. Derivation of Outcome Conditions

For calculating the SMA at which the binary is ejected, $a_{ej}$, we first use the fact that the energy released in one interaction between a single and a binary with SMA $a$ is given by $E_{bs} = (\Delta / \delta) \times E_{b}(a)$, where $E_{b}(a)$ is the internal energy of the binary before interaction (e.g., Samsung 2018). The energy $E_{bs}$ is “released” in the three-body center-of-mass (COM), which in the Newtonian limit is conserved from before to after the interaction. From momentum conservation it then follows that the binary receives a velocity kick, defined at infinity in the COM, of $v_{b}^\text{GW} = E_{bs}/(3m) = (1/6) (\Delta / \delta) G m/a$. When $a$ is such that $v_{b} > v_{\ast}$, then the binary escapes the cluster. By now defining $a_{ej} \equiv a(v_{b} = v_{\ast})$, it then follows that
\[
a_{ej} \approx \frac{G \Delta}{6 b f_{ed}^2} \times \frac{m}{v_{b}^2}. \tag{4}
\]
Note here that $a_{HB}/a_{ej} \approx 9 f_{ed}^2 \delta / \Delta = (63/2) f_{ed}^2$, where we have set $\delta = 7/9$ in the last equality. A single binary therefore has to decrease its SMA by 1–2 orders of magnitude through binary-single scatterings before a possible ejection can take place. As will be discussed and used later, several of the single objects interacting with the binary will also get ejected, as they likewise receive recoil kicks during the hardening process. As a result, for every single binary ejected there will also be $N^3$ single objects ejected. This number $N^3$ can be estimated by first comparing the SMA below which single ejections are possible, $a_{ej}^3 \approx 2 G \Delta / (3 \delta b f_{ed}^2) \times m / v_{b}^2$, where we have used $2 v_{b} = v_{\ast}$, with the binary ejection SMA $a_{ej}$ from Equation (4).

As seen, $a_{ej}^3/a_{ej} = 4$. Now considering that after $\Delta n$ binary-single interactions the binary SMA decreases by a factor $\delta^{\Delta n}$, it then follows that $N^3 \approx \ln(1/4) / \ln \delta \approx 5$, where we have made use of the fact that one single object is ejected in each scattering for $a_{ej} < a < a_{ej}^3$. The number $N^3$ is therefore a constant that does not depend on any properties of the system, as long all of the interaction steps are “available.” In this paper we use $N^3 = 4$, as this value is slightly closer to what is found in numerical simulations; however, the exact value does not play a large role—the important point is that it takes a constant value.

For $a_{GW}$, we start by calculating the probability that a binary with SMA $a$ merges before its next binary-single interaction, denoted here by $p_2(a)$. For this we assume that the eccentricity distribution of the binary follows that of a thermal distribution, $P(e) = 2e$ (e.g., Heggie 1975). In addition, we use the fact that the time in between binary-single interactions, $t_{bs}(a)$, is the inverse of the binary-single encounter rate, $t_{bs}(a) \approx (n_{bs} v_b)^{-1}$, where $n_{bs} \propto ma/v_b^2$ is the HB binary-single interaction cross section in the gravitational focusing limit (see, e.g., Spitzer 1987; Samsung et al. 2018). Under these assumptions it directly follows that $p_2(a) = (t_{bs}(a)/t_{GW}(a))^{2/7}$ (Samsung 2018), where we have used the fact that the inspiral time of a binary with a nonzero eccentricity is $\propto t_{GW}(a)(1 - e^2)^{2/7}$, where $t_{GW}(a)$ denotes the inspiral time for $e = 0$ (Peters 1964). This $p_2(a)$ is only the probability for merger during a single “interaction step” $k$, where we here introduce the notation $a_{k} = a_{HB}^k$. The total probability for a binary to merge in between its binary-single interactions from $a_{HB}$ to $a_{m}$, denoted by $p_2(a_{m})$, is therefore found by simply integrating from $k(a_{HB}) = 0$ to $k(a_{m})$. Using $da = -a \Delta da k$, the solution is found to be $p_2(a_{m}) \approx p_2(a_{m}) \times (7/(10 \Delta))$; (e.g., Samsung 2018; Samsung et al. 2019b), which can be written out as
\[
p_2(a_{m}) \approx A_c^{7/7} \times \frac{m a_{m}^{4/7} v_{b}^{2/7}}{n_{bs}^{7/7} a_{m}^{10/7}}, \tag{5}
\]
where we have assumed that $p_2(a_{m}) \gg p_2(a_{HB})$ and defined the constant $A_c = (7^{1/7} 85^2 G^3)/((10 \Delta)^{7/2} 9 \pi c^5)$. If we now set
$P_2 = 1$, then the corresponding $a_{GW} = a(P_2 = 1)$ can now be isolated and gives

$$a_{GW} = A_c^{1/5} \times \frac{m^{2/5} \nu_d^{1/5}}{n^{1/5}}.$$  \hspace{1cm} (6)

As is seen, this limit is surprisingly insensitive to the cluster parameters $\nu_d$ and $n$ (see also Antonini & Rasio 2016).

The last characteristic SMA we consider is $a_{dH}$, which is the value for which it takes the binary a Hubble time to reach from $a = a_{HB}$ through binary-single interactions alone. For calculating this, we start with the time it takes the binary to undergo one interaction at interaction step “k,” which can be approximated as

$$t_{bs}(a_k) \approx \frac{1}{a_m} \nu_d \sigma_{bs}(a_k).$$

The total time it takes to reach $a_m$, denoted by $\tau_m$, is found by integration $t_{bs}(a) \propto k(a_{HB}) = 0$ to $k(a_m)$. From this, one finds that $\tau_m \approx t_{bs}(a_m) / \Delta$, which also can be written as

$$\tau_m \approx (6\pi G \Delta)^{-1} \times \frac{1}{a_m} \nu_d \sigma_{bs}$$ \hspace{1cm} (7)

where we have assumed that $a_{HB} \gg a_m$. Setting this expression for $\tau_m$ equal to $t_H$, and isolating the corresponding $a_{dH} \equiv a_m(\tau_m = t_H)$, one now finds

$$a_{dH} \approx \left( \frac{6\pi G \Delta}{t_H} \right)^{-1} \times \frac{\nu_d}{\sigma_{bs}}$$ \hspace{1cm} (8)

which relates to $a_{dB}$ as $a_{dH} / a_{dB} \approx t_H / t_{bs}(a_{HB}) \Delta$.

### 2.2. Results

Having derived analytical expressions for the three characteristic scales $a_{ej}$, $a_{GW}$, and $a_{dH}$ in Section 2.1.2 above, we are now in a position to start exploring which cluster systems are likely to grow a population of 2G objects. In the sections below we study this by considering a few general relations and overview figures for a “static” cluster population. In Section 3 we use these results to model “time-evolving” populations.

#### 2.2.1. Outcome Regions

Figure 2 shows $a_m = max\{ (a_{ej}, a_{GW}, a_{dH}) \}$ with colored regions (blue, red, gray) as a function of cluster velocity dispersion $\nu_d$ and number density $n$ for $m = 30M_\odot$ (top) and $m = 1.4M_\odot$ (bottom). The three regions are separated by green dashed lines, where the point at which all of the three regions meet, a point we refer to as the “break point” (BP), is highlighted with a green circular dot. How the green dashed lines depend on the parameters $\nu_d$, $n$, and $m$, provide the key to understand which systems are likely to produce a sizable population of 2G objects. Below, we study this in more detail. Our expressions are written out for $b = 7/9$ if nothing else is stated.

We start by the line separating the two regions $a_{ej}$ (blue) and $a_{dH}$ (gray) to the left of the BP. By now setting $a_{ej} = a_{dH}$ and solving for the corresponding $n(ej, GW) = n(a_{ej} = a_{dH})$, one finds

$$n(ej, tH) = \frac{63 \nu_d^2}{4 \pi G^2 t_H} \times \frac{v_H^3}{m^2}$$ \hspace{1cm} (9)

where we used Equations (4) and (8). The next line is the one separating the regions $a_{ej}$ (blue) and $a_{GW}$ (red) to the right of the BP. Following the same procedure as above, we first set

$$a_{ej} = a_{GW}$$

and solve for the corresponding $n(ej, GW)$,

$$n(ej, GW) = \frac{1}{B_c G^3 c^5} \nu_d^{1/1} \frac{1}{m^3}$$ \hspace{1cm} (10)

where we have used Equations (4) and (6) and introduced the constant $B_c = (9\pi / 2^{109} 85) (20 / 63)^{3/2}$. Finally, the line separating $a_{GW}$ (red) and $a_{dH}$ (gray) to the right of the BP is found.
from setting $a_{GW} = a_{dH}$, from which we find

$$n(GW, t_H) = \left[ \frac{63}{4\pi} \frac{B_v c^3}{G v_d^4 t_H^4} \right]^{1/8} \times \frac{v_d}{m^{1/4}},$$  \hspace{1cm} (11)

where we have used Equations (6) and (8). The associated coordinates of the BP, denoted by $(v_d(BP), n(BP))$, can now be found from, e.g., setting $n(e|j, t_H) = n(e|j, GW)$, from which follows

$$v_d(BP) = \left[ \frac{63}{4\pi} \frac{B_v G c^5}{f_e^8 t_H^4 m} \right]^{1/8} \times m^{1/8},$$

$$n(BP) = \left[ \frac{63}{4\pi} \frac{B_v c^{11}}{G^{12} f_e^8 t_H^{11}} \right]^{1/8} \times m^{-11/8}. \hspace{1cm} (12)$$

As seen here, the BP coordinates $v_d(BP), n(BP)$ are $\propto m^{1/8}, m^{-11/8}$, respectively. Therefore, the location of the BP along the $v_d$-axis remains almost constant for reasonable changes in $m$, in contrast to the location along the $n$-axis, which can change by orders of magnitude. As a result, for 1G objects in the mass range $1M_\odot \lesssim m \lesssim 50M_\odot$, the BP will always be around $10 \sim 20$ km s$^{-1}$, which is slightly higher than the dispersion velocity of a typical GC. Since no configurations with $a_m = a_{GW}$ are possible for values of $v_d < v_d(BP)$ it then follows that GCs will in theory never be able produce binaries that only have the option of merging inside the cluster. The relevant value of $a_m$ for GCs is then either $a_{dH}$ or $a_{dH}$. Another feature linked to the BP is that clusters with $n \gtrsim n(BP)$ will (nearly) always produce and process binaries that undergo at least one IC due to the relative weak dependence on $v_d$ for the $n(GW, t_H)$ boundary. Regarding the dependence on $m$, one sees that the boundary quickly moves up for decreasing values of $m$, as $n(BP) \propto m^{-11/8}$. This makes it increasingly difficult for 1G objects with masses in the range $m \sim 1M_\odot$ to undergo more than 1 IC within a Hubble time for astrophysical cluster values compared to $m \sim 30M_\odot$. 1G objects, as is clearly seen in Figure 2. Before we study this in greater detail, we proceed below by exploring to what degree 3-body mergers and single-single (S-S) GW captures contribute to the in-cluster merger rate.

2.2.2. Three-body Mergers and Single-single GW Captures

Before moving on to how efficient a population of binaries is at producing a 2G population, we here address the potential importance of including the in-cluster merger contribution from S-S GW captures and 3-body mergers. As described in Section 2.1, “S-S GW captures” denote the process in which two initially unbound COs become bound through the emission of GWs (e.g., Samsing et al. 2020), while a “3-body merger” refers to COs merging during a chaotic 3-body interaction (Samsing et al. 2014).

We start by analyzing the contribution from 3-body mergers. For this, we first estimate for which part of the $(v_d, n)$ space the total integrated probability of producing a 3-body merger, $P_3$, is larger than the total probability for undergoing a 2-body merger, $P_2$. Following Samsing (2018), the probability for a binary-single interaction to produce a 3-body merger can be approximated by $p_3(a) \approx 2N'(R_m/a)^2/7$, where $N' \approx 20$ denotes the number of “temporary binary states” the chaotic triple interaction on average assemblies during one interaction, $R_m$ is the Schwarzschild radius of a BH with mass $m$, and $a$ is the SMA of the initial target binary. The total probability for a 3-body merger to form during one IC can now be found from integrating $p_3(a)$ from $a = a_{dH}$ to $a = a_m$ the same way we did for finding $P_2$ in Section 2.1.2. Following this approach, one finds that $P_3(a_m) \approx P_3(a_m) \times (7/(5\Delta))$, which can be written out in the following way

$$P_3(a_m) \approx \left[ \frac{5^{11/7} N' G^{5/7}}{5\Delta} \right] \times m^{5/7}a_m^{-5/7}, \hspace{1cm} (13)$$

where we have assumed that $p_3(a_m) \gg p_3(a_{dH})$. (Note here that we calculate these merger probabilities separately, i.e., we do not take into account the potential interplay between merger channels, including the S-S GW capture channel.) From this we see that $P_3(a_m)/P_2(a_m) \propto (n/v_d)^{2/7}$. This indicates that 3-body mergers will provide the greatest contribution relative to the 2-body mergers at high $n$ and low $v_d$, which is the regime in which $a_m = a_{dH}$, as seen on Figure 2. We therefore need to evaluate and compare $P_2$ and $P_3$ for $a_m = a_{dH}$. Using Equations (5), (13), and (4), this first lead us to

$$P_2(a_{dH}) \approx \left[ \frac{1}{f_e^2} \frac{B_v G^3 c^{11}}{m^{5/7}n^{2/7}} \right] \times \frac{v_d^{22/7}}{m^{9/7}},$$

and

$$P_3(a_{dH}) \approx \left[ \frac{42^{5/7} N' G^{5/7}}{(5/63)} \frac{1}{e^{10/7}} \right] \times \frac{v_d^{10/7}}{m^{3/7}}. \hspace{1cm} (15)$$

Now setting these two expressions equal to each other one finds

$$n(P_2, P_3) \approx \left[ \frac{(5/63)^{2/7} f_e^5}{42^{22/7} N' G^3} \right] \times \frac{v_d^6}{m^3}, \hspace{1cm} (16)$$

where $n(P_2, P_3)$ therefore represents the boundary in the $a_m = a_{dH}$ region for which $P_2 = P_3$. This boundary is shown in Figure 2 with the dotted line that encloses the black solid line hatched area. In this area $P_2 > P_3$. As is seen, for most systems, especially the one with a relatively low mass $m$ and moderate density $n$, 3-body mergers will not dominate the total in-cluster merger probability. We will therefore in our analytical models throughout this paper omit this contribution for simplicity and clarity.

We now move on to the S-S GW capture population. For this it is more easy to compare merger rates, $\Gamma$, than probabilities. In this case, the total rate of S-S GW capture mergers from a simple “$n\sigma v'v'$” estimate is given by Samsing et al. (2020),

$$\Gamma_s \approx \left[ \frac{4\pi G^2}{c^{10/7}} \left( \frac{85\pi}{24\sqrt{2}} \right)^{2/7} \right] \times \frac{N_r m^2}{v_d^{18/7}}. \hspace{1cm} (17)$$

where $N_r$ is the total number of single BHs. Note that we have here assumed that all of the single BHs, $N_r$, are distributed uniformly according to our model of a constant $v_d$, $n$; however, in reality, the single BHs naturally distribute according to some density and velocity profile. As a result, the real GW capture rate is generally smaller than the one presented in the above
Equation (17), as further discussed in Samsing et al. (2020). Regarding the merger rate from our considered binary-single interactions, one finds that this can be approximated by
\[
\Gamma_{23} \approx \frac{N_b(P_2(a_m) + P_3(a_m))}{\tau_m},
\]
where \(P_2 + P_3 \leq 1\) is the total number (probability) of 2-body and 3-body mergers forming during 1 IC, \(\tau_m\) is the time it takes to undergo 1 IC (see Equation (7)), and \(N_b\) is the number of CO binaries in the cluster that contribute to the merger rate. We have here included the 3-body mergers, as it turns out that the S–S GW captures only significantly contribute for low \(v_d\) and high \(n\), exactly where the 3-body mergers also contribute. This is seen in Figure 2, where the black dotted line enclosing the black solid area is where \(\Gamma_{as} = \Gamma_{23}\) for binary fraction \(N_b/N_e = 0.05\). The S–S GW captures are therefore not expected to provide a significant contribution in the regions we are interested in.

To conclude, we have here shown and argued that neither the 3-body mergers nor the S–S GW capture mergers contribute significantly to the in-cluster merger rate. We therefore only consider the 2-body merger contribution in the rest of this paper.

2.2.3. Interaction Cycles and In-cluster Mergers

The number of in-cluster GW mergers that can be produced over a Hubble time per 1G–1G binary, here denoted by \(N_{bh}(t_{H})\), serves as an approximate measure for how efficient a given cluster is at growing a 2G population. At this stage we approximate this number by the following product,
\[
N_{bh}(t_{H}) \approx N_c(t_{H}) \times P_M,
\]
where \(N_c(t_{H}) = t_{H}/\tau_m\) is the number of ICs a cluster can run through in a Hubble time, i.e., the number of binaries the cluster can process in time \(t_{H}\), and \(P_M\) is the probability for an in-cluster GW merger to form during one IC. In Figure 2 is shown with orange dashed lines and red solid lines the contours of \(N_c(t_{H})\) and \(P_M\), respectively, where for \(P_M\) we have here included the probability for 3-body mergers, i.e., \(P_M = P_2(a_m) + P_3(a_m)\). In short, our procedure for estimating \(N_c(t_{H})\) and \(P_M\) at a given point \((v_d, n)\), is first to calculate \(a_m\) from Equation (3), after which we use Equation (7) to find \(N_c(t_{H}) = \frac{t_{H}}{\tau_m}\), and Equations (5) and (13) to find \(P_M = P_2(a_m) + P_3(a_m)\).

As seen in Figure 2, for decreasing values of \(v_d\) the probability \(P_M\) decreases, which also follows from Equation (14) where \(P_2 \propto v_d^{-7/7}\), in contrast to the number of ICs, \(N_c(t_{H})\), that instead increases. Therefore, one can easily have a cluster with binaries where burning is efficient, i.e., where \(N_c(t_{H}) \gg 1\), but at the same time with a probability for merging during individual ICs is low, i.e., with \(P_M \ll 1\), and vice versa. How these two quantities “balance out” is clear in Figure 3, which shows in black solid lines \(N_{bh}(t_{H})\) from Equation (19). Surprisingly, the large changes in both \(P_M\) and \(N_c(t_{H})\) as \(v_d\) is varied almost cancel out, and \(N_{bh}(t_{H})\) is as a result almost flat across \(v_d\). To study this behavior further, we can write out \(N_{bh}(t_{H})\) in the region relevant for GC systems where \(P_2 \gg P_3\) and \(a_m = a_v\) using Equations (19) and (7) evaluated at \(a_m = a_v\), and the expression for \(P_2(a_v)\) given by

\[
\text{Equation (14), from which one finds}
\]
\[
N_{M}(t_{H}) \approx t_{H} \left[ \frac{4 \pi}{63} \frac{G^4}{c^6} \frac{\dot{m}^8}{R_{Ed}^3} \right] \times n^{8/7} m^{8/7} v_d^{1/7} \quad (20)
\]
\[
\approx 0.5 \left( \frac{n}{10^6 \text{pc}^{-3}} \right)^{8/7} \left( \frac{m}{1.4 M_{\odot}} \right)^{8/7} \left( \frac{v_d}{10 \text{ km s}^{-1}} \right)^{1/7}, \quad (21)
\]
where in the last line we have inserted values relevant for NS–NS mergers. This confirms the results we see in Figure 3, namely that \(N_{M}(t_{H})\) only depends weakly on \(v_d\) as \(N_{M}(t_{H}) \propto v_d^{-1/7}\). As a result, all systems with \(n > n(BP)\) will to leading order have \(N_{M}(t_{H}) \gtrsim 1\). From this follows that if the number of CO binaries is constant in time at a value \(N_b\), then
the number of in-cluster mergers for \( n \gtrsim n(BP) \) will be \( \gtrsim N_b \) after a Hubble time. For example, for our \( m = 30M_\odot \) case shown in the upper panel of Figure 3, the number of in-cluster mergers over a Hubble time per binary is of order 10 for \( \log n \approx 4 \sim 5 \ \text{pc}^{-3} \). If the number of BBHs in the cluster at any given time is a few, say \( \sim 5 \), then our model predicts that the total number of in-cluster mergers forming over a Hubble time is \( \sim 5 \times 10 = 50 \). Although this number of course fluctuates from cluster to cluster, we note that this number is consistent with what is found using numerical simulations (see Rodriguez et al. 2019, where 48 in-cluster mergers were reported for their example in Section IV.A). More generally, \( N'_M(t_H) \) provides an upper limit on the number of available 2G objects after a Hubble time produced per cluster binary, as only a small fraction of the in-cluster mergers, i.e., 2G objects, are actually retained by the cluster (Rodriguez et al. 2019). The remaining are either kicked out immediately as a result of GW kicks, or later dynamically through, e.g., a binary-single interaction. Considering the \( m = 1.4M_\odot \) case, we see both from Equation (20) and Figure 3 that \( N'_M(t_H) \) is only \( \gtrsim 1 \) for \( \log n \gtrsim 5\sim 6 \ \text{pc}^{-3} \), which is a very high density threshold for

astrophysical standards. Although such densities can be reached during core collapse (e.g., Kremer et al. 2019a) this phase will only be temporary relative to a Hubble time, and therefore will not give rise to the steady formation and interaction of CO binaries as otherwise required for our considered process to be effective. This provides a clear hint that clusters are not likely to be effective in populating the LMG through NS–NS mergers (see also Ye et al. 2020).

Lastly, in relation to the probability of observing a possible 2G population from a cluster, what matters is not only the number of 2G objects produced, but also how many of these that are present in the cluster compared to the number of remaining 1G objects. As described back in Section 2.1.2, a single IC will on average give rise to \( N_{1G}^0 + 2 \sim 6 \) ejected 1G objects (if in-cluster mergers and 2G objects are ignored), which naturally leads to a gradual reduction of this population over time. In our model considered so far, the number of in-cluster GW mergers relative to the number of (remaining) 1G objects after time \( t \) is therefore approximately

\[
\frac{N'_M(t)}{N_1(t)} \approx \frac{N_0 N_{1G}(t)}{N_1(0) - N_0 N_1(N_{1G}^0 + 2)}
\]

(22)

\[
\approx \frac{N_{1G}^0(t)}{f_b^{-1}(0) - N_{1G}^0(t)}
\]

(23)

where we have assumed that \( N_b \) remains constant, \( N_1(0) \) denotes the initial number of 1G objects, \( f_b(0) = N_b/N_1(0) \), and \( N_{1G}^0(t) \) denotes the total number of 1G objects ejected after time \( t \) per binary. We will explore this ratio and others in the sections below.

3. Populating the Black Hole Mass Gaps: Time-evolving Model

In this section we develop a simple time-dependent cluster model to study the evolution of both 1G and 2G objects as a function of time. As further described in the following sections, in this model we take into account both binary and single dynamical ejections, and in particular the growth of 2G objects as a result of in-cluster 1G–1G mergers. We (still) assume the cluster is described by a fixed set \( v_{\text{in}}, n \), and all objects have the same mass \( m \), which of course is a simplification of a real cluster. This in turn however enables us to put forward simple, general, and informative statements, based solely on characteristic mass, length, and timescales.

In the first section below we derive a set of evolution equations for \( N_1 \) and \( N_2 \), where \( N_i \) here denotes the number of objects of type \( \text{"i"} \). In Section 3.2 we solve these equations from which we put upper limits on the ratio \( N_2/N_1 \), illustrated for \( m = 1.4M_\odot \), (2G in the LMG) and \( m = 30M_\odot \) (2G in the UMG), for a grid of cluster systems described by \( v_{\text{in}}, n \).

3.1. Evolution Equations

We consider a cluster described by a constant \( v_{\text{in}}, n \) that initially has a population of \( N_1(0) \) 1G objects all with equal mass \( m \). In this cluster there is a (time-dependent) population of binaries that interact with the surrounding single population, which gives rise to dynamical ejections, exchanges, and in-cluster mergers. The absolute and relative number of \( N_1 \) (1G) and \( N_2 \) (2G) objects therefore changes over time through various dynamical mechanisms, which all depend on \( v_{\text{in}}, n, m \).
This configuration is described and illustrated in Figure 4. The question is, for what initial conditions of \( v_{i,k} n, m \) is the system able to produce a sizable population of 2G objects after a Hubble time? To answer this question, we start by writing out the following set of differential equations that we take to represent the evolution of \( N_1 \) and \( N_2 \):

\[
\dot{N}_1 = -(N_1^{eq} + N_1^{eq} + 2N_1^{eq}) - (N_1^{M} + 2N_1^{M}) + (R_1^{M}N_1^{M}),
\]

\[
\dot{N}_2 = -(N_2^{eq} + N_2^{eq} + 2N_2^{eq}) - (N_2^{M} + 2N_2^{M}) + (R_1^{M}N_1^{M}),
\]

(24)

where \( N_i^{eq} \) is the number of objects of type “i” (1G or 2G) that are ejected from the cluster as singles, \( N_i^{eq} \) is the number of binaries ejected from the cluster consisting of object types “(ij),” \( N_i^{M} \) is the number of \( (ij) \) binaries merging inside the cluster, and \( R_1^{M} \) is the retention fraction of 1G–1G mergers. As seen, both “ejections” (single and binary) and “in-cluster mergers” act as “sink terms,” except as for the term \( \propto N_1^{eq} \) that serves as the 2G “source term.” As we are studying the process of growing a 2G population through in-cluster GW mergers during successive ICs, we restrict ourselves in the following to describe systems that are able to undergo \( N_i \gg 1 \). Therefore, instead of evolving the above equations over, e.g., individual interaction steps “k,” or time, we evolve them over the number of ICs, \( N_c \). The “dot” over each \( N \) refers therefore to the change per IC.

The relevant terms for writing out our evolution equations from above can be written as

\[
N_i^{eq} \approx [N_0P_Mp_i^{eq}N_i^{eq}],
\]

(25)

\[
N_i^{eq} \approx [N_0P_Mp_b^{eq}],
\]

(26)

\[
N_i^{eq} \approx [N_0P_Mp_b^{eq}],
\]

(27)

where \( N_0 \) is the number of binaries, \( P_M (P_M = 1 - P_M) \) is the integrated probability that a given binary (does not) merge during a single IC, \( p_i^{eq} \) is the probability that object “i” is ejected after a binary-single interaction, \( N_i^{eq} \) is the total number of singles per binary ejected during one IC, \( p_i^{eq} \) is the probability that binary “(ij)” is ejected after a binary-single interaction, and \( p_b^{eq} \) is the probability that \( (ij) \) is in a binary at a random hardening step “k.” These terms can be further expanded as

\[
p_2^{eq} \approx p_1^{eq}p_1^{eq}[1 + B], \quad p_1^{eq} \approx 1 - p_2^{eq},
\]

(28)

\[
p_2^{eq} \approx p_1^{eq}p_1^{eq}[1 + B], \quad p_1^{eq} \approx 1 - p_2^{eq},
\]

(29)

\[
p_b^{eq} \approx p_b^{eq}B, \quad p_b^{eq} \approx 1 - p_2^{eq},
\]

(30)

\[
p_{21}^{eq} \approx 2w/3, \quad p_{12}^{eq} \approx 1 - p_{21}^{eq},
\]

(31)

\[
p_i^{eq} \approx N_2F/(N_1 + N_2), \quad p_i^{eq} \approx 1 - p_i^{eq},
\]

(32)

where \( p_i^{eq} \) is the probability that object type “2” (2G) is the incoming single object in a binary-single interaction at hardening step “k,” \( p_i^{eq} \) is the probability that a given binary-single interaction results in an end state characterized by a binary consisting of object pair \( (ij) \) that is unbound relative to the remaining single object “k” (often referred to an “exchange interaction”, see, e.g., Hut & Bahcall 1983), and \( B = 2Fw/(3 - 2w) \). The factor \( F \) is introduced to quantify the probability “enhancement” of a 2G object to interact with a binary compared to a 1G object. For example, the enhancement factor from standard gravitational focusing of having a 2G object to interact with a binary compared to a 1G object is \( F = (1 + 1 + 2)/(1 + 1 + 1) = 4/3 \). Similarly, \( w \) describes the “enhanced probability” that the outcome of a binary-single interaction involving a 2G object is \( \{12\} \), i.e., where \( \{12\} \) constitutes the binary components after the interaction. For this set of equations we have made four central assumptions. (1) All binary-single interactions involving objects \( \{ijk\} \) have the same outcome distributions irrespective of the initial configuration/permuation. (2) The probability to have interactions with \( > 1 \) 2G object is \( 0 \), which follows from our considered limit of \( N_2 \ll N_1 \). (3) Dynamical single and binary ejections associated with a given interacting binary are only \( > 0 \) if the binary in question does not merge before concluding its IC. (4) All interactions and ICs follow our “δ-model” illustrated in Figure 1. Now using these equations we can rewrite our evolution equations given by Equation (24) as follows:

\[
\dot{N}_1 = N_0B \times [+p_2^{eq}(A - P_M(A - B)) - (N_1^{eq} - P_MN_1^{eq})]
\]

\[
\dot{N}_2 = N_0B \times [-p_2^{eq}(A - P_M(A - B)) + (p_1^{eq}P_MB_{11}^{eq})],
\]

(33)

where \( N_i^{eq} = 2 + N_i^{eq} \) is here the total number of ejected objects over 1 IC, and \( A = [1 + B](p_{12}^{eq}N_1^{eq} + p_{21}^{eq}) \).

To summarize, our presented evolution equations given Equation (24) are completely general, and show simply which characteristic sink and source terms that are relevant for our problem. Other terms, such as strong binary-binary interactions (Zevin et al. 2019) and weak few-body scatterings (Hamers & Samsing 2019a; Samsing et al. 2019b, 2020), or more general mass-ratio dependent terms and corresponding GW kick prescriptions, can be included, but this is beyond this paper. The resulting terms shown in Equation (33) follow directly from simple combinatorics, and are constructed by calculating the (time-dependent) probability for 1G and 2G objects to interact and exchange into the interactions states shown in Figure 1, folded with the probability for dynamical ejections and in-cluster mergers during each IC. In the following sections we consider solutions to this coupled set of equations, from which we especially find a closed form solution to the upper limit on \( N_2/N_1 \) as a function of time.

### 3.2. Results

In the first section below, we study the evolution of \( N_1 \) and \( N_2 \) for two different cluster models, denoted \( cA \) and \( cB \), using the general set of evolution equations presented in the above Section 3.1. In the second section, we use these results to study the upper limit on the ratio \( N_2/N_1 \) evaluated at present day, i.e., at \( t = t_H \), for a grid of \( v_{i,k} n \) cluster systems.

#### 3.2.1. Time-evolving Populations

We study the evolution of \( N_1 \) and \( N_2 \) using Equation (33) for two distinct cases, \( cA \) and \( cB \). These two cases are described in the following.

**Case cA:** In this case we assume the weight factors \( F = 1 \) and \( w = 1 \), i.e., we keep track of the growing population of 2G objects, but we assume that in all dynamical aspects a 2G
object is indistinguishable from a 1G object. We are therefore able to explore the effect from pure “combinatorics” arising from the growing population of 2G objects that are free to exchange, merge, and be ejected in the same way as the 1G objects. Using Equation (33) with $F = 1, w = 1$ the evolution equations are in this case given by

$$N_1/N_b \approx -p_1^b (N_1^g - P_M N_1^g)$$
$$N_2/N_b \approx -p_1^b (N_2^g - P_M N_2^g) + P_M R_1^M (1 - 2p_2^b),$$

(34)

where we have used the fact that under these assumptions $A = N_2^g$ and $A - B = N_1^g$. In this case the number of 2G objects compared to 1G objects present in the cluster after a Hubble time represents approximately a lower limit, as in “reality” a higher number of 2G objects will be left in the cluster due to their higher mass (e.g., Sigurdsson & Phinney 1993).

Case $cB$: In this case we assume that $p_2^b = 0$ and $R_1^M = 1$, i.e., that the 2G objects are not participating in any interactions, and that the 1G objects as a result dynamically evolve through interactions, mergers, and ejections completely independent of the 2G objects. As a result, the number of 2G objects we here find after time $t_{H}$ represents the highest number possible, and the 1G population will also decrease to its lowest possible value. This case therefore represents the upper limit on how many 2G objects one can keep in a cluster after time $t_{H}$ compared to the 1G population. The evolution equations are in this case given by Equation (33) with $p_2^b = 0$ and $R_1^M = 1$,

$$\dot{N}_1/N_b = -(N_1^g - P_M N_1^g)$$
$$\dot{N}_2/N_b = +P_M.$$

(35)

This set of equations has a particular simple and interesting set of analytical solutions that we now explore before moving on. For this, we start by rewriting the above equations into a more general form to shorten the notations: \( \dot{N}_1 = -\alpha N_b, \dot{N}_2 = \beta N_b \), where we have defined

$$\alpha = N_1^g - P_M N_1^g$$
$$\beta = P_M.$$

(36)

To proceed, we now consider a specific model where the binary fraction stays constant such that $N_b = f_b \times N_1$. In this case, the solution to the above set of equations is easily found from simple integrations, from which it follows

$$N_1 = N_1(0) \times \exp(-\alpha f_b N_1),$$
$$N_2 = N_1(0) \times \beta/\alpha [1 - \exp(-\alpha f_b N_1)],$$

(37)

where $N_1(0)$ is the initial number of 1G objects, and $N_e = t/\tau_m$ is the number of ICs after time $t$. If we first consider the solution to $N_1$, we see that the population of 1G objects “decays” over time as if the cluster represents a giant “radioactive nucleus” with decay time $t_{d}$, given by

$$t_d \approx \frac{\tau_m}{\alpha f_b},$$

(38)

where the time for undergoing one IC, $\tau_m$, is given by Equation (7). For example, for $a_m = 10^5$ the decay time is $t_d \propto \sqrt{\omega f_b/(nm^2 f_b)}$, where we have used Equation (4). One consequence of this model is that the decay rate, and thereby the number of 1G objects $N_1$ after a Hubble time, depends exponentially on the binary fraction $f_b$. The binary fraction is at the moment unknown observationally, but numerical simulations of GCs using Monte Carlo techniques have shown that it very likely stays constant with only a small scatter around 1%–5% (see, e.g., Figure 2 in Samsung et al. 2020). As a result, a significant fraction of present-day GCs likely have many of their 1G objects left in their core, where the remaining fraction have lost their BHs through binary-single “evaporation.” This “evaporation effect” will lead to a characteristic change in BBH merger rates as a function of redshift, similar to what is found for the set of GCs that “evaporates” through tidal heating or direct tidal disruptions (e.g., Fragione & Kocsis 2018). Considering now $N_2$ for our model, we see that at early times $N_2 \approx N_1(0)f_b P_M N_e$, where we have used $\exp(-ax) \approx 1 - ax$. This is expected, as this simply equals the number of mergers per IC evaluated for the initial $N_1(0)$ population $(N_1(0)f_b P_M)$.
times the number of ICs \(N_c\). Note that this is similar to Equation (19), where we studied how effective a population consisting of a single binary \("1 = N(0)f_b\)" is at growing a 2G population. As \(N_c\) increases toward infinity, the \(N_2\) population reaches a maximum “freeze-out value”, \(\max(N_2)\), given by

\[
\max(N_2) = N(0)(\beta/\alpha), \quad N_c \to \infty,
\]

which interestingly does not depend on the binary fraction, although how fast \(N_2\) reaches \(\max(N_2)\) does. As seen, within a factor of unity, \(\max(N_2)\) is simply given by the total number of binary mergers one would get if one turned the initial \(N(0)\) population into a total of \(N(0)/2\) binaries. Finally, if we now consider the number of 2G objects relative to 1G objects, one finds using Equation (37) that

\[
N_2/N_1 = (\beta/\alpha)[\exp(\alpha f_b N_c) - 1].
\]

We see here that this ratio always increases, i.e., in this case there is no “freeze-out” value. This of course originates from the fact that \(N_1\) keeps decreasing, whereas \(N_2\) keeps increasing until it asymptotically reaches its value \(\max(N_2)\). Considering the limit where \(N_2/N_1 = 1\), we can solve for the corresponding characteristic \(N_c\) scale, denoted here by \(N_c^{\text{EFL}}\),

\[
N_c^{\text{EFL}} = \frac{\ln(1 + \alpha/\beta)}{\alpha f_b},
\]

which equals the number of IC cycles, or time \(t_c^{\text{EFL}} \approx N_c^{\text{EFL}} \times \tau_m\), it takes for \(N_2\) to be similar to \(N_1\).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6.png}
\caption{Results from our considered case \(cB\) described in Section 3.2.1, where the number of 1G and 2G objects as a function of time is given by Equation (37), and their ratio \(N_2/N_1\) by Equation (40). We here consider solutions to \(t = t_H\) for a model described by \(f_b = 0.01, R_f^0 = 1, N_f^0 = 6, \) and \(N_f^0 = 4,\) where the upper and lower plots correspond to \(m = 30 M_\odot\) and \(m = 1.4 M_\odot\), respectively. The colored contours show values of \(N_2/N_1\), where the area covered by the red line contours highlights where \(0.01 < N_2/N_1 < 1\), i.e., it is the region that gives rise to both consistent \((<1.0)\) and interesting \((>0.01)\) outcomes for growing a 2G population. In the yellow “+” hatched area, our formalism evaluated at \(t = t_H\) breaks down as \(N_2\) is here > \(N_1\), in the gray area our \(N_2\) averaging approach breaks down as \(N_1\) is here <1. In the yellow “X” hatched area \(N_1/N_0(0) < 10^{-4}\); therefore, if a system is located within this area it will “evaporate” within a Hubble time if its initial number of BHs is \(\leq 10^4\). The “+” hatched area is where \(N_2/N_1 < 0.01\) and \(N_1 > 1\), and highlights therefore systems that clearly undergo several ICs, but still end up with a relative small 2G population. The green separation lines are described in Section 3.2.1. Results related to this figure are described in Section 3.2.2.}
\end{figure}

\subsection*{3.2.2. Upper Limits on 2G Objects}

Figure 6 shows results related to the ratio \(N_2/N_1\) given by Equation (40) evaluated at \(t = t_H\), as further described in the figure caption. As described in Section 3.2.1, this case represents in our model an upper limit on \(N_2/N_1\). Considering first the upper plot showing the \(m = 30 M_\odot\) case, we see that for a GC with \(V_f \sim 10 \text{ km s}^{-1}\) a population of 1G objects can over a Hubble time turn into a population with \(N_2/N_1 > 0.1\) if \(n \geq 10^4 \text{ pc}^{-3}\). Although this is an upper limit, it greatly
illustrates that the length, mass, and timescales associated with a typical cluster hosting BHs of mass \(\sim 30M_\odot\) in the core is able to populate the upper mass gap through successive mergers of its 1G population. Considering now the lower plot showing results for the 1.4\(M_\odot\) case, we see that for \(v_d \sim 10\) km s\(^{-1}\) the density has to be \(\gtrsim 10^8\) pc\(^{-3}\) to even grow a 2G population with \(N_2/N_1 > 0.01\), and \(\gtrsim 10^9\) pc\(^{-3}\) for \(N_2/N_1 > 0.1\). This further hints that populating the lower mass gap in clusters by NS–NS mergers is highly challenging. In addition, we have for our presented upper limit estimates assumed that the entire CO population consists of only NSs. However, this assumption is highly optimistic as NSs will not segregate and form their own subcluster in the same way as BHs, as the NS mass is close to that of the ordinary stars in the cluster (e.g., Ye et al. 2020). As a result, NSs will exchange and interact frequently with the stellar population, which introduces “impurities” in the IC illustrated in Figure 1. The probability that two NSs merge inside the cluster is therefore significantly smaller than what we have assumed in our considered cB scenario. In Figure 7 we show how these results depend more broadly on the mass \(m\), where we show \(N_2/N_1\) from case cB as a function of \(m\) for \(n = 10^7\) pc\(^{-3}\) (top plot) and \(n = 10^5\) pc\(^{-3}\) (bottom plot), and two different binary fractions, as further described in the figure caption.

Finally, we note that the real “bottle neck” in populating the lower mass gap is not directly related to the probability \(P_m\) per IC for a NS population to undergo NS–NS mergers inside their cluster. Instead, it is the time it takes for a NS binary to undergo one IC, \(\tau_m\), that simply is too long for a standard cluster. This is clear from Figure 6, as the gray area, where \(N_c \sim 1\), sets the lower limit at \(n = 10^5\) pc\(^{-3}\) for 10 km s\(^{-1}\). In the limit where \(a_m = a_e\), the number of ICs evaluated at \(t_H\), 

\[N_c(t_H) \approx t_H \left[ \frac{\pi G^2 \Delta^2}{\delta^2 f_{\Delta}} \right] \frac{nm^2}{v_d^3}\]  

\[\approx 0.8 \left( \frac{n}{10^5\text{ pc}^{-3}} \right) \left( \frac{m}{1.4M_\odot} \right)^2 \left( \frac{v_d}{10\text{ km s}^{-1}} \right)^{-3},\]  

and is indeed just around unity for NS–NS binaries for our chosen normalizations. It is furthermore seen that \(N_c(t_H)\) rapidly decreases with mass \(m\) as \(\propto m^2\). However, as seen in Figure 3, if the system is in the area for which \(N_c(t_H) > 1\), the number of in-cluster mergers a given binary can produce within a Hubble time, \(N_m(t_H)\), is less sensitive to \(m\), as \(N_m(t_H) \propto P_m \propto m^{8/7}\). All in all, the limit for which \(N_c(t_H) = 1\) plays therefore a crucial role in determining which systems are able to produce a significant 2G population. We conclude our study below.

4. Conclusions

We have in this paper studied the formation of 2G objects formed through 1G–1G in-cluster mergers in dense clusters. We have in particular explored the possibility of populating the LMG (3–5\(M_\odot\)) and the UMG (\(\gtrsim 45M_\odot\)) through the mergers of BNSs and BBHs, respectively. Understanding which cluster systems are able to populate these two mass gaps has wide implications for both GW astrophysics and stellar physics. For example, if nature is proven not to be able to create mass-gap BHs through normal stellar evolution, then current and future measures of the BH mass spectrum, through, e.g., GW observations, will give us insight into the formation mechanisms of BBH mergers in clusters. That said, if observations hint that stellar clusters do not contribute significantly to the observed GW merger rate, e.g., through independent measures of the fraction of eccentric BBH mergers (e.g., Samsing 2018), then an observed population of mass-gap objects will hint that our single stellar models need to be revised. For these reasons, several new studies have discussed the possibility of dynamically populating these mass gaps (e.g., O’Leary et al. 2016; Fishbach et al. 2017; Gerosa & Berti 2017; Yang et al. 2019; Antonini et al. 2019; Gerosa & Berti 2019; Samsing & Illan 2019; Rodriguez et al. 2019; Gerosa et al. 2020; Safarzadeh et al. 2020; Gayathri et al. 2020; Kimball et al. 2020; Doctor et al. 2020; Baibhav et al. 2020).
Through a fully analytical approach we have here studied how efficient a cluster, described by a constant $v_{\text{th}}$, can turn its initial population of $N_1$ 1G objects into a sizable population of $N_2$ 2G objects through in-cluster GW mergers. We have in particular explored the upper limit on the ratio $N_2/N_1$ evaluated after a Hubble time, as a function of $v_{\text{th}}$, $n$, and $m$ (Section 3.2.2). Our limit is based entirely on dynamics and complements therefore greatly the recent study by Gerosa & Berti (2019), where the limit was derived from considering the magnitude of GW kicks. From our analysis we have reached the following conclusions.

Populating the LMG through in-cluster mergers of NSs is a very slow process for any relevant astrophysical cluster; for example, even in the highly idealized case of a GC core populated entirely by NSs, the number density $n$ has to stay at values $>10^6$ pc$^{-3}$ over an entire Hubble time to reach $N_2/N_1 \sim 0.1$ as shown in Figure 6. As discussed in Section 3.2.2, not only is this density much higher than what is found for real clusters, but also NSs are likely to mix with other stars due to their similar mass, which reduces their in-cluster merger probability further. In fact, our results show that what really limits an NS-rich core in undergoing enough in-cluster mergers to populate the LMG is actually the timescale for interactions, and not how the NSs exactly merge inside their cluster. This is seen in Figure 6, where for an NS-dominated core (bottom plot) a density of $n \gtrsim 10^5$ pc$^{-3}$ for $v_{\text{th}} \sim 10$ km s$^{-1}$ is required to move above the gray area, i.e., for a BNS to undergo at least 1 IC. In our described “standard picture” of dynamically assembled in-cluster mergers (Section 2.1), an efficient production of LMG objects is therefore highly unlikely. If clusters for some reason are still observed to effectively produce LMG objects through dynamics, then more “exotic” dynamical pathways have to be evoked. Alternatively, it could be that some clusters start out with a high BNS fraction (see Figure 7) that would lead to a relative high number of 2G objects after a Hubble time. However, in that case, there would still be problems related to how fast this 2G population can be dynamically paired up with other COs to undergo observable GW mergers. Therefore, observing GW sources with at least one LMG object formed in a cluster near the gray area in Figure 6 (bottom) seems therefore highly unlikely.

Populating the UMG is in comparison much easier, e.g., in Figure 6 (top) it is clearly seen that reaching values of $N_2/N_1 \sim 0.1$ only requires clusters with a central density of $\sim 10^4$ pc$^{-3}$. This is much more reasonable, which leads us to conclude that populating the UMG in clusters is possible, at least dynamically, without introducing any nonstandard pathways. This effective change of the black hole mass function through in-cluster GW mergers will happen as a function of cosmic time. This can, e.g., be probed with third-generation GW observatories with great implications also for constraining the formation and evolution of stellar clusters (e.g., Romero-Shaw et al. 2021).

Finally, we note that a few studies that were completed while our present study was underway point toward similar conclusions. For example, in Ye et al. (2020) it was shown using a fully numerical approach that the rate of BNS mergers originating from GCs is low, while both Rodriguez et al. (2019) and Baibhav et al. (2020) illustrated that populating the UMG definitely seems possible. However, other studies keep the question open as to what degree the LMG can be populated in clusters (e.g., Gupta et al. 2020). We have in our work presented the first set of closed form solutions that encapsulate the correct scalings and relations of the problem. We are currently working on a self-consistent hybrid scheme involving both analytical and numerical techniques that will enable us to evolve clusters with a more realistic mass distribution.

The authors thank the Yukawa Institute for Theoretical Physics at Kyoto University and the organizers of the workshop YKIS2019 “Black Holes and Neutron Stars with gravitational Waves,” where many useful conversations took place. It is also a pleasure to thank Kyle Kremer for enlightening discussions. J.S. acknowledges support from the Lyman Spitzer Fellowship and the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska–Curie grant agreement No. 844629. K.H. acknowledges support from the Lyman Spitzer Fellowship.
