Hunting for the fastest stars in the Milky Way

Marchetti, T.

1 Introduction

"Quod tertio loco a nobis fuit observatum, est ipsiusmet LACTEI Circuli essentia, seu materies, quam Perspicilli beneficio adeo ad sensum licet intueri, ut et altercationes omnes, quæ per tot sæcula philosophos excruciarunt, ab oculata certitudine dirimantur, nosque a verbosis disputationibus liberemur. Est enim GALAXIA nihil aliud, quam innumerarum Stellarum coacervatim consitarum congeries: in quamcumque enim regionem illius Perspicillum dirigas, statim Stellarum ingens frequentia sese in conspectum profert, quarum complures satis magnæ ac valde conspicuæ videntur; sed exiguarum multitudo prorsus inexplorabilis est."
(Galileo Galilei, Sidereus Nuncius, 1610)

"What was observed by us in the third place is the nature or matter of the Milky Way itself, which, with the aid of the spyglass, may be observed so well that all the disputes that for so many generations have vexed philosophers are destroyed by visible certainty, and we are liberated from wordy arguments. For the Galaxy is nothing else than a congeries of innumerable stars distributed in clusters. To whatever region of it you direct your spyglass, an immense number of stars immediately offer themselves to view, of which very many appear rather large and very conspicuous but the multitude of small ones is truly unfathomable." [1]

It was the year 1610 when, using the telescope he constructed, Galileo Galilei first showed that the bright band on the sky whose origin and composition fascinated ancient cultures is a collection of multiple stars, whose majority cannot be resolved by the naked eye. This was the first step towards a modern scientific approach to the study of the Milky Way (MW), the Galaxy we are living in. Today, with the help of large ground- and space-based tele-

[1] English translation from Albert Van Helden, University of Chicago Press, Chicago, Illinois, 1989.

scopes, we have made huge steps forward to understand our Galaxy, but still, we are far from a comprehensive, complete and self-consistent picture, and many questions are still open. What are the accretion and evolutionary history of the MW? How do stars behave in the proximity of the central massive black hole (MBH)? What is the shape and extent of the dark matter halo? More than four hundred years after Galileo Galilei's breakthrough discovery, we are still looking up staring at the night sky, building new telescopes and satellites to better understand our Galaxy. In the light of these open questions, we present here our work on searching for the fastest objects in the MW: stars whose speed is so high that they are flying away from it on unbound trajectories. We show how these remarkable objects can help us decipher the Galaxy, giving us insights into its structure, its building components, and on some of its most energetic phenomena.

The MW is a barred spiral galaxy, and our Sun is only one of the hundreds of billions of stars orbiting inside it. Jan Oort in 1927 first discovered that the majority of these stars rotate coherently around the Galactic Centre (GC) in the shape of a flattened disk (Oort 1927). The principal stellar components of the Galaxy are the central box/peanut bulge, the stellar disk (composed of the thin and thick disks), and a diffuse stellar halo. The MW is embedded in a vast dark matter halo, which constitutes most of the MW mass, and extends up to hundreds of kpc from the GC (Bland-Hawthorn & Gerhard 2016).

Thanks to the exquisite quality of the recent imaging of the centre of the galaxy M87 with the Event Horizon Telescope (Event Horizon Telescope Collaboration et al. 2019), it has been definitely proven that MBHs exist at the centre of galaxies. In our MW, the location of the MBH coincides with the radio source Sagittarius A* (often abbreviated as Sgr A*, Balick & Brown 1974; Reid et al. 2009). The Sun is located at a distance of 8.127 kpc from the GC (Gravity Collaboration et al. 2018). Observations have shown the presence of several dozens of main sequence B-type stars orbiting around Sgr A*, the so-called S stars (Ghez et al. 2003). The orbits of these stars represent the best proof for the existence of our MBH (and provide tight constraints on the enclosed mass, Gillessen et al. 2009, 2017). S stars challenge our knowledge of how stars form in this extreme environment: the tidal forces of the MBH are predicted to be too strong to permit star formation within 1 arcsecond of the GC (Morris 1993).

In this introduction, we will discuss, among others, how high velocity stars can provide valuable information on the dynamics and origin of S stars, and how they can constrain global properties of the MW. This chapter

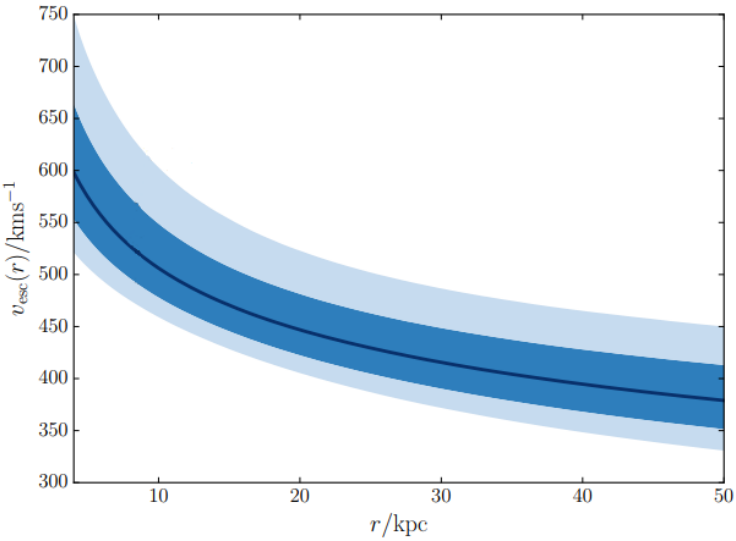

Figure 1.1: Escape speed from the Galaxy as a function of Galactocentric distance. Adapted from Williams et al. (2017).

is organized as follows. In Section 1.1 we introduce the two main classes of high velocity stars that will be studied in this thesis: runaway stars and hypervelocity stars. We will give a theoretical introduction to the acceleration mechanisms, and we will present the current status of the observations. Section 1.2 gives an overview of the European Space Agency (ESA) satellite *Gaia*, which has provided the largest stellar catalogue of the Galaxy ever produced. We use this dataset in three chapters of this thesis. In Section 1.3 we will cover the main methods used in this thesis. Finally, Section 1.4 provides an overview of the content of each of the following scientific chapters.

1.1 High velocity stars

Fast moving stars are intriguing for several reasons. The mechanisms leading to the acceleration of a star above its original velocity can give insights into multiple astrophysical processes, including but not limited to stellar and binary evolution, dynamics in the proximity of (massive) compact objects, and mergers between galaxies. In this Section we will introduce the main classes of high velocity stars. Typical velocities of stars can be com-

pared to the *escape speed* from the Galaxy, which defines the minimum velocity that a star needs to have in order to be *unbound* from the MW. Fig. 1.1 shows a recent result from Williams et al. (2017), showing the derived escape speed across a range of ~ 50 kpc from the GC, inferred using a variety of different kinematic tracers. The value at the Sun position is found to be 521^{+46}_{-30} km s^{-1}, falling to ~ 380 km s^{-1}at a Galactocentric distance of 50 kpc. In a more recent study, Monari et al. (2018) find a slightly higher value at the Sun position, 580 ± 63 km s^{-1}. A measurement of the escape speed can be converted into an estimate of the total mass of the MW (e.g. Smith et al. 2007; Piffl et al. 2014; Monari et al. 2018).

If we consider the encounter of two individual stars, the highest speed that can result is set by the escape velocity from their surface, since higher velocities would require the two stars to orbit at a distance smaller than their physical size (Leonard 1991):

$$v^*_{\rm esc} = \sqrt{\frac{2Gm_*}{r_*}} \simeq 618 \left(\frac{m_*}{M_\odot}\,\frac{R_\odot}{r_*}\right)^{1/2} \text{km /s}, \tag{1.1}$$

where G is the gravitational constant, m_* is the mass of the star, and r_* is its radius. Because of the approximately linear relation between m_* and r_* for stars on the main sequence, it follows that $v^*_{\rm esc} \simeq 600$ km s^{-1}in the mass range $m_* \in [0.4, 4]\ M_\odot$. Higher velocities can be achieved for compact objects such as white dwarfs and neutron stars. It turns out that equation (1.1) is an overestimate of the value of the escape velocity from a star: more precise calculations including binary evolution and mass transfer result into lower values of $v^*_{\rm esc}$.

1.1.1 Runaway stars

The term *runaway star* has first been coined by Blaauw (1961) to refer to the young, O and B-type stars observed out of the Galactic plane. Two main mechanisms have been introduced to predict the excess of velocity with respect to the Galaxy at their location. Blaauw (1961) proposed that runaway stars form as the result of a supernova explosion in a binary system. The more massive star in the binary evolves faster, transferring mass to the companion. When the donor explodes as a supernova, it can eject the companion star with a high velocity, forming a runaway star. The other proposed mechanism is dynamical encounters between stars in a dense stellar system (Poveda et al. 1967). In systems such as a young open cluster, interaction between binaries can lead to the ejection of one star from the cluster.

Tracing back the orbit of known runaway star candidates to their natal cluster, both these mechanisms have been observed to take place in the MW (Hoogerwerf et al. 2001). Maximum ejection velocities for both channels are typically $\lesssim 300 - 400$ km s^{-1}(e.g. Leonard & Duncan 1990; Portegies Zwart 2000; Przybilla et al. 2008; Gvaramadze et al. 2009; Renzo et al. 2019), even if values up to ~ 1000 km s^{-1}are possible (Leonard 1991; Tauris 2015), but should be extremely rare for runaway stars (Brown 2015).

1.1.2 Hypervelocity Stars

The first observation of a hypervelocity star

With previous results for the ejection velocity of runaway stars in mind, it was a great surprise when, in 2005, a B-type star was observed in the outer halo of the MW with a heliocentric radial velocity of ~ 830 km s^{-1}(Brown et al. 2005, 2014). This value, once corrected for the motion of the Sun and the local standard of rest (LSR), corresponds to a lower limit on the total velocity of the star of 673 km s^{-1}(Brown et al. 2014), which is sufficiently high to escape the gravitational field of the MW at the star's position. The authors, targeting blue horizontal branch stars to trace the stellar halo, found this star to be a 6σ outlier from the radial velocity distribution. This *unbound* star, SDSS J090745.0+024507, is the first hypervelocity star (HVS) observed, and was referred to as HVS1. As a hint of its puzzling origin, the radial velocity vector of HVS1 points at $\sim 175°$ from the GC, suggesting an origin in the central region of our Galaxy. This intriguing possibility will now be further discussed.

The Hills mechanism

One possible way to explain the surprising velocity of HVS1 involves the interaction with a massive compact object. According to the Hills mechanism, the tidal field of the MBH in the centre of our Galaxy can disrupt a binary system passing sufficiently close (Hills 1988). This results in one of the stars starting to orbit around the MBH, with the other one being ejected with an incredibly high velocity, of the order of thousands of km s^{-1}. Following Brown (2015), we will now derive with a simple calculation an estimate of the ejection velocity of the HVS, showing how the Hills mechanism can easily explain the acceleration of stars to unbound velocities.

A stellar binary with total mass m_b and semi-major axis a gets disrupted by the gravitational field of a MBH of mass M, if the encounter happens

at a distance closer than the *tidal radius* $r_\bullet$. This characteristic distance is defined as the distance within which tidal forces from the MBH dominate over the binary binding force:

$$r_\bullet = a\left(3\frac{M}{m_b}\right)^{1/3} \simeq 14\,\mathrm{AU}\left(\frac{a}{0.1\,\mathrm{AU}}\right)\left(\frac{M_\odot}{m_b}\right)^{1/3}\left(\frac{M}{10^6\,M_\odot}\right)^{1/3}. \tag{1.2}$$

We can compare this characteristic scale to the Schwarzschild radius of a MBH:

$$r_{\mathrm{MBH}} = \frac{2GM}{c^2} \simeq 0.02\,\mathrm{AU}\left(\frac{M}{10^6 M_\odot}\right), \tag{1.3}$$

where c is the speed of light. We can see that, for MBHs with $M > 10^8\ \mathrm{M}_\odot$, stars fall inside the event horizon before reaching the tidal radius (Hills 1988). This is not the case in our Galaxy, where $M \simeq 4.3\cdot 10^6\ \mathrm{M}_\odot$ (Gillessen et al. 2017).

The typical orbital velocity of stars in an equal mass binary is:

$$v_b = \sqrt{\frac{Gm_b}{a}} \simeq 94\ \mathrm{km\ s^{-1}}\left(\frac{m_b}{M_\odot}\right)^{1/2}\left(\frac{0.1\ \mathrm{AU}}{a}\right)^{1/2}. \tag{1.4}$$

For example, $v_b \simeq 100$ km s^{-1}for a binary consisting of two 3 $\mathrm{M}_\odot$ stars at $a = 0.5$ AU. At the moment of the disruption of the binary, the binary orbital velocity is:

$$v = \sqrt{\frac{GM}{r_\bullet}} = v_b\left(\frac{M}{m_b}\right)^{1/3} \simeq 10^4\ \mathrm{km\ s^{-1}}. \tag{1.5}$$

This velocity is equal to few percent of the speed of light, and is consistent with observations of S stars in the GC (see for example Ghez et al. 2005).

When the binary gets disrupted, the stars experience a change in specific kinetic energy δE that we can compute as:

$$\delta E = \frac{1}{2}(v + v_b)^2 - \frac{1}{2}v^2 \simeq v v_b. \tag{1.6}$$

Using energy conservation, we can therefore estimate the resulting velocity of the star ejected from the binary as:

$$v_{\mathrm{ej}} = \sqrt{2 v v_b} \simeq 10^3\ \mathrm{km\ s^{-1}}. \tag{1.7}$$

Equation (1.7) shows that the Hills mechanism is able to predict ejection velocities in the GC up to thousands of km s^{-1}. These incredibly high velocities allow HVSs to travel across the whole MW on unbound trajectories.

Besides explaining the extreme velocities of the observed HVS, the Hills mechanism also provides a possible solution to the puzzling origin of the S stars in the GC: these stars are the binary companions of the ejected HVS, bound to the central MBH after the disruption. The observed orbit and eccentricity distributions of S stars are consistent with predictions from the Hills mechanism (Gillessen et al. 2009; Madigan et al. 2014), if the relaxation time is shorter than the stellar age (Habibi et al. 2017).

After being ejected in the GC, HVSs travel through the Galaxy on almost radial trajectories. A star moving at ~ 1000km s^{-1}travels a distance of ~ 1 kpc in ~ 1 Myr, a small fraction of the typical main sequence lifetime of a star. The initial velocity $v_{\rm ej}$ will then decrease because of the deceleration induced by the Galactic potential, which acts as a high-pass filter: only the stars with sufficiently high velocity at the ejection can travel to distances large enough to be observable (Kenyon et al. 2008). For example, stars with $v_{\rm ej} > 700$ km s^{-1}can reach the Sun position, stars with $v_{\rm ej} > 800$ travel to the edge of the stellar disk, and only stars with $v_{\rm ej} > 800$ can get to the virial radius of the MW, around 250 kpc from the GC. The radial motion of HVSs is deflected by the non-spherical components of the Galactic potential, namely the stellar disk, a possible triaxiality of the dark matter halo, and the presence of satellite galaxies orbiting the MW (Kenyon et al. 2018).

In addition to the population of unbound HVSs, the Hills mechanisms naturally predicts the existence of *bound HVSs*: stars ejected according to the same three-body interaction in the GC, but with an initial velocity not sufficient to escape from the gravitational field of the whole MW (Bromley et al. 2006; Kenyon et al. 2008). The trajectories of these stars do not follow straight lines anymore, and they can cross the stellar disk multiple times during their lifetime.

HVS observations

Following the first detection, a dedicated spectroscopic survey with the MMT telescope was performed to find HVS candidates (Brown et al. 2014). The survey targeted young stars in the outer halo of the MW, which are not expected to be found so far from an active star forming region (such as the GC), unless they traveled there with an extremely high velocity. The survey identified 21 unbound late B-type HVSs, with masses in the range $[2.5,\ 4]$ $M_\odot$, at distances $50 - 120$ kpc from the GC. All these stars are unbound

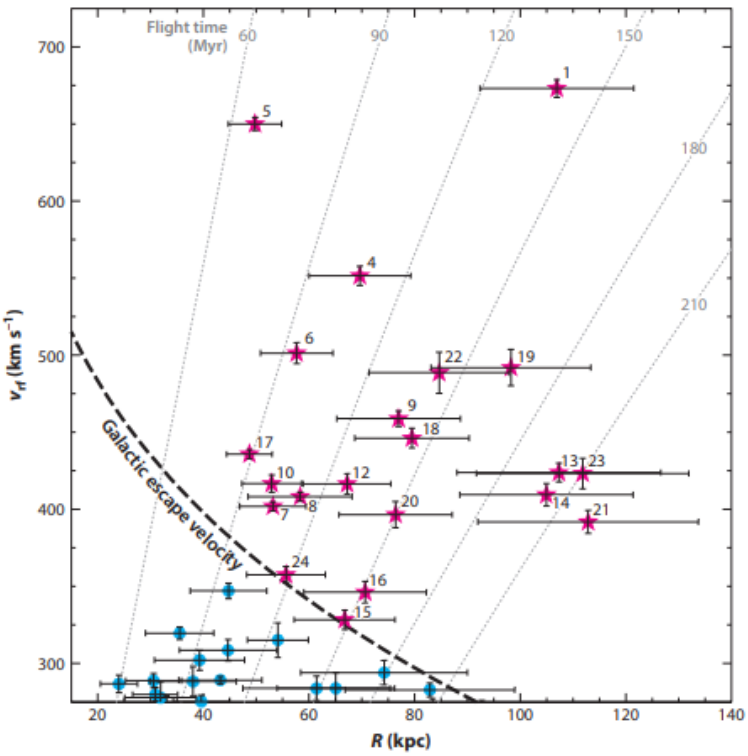

Figure 1.2: Total velocity as a function of Galactocentric distance for the HVS candidates discovered in the outer halo of the MW by the MMT HVS survey. Magenta stars mark the unbound candidates, while blue dots the bound ones. The dashed line marks the escape velocity from the Galaxy. From Brown (2015).

from radial velocity alone, and are moving outward (consistent with the prediction from the Hills mechanism). Fig. 1.2 shows the total velocity in the Galactic rest-frame as a function of distance from the GC for the stars found in the survey (Brown 2015). The dashed line is a choice for the escape speed from the Galaxy (Kenyon et al. 2008). Magenta stars are the unbound HVSs, while blue dots are the bound HVS candidates.

In addition to the population of young stars in the outer halo, many works focused on finding late-type, low mass HVS candidates in the Solar neighbourhood and the inner Galactic halo. For example, Palladino et al. (2014) discovered 20 HVS candidates in the G and K samples of the Sloan Extension for Galactic Understanding and Exploration (SEGUE), and Li et al. (2015) found 19 F, G, and K type candidates using LAMOST data. Most of the known late-type HVSs are likely to be bound to the MW, or not to originate from the GC (e.g. Zheng et al. 2014; Hawkins et al. 2015; Ziegerer et al. 2015, 2017; Boubert et al. 2018). Chemical tagging with high resolution spectroscopy can help to narrow down the ejection location of HVS candidates, by determining their precise chemical composition (e.g. Hawkins & Wyse 2018).

The search for HVSs is complicated by the fact that HVSs are extremely rare objects, with an ejection rate from the GC between 10^{-5} and 10^{-4} yr^{-1} (Brown et al. 2015). The advent of new astrometric and spectroscopic surveys will change dramatically our view on the fastest stars in our Galaxy (see Section 1.2).

Alternative ejection mechanisms for HVSs

In addition to the Hills mechanism, discussed in Section 1.1.2, other ejection scenarios have been proposed to explain the unbound velocities of observed HVSs. Yu & Tremaine (2003) first discussed the chance that HVSs could be ejected following the interaction between a single star and a massive black hole binary (BHB) in the GC. The possibility of a intermediate mass black hole orbiting around Sagittarius A^* cannot be excluded by observations in the GC, with current upper limits on its mass around 10^4 $\mathrm{M}_\odot$ (Gillessen et al. 2017). The presence of a fixed, preferential plane in the geometry of the encounter (the plane of the BHB) introduces an anisotropy in the expected spatial distribution of HVSs, which is flattened along the inspiral plane of the BHB. The degree of flattening is expected to decrease as the BHB hardens, leading to a more isotropic distribution (Sesana et al. 2006). HVSs produced by these mechanisms might be slower compared to the Hills mechanism, depending on the system parameters (Rasskazov

et al. 2019).

Recently it has been proposed that the known B-type HVS could be runaway stars ejected from the Large Magellanic Cloud (LMC), the most massive satellite galaxy orbiting the MW (Boubert et al. 2017a). The LMC is an active star forming region, so runaway stars ejected from supernova explosions in binary systems, summing their velocity to the orbital velocity of the LMC, can easily become unbound to the Galaxy. Recently, a HVS has been shown to originate almost from the centre of the LMC (Erkal et al. 2019), suggesting the presence of a MBH (Boubert & Evans 2016).

Other proposed mechanisms to produce HVSs include tidal interaction between dwarf galaxies infalling in the gravitational field of the MW (Abadi et al. 2009), which might accelerate stars to unbound velocities. Also, massive globular clusters infalling towards the centre of the Galaxy, interacting with the MBH or with a BHB, can produce a population of high velocities stars, with an unbound tail (Capuzzo-Dolcetta & Fragione 2015; Fragione & Capuzzo-Dolcetta 2016). Another possibility is the scatter between single stars and stellar black holes in the proximity of Sgr A* (O'Leary & Loeb 2008).

Different mechanisms predict different spatial and velocity distributions, therefore a large sample of HVSs can be used to investigate the dynamical processes responsible for the acceleration of these stars to unbound velocities.

HVSs as tools to investigate the Milky Way

HVSs are a unique probe to study our Galaxy as a whole. HVSs are predicted to originate in the centre of the MW, and then, because of their extremely large velocities, travel through the Galaxy on unbound trajectories. Therefore they provide a connection between the inner center and the outskirts of the Galaxy. The GC is difficult to observe because of dust extinction and stellar crowding, so HVSs can be used to probe the stellar population in the proximity of the quiescent MBH. A large sample of HVSs, for example, can be used to constrain the mass function and metallicity distribution in the inner parsec of the Galaxy. On the other hand, HVS trajectories are affected by the way the mass is distributed in the MW, therefore they can be used as probes of the Galactic Potential (e.g. Gnedin et al. 2005; Sesana et al. 2007; Yu & Madau 2007; Perets et al. 2009). In particular, the mass and orientation of the halo are still a matter of debate, and there is no general consensus on its shape (e.g. Wang et al. 2015; Bovy et al. 2016; Posti & Helmi 2019). Gnedin et al. (2005) first proposed HVSs to study the dark

matter halo of the MW. The authors show how precise proper motions of the first HVS candidate, SDSS J090745.0+024507, can provide constraints on the triaxiality of the halo, as predicted from cosmological simulations of structure formation. A recent work from Contigiani et al. (2019) shows how a sample of ~ 200 HVSs can be used to nail down the Galactic halo potential parameters with percent precision. In particular, HVSs are found to be extremely sensitive to the axis-ratio of the spheroidal, because of the spherical symmetry of the ejection in the Hills mechanism. A joint constraint on both the GC and the dark matter halo was first performed by Rossi et al. (2017), but tight constraints have been hampered by the low number of known HVSs. Recently, HVSs have also been proposed to constrain the Solar parameters, relying on the condition of zero azimuthal angular momentum (Hattori et al. 2018b).

1.2 The ESA mission *Gaia*

The ESA satellite *Gaia* was launched on 9 December 2013 from the European spaceport in French Guiana, and a few weeks later it arrived at the Lagrangian point L2 for a planned 5 years operations (Gaia Collaboration et al. 2016b). The goal of *Gaia* is to provide the largest three dimensional stellar catalogue ever produced of the Galaxy, providing positions, parallaxes, and proper motions for more than 1 billion sources, and radial velocities for a subset of bright stars. Here we outline the main contents of its data releases.

1.2.1 The first *Gaia* data release

The first data release (DR1) of the ESA satellite *Gaia* was delivered to the general public on the 14th of September 2016, and is based on observations collected between the 25th of July 2014 and the 16th of September 2015, for a total of almost 14 months (Gaia Collaboration et al. 2016b,a). Here we summarize the main contents of *Gaia* DR1:

- Coordinates (right ascension α and declination δ) and magnitudes in the *Gaia G* band for 1142679769 sources;
- The five parameters astrometric solution (positions, parallax ϖ, and proper motions μ_α, μ_δ) for 2057050 sources.

The presence of parallaxes and distances for more than 2 million stars was possible thanks to a joint Tycho-Gaia astrometric solution (TGAS),

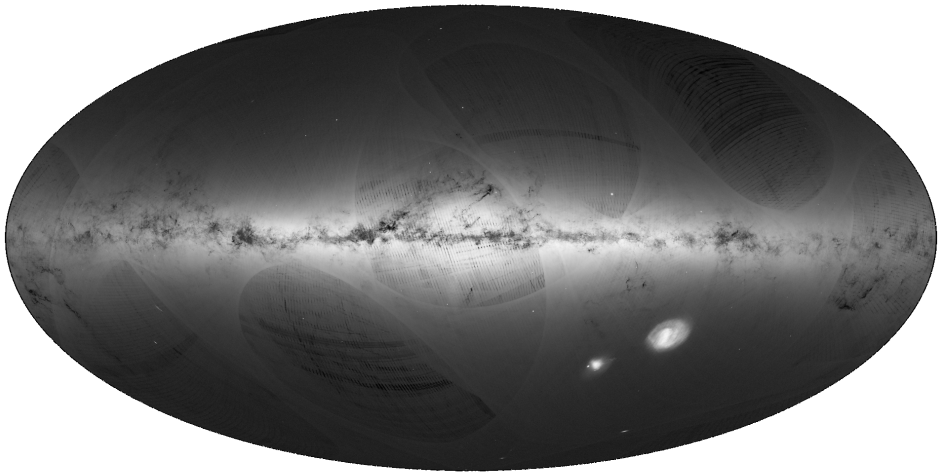

Figure 1.3: First full sky map released by *Gaia*, using data from DR1 (credits: ESA).

performed on the sources in common between *Gaia* and the Tycho-2 Catalogue (Michalik et al. 2015; Lindegren et al. 2016).

Fig. 1.3 shows the first full sky map made using data from *Gaia* DR1. A quick look at the map reveals the presence of characteristic arches and patterns in the density distribution. Those are a unique imprint of the *Gaia* scanning strategy on the sky, and disappeared in future data releases.

1.2.2 The second *Gaia* data release

The second data release (DR2) of *Gaia* happened on the 25th of April 2018, containing observations collected between the 25th of July 2014 and the 23rd of May 2016, spanning a period of 22 months (Gaia Collaboration et al. 2018a). DR2 represents a huge improvement over DR1, both in terms of number of sources observed, and of quality of the measurements. It contains:

- Position and *Gaia G* band magnitude for 1692919135 stars;
- Magnitudes in the *Gaia* blue pass (BP) $G_{\rm BP}$ and red pass (RP) $G_{\rm RP}$ band for 1381964755 and 138551713 sources, respectively;

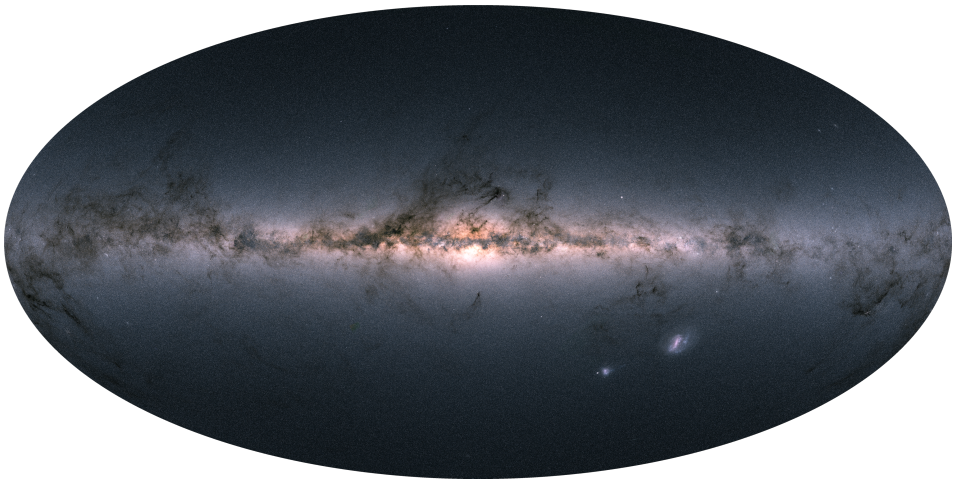

Figure 1.4: Full sky map released by *Gaia* DR2 (credits: ESA).

- The five parameters astrometric solution for 1331909727 sources;
- Radial velocity for 7224631 stars with $4 \lesssim G \lesssim 13$ and with effective temperatures $3550 \lesssim T_{\rm eff} \lesssim 6900$ K;
- Effective temperature for 161497595 stars;
- Extinction and reddening for 87733672 objects;
- Radius and luminosity for 76956778 sources.

Figure 1.4 shows the full sky map for the ~ 1.7 billion sources in *Gaia*, obtained combining the magnitudes in the G, $G_{\rm BP}$ and $G_{\rm RP}$ passbands. Comparing this to Figure 1.3 shows how all the arches due to the scanning law of the satellite have now disappeared, thanks to the longer baseline and more homogeneous sky coverage.

1.2.3 Future *Gaia* data releases

The third data release (DR3) of *Gaia* is currently planned to be split into two different releases. An early data release (EDR3) is expected in the third quarter of 2020 and will contain updated parallaxes and proper motions, with uncertainties reduced by the longer baseline (34 months of data). *Gaia* DR3 is expected in the second half of 2021 and will contain astrophysical parameters and radial velocities for all the spectroscopically well behaved

sources. The final *Gaia* data release, which has not been announced yet, will consist of the full photometric, astrometric, and radial velocity catalogues[2]. This will be the largest and most precise stellar catalogue ever produced and will allow understanding the history of the MW and its stellar population with unprecedented detail.

1.2.4 Warnings and caveats while using *Gaia* data

Gaia is the largest stellar catalogue ever produced, and the most recent data release (DR2) has provided astrometric measurements for more than 1.3 billion sources. There are known issues with *Gaia* astrometry and radial velocities, which have not been corrected for during the raw data reduction. Taking this into account while analyzing the data is essential. As an example, a possible wrong determination of the parallax can severely affect the distance determination, and therefore the total velocity of a star. Lindegren et al. (2018a) pointed out the existence of a global zero point in parallax of -0.029 mas, derived looking at the parallax distribution of distant quasars. This offset is expected to be different for bright sources, and asteroseismic and spectroscopic observations report a global offset of -0.05 mas for $G < 14$ (Zinn et al. 2019; Khan et al. 2019). Parallax uncertainties are also affected by systematics, which can be included inflating the quoted measurement errors by a magnitude-dependent factor (Lindegren et al. 2018a). Spurious astrometry from *Gaia* DR2 can be filtered out using the renormalised unit weight error (Lindegren et al. 2018a). In a recent paper, Boubert et al. (2019) show that *Gaia* spectra for stars in crowded regions could be contaminated by the light coming from nearby sources, causing a shift in the radial velocity measurement. The authors propose further quality cuts to select a clean sample of *Gaia* stars with reliable astrometric and spectroscopic measurements.

1.2.5 *Gaia* and HVSs

The advent of the exquisite astrometric data provided by *Gaia* has revolutionized our knowledge on high velocity stars. The combination of *Gaia* with ground-based spectroscopic surveys has enabled the determination of precise and accurate total velocities for millions of stars. Marchetti et al. (2017) first attempted to find HVS candidates in *Gaia* DR1/TGAS, using a data mining routine based on machine learning. Boubert et al. (2018) revisited the origin of previously known unbound objects with the updated *Gaia*

[2] `https://www.cosmos.esa.int/web/gaia/release`

DR2 astrometric information. The authors found that, apart from one star (LAMOST J115209.12+120258.0), all the high velocity late-type candidates are actually bound to the Galaxy, including the ones identified in Marchetti et al. (2017). For what concerns the late B-type HVS, the new *Gaia* proper motions confirm the GC origin for the fastest objects (Brown et al. 2018). Marchetti et al. (2018a) computed total velocities for all the $\sim$ 7 million stars with a radial velocity determination from *Gaia* DR2, finding 20 stars with high probabilities of being unbound, but no HVS candidates from the GC (in agreement with predictions from Marchetti et al. 2018b). Bromley et al. (2018) supported these findings, extending the search to stars with precise parallaxes and high tangential velocity. Hattori et al. (2018a) suggested that this sample is composed of old and metal-poor stars, a result confirmed by Hawkins & Wyse (2018) using high resolution spectroscopy. Thousands of HVSs with precise proper motions are expected to be contained in the *Gaia* catalogue (Marchetti et al. 2018b), but these stars are predicted to be too faint to have a radial velocity from *Gaia*, a fact that has so far prevented their discovery.

1.3 Methods used in this thesis

In this section we will quickly describe some of the methods used in this thesis to analyze and derive properties from the *Gaia* data: Bayes' theorem, which is the basic concept behind Bayesian statistics, and machine learning, which will be used in Chapter 3 to identify HVS candidates.

1.3.1 Bayes' Theorem

Bayes' theorem is a direct consequence of the law of conditional probability. Indicating with $P(A)$ and $P(B)$ the probabilities of two independent events A and B, we can write the conditional probability:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1.8}$$

We can now express equation (1.8) in a Bayesian fashion. To do that, we consider the case in which we want to fit some model parameter θ given the data $\mathbf{x}$. Equation (1.8) then becomes:

$$P(\theta|\mathbf{x}) = \frac{P(\mathbf{x}|\theta)P(\theta)}{P(\mathbf{x})}. \tag{1.9}$$

This is the most general form of Bayes' theorem. The term $P(\theta|\mathbf{x})$ is called *posterior* probability, and represents the probability distribution of the parameter θ given the data $\mathbf{x}$. The term $P(\mathbf{x}|\theta)$, called *likelihood* probability, is the probability of observing the data $\mathbf{x}$ given a certain model parametrized by θ. The term $P(\theta)$ is the *prior* probability, which represents our prior knowledge on the parameter θ. The advantage of Bayesian statistics is that we can incorporate this prior knowledge on the model parameters, which might come from other experiments. Finally, the term $P(\mathbf{x})$ is called the *model evidence*, and is a normalization factor that is usually not considered (one is interested in relative probabilities), so that equation (1.9) is just expressed as a proportionality.

A common approach is to determine the likelihood using the chosen physical model, assume a prior on the parameters, and then sample the posterior distribution with a Monte Carlo Markov Chain (MCMC) algorithm, such as the affine-invariant ensemble sampler emcee (Foreman-Mackey et al. 2013).

1.3.2 Machine Learning

Machine learning is a data-driven approach to science, in which the algorithms learn from existing data, to make predictions on new data. The machine learning approach is generally divided into two classes: supervised and unsupervised learning. Supervised learning algorithms rely on a training set: a set of data for which ones know the features (the properties used for the training) and the desired output. The goal of a supervised learning algorithm is to learn from the data what is the function that best maps inputs into outputs. In regression algorithms, the output is a single real number, while the goal of classification algorithms is to assign each data-point to a particular class so that the output of the algorithm is the probability that each input belongs to a given class. Unsupervised learning algorithms, on the other hand, do not need a training set for the learning process, but their goal is to find hidden structures in the data. The most common unsupervised learning algorithms are clustering algorithms, that aim to find clustering in a high dimensional space.

The training set comprises of m training examples, each one with n features: $x^{(i)} \in \mathbb{R}^n$, where the superscript (i) refers to the i-th training point. In a supervised learning algorithm, each training example $x^{(i)}$ corresponds to a label $y^{(i)}$, with $y^{(i)} \in \mathbb{R}$ for regression problems, and $y^{(i)} \in \{0, 1, \ldots, M\}$ for a classification problem with M distinct classes. The *hypothesis* function $h_\Theta(x^{(i)})$ represents our best estimate of $y^{(i)}$, which we call $\hat{y}^{(i)}$. For ex-

ample, in multivariate linear regression, we compute the hypothesis for a single data point as:

$$\hat{y}^{(i)} \equiv h_\Theta(x^{(i)}) = \theta_0 + \theta_1 x_1^{(i)} + \cdots + \theta_n x_n^{(i)}, \tag{1.10}$$

where $\Theta = (\theta_0, \ldots, \theta_n) \in \mathbb{R}^{n+1}$ is the parameter vector. In classification algorithms, the output of the hypothesis can be interpreted as the probability that the data point belongs to a certain class. So, for example, in multivariate logistic regression the hypothesis is computed applying a sigmoid function to equation (1.10).

The goal of a supervised machine learning algorithm is to find the values of the parameter vector Θ which minimize the cost function $J(\Theta)$, which is often defined as:

$$J(\Theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right)^2, \tag{1.11}$$

which is the sum over all the training examples of the squared difference between the true labels y and the predicted labels $\hat{y}$. The search for the global minimum of the cost function is usually performed in an iterative fashion using the gradient descent optimization algorithm, but more advanced techniques have been proposed to achieve faster convergence (e.g. Robbins & Monro 1951; Duchi et al. 2011; Singh et al. 2015).

Artificial neural networks are supervised learning algorithms (see Haykin 2009, for an exhaustive description of neural networks). In chapter 3 of this thesis, we will make use of a neural network for a binary classification problem. The advantage of neural networks is their ability to learn highly non-linear mapping functions, more complex than the form presented in equation (1.10). Neural networks are often employed because of their ability to generalize: to provide reasonable outputs for inputs not encountered during the training session. A natural drawback is that overfitting can prevent the algorithm to generalize to new data-points. Overfitting can be avoided in several ways, both splitting the original training set into separate datasets that can be used to tune and test the algorithm, and applying different techniques of regularization by modifying the cost function in equation (1.11).

Neural networks are often used for pattern classification, image recognition, and in general high-dimensional problems with a large number of features. In astronomy, these algorithms are getting popular in different fields, for example for estimating redshifts or for galaxy classification (e.g. Dai & Tong 2018; Stivaktakis et al. 2018; Carrasco-Davis et al. 2018).

1.4 Thesis content

This thesis focuses on the search for the fastest stars in our Galaxy. We combine modelling, observations, and data mining techniques to identify and characterize these rare objects in the largest and most precise stellar catalogue ever produced: the data released from the ESA satellite *Gaia*.

In **Chapter 2** we create mock catalogues of HVSs to predict the properties of the HVS population in *Gaia*. We build three mock catalogues, adopting different assumptions on the ejection mechanism, including the Hills mechanism and the interaction between a single star and a massive black hole binary. In all cases, we find hundreds to thousands of HVSs to be contained in the final *Gaia* data release with precise proper motions, representing a huge improvement over the few tens of known candidates. We show how their identification is not trivial since the bulk of the population is expected to be too faint to have a radial velocity measurement from *Gaia*. Therefore new, advanced data mining techniques need to be implemented to search for these rare objects.

In **Chapter 3** we develop, implement and apply a novel data mining routine based on machine learning techniques, to identify HVS candidates in the *Gaia* DR1 TGAS subset. We choose to use an artificial neural network, trained on mock populations of HVSs created in Marchetti et al. (2018b), as presented in Chapter 2. Because of the missing radial velocity information, we choose to use the 5 parameters astrometric solution for the training process. The application to the TGAS subset results in the identification of 80 stars with high probabilities of being HVSs. Subsequent spectroscopic follow-ups with the Isaac Newton Telescope in La Palma and cross-match with spectroscopic surveys of the MW resulted in radial velocities for more than half of the candidates. We discovered one possibly unbound HVS, 5 bound HVSs, and 5 runaway star candidates with median velocities up to ~ 780 km s^{-1}.

Chapter 4 focuses on characterizing the high velocity tail of the velocity distribution of stars in the MW, using the subset of ~ 7 million stars with a radial velocity measurement from *Gaia* DR2. We derive distances from *Gaia* parallaxes using a Bayesian approach, and we then compute total velocities for the whole sample of stars. Focusing on the subset of stars with reliable astrometric measurements from *Gaia*, we identify 125 stars with

predicted probability $> 50\%$ of being unbound from the MW, and 20 with a probability $> 80\%$. Thanks to the precise full phase information given by *Gaia*, we can trace back in time these stars in the Galactic potential to identify their ejection location. We discover 7 stars coming from the stellar disk, consistent with being runaway stars. Surprisingly, the remaining 13 stars cannot be traced back to any star forming region. These objects have a preferred extragalactic origin, and they could be the result of the tidal disruptions of satellite galaxies from the gravitational field of the MW, or might be runaway stars originating in MW satellite galaxies, such as the LMC.

In **Chapter 5** we use the sample of ~ 20 unbound late B-type HVSs from Brown et al. (2014) to give joint constraints on the GC binary population and on the dark matter halo of the MW. We model the ejection velocity distribution of HVSs adopting the Hills mechanism, and we compare the resulting observed velocity distribution to the HVS data using a Kolmogorov-Smirnov (KS) statistical test. We find that assuming typical values observed in Galactic star forming regions for the binary properties in the GC, a good fit is achieved for dark matter haloes that result into an escape velocity from the GC to 50 kpc lower than 850 km s^{-1}. For realistic choices of the mass profile, these haloes are consistent with MW circular velocity data out to ~ 100 kpc, and with predictions from the concordance Λ CDM cosmological model. The discovery of hundreds of HVSs will break degeneracies between the GC and potential parameters, allowing a systematic study of these two different but complementary environments.