Structure in Star Forming Regions

Paweł Lung Lee

Department of Physics & Astronomy
The University of Sheffield

A dissertation submitted in candidature for the degree of Doctor of Philosophy at the University of Sheffield

29th of February 2016

"I don't need time, I need a deadline" – Duke Ellington

Declaration

I declare that no part of this thesis has been accepted, or is currently being submitted, for any other degree, diploma, certificate or any other qualification in this University or elsewhere.
This thesis is the result of my own work unless otherwise stated.
The following chapters have been based on publications:

- Chapter 2 – Lee & Goodwin, in preparation
- Chapter 5 – Lee & Goodwin, submitted, MNRAS

Acknowledgements

I would like to dedicate this thesis and all my academic adventures to Jerzy Moll – my late grandfather, who instilled in me curiosity and love for the surrounding world and its mysteries.

A special mention goes to my Mum and my Grandma who offered moral and material support throughout my entire PhD. This thesis is as much your work as it is mine. A big thank you also goes to my family from Barcelona, especially Basia, Mònica, and Luis.

This thesis would not be possible without my supervisor – Simon Goodwin who guided me through the world of star formation. Thank you for your passion, technical expertise, sense of humour and most importantly patience.

Thanks to all my 'partners in crime' – Jen and Marv, Saida, Andy, Liam, Krisada, Chrises, Davids, Ana, Rik, and all the other exceptional people from the panda department. Last but not least, I would like to thank all my other friends, especially Jools, Arhn+Neko+Qzak, as well as Lance, ChernYean, Jobi, Pete, Shiv, and all those who held their fingers crossed for me.

Abstract

Stars form in clumpy, highly substructured environments. In this thesis I set up N-body simulations of substructured star forming regions and investigate the impact that the substructure has on the survival of the star forming regions.

I also present a broad range of methods used in other fields to quantify and identify structure. I discuss their strengths and shortcomings and assess their suitability for use in astronomical contexts.

I use the $\mathcal{Q}$ and Λ methods to compare the distributions of class I and class II sources in observed star forming regions to learn more about the dynamical evolution of systems and infer whether they are bound or unbound.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Star formation is one of the main fields of interest in contemporary astrophysics. Electromagnetic radiation is the primary source of information about the baryonic Universe and most of the radiation that astronomers observe comes directly or indirectly from stars.

Stars are the building blocks of galaxies and act as cosmic furnaces that enrich the interstellar medium by producing heavier elements necessary for the existence of life. They may also harbour planetary systems that could support life. The environment and interactions with other stars also shape the fates of planetary systems, therefore understanding of star formation and the conditions in which stars form is crucial, if we want to learn more about planet formation and the origins of life.

The complexity of star formation lies in the great range of physical scales involved and multitude of processes whose understanding is required, if we wish to describe accurately the process of star formation. The main ones are gravity, chemistry, hydrodynamics, turbulence, magnetic fields and radiative transfer.

1.1 Star formation

The idea that stars form when cold gas fragments gravitationally is not new; it can be traced back to 18th century philosophers Swedenborg and Kant who postulated that the Sun formed from a cloud of gas that collapsed under its own gravity.

Until recently star formation was believed to be a slow, quasi-static process (Shu et al., 1987), however, advances in technology (increased computational power and bigger telescopes) allowed astronomers to better appreciate the impact of turbulence which points towards much shorter timescales of the order of star forming region's crossing time (a few Myr) (Elmegreen, 2000).

1.1.1 Molecular Clouds

In order for gas to cool down sufficiently for stars to form (10s of K), it needs to be shielded from external sources of heating hence star formation can only occur at high column densities. First stars formed in the early, metal free Universe, in which the the conditions were vastly different; high densities mean that exotic forms of heating such as cosmic rays and self-annihilation of WIMP dark matter become important (Bromm, 2013). Furthermore, the lack of heavier elements made cooling more difficult and as a result the temperatures are much higher (300K; Larson, 2000).

In the present day Universe the conditions that allow stars to be formed are usually found in giant molecular clouds (GMCs). The sizes of GMCs range from 10s of pc to ~ 100 pc (Kennicutt & Evans, 2012) and they can have masses between 10^2 $M_\odot$ (Magnani et al., 1985; Heyer et al., 2001) for the small clouds on the outskirts of the Milky Way, to 10^7 $M_\odot$ for the biggest GMCs found close to the galactic centre (Oka et al., 2001). Typical surface densities of molecular clouds in the Galaxy are $\Sigma_{\rm GMC} \sim 100$ $M_\odot\,{\rm pc}^{-2}$ (Heyer et al., 2009). Observations indicate, however, that this value could be different in other galaxies – as low as 50 $M_\odot\,{\rm pc}^{-2}$ in the LMC (Hughes

et al., 2010) and higher than 180 $M_{\odot}\,pc^{-2}$ in M51 (Colombo et al., 2014), and it is also dependent on the galactocentric radius at which the GMC is found (Adamo et al., 2015).

Since GMCs contain mostly molecular hydrogen (H_2) they are very difficult to observe since H_2 has no excited states at <200K. Fortunately, there are other molecular species, known as tracers, that are observable at such low temperatures and whose density is proportional to the density of H_2. CO is an example of a commonly used tracer. In high density areas where CO lines saturate or CO freezes-out, other tracers such as HCN, NH_3 or CS can be used (Evans, 1999).

The dust grains are kept in thermal and radiative equilibrium with the gas by constant collisions with the gas particles and will emit as a 'grey body'. If we assume that dust constitutes approximately 1 per cent of the total mass of the cloud (100 to 1 gas-to-dust ratio), we can use dust emission as a tracer of gas column density (Evans, 1999; Longmore et al., 2014).

In addition to mapping out the density of the gas, the emission lines from the molecular tracers provide information about the gas velocities. The width of the lines provides us with a velocity dispersion that measures both the thermal and non-thermal motions, while the ratios of different emission lines allow us to deduce the kinetic temperature of the gas. Larson (1981) found a set of empirical power law relationships, known as the *Larson Relations*, that describe the structure of GMCs. One of the relations shows a dependence of the velocity dispersion (σ, km/s) on the size of the region (L, pc):

$$\sigma \propto L^{\beta} \tag{1.1}$$

where β = 0.38 - 0.5 (McKee & Ostriker, 2007). This behaviour is interpreted as a signature of turbulence which explains why linewidths corresponding to supersonic

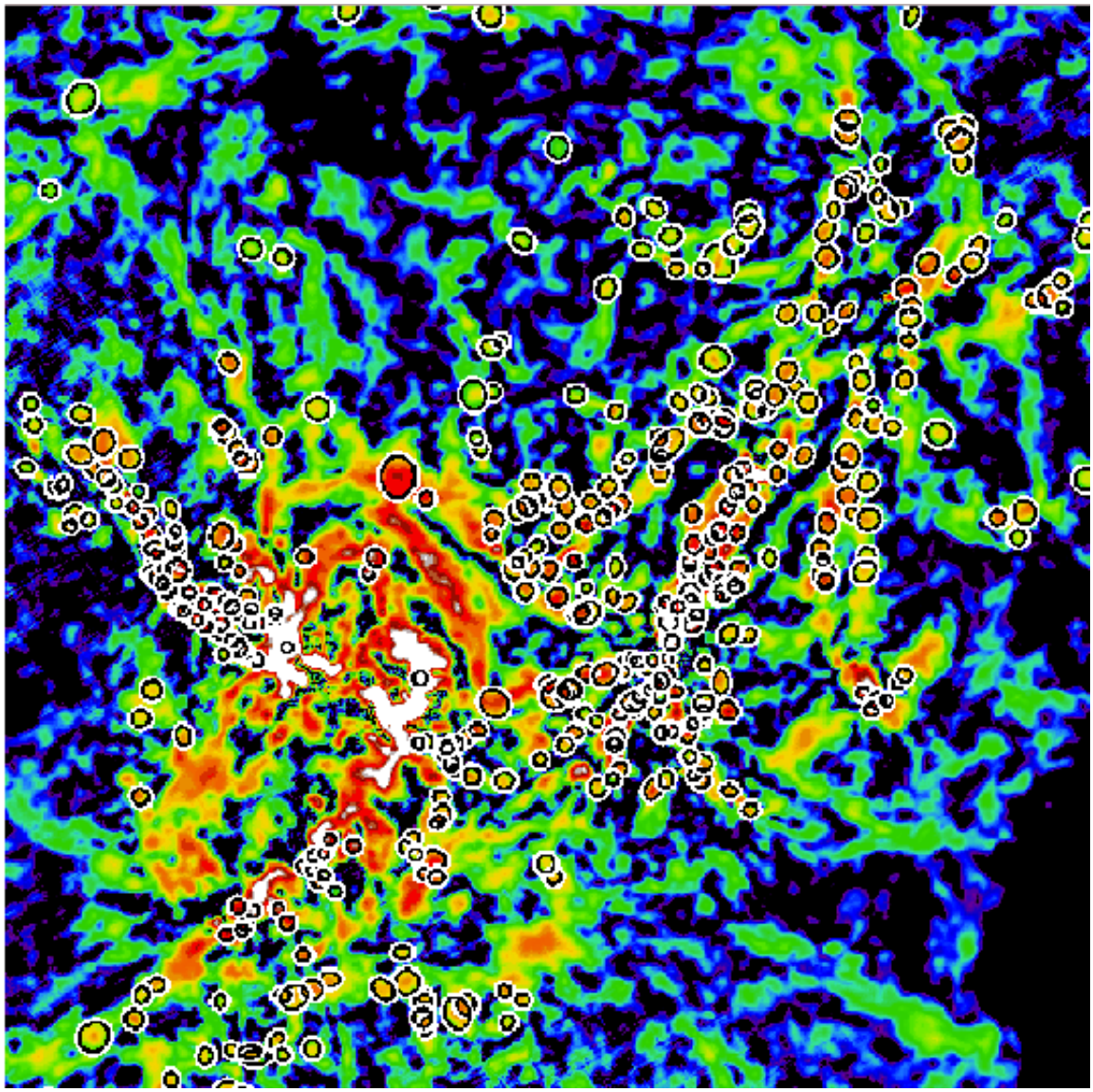


Figure 1.1: A *Herschel* image of the Aquila cloud (Men'shchikov et al., 2010). Colours correspond to gas column density with white being the densest and black being the least dense. The ellipses correspond to starless cores identified through source extraction. Notice that the cores lie along filaments of enhanced density.

velocities with Mach number (ratio of the observed velocity to the speed of sound) $\mathcal{M} \sim 5 - 20$ are observed in GMCs (Williams et al., 2000; Ballesteros-Paredes et al., 2007).

Turbulence is a dissipative process whereby energy cascades to increasingly smaller scales via eddies until it reaches a dissipation scale. The energy spectrum follows a power-law relationship:

$$E(k) = k^{-n} \tag{1.2}$$

where k is the wavenumber (the inverse of scale). The Kolmogorov (1941) spectrum for transonic turbulence predicts an exponent $n = 5/3$, which is consistent with $\beta = 0.38$. In the supersonic regime, however, the Kolmogorov-Burgers model of turbulence (Boldyrev, 2002) predicts a value of $n \sim 1.75$ which is consistent with $\beta \sim 0.5$ (McKee & Ostriker, 2007).

The turbulent nature of the GMCs also is the basis for the theory that star formation is a quick process, since supersonic turbulence is a transient phenomenon and unless driven, will die away on the scale of the system crossing time (Ballesteros-Paredes, 2006).

Eqn. 1.1 indicates that on large scales the turbulence is supersonic which causes shocks in the gas. The shocks can result in formation of overdense sheets that fragment into filaments (Bhattal et al., 1998). Observations suggest that filaments are ubiquitous in GMCs (see fig. 1.1) and the overdense clumps within them are the loci of star formation (Williams et al., 2000; André et al., 2014).

1.1.2 Cores

The density enhancements in GMCs can lead to formation of bound prestellar cores with sizes $\sim 10^4$ AU (Ward-Thompson et al., 2007). A core can collapse further to

form a stellar system, if it is dense enough for gravity to overcome the thermal and magnetic support (Crutcher, 1999).

The critical mass above which a gaseous object will undergo collapse is known as the Jeans mass:

$$M_{\rm Jeans} = \frac{\pi}{6} \frac{c_{\rm s}^3}{G^{3/2} \rho^{1/2}} \tag{1.3}$$

where ρ is the gas density, $c_{\rm s}$ is the sound speed and G is the gravitational constant. Prestellar cores with masses greater than $M_{\rm Jeans}$ will collapse to form protostars.

A related quantity is the Jeans length ($R_{\rm Jeans}$) which is equal to the radius of a sphere of mass $M_{\rm Jeans}$:

$$R_{\rm Jeans} = \left(\frac{M_{\rm Jeans}}{\frac{4\pi}{3}\rho} \right)^{1/3} \tag{1.4}$$

Objects larger than the local Jeans length will undergo fragmentation. Considering the conditions in the GMCs ($c_{\rm s} \sim 0.2$ km s^{-1} at $T = 10$ K and $\rho \sim 10^{-19}$ g cm^{-3}) a typical Jeans mass is ~ 1 M$_\odot$ and Jeans length is ~ 0.1 pc.

It is important to keep in mind that as the core collapses the value of $M_{\rm Jeans}$ will change since it depends on the density of gas and the sound speed/temperature. Assuming a barotropic equation of state (one where pressure depends only on density; $P(\rho) = \rho^\gamma$):

$$c_{\rm s}^2 = \frac{P(\rho)}{\rho} = \rho^{\gamma - 1} \tag{1.5}$$

where γ is the polytropic index. Since the Jeans mass of a fragment depends on the temperature and density (eqn. 1.4; Low & Lynden-Bell, 1976; Masunaga & Inutsuka, 2000):

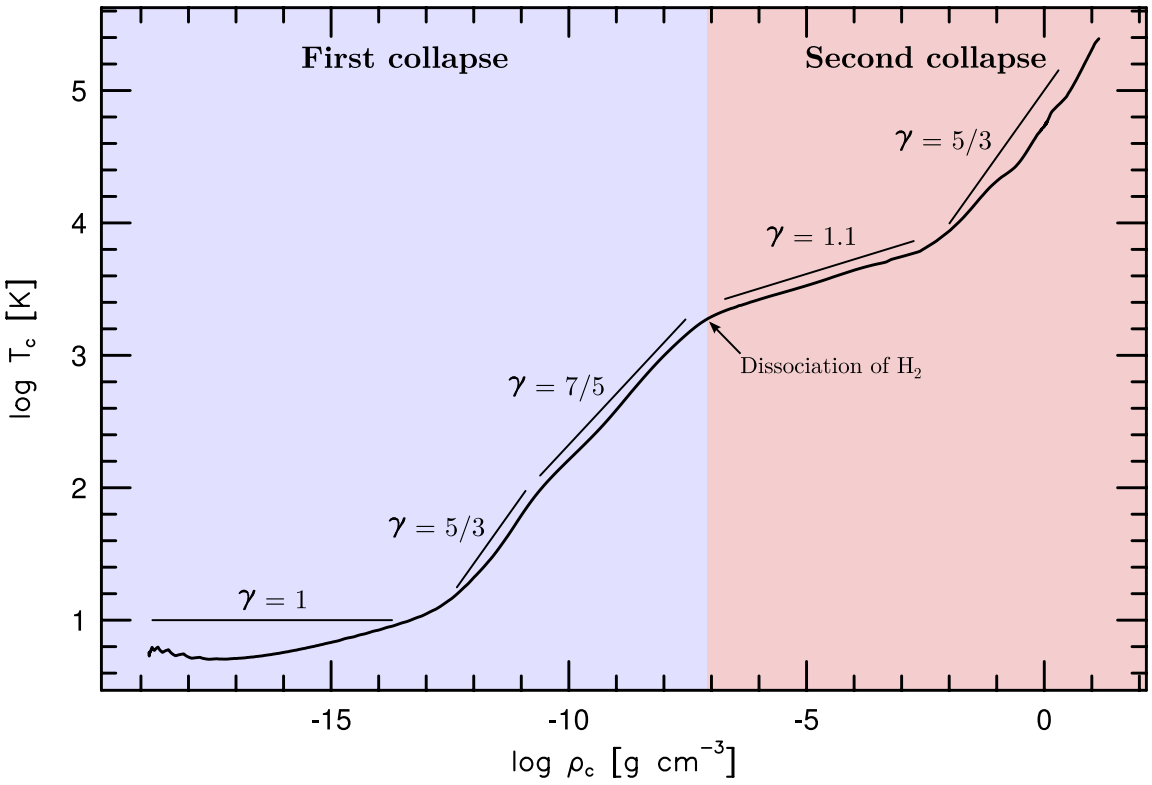


Figure 1.2: Central temperature T_c as a function of central density ρ_c for a collapsing core as simulated by Masunaga & Inutsuka (2000). The boundary between the first and second collapse corresponds to the value of $T_c \sim 2000$K, at which the H_2 dissociation commences.The plot also shows an approximate fit to each stage of the collapse. The slope of each segment is $\gamma - 1$ for that stage's γ. Adapted from Masunaga & Inutsuka (2000) by Rawiraswattana (2012).

$$M_{\text{Jeans}} \propto \rho^{3/2\gamma-2} \tag{1.6}$$

Fig. 1.2 shows the dependence of central core temperature on the density; the slopes of the segments correspond to the value of $\gamma - 1$ during that stage. Initially the collapse is isothermal ($\gamma = 1$) since the radiative cooling is efficient. The gas is optically thin and excess energy is radiated away as blackbody radiation from dust (Masunaga et al., 1998). This stage happens roughly on a free fall timescale, t_{ff},

$$t_{\text{ff}} = \left(\frac{3\pi}{32G\rho}\right) \tag{1.7}$$

The increase in density due to collapse results in a decreased Jeans length and causes fragmentation. In the long run, however, the increasing density makes the opacity increase and results in decreased efficiency of the cooling. Once the density reaches $\rho \sim 10^{-13}\text{g}\,\text{cm}^{-3}$ the collapse becomes adiabatic ($\gamma = 5/3$) and the temperature starts going up (Larson, 1969; Masunaga & Inutsuka, 2000).

During this adiabatic heating phase, H_2 behaves like a monatomic gas ($\gamma = 5/3$) until the temperature reaches $\sim$ 100K, at which point the excitation of the rotational states becomes possible and the behaviour changes to that of a diatomic gas ($\gamma = 7/5$).

At this stage of the collapse, the Jeans mass reaches its minimum at $\sim 10^{-2}\text{M}_\odot$. This point is known as the *opacity limit for fragmentation* and corresponds to the smallest mass an object that formed through fragmentation can have (Rees, 1976; Low & Lynden-Bell, 1976). This value is remarkably close to the planet - brown dwarf boundary as defined by the deuterium burning limit (13 M_J, where M_J is a Jupiter mass $\sim 10^{-3}\,\text{M}_\odot$). The opacity limit for fragmentation creates a more intuitive distinction between these two types of objects that also provides insight into the physics and origin of the formation mechanism – brown dwarfs form by fragmentation, planets form via core accretion (Whitworth, 2000; Basri & Brown,

2006).

Once the core reaches the opacity limit it is usually referred to as a protostar. It is close to hydrostatic equilibrium and slowly contracts adiabatically on the Kelvin-Helmholtz timescale, $t_{\rm KH}$,

$$t_{\rm KH} = \frac{GM^2}{RL} \tag{1.8}$$

where L is the luminosity of the protostar.

Once the temperature of the core reaches $\sim$ 2000 K molecular hydrogen dissociates. This acts as a heat sink and makes rapid further collapse possible. Moreover, the dust evaporates around this temperature thus reducing the collective opacity (Masunaga & Inutsuka, 2000). It is, however, not clear whether during this *second collapse* the core might fragment further (Bate, 1998).

During this stage, the gas is approximately isothermal with an adiabatic index $\gamma \sim 1.1$ (see Figure 1.2). The second collapse results in a decrease in the Jeans mass and it might explain the formation of very close binaries (Bonnell & Bate, 1994; Masunaga & Inutsuka, 2000; Whitworth & Stamatellos, 2006).

The dissociation is followed by a phase of Kelvin-Helmholtz contraction which continues until the temperature is high enough for hydrogen burning to be ignited, at which point the main sequence is reached.

1.1.3 Stages of star formation

Based on observations of the spectral energy distribution (SED) the pre-main sequence evolution of stars is divided into four classes (Lada, 1987; Evans et al., 2009; Dunham et al., 2014). When analysing the shape of the SED, the spectral index (α) is used to discriminate between the classes (Lada, 1987),

$$\alpha = \frac{\mathrm{d}\log(\lambda F_\lambda)}{\mathrm{dlog}\,\lambda} \tag{1.9}$$

where F_λ is the flux density (energy per unit area per unit wavelength) at wavelength λ. The index α is calculated for wavelengths $\lambda > 2$ µm and is the slope of the SED longward of 2 µm (see fig. 1.4).

Class 0

During the earliest stage of star formation the core starts collapsing to form a protostar. This phase is short-lived and lasts roughly for the duration of a free-fall timescale. The temperature of the core is a few 10s of K and it emits as a blackbody peaking in the sub-mm ($\lambda_{\mathrm{peak}} \lesssim 100$ µm; first panel of fig. 1.4). Since the source is heavily embedded it is not often directly detectable in the infrared. Radio observations of class 0 objects show collimated CO outflows stronger than in class I sources (Bontemps et al., 1996; Evans et al., 2009).

Class I

The next stage starts when roughly half of the core mass has been accreted onto the central object. The luminosity and gravitational energy from the collapse increases the temperature of the envelope to >70 K. The SED is a combination of the blackbody radiation from the protostar and the contribution from the envelope. Most of the energy radiated by the protostar is reprocessed by the surrounding gas and results in a very large IR excess (panel 2 of fig. 1.4). Because of this excess class I sources have a positive spectral index ($\alpha \geq 0.3$).

Flat Spectrum

Greene & Wilking (1994) identified a class with a flat IR SED ($-0.3 \leq \alpha < 0.3$). The exact nature of these objects is not clear, however, a bolometric temperature between 350 K and 950 K (Evans et al., 2009) suggests that at least some of them could belong to an intermediate phase between class I and class II (Calvet & Hartmann,

1994; Dunham et al., 2014).

Class II

Once most of the envelope has been accreted the protostar becomes a class II also known as classical T Tauri star (CTT). Compared with class I most of the energy now comes from the central object and the peak of the SED has shifted towards the optical. Due to light being re-radiated by the disc, there is still a pronounced, albeit smaller, infrared excess (fig. 1.4) corresponding to a spectral index between -1.6 and -0.3. Observationally T Tauri stars are characterised by strong Hα emission as well as coronal X-ray signatures (Soderblom et al., 2014) in addition to an IR excess. This phase lasts until the disc is depleted, which takes a few Myr (Armitage et al., 2003).

The rate of accretion throughout this phase is not constant – eventually matter will start building up in the inner part of the disc. Once enough material has been accumulated, a period of rapid accretion commences and the object undergoes a brightening of the order of several magnitudes that subsequently decays over the course of several decades (Greene et al., 2008). This scenario is believed to be the explanation for the behaviour of a class of objects known as FU Orionis stars.

Class III

Once most of the material in the disc has been depleted the star reaches class III stage. The SED might still exhibit a small IR excess due to traces of a disc but the accretion signatures have nearly disappeared hence the name weak-lined T ; stars (WTT). During this phase the star is very similar to a main sequence star, but still emits strongly in X-ray (Preibisch & Kim, 2005). Since it has not fully contracted it is also more luminous than a main sequence star of the same colour i.e. it lies above the main sequence on the H-R diagram. Fig. 1.3 shows the main sequence (black line) and the pre-main sequence tracks of 0.1 $M_\odot$, 1.0 $M_\odot$ and 5.0 $M_\odot$ stars. Objects with masses $<0.5\,M_\odot$ are almost fully convective and they simply contract towards the main sequence on what is known as the Hayashi track. More massive stars (0.5

$M_\odot$ – 3.0 $M_\odot$ follow the Hayashi track at first, however, once they become partially radiative they start following the Henyey track to the left. Stars with masses over 3 $M_\odot$ are fully radiative and start on the Henyey track (Palla, 2012). Once an object starts burning hydrogen in its core, the object becomes a zero-age main sequence star.

1.1.4 Pre-main-sequence lifetimes

It can take up to 10^8 yrs for a star to reach the main sequence, however, based on evolutionary models, the discs surrounding classical T Tauri stars have a median estimated half-life $\sim$ 2 Myr (Evans et al., 2009). The majority of observed pre-main-sequence objects fall into this category – even if we consider all the protostellar (class 0 and class I) sources together, CTTs are still $\sim$ 10 times more common.

The relative populations of the classes can be used to estimate the lifetimes of the stages leading up to CTTs. If we assume a constant star formation rate, it implies a tenfold difference in the expected lifetimes. By compiling information from several surveys Dunham et al. (2014) estimated the protostellar lifetime $\sim$ 0.5 Myr, which is consistent with observations of core kinematics (André et al., 2007). Using the same method a flat spectrum lifetime $\sim$ 0.4 Myr was derived, one should however keep in mind that the exact nature of these sources is unclear. The relative scarcity of class 0 is consistent with the stage being very short lived (0.05 Myr) due to the rapid infall of the gas (Ward-Thompson et al., 2007; Evans et al., 2009).

1.1.5 Initial mass function

The definition of star as an object capable of hydrogen burning in its core sets the lower stellar mass limit at $\sim$0.08 $M_\odot$. On the other end of the mass spectrum stars as massive as 300 $M_\odot$ are thought to exist (Crowther et al., 2010). The distribution

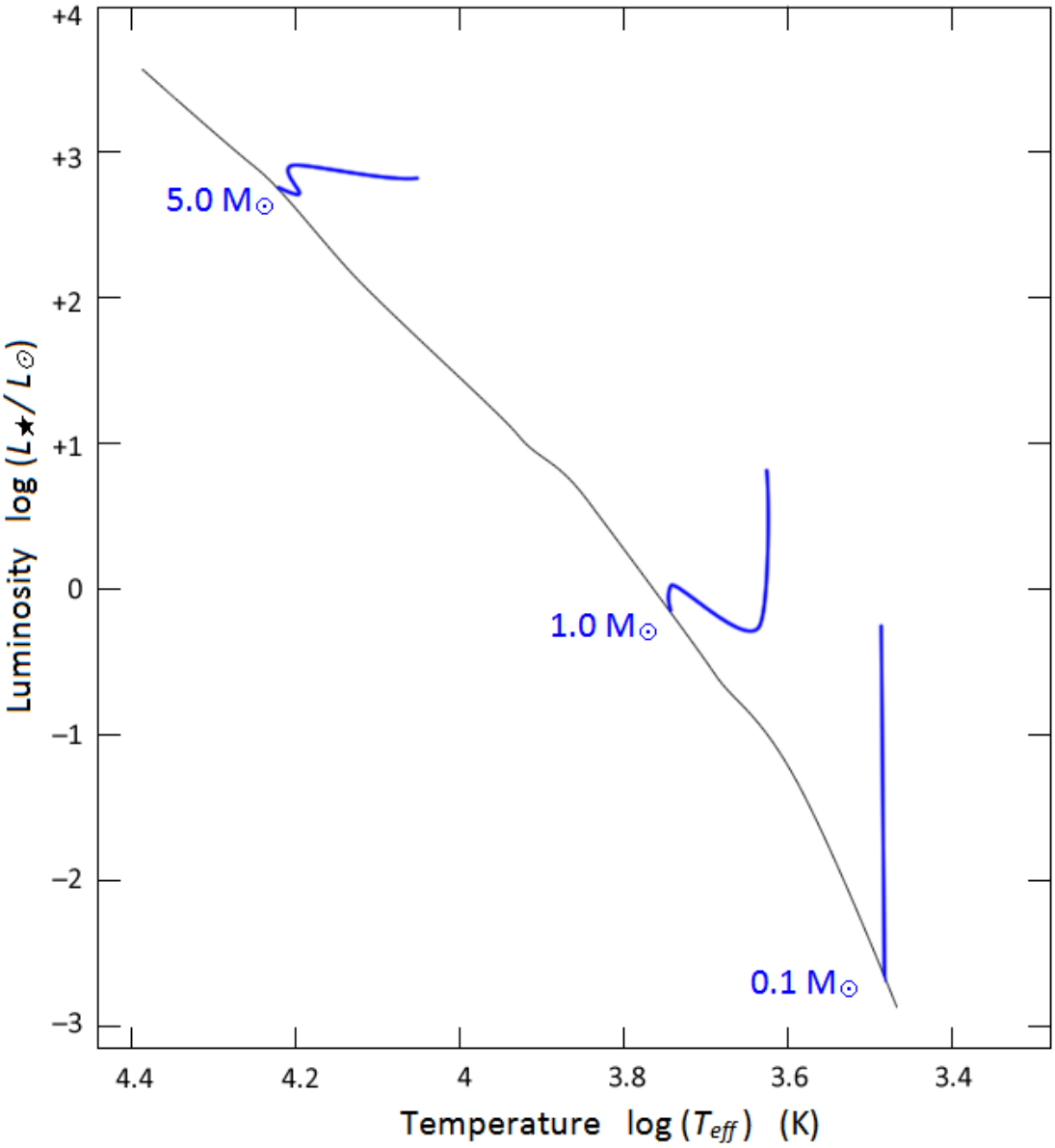


Figure 1.3: Hertzsprung-Russell diagram. The main sequence is represented by the black line, the blue lines are pre-main sequence evolutionary tracks for stars of masses 0.1 $M_\odot$, 1.0 $M_\odot$ and 5.0 $M_\odot$. Adapted from (Palla, 2012).

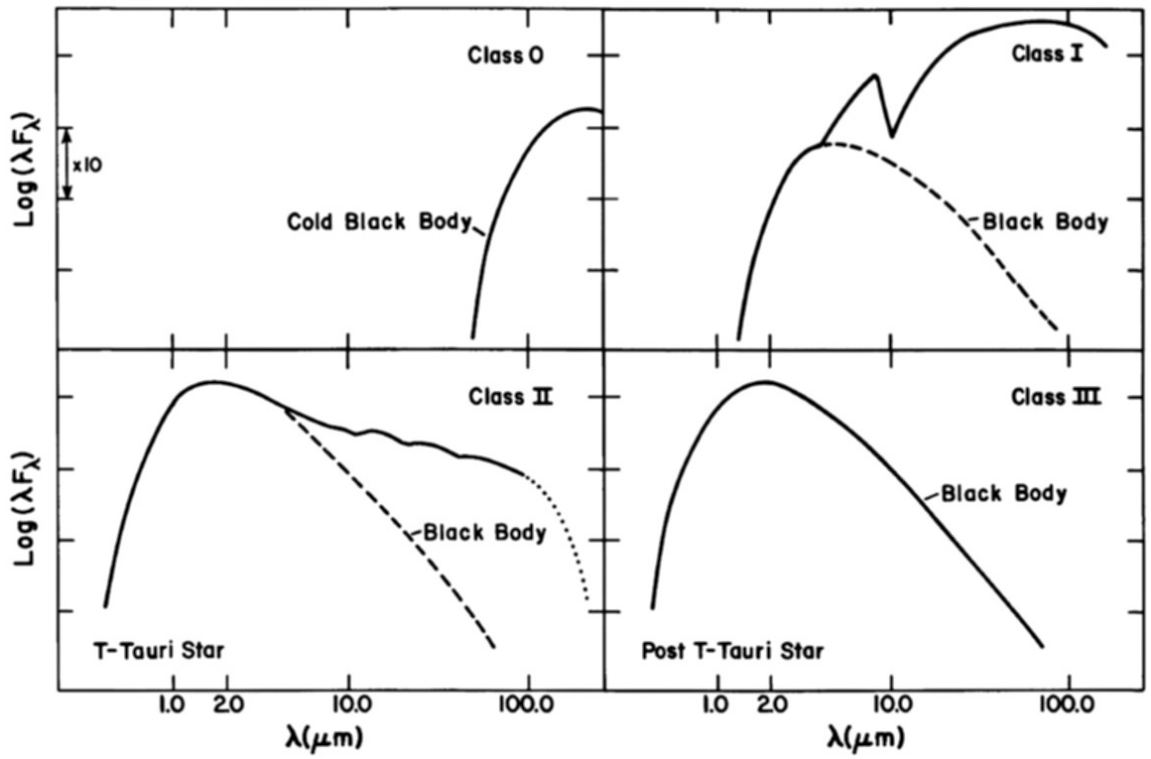


Figure 1.4: A comparison of spectral energy distribution shapes of young stellar objects of class 0 to class III. Class 0 sources are deeply embedded protostars, whose SED resembles that of a blackbody and peaks in the sub-mm (top left). Class I objects have accreted ~ 50 per cent of the surrounding envelope. Their SED consists of a warmer black body corresponding to the central object and an IR excess due to the envelope (top right). During the class II phase most of the envelope has been cleared, except for a disc which still results in an IR excess (bottom left). Class III sources are very similar to main sequence stars but they might also exhibit a very small IR excess due to a debris disc. Adapted from Lada (1987).

of masses is referred to as the mass function. Of particular interest is the distribution of masses at birth – the initial mass function (IMF) and its apparent universality.

Salpeter (1955) described the initial mass function using a power-law of the form:

$$dN \propto m^{-\alpha}\, dm \tag{1.10}$$

and observationally obtained a value for the exponent $\alpha = 2.35$. Sixty years later this slope is still applicable for stars above 1 $M_\odot$.

Remarkably, the shape of the initial mass function does not change across a range of different environments and populations (Bastian et al., 2010). Since Salpeter's original publication there has been a multitude of attempts at explaining this universality, however, most of the simulations simply manage to reproduce the shape but lack predictive powers (see Bonnell et al. 2007 and the references therein).

More recently Chabrier (2003) proposed a different fit that uses a two-piece description consisting of a log-normal distribution at the low-mass end and a Salpeter slope for the higher masses. Another prominent contemporary formulation by Kroupa (2002) consists of three power laws with different slopes: one for substellar regime, one for average-mass stars, and a high mass slope similar to Salpeter (1955).

The work included in this thesis uses the L_3 parametrisation of the IMF proposed by Maschberger (2013). This approach uses fewer free parameters than the other formulations and its probability distribution function is given by the following formula:

$$p_{L3}(m) = A \left(\frac{m}{\mu}\right)^{-\alpha} \left(1 + \left(\frac{m}{\mu}\right)^{1-\alpha}\right)^{-\beta} \tag{1.11}$$

where:

$$A = \frac{(1-\alpha)(1-\beta)}{\mu} \frac{1}{G(m_u) - G(m_l)} \tag{1.12}$$

and $G(m)$ is the auxiliary function defined as:

$$G(m) = \left(1 + \left(\frac{m}{\mu}\right)^{1-\alpha}\right)^{1-\beta}. \tag{1.13}$$

Parameters m_l and m_u set the lower and upper mass limits respectively, while α, β and μ determine the shape of the IMF. The following values recommended by Maschberger (2013) were used: α=2.3, β=1.4, μ=0.2 $M_\odot$, m_l=0.01 $M_\odot$, m_u=150 $M_\odot$.

In order to sample a mass from the IMF one only needs to generate a uniform random number $\mathcal{R}$ in the range between 0 and 1 and substitute it into the equation:

$$m(u) = \mu \left([\mathcal{R}(G(m_u) - G(m_l)) + G(m_l)]^{\frac{1}{1-\beta}} - 1 \right)^{\frac{1}{1-\alpha}}. \tag{1.14}$$

Figure 1.5 compares the mass probability distribution function of the Maschberger (2013) IMF (solid black line), Chabrier (2003) (dashed green line) and Kroupa (2002) (dashed blue line). The shape of the Maschberger formulation is a very good fit to the data and is similar to the other two.

1.1.6 Binary and multiple systems

When the prestellar core collapses it is very likely that it will fragment into more than one protostar and the observations indeed indicate that multiple systems are very common (Duchêne & Kraus, 2013). It is too difficult to form multiples dynamically to explain the large numbers of observed multiple systems, hence they must have formed as such (Clarke & Pringle, 1991; Goodman & Hut, 1993).

Not all stars have the same likelihood of having companions – the frequency of

multiplicity depends on the primary mass. While virtually all the O stars have a companion; the frequency falls to only 30 per cent of M-dwarfs. (Duchêne & Kraus, 2013; Ward-Duong et al., 2014).

Pre-main-sequence stars are young and have had fewer opportunities to interact, hence they tend to have higher multiplicity frequency than field stars. Furthermore, dynamically young associations have higher binary fractions than denser, more dynamically evolved regions. The differences between the Taurus-Auriga complex and the Orion Nebula Cluster are, however, so dramatic (Taurus has twice the binary fraction of the ONC) that they beget the question of whether the dynamical processing is sufficient to account for them or perhaps the populations were initially different (Parker et al., 2009; Goodwin, 2010; King et al., 2012).

1.2 Stellar Clusters

Observations suggest that most stars are not born in isolation but in star forming regions (SFRs), which can have vastly different masses and appearances: from low surface density associations such as Taurus (10^2 $M_\odot$; ~ 5 stars pc^{-3}, Kenyon et al. 2008) to massive clusters such as Westerlund 1 (10^5 $M_\odot$; $\sim 10^4$ stars pc^{-2}; Clark et al. 2005). It is important to study the environment in which the stars form since it can affect their future evolution; frequent interactions in a cluster can result in disc disruption, planet ejection and a change in the binary properties.

Traditionally SFRs were thought of as dense *clusters* (Lada & Lada, 2003). More recently, however, there has been a shift towards a hierarchical picture of star formation. Bressert et al. (2010) compared the stellar surface densities of nearby star forming regions and concluded that the number of stars that form in 'clusters' is very strongly dependent on the definition of 'cluster' and only ~ 25 per cent of stars form in environments dense enough to actually affect their evolution. The encounter

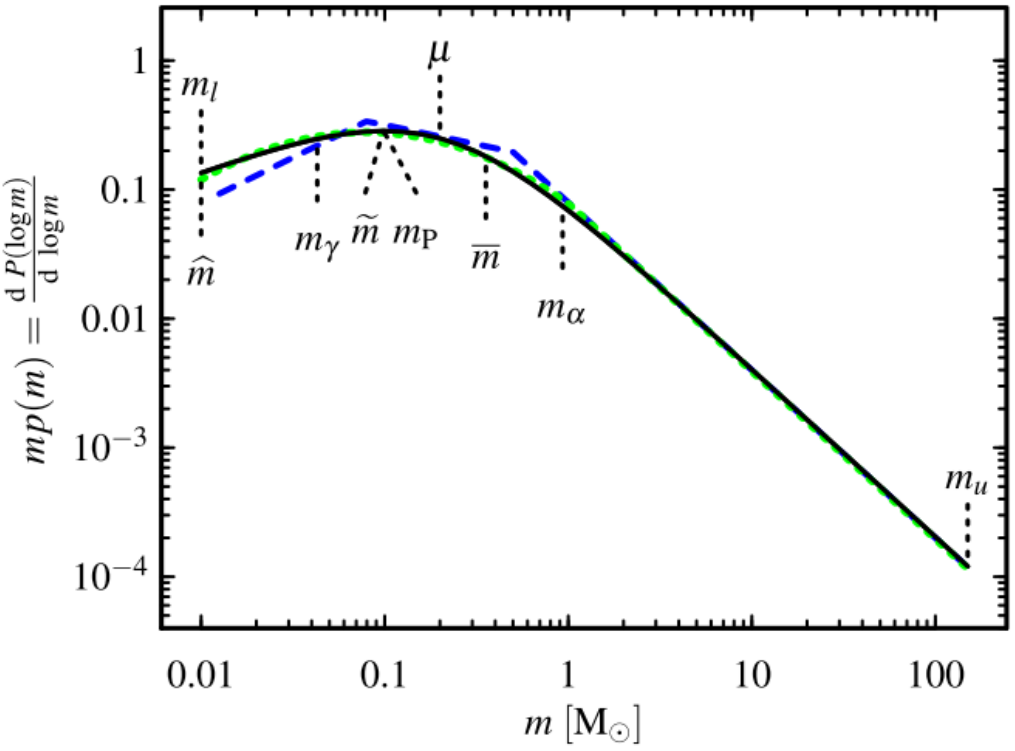


Figure 1.5: Mass probability distribution function against stellar mass. Black line corresponds to Maschberger (2013), green dashed line to Chabrier (2003), blue dashed line to Kroupa (2002). Adapted from Maschberger (2013)

timescale of a star can be approximated as:

$$t_{\rm enc} \sim \frac{1}{\pi d^2} \cdot \frac{1}{v} \cdot \frac{1}{n} \tag{1.15}$$

where d is the distance at which the encounter occurs, v is the star's velocity, and n is the number density. The separation at which an encounter can disrupt the disc and affect the evolution is 200 AU $\sim 10^{-3}$ pc. A typical value of velocity of a few km s^{-1} gives:

$$t_{\rm enc} \sim 10^5 \cdot \frac{1}{n} \tag{1.16}$$

For the density of Taurus ($n \sim 10$ pc^{-3}) the encounter timescale is $\sim 10^4$ Myr, while for Westerlund 1 it is ~ 10 Myr. Given the approximate ages of ~ 5 Myr, one can conclude that significant encounters in Taurus are rare and unlikely to have affected the evolution of the stars. On the other hand, the dense environment of Westerlund 1 makes significant encounters fairly frequent and they are likely to have an effect on the stellar evolution.

In the hierarchical paradigm there is no preferred scale for star formation, instead there is a hierarchy of structures – SFRs may form double clusters (de la Fuente Marcos & de la Fuente Marcos, 2009), but they may also consist of sub-clusters perhaps containing even smaller clumps (Elmegreen, 2006; Allison et al., 2009a). This can be interpreted as an imprint of the scale-free nature of turbulence, especially given the correlation between the structure in the star forming region and the underlying gas distribution (Gutermuth et al. 2011; see sec. 1.1.1).

If the star forming region is dynamically cold, it will rapidly collapse to form a bound cluster (Elmegreen, 2006). On the scale of a crossing time the mixing will erase the substructure and result in mass segregation (Allison et al., 2009a). If a region is unbound it will not undergo collapse, instead it will become an OB

association – a loose group of comoving stars. These associations are unbound and slowly drift apart, therefore they retain their initial structure and do not undergo mass segregation (Goodwin & Whitworth, 2004; Parker et al., 2014). Analysing the mass segregation and the amount of substructure in a region can provide a lot of insight into its dynamical history and whether it formed as an association (Parker et al., 2014; Wright et al., 2014).

1.2.1 Dynamical Evolution

Over time stars in a bound cluster will undergo relaxation – a process whereby their energies are changed in an attempt to reach dynamical equilibrium. The associated timescale known as the relaxation time ($t_{\rm relax}$) is the time it takes for a star's velocity to change significantly ($|\delta v| \sim |v|$ i.e. for it to lose all information about the initial conditions):

$$t_{\rm relax} \sim \frac{N}{8\,lnN} t_{\rm cr} \tag{1.17}$$

where N is the number of stars in the system and $t_{\rm cr}$ is the crossing time(Spitzer, 1969).

In the process known as two-body relaxation a star can exchange a portion of its energy with another star during a (close) encounter. Through the interaction typically the energy will be transferred from the more energetic star to the less energetic ones. In an idealised scenario of an ensemble of particles in a box with perfectly elastic walls this process will continue until the system reaches a Maxwellian distribution. In reality, however, the low-mass stars may occasionally be ejected from the system, if their subsequent velocity exceeds the escape speed (this phenomenon is called *evaporation*). On the other end of the spectrum, the encounters make the high mass stars lose energy and sink deeper into the cluster potential well. This gives

rise to dynamical mass segregation – the most massive stars are preferentially found close to the centre of the cluster (Spitzer, 1969).

The speed of mass segregation is dependent on the mass of the star (it scales as $\frac{M}{\langle M \rangle}$, where M is the mass of the star and $\langle M \rangle$ is the mean stellar mass.) and happens quickly (on the scale of several crossing times) for the massive stars (Allison et al., 2009a). High mass stars attract the other ones more strongly hence their interactions are more frequent (gravitational focussing) and they segregate on a shorter timescale.

The above picture holds for a fairly smooth background potential with local perturbations. In a more substructured environment such as a collapsing SFR, however, bulk flows due to infalling clumps of stars will cause a substantial change to the potential experienced by the stars. Lynden-Bell (1967) demonstrated that these changes to the potential cause a period of *violent relaxation*, which lasts until the distribution has been smoothed out. Since the violent relaxation happens on a short timescale (of the order of crossing time), the presence of substructure in a region implies that it is dynamically young and has not had enough time to erase it (Parker et al., 2014).

1.2.2 Gas expulsion

During the early stages of their evolution star forming regions are embedded in the natal gas, only a fraction of which is turned into stars. Over time, feedback from the most massive stars (ionisation, radiation pressure, supernovae, winds etc.) will result in the removal of gas on the timescale of a few Myr. Loss of mass from the system is dynamically important as it reduces the potential energy felt by the stars and will often lead to the destruction of the system.

Assuming that initially the gas and stars are well mixed and close to equilibrium, the effective star formation efficiency (eSFE) can be used to describe the dynamical state of the cluster post-gas-expulsion (Verschueren & David, 1989; Goodwin &

Bastian, 2006; Goodwin, 2009). Under these conditions the cluster can lose up two thirds of its initial mass and still retain some stars (Goodwin, 1997; Baumgardt & Kroupa, 2007).

In reality, however, SFRs are not virialised Plummer spheres – the distribution can be clumpy and dynamically cold (Goodwin & Whitworth, 2004). Smith et al. (2011) showed that eSFE is not a good predictor of the cluster's fate if the cluster is substructured. I pursue this idea further in chapter 5 where after the expulsion the star forming region is considered as a whole.

1.3 Thesis Outline

This thesis follows the structure outlined below. Chapter 2 contains my analysis of the structure of star forming regions contained in the Gutermuth et al. (2009) survey. I use the difference between the values of $\mathcal{Q}$ of class I and class II sources to infer whether a cluster is bound or unbound. In chapter 3 I present a range of methods that can be used to find and quantify structure and I discuss their suitability for being applied to the SFRs. Chapter 4 is an outline of the principles behind N-body simulations and a discussion of the algorithms. Finally, in chapter 5 I present the results of N-body simulations performed on a simple model of substructured initial conditions in order to investigate the influence of the virial ratio of the sub-clusters on the fate of the system.

Chapter 2

Analysing the Structure of Star Forming Regions

In this chapter I discuss methods that allow us to quantify the substructure in star forming regions. Furthermore, I present the results obtained by applying these methods to the Gutermuth et al. (2009) data set.

2.1 The data

Gutermuth et al. (2009) conducted a Spitzer survey of 36 nearby star forming regions. They performed observations using the mid infrared IRAC instrument operating in high dynamic range mode and the far infrared photometer MIPS. IRAC allows for simultaneous imaging in four bands: 3.6, 4.5, 5.8, and 8.0 µm, while MIPS operates in three bands: 24, 60, and 160 µm.

Source extraction and aperture photometry methods as outlined in Gutermuth et al. (2008) were then performed on the images and the fluxes of the objects were obtained. The ratios of fluxes in different bands were used to classify Class I and Class II sources.

Objects with stellar photospheres have $[3.6] - [4.5]$ and $[5.8] - [8.0]$ colours close to zero and therefore lie close to the origin of a colour-colour diagram (Hartmann et al., 2005). Class I and class II sources, however, are characterised by an infrared excess due to the presence of an envelope or a disk, hence they will exhibit redder infrared colours (see sec. 1.1.3).

Gutermuth et al. (2009) used the following colour cuts to identify class I sources:

$$\begin{aligned} [4.5] - [5.8] &> 0.7 \\ [3.6] - [4.5] &> 0.7 \end{aligned} \tag{2.1}$$

where [3.6] is the objects magnitude in the 3.6 µm band, [4.5] is the magnitude in the 4.5 µm band, and [5.8] is the magnitude in the 5.8 µm band. Note: the difference in magnitudes is equivalent to the ratio of fluxes. Remaining objects that satisfied the following criteria were then classified as class II sources:

$$\begin{aligned} [4.5] - [8.0] &> 0.5 \\ [3.6] - [5.8] &> 0.35 \\ [3.6] - [4.5] &> 0.15 \end{aligned} \tag{2.2}$$

MIPS 24 µm fluxes were then used to double check the data. Objects with insufficient 24 µm excess that had been previously classified as class I were moved to the class II category. MIPS data was also used to identify transition disks - objects with a substantial gap between the PMS star and the disk which show photospheric signatures in IRAC colours but exhibit an excess at 24 µm; such sources are classified as class II for the purpose of the survey.

Gutermuth et al. (2009) also provide column density maps of gas in the regions,

however, we do not use them since we are interested in the relative distributions of class I and class II sources.

Fig. 2.1 shows plots of the spatial distribution of protostars in two regions from the survey – AFGL490 and Serpens. Red points represent class I sources, while black points correspond to class II objects.

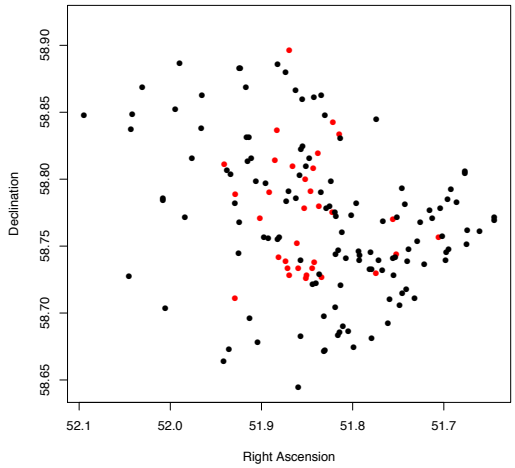
We want to investigate the clumpiness and segregation of the two classes in the regions which is impossible to quantify by visual inspection – the regions in both panels look rather 'messy'.

Further problems are caused by the limited field of view which makes it impossible to identify the 'centre' of the region (if that even makes sense), hence the need for methods that are not biased by the location of the centre of mass.

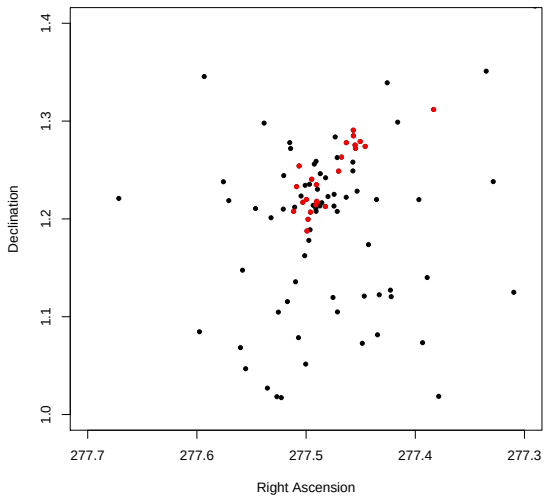
2.2 Minimum Spanning Tree

The methods described in this chapter use a graph theory concept known as minimum spanning tree (MST). A graph is a set of points known as vertices connected by line segments called edges. A graph that connects all the vertices and whose edges do not form any closed loops is called a spanning tree. For a set of vertices it is possible to create a whole family of spanning trees, the shortest of which is known as the minimum spanning tree. By definition the length of the MST is unique for a distribution and in the astronomical problems where MST might be applied (all the edges have different lengths) there is only one way of drawing it. Fig. 2.2 shows an MST of a sample distribution: the thick lines are edges of the MST and the thinner lines are the remaining edges of the underlying distribution.

In order to calculate the distances between the points I converted the right ascension from sexagesimal system to decimal degrees. It is also important to point out that in order to calculate the angular separation between two objects (ω) we need to



(a) AFGL490



(b) Serpens

Figure 2.1: Distribution of class I (red points) and class II (black points) objects in AFGL490 and Serpens regions. Both declination and right ascension are given in decimal degrees.

calculate the great circle distance:

$$\omega = \arccos\left(\sin\delta_1 \sin\delta_2 + \cos\delta_1 \cos\delta_2 \cos\Delta\alpha\right) \tag{2.3}$$

Where δ_1 and δ_2 are the declinations of objects 1 and 2 respectively and $\Delta\alpha$ is the difference in right ascensions between object 1 and object 2. If the difference between the declinations is small, this can be approximated by the formula for objects at the same declination: $\omega = (\delta^2 + \alpha^2 \cos^2(\frac{\delta_1+\delta_2}{2}))^{0.5}$, where α is the right ascension of the object.

2.2.1 Prim's Algorithm

While there are many algorithms that can be used to construct an MST, the work included in this thesis uses Prim's algorithm (Jarník, 1930; Prim, 1957; Dijkstra, 1959). We need to prepare two lists of points, T containing the points that have already been added to the tree and L containing the points yet to be added to the tree. Initially T is empty and all the points in the set are in L. The algorithm is started by picking an arbitrary point and adding it to the tree (we remove this entry from the L and add it to T). Now we look for the shortest edge connecting any point in T with any point in L, we add this edge to the tree and move the point from L to T. We iterate this process until all the points have been moved from L to T

2.3 MST Λ parameter

Since the MST length is a measure of separation of objects within a population, we can use it to compare the distribution of a subset of N special objects (e.g. class I sources or most massive stars within the region) with the rest of the cluster (Allison et al., 2009b).

First we need to obtain the MST length of the subset containing only the N special points, $l_{\rm special}$. Then we create a subset of N points chosen randomly from the *whole* population and calculate the associated MST length. We repeat this process a large number of times and get the median MST value, $\langle l_{\rm norm} \rangle$, along with the variance, $\sigma_{\rm norm}$.

Allison et al. (2009b) proposed a Λ parameter that used MSTs to determine the degree of mass segregation in star forming regions:

$$\Lambda = \frac{\langle l_{\rm norm} \rangle}{l_{\rm special}} \pm \frac{\sigma_{\rm norm}}{l_{\rm special}} \tag{2.4}$$

where the variables have the same meaning as above and the subset of interest consisted of the N most massive stars.

A value of $\Lambda \sim 1$ means that the typical separation of the most massive stars is similar to that of all the stars in the region i.e. there is no mass segregation. Λ significantly greater than 1 means that the typical separation of stars in the region is greater than that of the most massive ones i.e. the region is mass segregated. The higher the value of Λ the stronger the mass segregation. Conversely, Λ significantly smaller than unity is an indicator of inverse mass segregation.

While Allison et al. (2009b) used the parameter to look for mass segregation in star forming regions it is possible to choose another group of stars whose distribution we want to investigate e.g. class I sources within a region, instead of N most massive stars.

2.4 $\mathcal{Q}$-parameter

Cartwright & Whitworth (2004) proposed a $\mathcal{Q}$-parameter as a statistical means to quantify structure and distinguish between 'clumpy' and smooth regions:

$$\mathcal{Q} = \frac{\bar{m}}{\bar{s}} \tag{2.5}$$

where $\bar{m}$ is the normalised mean MST edge length and $\bar{s}$ is the normalised correlation length defined as the average separation between objects in the region divided by the region radius. The mean edge length is given by $\bar{l} = \frac{l_{\rm total}}{N_{\rm total}-1}$, where $l_{\rm total}$ is the MST length and $N_{\rm total}$ is the number of vertices. Cartwright & Whitworth (2004) pointed out, however, that if we wish to compare regions with different N and size, the value of $\bar{l}$ should be normalised. Instead they used a normalised mean edge length, $\bar{m}$, defined as:

$$\bar{m} = \bar{l} / \frac{(N_{\rm total} A)^{0.5}}{N_{\rm total} - 1} \tag{2.6}$$

where A is the area of the region.

Fig. 2.3 taken from Cartwright & Whitworth (2004) illustrates how the value of $\mathcal{Q}$ for computer generated regions is related to the values of fractal dimension F and radial density exponent α. For $\mathcal{Q} \leq 0.8$ the plot shows F against $\mathcal{Q}$; the values of F from 1.5 (highly fractal) for $\mathcal{Q}$=0.45 to 3.0 (uniform density) for $\mathcal{Q}$=0.8 can be read from the left-hand axis. The right-hand side of the figure ($\mathcal{Q} \geq 0.8$) shows α, against $\mathcal{Q}$. The values of α increase from 0 (uniform density) for $\mathcal{Q}$=0.8 to ~ 3 (density $n \propto r^{-3}$) for $\mathcal{Q}$=1.5.

There is a slight kink in the plot at $\mathcal{Q}$=0.8 due to the fact that the $F = 3.0$ fractal has the points distributed uniformly in a grid-like fashion, while the stars in the $\alpha = 0$ cluster are distributed randomly and small density fluctuations may happen.

The $\mathcal{Q}$-parameter allows us to distinguish between clumpy ($\mathcal{Q} < 0.8$) and centrally concentrated systems ($\mathcal{Q} > 0.8$) and quantify the degree of substructure. It has to be pointed out, however, that the method is not very sensitive at $\mathcal{Q} \sim 0.8$ due to the

steep gradient in that section of the plot.

Since the method uses region radius and region area calculated as a circle for normalisation and the value of $\mathcal{Q}$ is correlated with the radial density distribution exponent, it is well suited for analysis of circular clusters. Bastian et al. (2009) pointed out that there is a correlation between the elongation of the region and the $\mathcal{Q}$-parameter. Cartwright & Whitworth (2009) revisited the method and proposed a correction, however, they also pointed out that for regions with aspect ratio < 3 there is no need to apply it, since the effects of elongation have little impact.

2.5 Results

Some of the regions had class I sources mapped over a larger field than class IIs due to different instrument coverage. Those extra sections of the field were disregarded in order to keep our analysis consistent. Furthermore, we discarded 5 regions with the fewest sources ($N_{\rm tot} < 20$) because a meaningful analysis of these regions was not possible.

We used the methods outlined above to compare the distributions of sources in the Gutermuth et al. (2009) data. For each region we obtained the value of the $\mathcal{Q}$ parameter for class I and class II objects ($\mathcal{Q}_{\rm classI}$ and $\mathcal{Q}_{\rm classII}$ respectively) and the whole region ($\mathcal{Q}_{\rm tot}$). Additionally, we calculated the value of Λ for every region using class I sources as the smaller special subset. The results are presented in tab. 2.1.

2.6 Discussion

Parker et al. (2014) demonstrated how $\mathcal{Q}$ evolves with time for systems with different initial virial ratios. If a system is bound (low virial ratio), it will relax on the scale of a crossing time and erase the structure; $\mathcal{Q}$ will go up as time progresses (Parker

et al., 2015). For an unbound system, however, $\mathcal{Q}$ will stay roughly constant – the substructure is 'frozen in' as the system expands – an unbound system has far fewer interactions between stars and on average its members tend to move away from each other.

We observe a range of $\mathcal{Q}$s in Gutermuth et al. (2009) data which means that SFRs form with different values of $\mathcal{Q}$ and/or the $\mathcal{Q}$ changes as the regions evolve. Assuming that class II objects are older than class Is we can postulate that SFRs form at low $\mathcal{Q}$ – if we see high values of $\mathcal{Q}$, it is an indicator of dynamical evolution.

Measuring the current value of $\mathcal{Q}$ can provide clues about the dynamical age of a system and whether it is bound or not. A low value of $\mathcal{Q}$ indicates that a system is dynamically young; it could be unbound but could also be bound and too young to have erased the substructure. On the other hand, a high $\mathcal{Q}$ suggests that the system is bound and dynamically old (has evolved) or simply formed at high $\mathcal{Q}$. However, $\mathcal{Q}_{\mathrm{classI}}$ tends to be low (tab. 2.1) which suggests that all stars form at low $\mathcal{Q}$. This is not unexpected given the clumpy appearance of the gas in star forming regions.

We applied this line of thought to 10 Gutermuth et al. (2009) regions with the 'best data' – i.e. reasonably large number of class I sources, sufficiently large total number of objects and no discrepancies between the class I and class II coverage (these objects have an entry in the last column of the Table 2.1).

We found that for the 4 regions with the highest $\mathcal{Q}_{\mathrm{total}}$ the $\mathcal{Q}$ parameter of class II sources was also higher than that of class I protostars. The value of $p = 0.003$ calculated using the paired t-test comparing the values of $\mathcal{Q}$ of class I and class II indicates that the difference is indeed significant. This is consistent with class II objects being older and more evolved than class Is, hence they will have had more time to relax and delete their structure. Based on our previous assumptions we classify these regions as bound (B in the last column of tab. 2.1).

The other 6 'good' SFRs have a lower $\mathcal{Q}_{\mathrm{total}}$ which indicates that they are dy-

namically younger or unbound (U in the last column). The p value for paired t-test comparing $\mathcal{Q}$ of class I and class II distributions is much higher($p = 0.563$). This result is however not unexpected since between class I and class II the value of $\mathcal{Q}$ goes down, while for others it goes up. The SFRs for which the value of $\mathcal{Q}_{\rm classII}$ is greater than $\mathcal{Q}_{\rm classI}$ yet the overall value of $\mathcal{Q}$ is still quite low (CepC, S140, OphL1688) could be young objects in the process of erasing its structure that are still too young to have erased it all. This result is, however, inconclusive due to small number statistics.

Interestingly, in some regions $\mathcal{Q}_{\rm classI}$ is clearly larger than $\mathcal{Q}_{\rm classII}$ (AFGL 490, IRAS20050+2720). This suggests that the younger population is distributed more smoothly than the older one and could be caused by the the fact that the gas from which the stars form has moved between the formation of the objects that are now class II and ones that are class I. Parker et al. (2014) also showed that for simulated regions the value of $\mathcal{Q}$ can sometimes temporarily drop due to stochastic effects. Furthermore, this result could be caused by the way that $\mathcal{Q}$ is calculated – the area of the cluster is used to normalise the result and in some of the cases the area covered by class I sources could be different than the area covered by class IIs.

Since the way that $\mathcal{Q}$ is calculated is not stochastic in nature, the only uncertainty comes from astrometric errors and incompleteness of the sample. For systems with fewer stars, however, the method suffers from low number statistics and is not very reliable (Parker et al., 2014). This means that the results presented above are not conclusive – they are am interpretation of the differences between the values of $\mathcal{Q}$ for class I and class II sources, which in an idealised case could allow us to infer whether a cluster is bound or unbound.

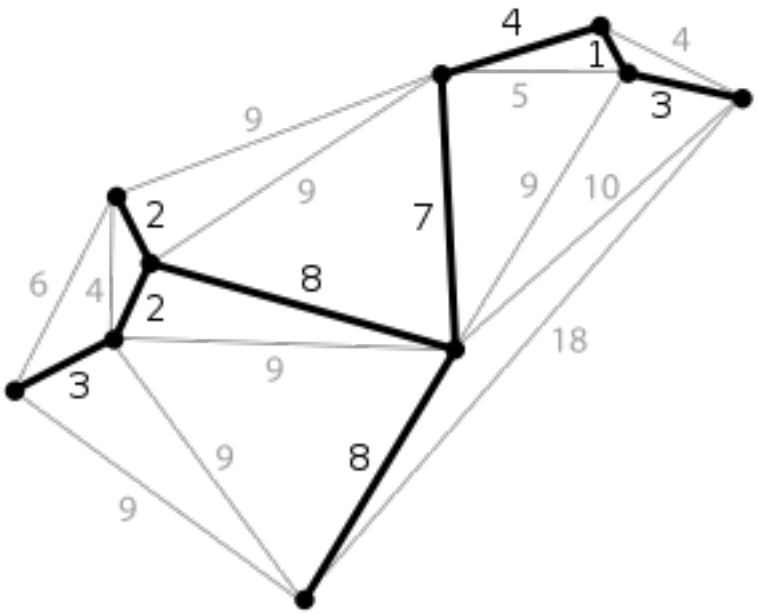


Figure 2.2: Example MST. Numbers give lengths of the edges, the thin grey lines represent the edges of the underlying network, while the thicker black lines are the shortest tree that joins all the points – the minimum spanning tree. Taken from wikipedia.org.

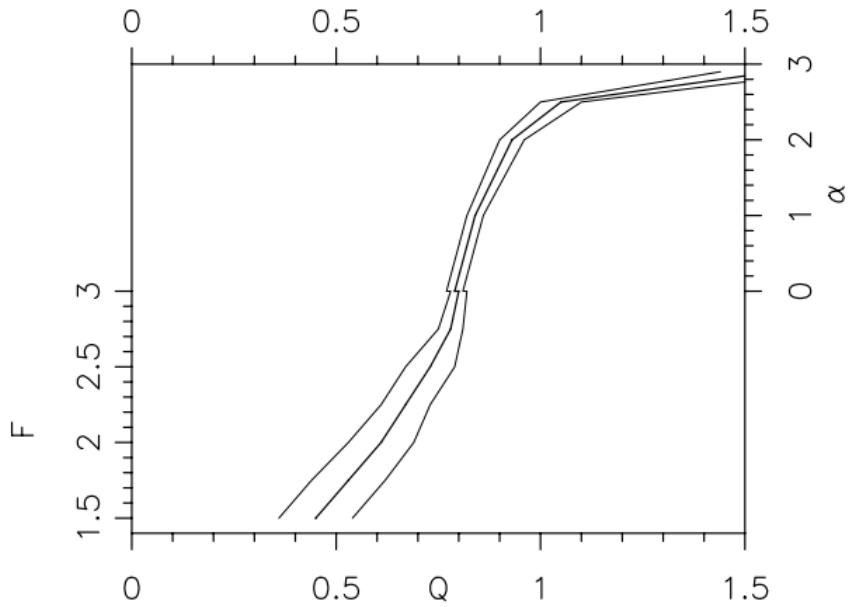


Figure 2.3: Fractal dimension F and spherical density exponent α as a function of $\mathcal{Q}$ for simulated star clusters. For values of $\mathcal{Q} \leq 0.8$ the fractal dimension F can be read on the left-hand y-axis. The values of the spherical density exponent, α,are shown on the right-hand axis for $\mathcal{Q} \geq 0.8$. Taken from Cartwright & Whitworth (2004)

Table 2.1: The properties of all the regions included in Gutermuth et al. (2009). $N_{\rm tot}$ is the total number of objects in the region, $N_{\rm classI}$ is the number of class I sources, $N_{\rm classII}$ is the number of class II objects, $\mathcal{Q}_{\rm tot}$ is the value of the $\mathcal{Q}$ parameter calculated for the entire region, $\mathcal{Q}_{\rm classI}$ is the $\mathcal{Q}$ parameter of the class I sources, $\mathcal{Q}_{\rm classII}$ is the $\mathcal{Q}$ parameter of the class II sources, the next column is the value of Λ and the flag in the last column indicates whether I classified the object as bound (B) or unbound (U).

Region name	$N_{\rm tot}$	$N_{\rm classI}$	$N_{\rm classII}$	$\mathcal{Q}_{\rm tot}$	$\mathcal{Q}_{\rm classI}$	$\mathcal{Q}_{\rm classII}$	Λ	B/U
AFGL490	157	32	125	0.74	0.81	0.71	1.45 ± 0.08	U
IRAS06046-0603	26	3	23	0.92	0.49	0.91	1.26 ± 0.06	
BD+404124	34	8	26	0.82	0.68	0.82	2.11 ± 0.07	
IRAS20050+2720	177	53	124	0.74	0.84	0.72	1.11 ± 0.05	U
CB34	23	7	16	0.91	0.66	0.87	1.48 ± 0.02	
L1206	25	3	22	0.63	0.49	0.81	0.65 ± 0.06	
S106	79	11	68	0.76	0.66	0.78	1.27 ± 0.06	
CepA	94	9	85	0.76	0.68	0.75	1.58 ± 0.11	
L1211	59	12	47	0.75	0.62	0.74	2.57 ± 0.11	
CepC	114	22	92	0.61	0.50	0.61	1.30 ± 0.08	U
L988-e	75	8	67	0.84	0.76	0.80	2.52 ± 0.09	
S140-North	24	13	11	0.74	0.80	0.60	1.81 ± 0.07	
ChaI	38	4	34	0.62	0.58	0.59	0.99 ± 0.38	
LkHalpha101	102	8	94	0.84	0.82	0.88	0.75 ± 0.07	
S140	48	15	33	0.61	0.49	0.62	1.60 ± 0.10	U
CrA	34	9	25	0.77	0.81	0.66	1.74 ± 0.17	
LupusIII	39	2	37	0.86	0.40	0.84	0.88 ± 0.12	
S171	48	11	37	0.92	0.84	0.81	0.90 ± 0.12	
GGD12-15	118	16	102	0.89	0.61	0.87	1.93 ± 0.06	B
MWC297	23	3	20	0.69	0.49	0.70	1.08 ± 0.05	
Serpens	97	24	73	0.91	0.64	0.90	3.00 ± 0.13	B
GGD17	53	11	42	0.91	0.66	0.86	3.02 ± 0.07	
MonR2	230	29	201	0.82	0.80	0.81	1.54 ± 0.07	U
TauL1495	38	9	29	0.45	0.58	0.47	0.97 ± 0.15	
GGD4	31	7	24	0.58	0.70	0.50	1.06 ± 0.04	
NGC1333	133	36	97	0.91	0.72	0.91	1.06 ± 0.13	B
IC348	158	14	144	0.89	0.75	0.91	0.97 ± 0.15	B
NGC7129	57	8	49	0.93	0.75	0.93	1.16 ± 0.09	
VYMon	48	7	41	0.96	0.72	0.96	1.10 ± 0.07	
IC5146	149	7	142	0.77	0.55	0.78	0.91 ± 0.12	
OphL1688	146	32	114	0.72	0.60	0.75	1.20 ± 0.11	U

Chapter 3

Finding structure

The star forming regions are not smooth but have a clumpy substructure that might affect their evolution (chapter 6). Even visual inspection of Taurus and Orion Nebula Cluster reveals, however, that the degree to which these systems are substructured is different. In order to properly assess its impact, we need a way of identifying and quantifying the substructure.

In this chapter I will describe methods that can be used to find structure in data. The presented methods cover a range of approaches – some are density based, while other use graphs and dendrograms. Since many of the methods presented in this chapter have not been previously used in the context of star forming regions, I will also assess their suitability and identify their shortcomings. It is important to emphasise that in this chapter *clusters* should be understood as groupings of stars; the 'clusters' found by the structure finding algorithms will not necessarily satisfy the formal definition of a stellar cluster (Gieles & Portegies Zwart, 2011), or even be physical groupings.

3.1 MST methods

The minimum spanning tree contains information about the structure and the lengths of the edges in a graph and can be used as a measure of the 'density'; it has also been successfully used to quantify the mass segregation in star forming regions (see sec. 2.3). Gutermuth et al. (2009) proposed a way of cutting the MST in an attempt to look for sub-clusters in nearby star forming regions.

First, a minimum spanning tree that includes all the identified young stellar objects in the region is constructed. Then one needs to plot the cumulative distribution function of the edge lengths in the MST. The next step is to fit two straight lines to the distribution; one fits the steep part of the curve (short edges), the other one the shallow part (longer edges). The points around the 'bend' of the function are omitted from the fit since they do not adhere to the proposed distribution and would affect the quality of the fit at the shallow end.

The idea behind this fit is that the slopes correspond to two regimes – the steep part of the distribution contains the short edges found in the sub-clusters, while the longer edges typically connect the background points. The point where the two lines meet marks the *critical cutting length*; edges shorter than that value belong to connections within clusters. The longer edges are removed from the graph; this cuts the tree and the result is no longer a complete graph but a collection of smaller subtrees which are interpreted as sub-clusters.

The length of an edge in the MST graph is related to the local surface density of the objects, therefore using the cutting length method will indeed identify the densest regions of the distribution. While using the critical length approach helps remove the background stars, it can also affect real structure and mistakenly split a single dominant cluster into smaller ones (see fig. 3.3), while in other cases two nearby objects could get blended into one.

By using a single cutting length for the entire observed SFR one assumes that all the sub-clusters have similar densities. Should well defined clusters of different densities be present in the region, this method will fail to identify them. The fact that in the fitting process the central part of the CDF is disregarded also raises concerns, especially given the fact that the cutting length falls in the region of the plot that is not being fitted by definition.

I used this method to analyse a set of four two dimensional test cases generated on a 10 pc×10 pc field (the units are arbitrary) shown in fig. 3.1.

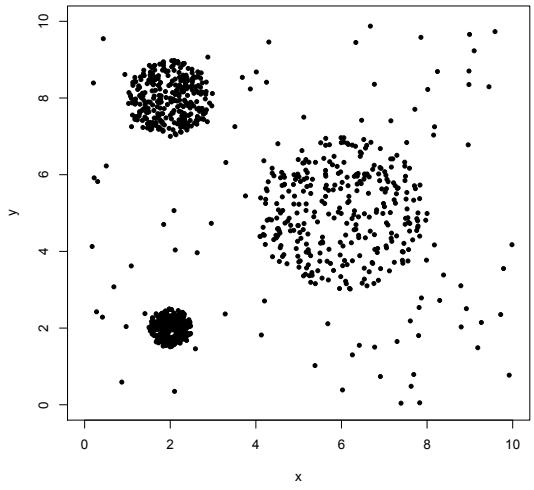
a) Clusters of different densities – three clusters consisting of 300 stars each, with radii 0.5pc, 1pc and 2pc; plotted against a random background of 100 stars.

b) A single cluster of 500 stars on a 500 star random background.

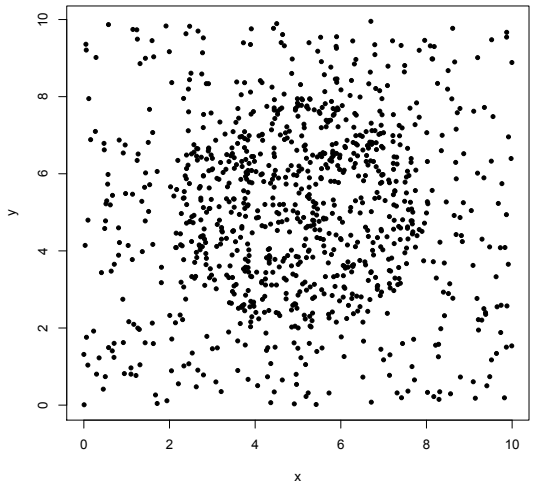
c) A random field of 1000 stars.

d) A random distribution with Binaries. First I generated a random distribution of 680 stars, then I added binary companions to 320 randomly chosen stars. The separation of the binaries was drawn from a gaussian ($\mu = 0.13$pc, $\sigma = 0.08$pc) and the orientation was a random number in range $[0, 2\pi]$.

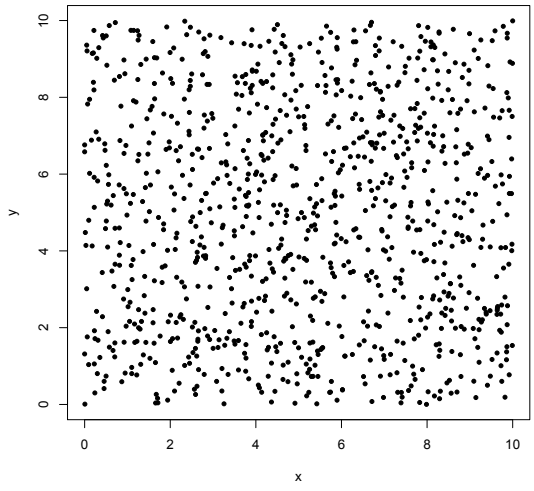
The above cases are intended as simple tests that allow us to reject methods that fail to pass them. More realistic test scenarios could be generated in three dimensions and then be projected onto a plane. They could also include filamentary and fractal structure. Performing such test, however, is not necessary if the methods do not perform well in the idealised cases outlined above.



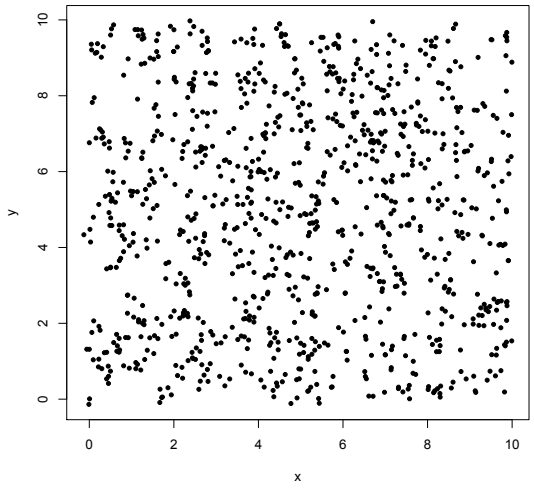
(a) Three clusters of different densities



(b) A single cluster



(c) Random distribution



(d) Random distribution with binaries

Figure 3.1: Plots of distributions used to test the algorithms presented in this chapter. The fields are 10 pc×10 pc and contain 1000 points each.

Fig. 3.2 shows the cumulative distribution functions of the MST edge lengths for each of the above cases. The green line represents the multiple cluster case, the black line corresponds to the single cluster, blue to the random distribution, while the red one illustrates the random distribution with binaries.

Of particular interest is the similarity between the single cluster and the random distribution with binaries. The critical cutting length method has trouble distinguishing between these two cases, since both have two typical lengths – the longer one due to the random distribution and a shorter one due the connections within the cluster in the single cluster case, or the binaries in the case of a random distribution with binaries. This example illustrates how important for any method whose goal is find or quantify structure to be unbiased by binaries.

Fig. 3.3 shows the test regions after the the cutting length method was applied to them. The red symbols are the points that were discarded from the distribution because the edges leading to them exceeded the critical cutting length. The black symbols are the points that were identified as the cluster members and the lines joining them represent the MST edges that were shorter than the cutting length. In the triple cluster case (top left panel) the method identified only the most compact cluster and split the less dense ones into smaller fragments. In the single cluster case (top right panel) applying the critical cutting length method identified the cluster, however, it also classified points from the random background as members of the cluster. The case of a random distribution (bottom left) resulted in a false detection of structure that is not actually present in the data. Similarly, applying the method to the random distribution with binaries (bottom right panel) returned a cutting length that identified clustering where it is actually absent.

Based on the above results, I concluded that this method is not suitable for identifying structure in star forming regions (for an in-depth discussion of the shortcomings of a related method see Parker & Goodwin 2015).

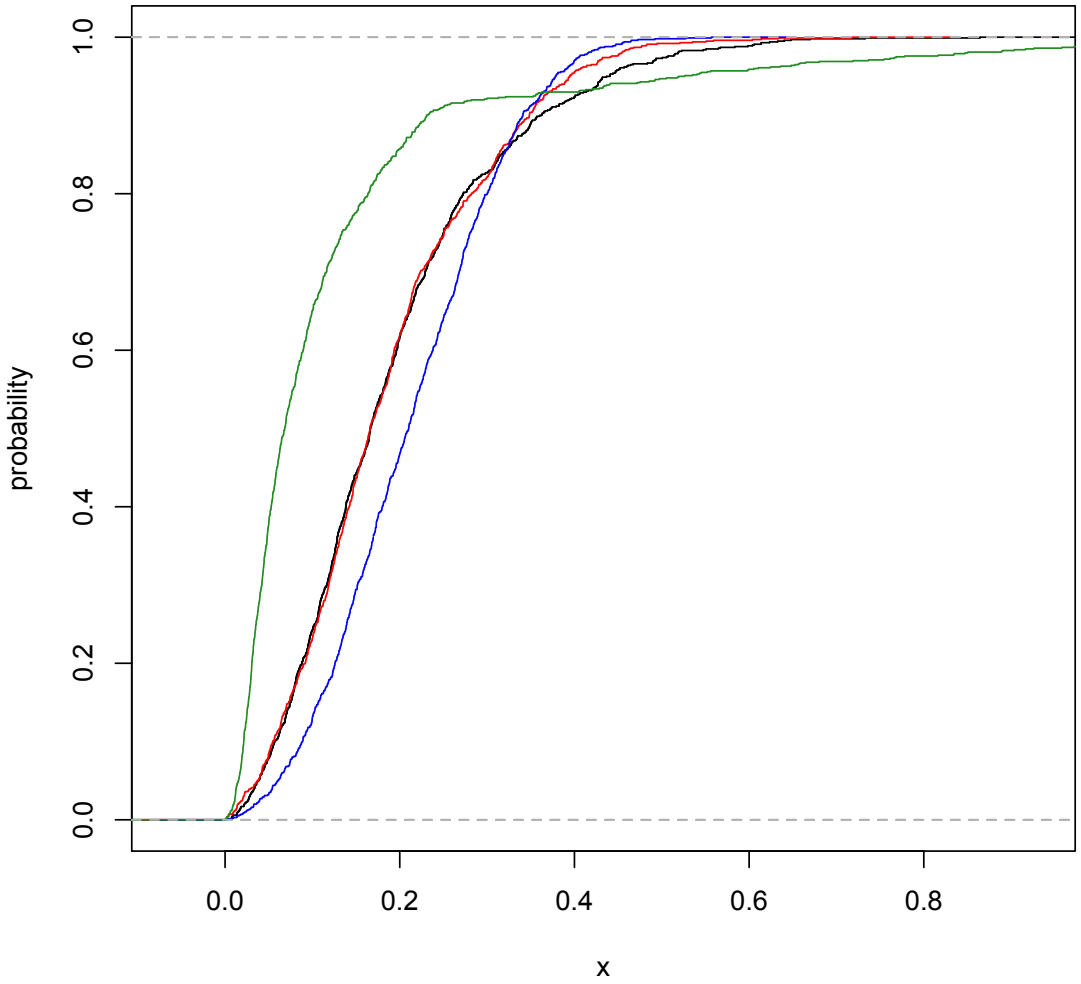
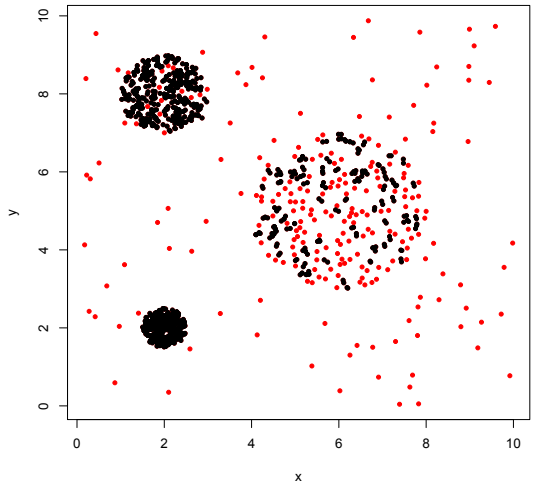
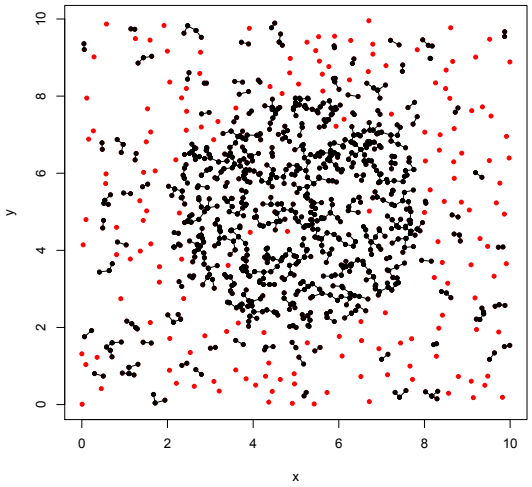


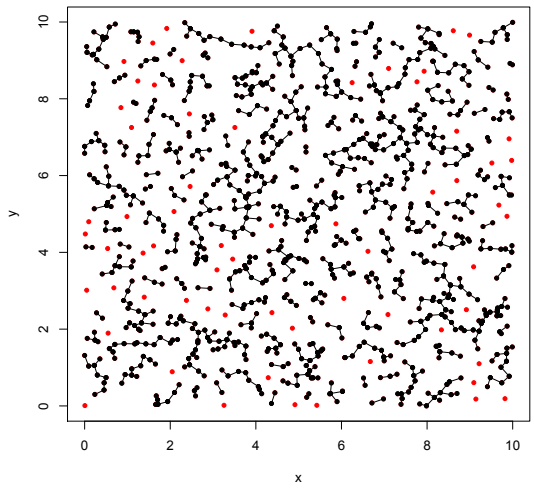
Figure 3.2: Cumulative distribution functions of MST edge lengths for the test cases in fig. 3.1. Three cluster case is represented by the green line, the single cluster by the black line, the random distribution by the blue line and the random distribution with binaries by the red line.



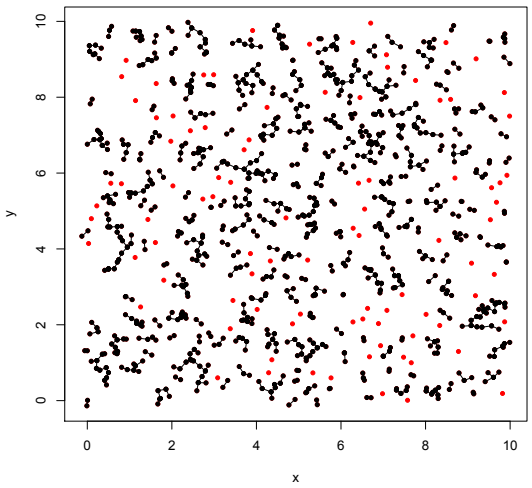
(a) Three clusters of different densities



(b) A single cluster



(c) Random distribution



(d) Random distribution with binaries

Figure 3.3: Plots of the test regions after cutting the MST. The black symbols connected by black lines represent the points that were classified as cluster members i.e. the edges joining them are shorter than the critical cutting length. The red points are the outliers that were removed by the cutting length method.

3.2 Graph Methods

There is a wide range of graph theory based methods that are used to quantify structure. Unfortunately their implementation in astronomical context is not trivial. These methods are primarily used to analyse social, economical and technological networks, in which connections are pre-defined; in star forming regions, however, it is not obvious how the stars should be connected to each other.

The MST is a special type of graph which is not always the best choice. We need to choose a suitable underlying graph; ideally, the 'connectivity' of a point should depend on the proximity to other points, but it also should not have a pre-defined length scale.

A way of achieving this is by connecting each point to its n nearest neighbours. In the previous examples the distance to the n-th nearest neighbour is larger in the more diffuse cluster, the stars in the clusters will be connected in a similar way. To create the underlying network we connect each point to its $N_{\rm n}$ closest neighbours to create a *directed graph* i.e. a graph where the edges have a direction: if point i is connected to point j, it does not necessarily mean that point j is connected to point i. Then we choose all the bidirectional connections ($i \rightarrow j$, $j \rightarrow i$) and discard all the one way connections leaving only the mutual nearest neighbours. Now we can construct an adjacency matrix for the graph – an $N \times N$ matrix (N is the total number of points), whose entry a_{ij} represents the number of connections between points i and j (in this case 0 or 1).

3.2.1 Clustering Coefficient

A commonly used measure of clustering is 'transitivity', also referred to as the *global clustering coefficient*, C (Costa et al., 2007). It is defined as the ratio of the number of closed triples of points to the total number of connected triples.

$$C = \frac{3N_{\triangle}}{N_3} \tag{3.1}$$

where $N_{\triangle}$ is the number of triangles in the graph and N_3 is the number of connected triples. Points P_1, P_2 and P_3 form a connected triple, if P_1 is connected to P_2 and P_2 is connected to P_3. For a triple to be considered a triangle the additional requirement of P_1 being connected to P_3 has to be satisfied. A social network equivalent of a triangle would be two of my friends independently being friends with each other. Note the factor of 3 used to account for the fact each triangle can be interpreted as 3 closed triples ($P_1P_2P_3$, $P_2P_3P_1$ and $P_3P_1P_2$).

The adjacency matrix can then be used to identify triples, triangles and other types of structure within the graph (if there is a connection between point i and j, then $a_{ij} = 1$; if the points are not connected $a_{ij} = 0$). Points i, j and k form a triple when $a_{ij} = a_{jk} = 1$ i.e. there is a connection between points i and j and between points j and k . For these three edges to form a triangle an additional requirement of $a_{ik} = 1$ needs to be satisfied, i.e. there is also an edge that connects points i and k. In order to calculate the clustering coefficient in an efficient way I used the `igraph` package in R and C++ (Csardi & Nepusz, 2006; R Core Team, 2012).

A related concept is the *local* clustering coefficient, C_i, defined as the ratio of the number of edges connecting the neighbours of the i-th point to the number of possible connections between the neighbours:

$$C_i = \frac{2l_i}{k_i(k_i - 1)} \tag{3.2}$$

where k_i is the number of neighbours of the i-th point and l_i is the number of connections between them (Watts & Strogatz, 1998). The local clustering coefficient can also be interpreted as the ratio of the number of triangles that a point belongs to to the number of triples. The average of the local clustering coefficients is known as

the *network clustering coefficient*, $\widetilde{C} = \frac{1}{N}\sum C_i$ and it is not the same as the global clustering coefficient.

In order to choose the most suitable value of N_n I investigated how the transitivity coefficient changes with N_n. I generated 1000 underlying graphs for values of N_n in range [1,1000] for each of the four test cases from the previous section. The results are shown in fig. 3.4; green line corresponds to the three clusters case, black to the single cluster, red to random distribution with binaries and blue to random distribution.

In the limit of $N_\mathrm{n} \to N$ the underlying distribution is a complete graph i.e. each point is connected to every other point and C tends to unity. On the opposite extreme are the triangle-free graphs; trivially, $C = 0$ for $N_\mathrm{n} = 1$.

A striking feature of the plot is the very well pronounced local maximum of the three clusters case (green line) that occurs around the value $N_\mathrm{n} = 300$. This represents the points in the dense regions preferentially forming triples – among 300 nearest neighbours of a point in the clump most are also in the same clump because each cluster has $N = 300$. For higher values of N_n the transitivity decreases again, because not all triples are now closed – some of the connections are now reaching to points in two distinct clumps that are less likely to be connected to each other and form triangles.

The random distribution with binaries (red line) exhibits a very small spike at lower values of N_n. This is caused by the presence of binaries – every binary member's nearest neighbour is the other member; if a point is connected to a binary member it is also likely connected to the other one and together they form a triangle. If one naively tried to identify the best connected points by simply using $N_\mathrm{n} = 2$, the result would be misleading – the binary members are the best connected ones, however, that does not imply that there is an underlying structure associated with this increased transitivity.

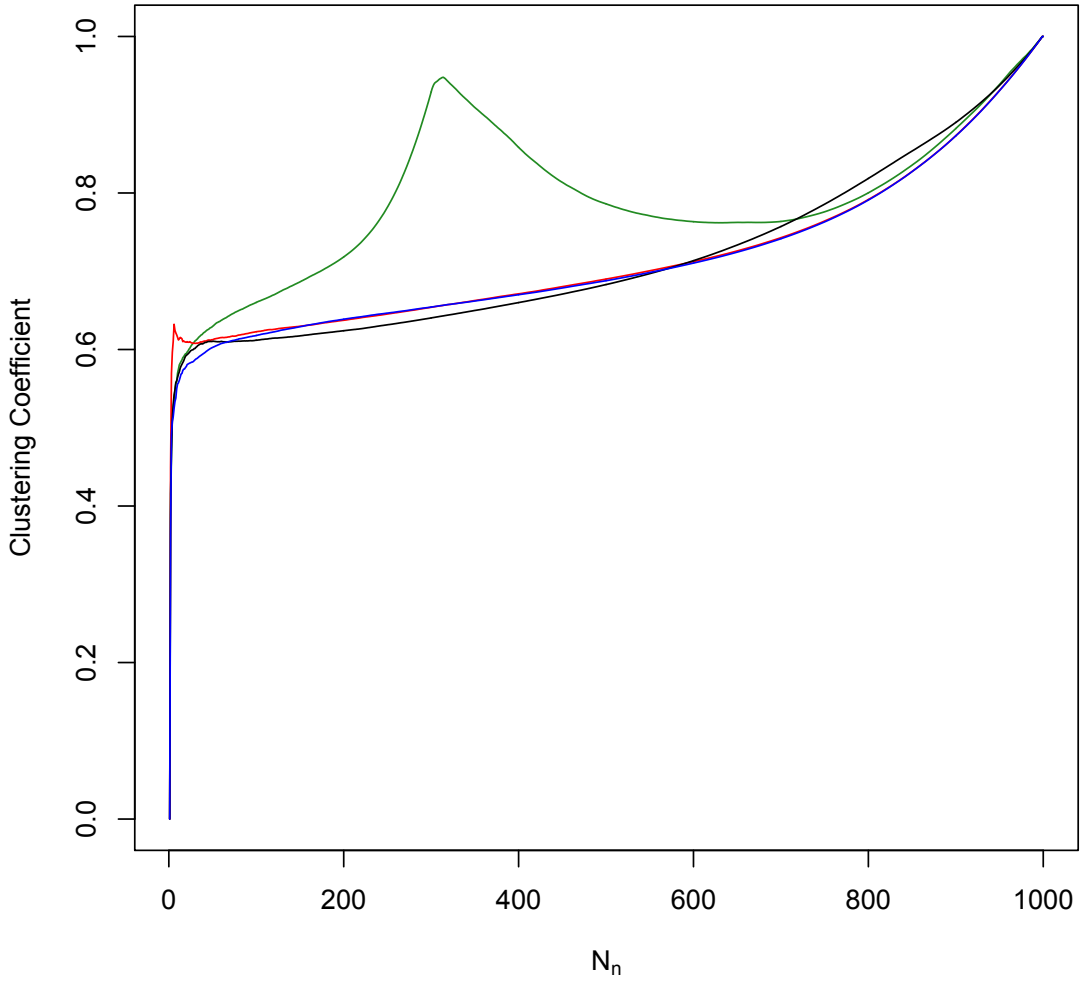


Figure 3.4: Plot of global clustering coefficient against N_n for the multiple clusters case (green line), single cluster case (black), random distribution with binaries (red) and random distribution (blue).

3.2.2 Cliques

A clique is an extension of the triangle concept – it is a subset of points in a graph that are all connected to each other i.e. all the neighbours of a given point in the clique are also each other's neighbours. To explore possible applications of cliques in identifying the structure in star forming regions I used the same test cases as in the previous sections i.e. I constructed an underlying graph in which each point is connected to its N_{n} nearest neighbours and chose the bidirectional connections (i is connected to j and j is connected to i) for the distribution.

When looking for cliques in a network it is important to keep in mind a property of cliques that any subset of the points in a clique is also a clique. This gives rise to the notion of a *maximal* clique i.e. a clique that is not a subset of another larger clique. This is not necessarily the same as the *maximum* clique which is the largest clique in the graph.

To find a maximal clique we start with a single point and add a connected point to the clique, then we look for a point connected to both of the members. We repeat this process until there are no more points that are connected to every point in our clique. However, a single point can belong to more than one maximal clique as shown in fig. 3.5. There are two maximal cliques in this network – one containing points 1, 2, 3, 4 and another one consisting of points 1, 2, 3, 5. This illustrates a problem for the clique finding algorithm as at each step we need to keep track of all the points that could be added to the clique. This process can become very complicated for large data sets, therefore I used the implementation of the algorithm provided in the `igraph` library in C and in R (Csardi & Nepusz, 2006).

In order to identify regions of the graph where the points are particularly well connected, I investigated how the clique membership depends on N_{n}. When $N_{\mathrm{n}} = N$ the graph is complete and all the points belong to one maximal clique. On the other

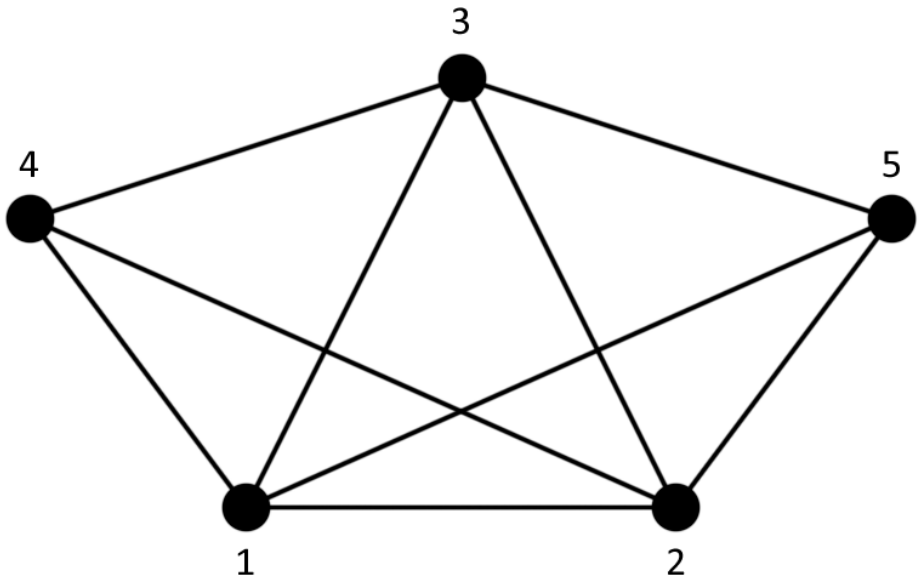


Figure 3.5: Two maximal cliques that differ only by one point. Points 1, 2, 3, and 4 form a maximal clique, but so do point 1, 2, 3, and 5.

end of the spectrum, for $N_{\mathrm{n}} = 2$ the only cliques present (other than the trivial cases of individual edges and points) are triangles.

The background points will form small cliques too, but are less likely to form large ones, therefore we discard cliques with $n < 10$ in order to only keep the cluster members. This method, however, is flawed too – a single cluster could consist of several maximal cliques, some of which might contain fewer than 10 objects. This would result in some of the cluster members being discarded.

The parameter N_{n} sets the maximum size of a clique, therefore it should be higher than the cutoff value of 10. Setting it to too high a number, however, will result in connections between unrelated regions. Fig. 3.6 shows the cliques extracted from the distribution of three clumps used in the previous sections. Black points are the members of maximal cliques with 10 or more members, red points belong to the original distribution but were discarded since they do not belong to a clique that

satisfies the above requirements.

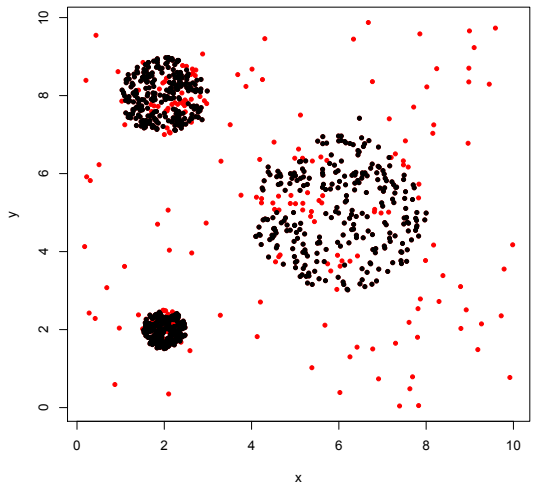
For an underlying network with $N_{\mathrm{n}} = 25$ (top left panel) the clique extraction discards all the background points, but it does not identify all the members of the clusters – it discards the points that only belong to maximal cliques with fewer than 10 members. For a network with $N_{\mathrm{n}} = 30$ (top right panel), almost all the cluster members are identified correctly and the background points are still discarded. When the N_{n} value is further increased to 35 the clusters members are identified correctly, however, there are also false positives – some of the background points start forming cliques with 10 or more members.

In comparison with the critical cutting method (panel (a) of fig. 3.3) the results produced by the cliques based method are more satisfactory. Three clusters of different densities are identified correctly because there is no implicit density cutoff, instead the method relies on local connectivity and the difference in densities between the clusters and the background is the deciding factor. Among N_{n} nearest neighbours of a background point many will be cluster members; on the other hand, a cluster member is far less likely to have a background point as one of its N_{n} nearest neighbours.

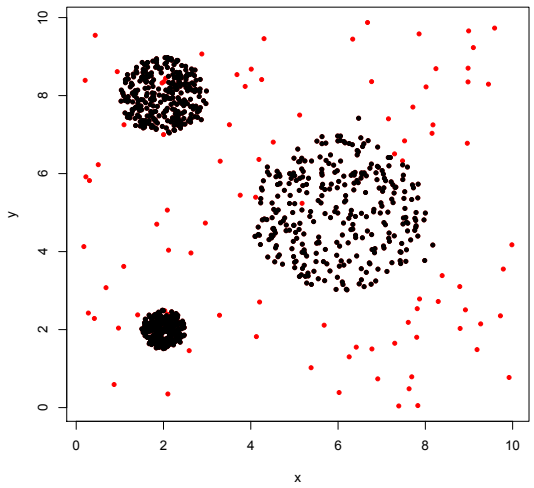
Clique extraction, however, suffers from a major shortcoming that makes it not suitable for detecting substructure – the choice of the parameter N_{n} and the minimum clique size n is arbitrary and based on visual inspection. Furthermore the method does not provide a meaningful statistic that would describe the extent of clustering in the region.

3.2.3 Small-world networks

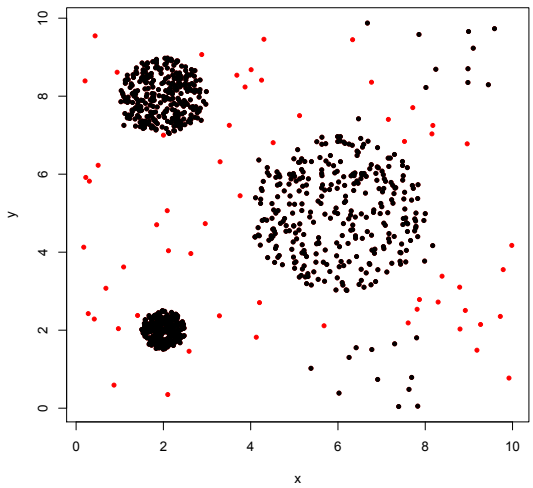
Small-world networks are a concept related to the transitivity and clustering, they are locally clustered, but have a short average path length which is a property of



(a) $N_{\mathrm{n}} = 25$



(b) $N_{\mathrm{n}} = 30$



(c) $N_{\mathrm{n}} = 35$

Figure 3.6: Plots of the distribution with three clusters following a clique extraction. The black points are members of cliques with $10 \geq$ members, the red points are members of the original distribution that do not belong to such cliques. Top left panel shows the result of clique extraction for $N_{\mathrm{n}} = 25$; top right panel for $N_{\mathrm{n}} = 30$; the bottom panel for $N_{\mathrm{n}} = 35$.

random graphs (Watts & Strogatz, 1998). Examples of small-world networks range from connections between actors (Six Degrees of Kevin Bacon) or scientists (Erdős number) to power grids and neural networks.

Humphries & Gurney (2008) proposed a measure of *small-world-ness* (S) which makes it possible to quantitatively decide whether a graph is a small-world network. In order to calculate S of a graph g we need to introduce two parameters:

$$\gamma_g = \frac{C_g}{C_{\text{random}}} \tag{3.3}$$

where C_g is the clustering coefficient of the graph g and C_{random} is the clustering coefficient of a random graph with the same number of edges and points.

$$\lambda_g = \frac{L_g}{L_{\text{random}}} \tag{3.4}$$

where L_g is the average path length of the graph g and L_{random} is the average path length of a corresponding random graph. For a small-world network typically $L_g \geqslant L_{\text{random}}$.

A random graph with a given number of edges and vertices, where each possible vertex pair has an equal probability of having an edge connecting them is known as an Erdős-Rényi (Erdős & Rényi, 1959) graph. To create an Erdős-Rényi graph we choose the number of edges (m) and points (N) and generate m pairs of random integers in range [0,N], for each pair we draw an edge between the corresponding points. One of the properties of the Erdős-Rényi graphs is that they have fewer closed loops than real-life networks (their transitivity is lower), therefore for a small-world network $C_g \gg C_{\text{random}}$.

For each value of N_{n} in the range [1,50] I generated 100 Erdős-Rényi graphs with different random number seeds in order to obtain the mean values of γ_g and λ_g. The small-world coefficient is defined as the ratio of these parameters:

$$S = \frac{\gamma_g}{\lambda_g} \tag{3.5}$$

By combining the above equation with inequalities $L_g \geqslant L_{\mathrm{ER}}$ and $C_g \gg C_{\mathrm{ER}}$, one can deduce that a graph is small-world network if $S > 1$. Humphries & Gurney (2008) show that using the average local clustering coefficient instead of the global coefficient gives slightly different results, however, the behaviour of the coefficient S does not change.

Fig. 3.7 shows small-world coefficient S as a function of N_{n} for the range [1,50] for the test cases. In multiple cluster case (green line) the value of S drops very quickly and stays very close to zero for the rest of the range. This is caused by the increase of the average path length as connections between the clusters start appearing. The sharp initial decline is similar for the single cluster (black line), the random distribution (blue line), and the random distribution with binaries. Between 7 and 10, however, the value of S rises again for these distributions, because in this range γ_g increases faster than λ_g for these distributions. Beyond the range of the plot the values of the S coefficient fall monotonically for all four test cases. While this figure shows that the behaviour of the small-world coefficient is different for the case with three clusters, it still fails to distinguish between the single cluster case and the random distribution based cases, which disqualifies this coefficient as a useful diagnostic of clustering.

3.2.4 Rich-club coefficient

In social networks in addition to cliques and small-world networks a phenomenon known as a *rich-club* can also be observed. Let us define the prominence of a point as its degree i.e. the number of edges connected to it. A rich-club is defined as a subset of the network consisting of the most prominent points that are also well

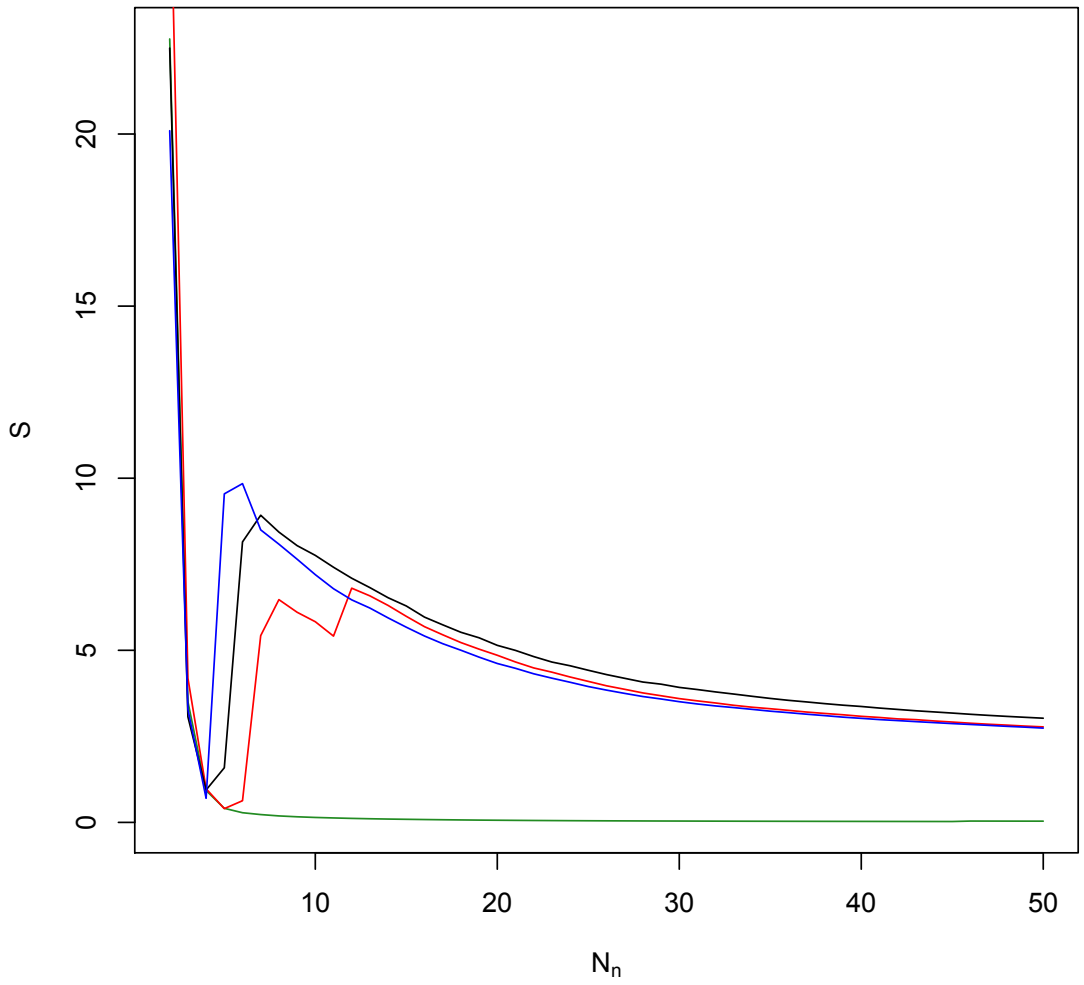


Figure 3.7: The dependence of the small world coefficient S on the number of nearest neighbour connections in the underlying distribution(N_{n}). The green line correspond to the distribution with three clusters, the black one to the single cluster case, the red one to the random distribution with binaries, and the blue one to the random distribution.

connected to each other. Zhou & Mondragon (2005) proposed a rich-club coefficient, $\phi(k)$, that quantifies how well the points with a degree $\geq k$ are connected to each other:

$$\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k} - 1)} \tag{3.6}$$

where $E_{>k}$ is the number of edges connecting the members of the club and $N_{>k}$ is the number of points with degree greater than k (rich-club members).

To choose a value of k, one needs to construct a cumulative distribution function of the degrees of the points in the graph and find a value for the upper 20th percentile. However, the discrete nature of the degree distribution will often result in the same value being shared by several points and therefore more than 20 per cent of the points might belong to such a rich club.

I applied this method to the previously described underlying distributions that join each point to its $N_{\rm n}$ nearest neighbours for the four test cases. Fig. 3.8 shows the dependence of $\phi(k)$ on N_n. Interestingly, the distribution with three clusters (green points) behaves differently to the single cluster, binaries and random distribution (black, red and blue symbols respectively). For $N_{\rm n} < 150$ the green points have substantially lower values and the curve they trace is shallower until ~ 300 where it becomes steeper until it meets the other lines at $N_{\rm n} \sim 500$. The shallow shape between 150 and 300 is due to the fact that the most connected points will lie in different clusters and will be relatively poorly connected to each other. For $N_{\rm n} > 300$ the connections between clusters start appearing and the amount of connections between the members of the rich-club increases rapidly. The test case is set up from three clusters with the same N, therefore ϕ is showing a particular mode of clustering particular to this (not very realistic) distribution. If the clusters had different number of members, the distribution would be much smoother and similar to the other cases.

Therefore, I concluded that this method is not well suited for analysing this kind of underlying graph.

3.2.5 Weighted rich-club coefficient

The rich-club coefficient also has a weighted form which can be applied to a different kind of underlying network (Opsahl et al., 2008). Instead of finding N_{n} nearest neighbours of each point, let us consider a complete graph i.e. one in which each point is connected to all the other points. Each of the edges has a weight w_i dependent on its length (r_i). We want the shorter edges to have a higher weight therefore we choose an inverse square law: $w_i = r_i^{-2}$.

We define the weight of a point (W_i) as the sum of all the edge weights originating in this point. We then construct a CDF of point weights and make a cutoff to identify the rich-club of k best connected points. We define s as the weight of the point with the lowest weight in the rich-club.

The next step is to find all the connections between the members of the rich club. From basic combinatorics, if each point is connected to every other point, there are $n = k(k-1)/2$ edges in a graph of k points. The weighted rich-club coefficient is defined as the ratio of the sum of the weights of edges connecting the rich-club points to the sum of k points with the highest weights in the entire distribution.

$$\phi^{\mathrm{w}} = \frac{W_{\geq s}}{\sum_{i=1}^{n} w_i^{\mathrm{rank}}} \tag{3.7}$$

The parameter $W_{\geq s}$ is the total weight of the points in the rich club and w_i^{rank} is the edge with the i-th highest weight.

Fig. 3.9 shows the dependence of the weighted rich club coefficient (ϕ^{w}) on the size of the rich-club (k) for the previously described test cases. Red symbols correspond

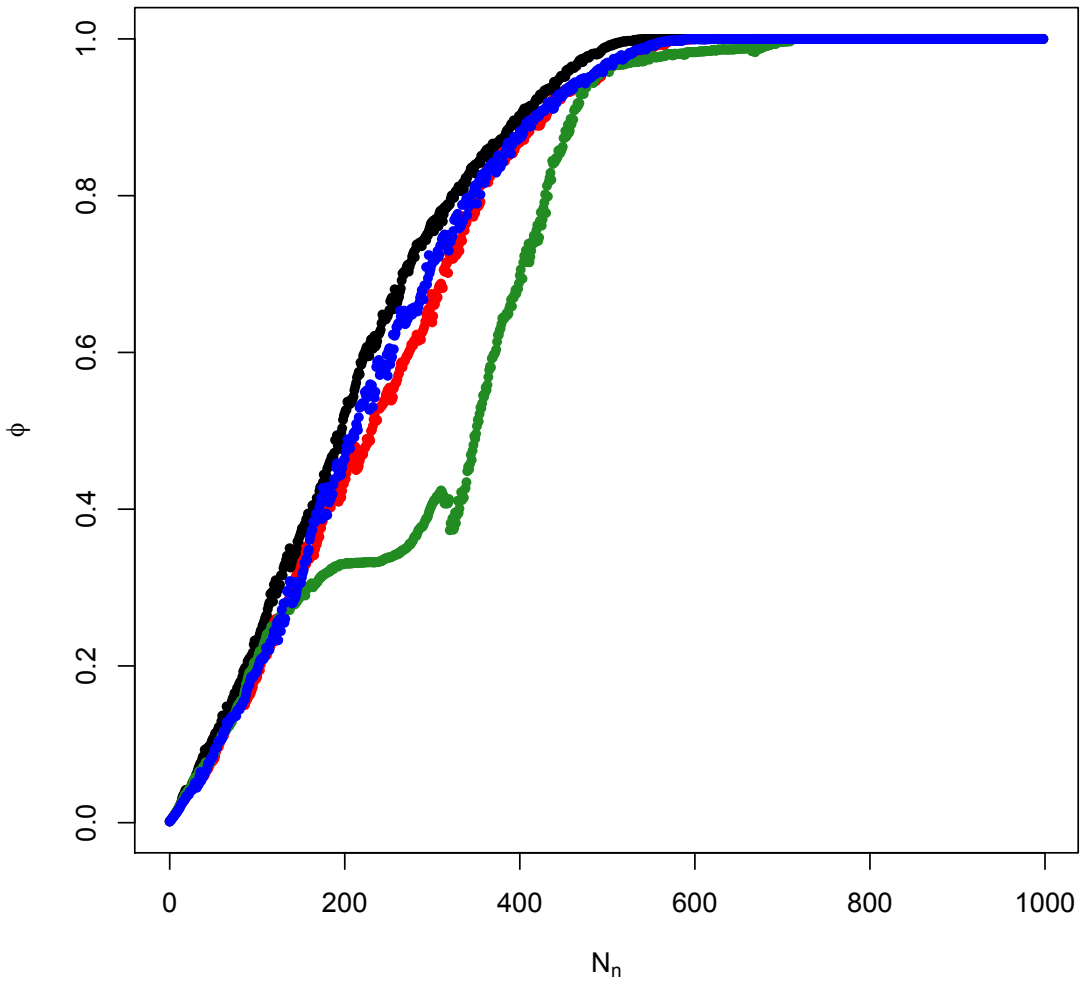


Figure 3.8: The rich-club coefficient (ϕ) as a function of the number of nearest neighbours connections in the underlying distribution (N_{n}) for the test cases. The green points represent the distribution with three clusters, the black ones the single cluster case, the red ones the random distribution with binaries, and the blue ones the random distribution.

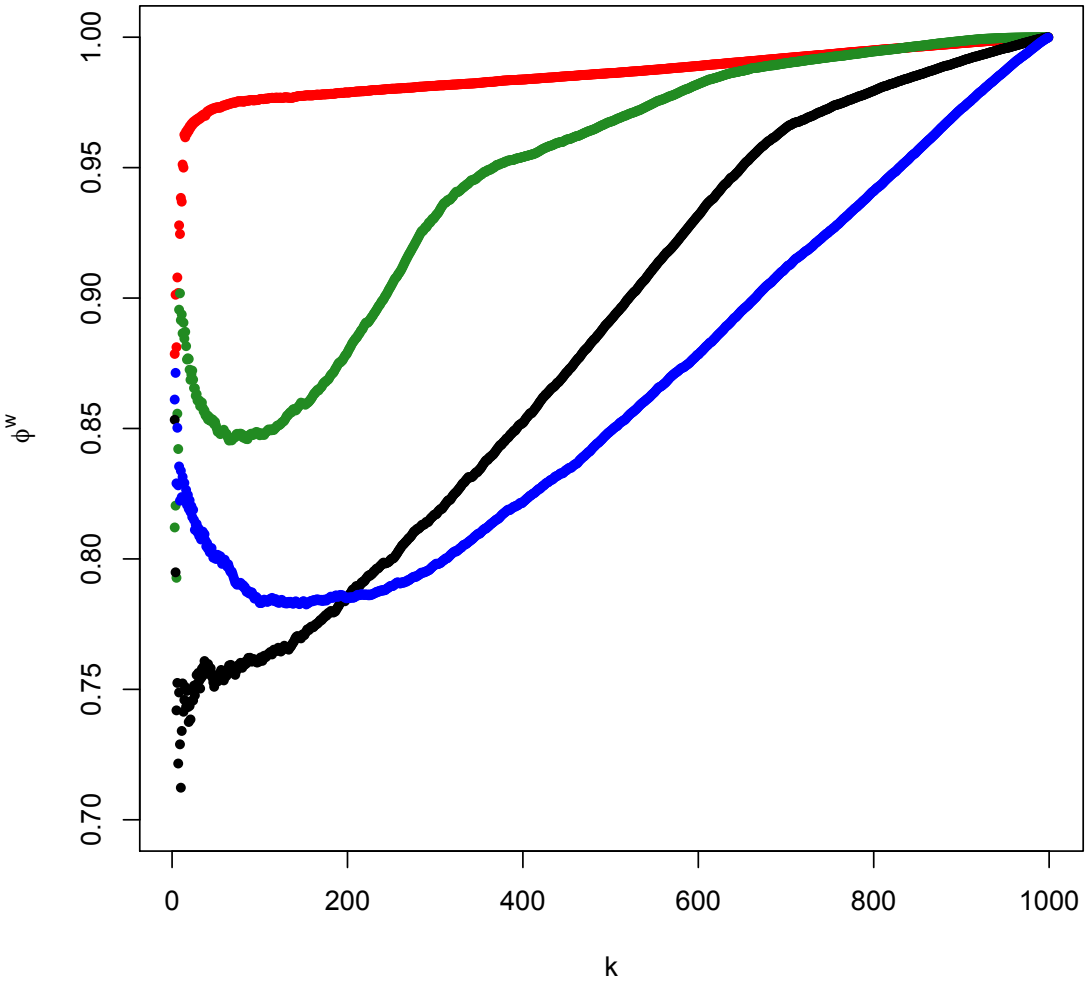


Figure 3.9: The dependence of the weighted rich-club coefficient (ϕ^{w}) on the size of the rich club (k) for the test cases. The green points correspond to the distribution with three clusters, the black ones to the single cluster case, the red ones to the random distribution with binaries, and the blue ones to the random distribution.

to the random distribution with binaries, green points to the three clusters case, black to a single cluster and blue to a random distribution). As k tends to 1000 all the curves approach unity, however, one should bear in mind that it conceptually makes little sense to define the rich-club as a subset of more than 50 per cent of the region.

Remarkably, the random distribution with binaries boasts the highest values of ϕ^{w}, while the single cluster case has the lowest rich-club coefficient. This method is obviously biased by binaries – points that belong to the binaries will typically have higher values of W_i and will be well connected to other binary members. Given the ubiquity of binary systems this method can not be successfully applied to SFRs.

Opsahl et al. (2008) proposed a parameter ω which compares the network against a null-case constructed by reshuffling the weights of the edges and calculating the weighted rich-club coefficient for the new distribution. To find a typical value of ϕ for a randomised network this process is repeated 1000 times and then the average ($\phi^{\mathrm{w}}_{\mathrm{null}}$) is taken. Dividing ϕ^{w} by the average null-model result provides a normalisation factor:

$$\omega = \frac{\phi^{\mathrm{w}}}{\phi^{\mathrm{w}}_{\mathrm{null}}} \tag{3.8}$$

To create a null model we need to reshuffle the weights of the edges, however in the case where the weights depend on Euclidean distances in a spatial distribution such a randomisation means that the edge weights of the reshuffled graph no longer correspond to the physical distances between the points. Instead I chose to pick k points at random and find the weights of the edges that connect them. This allowed me to calculate the value of ϕ for this subgraph by summing the total weight of these connections. This approach make calculating $\phi^{\mathrm{w}}_{\mathrm{null}}$ analogous to finding the normalisation factor in the MST Λ method (see eqn. 2.4).

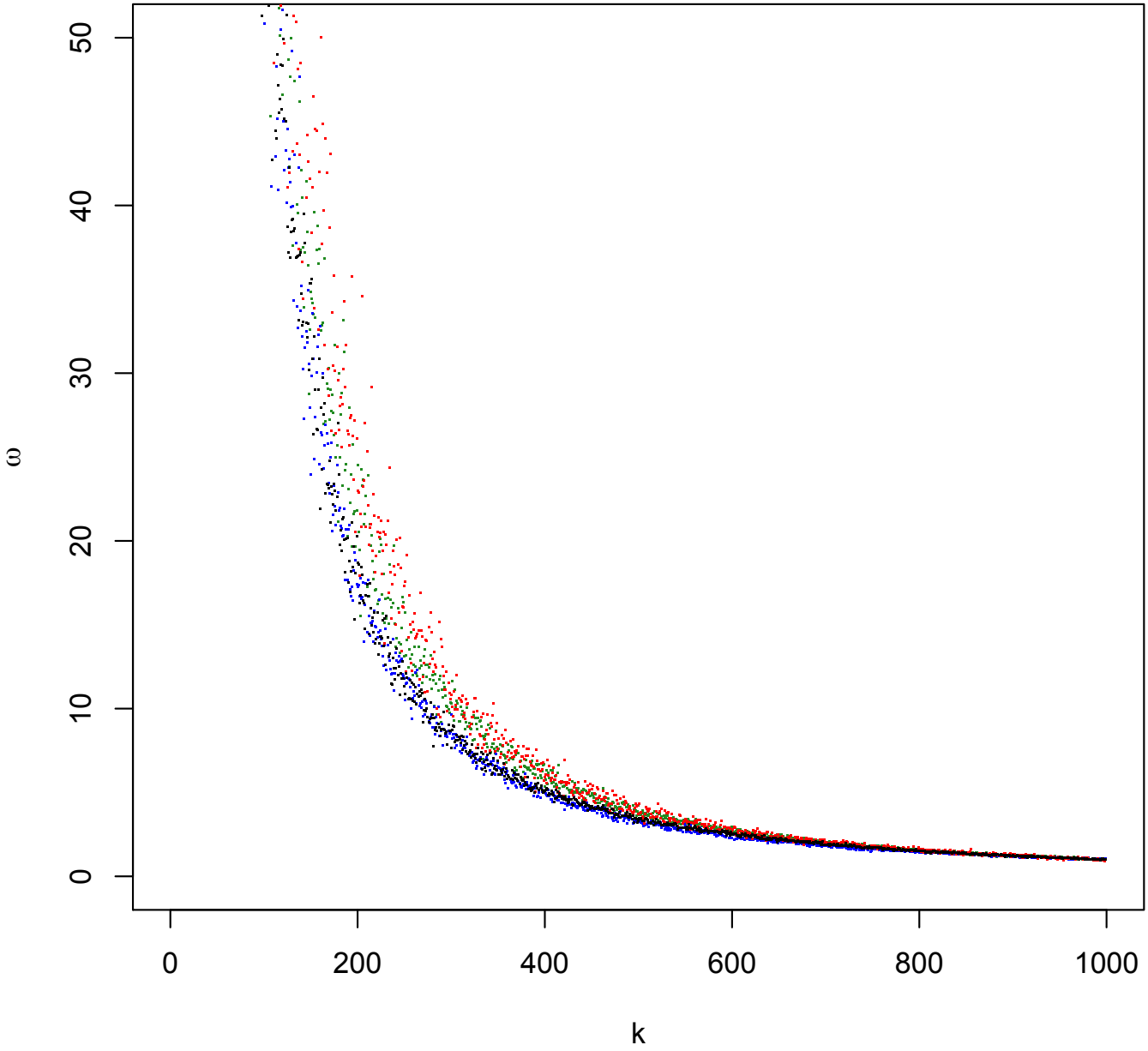


Figure 3.10: The null case normalised weighted rich-club coefficient (ω) as function of rich club size (k) for the previously outlined test cases. The green symbols represent the case with three clusters, the black ones represent the single cluster case, the red ones correspond to the random distribution with binaries, and the random distribution is represented by the blue ones.

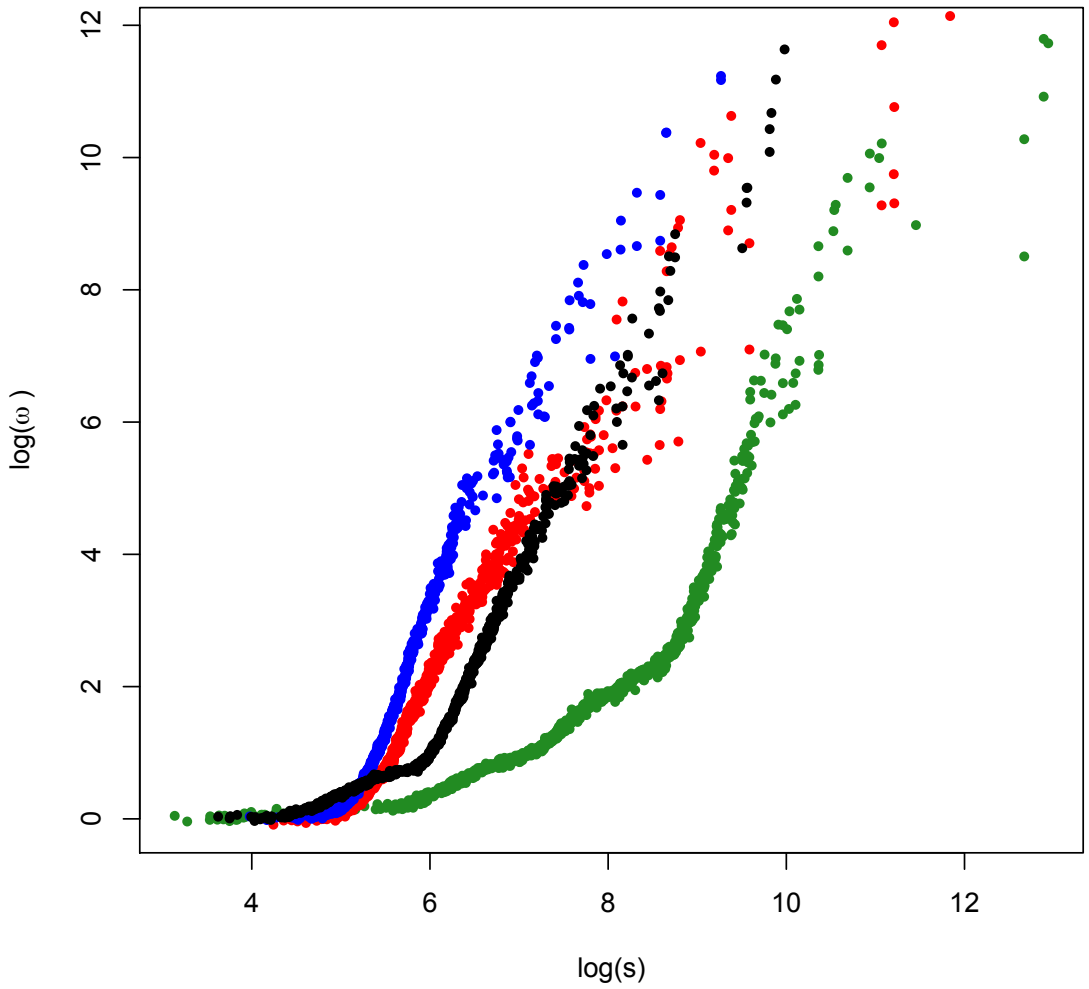


Figure 3.11: The logarithm of the null case normalised weighted rich-club coefficient (ω) as a function of the vertex strength(s) for the test cases. The multiple clusters case is represented by the green symbols, the single cluster case by the black ones, the random distribution with binaries by the red ones, and the random distribution by the blue ones.

Fig. 3.10 shows the dependence of the null model normalised weighted rich-club coefficient ω on the size of the rich club (k). The vertical scatter is caused by the random nature of the null case sampling. All the test cases show a sharp decay with increasing k. The red points (random distribution with binaries) lie slightly higher than the the green points (three clusters) which in turn take higher values than the black (single cluster) and blue ones (random distribution). Overall, however, there is little quantifiable information that can be extracted from this plot.

An alternative presentation of the results of the same analysis is shown in fig. 3.11. Parameter ω is plotted against the cutoff strength (s) on a log-log scale. The prominent points in the three clusters region (green symbols) have higher absolute strength, but as the previous plot demonstrated they are not necessarily well connected to each other. The strength of the point can be interpreted as a measure of local density and this plot indicates that the points in the triple cluster case are typically found in denser environments than the single cluster case, which is denser than the random distribution based ones. This behaviour, however, is specific to the test cases; in practice the clusters could have a range of densities and the one with the most subclusters will not necessarily have the strongest points.

3.2.6 Discussion

Methods of quantifying structure based on graph theory have been studied extensively and are frequently applied in the context of social and computer networks. The results they produce when applied to star forming regions, however, are rather unsatisfactory.

The principal difference is that the connections between computers or people are well defined and can be described in a binary fashion – either the connection exists or it does not. I tried to circumvent this problem by using an inverse square edge weighting for an underlying distribution wherein each point is connected to every

other point in the distribution. By doing so I made the notion of connectivity more 'fuzzy' and reduced the power of the methods.

Furthermore, many methods described in this section are strongly affected by binaries and short connections; in a social network the most important connections are to ones immediate neighbours – friends of ones friends have a lesser impact. In the case of a star forming region it is important to treat the system as a whole and not get fixated on the short links since the more distant objects can still dramatically impact the cluster's evolution.

3.3 Hierarchical Clustering

Hierarchical clustering encompasses a family of methods in the field of cluster analysis used to construct a hierarchy of structures in the data. This approach could be appropriate for the analysis of SFRs given their hierarchical nature; clusters may consist of smaller subclusters and might themselves be parts of a structure on a larger scale (see sec. 1.2).

There are two strategies for constructing the hierarchy of clusters: bottom up (agglomerative) and top down (divisive). The agglomerative methods start with each point belonging to a separate cluster at each step two clusters with the smallest intergroup *dissimilarity* are merged into one. The process is repeated until only one cluster encompassing all the data remains. Conversely, the divisive methods start with a single cluster that is divided into two parts. The resulting subclusters have largest possible dissimilarity. At each subsequent step one of the clusters is divided again until each point is in a separate cluster

While the two approaches are quite similar, the results they produce might differ. The work in this thesis focuses on the agglomerative paradigm, since in addition to being computationally less demanding, it also ensures that objects with low dissimi-

larity are clustered together.

3.3.1 Measures of dissimilarity

Dissimilarity is a measure of 'distance' between objects. In case of the spatial clustering of stars it is appropriate to use a Euclidean metric. However, the concept of dissimilarity encompasses more than just physical distance – it can also be applied to families of languages, gene groups, etc.

While finding the distance between two individual points is trivial, finding the dissimilarity between two clusters is more challenging - we need to assume an appropriate definition of intercluster dissimilarity and keep it consistent throughout the analysis. In the following subsections I will outline methods of calculating the dissimilarity between sets based on an example of two sets Q and R shown in fig. 3.12.

Single-linkage clustering

Single-linkage clustering (SLINK) is perhaps the most intuitive way of finding the distance between clusters Q and R. The dissimilarity between clusters Q and R is simply the smallest distance between an object in Q and an object in R:

$$d_{\mathrm{SLINK}}(Q, R) = \min_{i \in Q, j \in R} d_{ij} \tag{3.9}$$

where i is a point in set Q and j is a point in set R (panel a) of fig. 3.12). This method (also referred to as nearest neighbour linkage) has the property of *chaining* – whenever two clusters come too close to each other they will merge. The resulting cluster might contain members that are very far away from each other, therefore the method excels at finding elongated clusters.

Complete-linkage clustering

Conversely, we can choose complete-linkage clustering (CLINK) which defines the dissimilarity between two sets as the maximum separation between members of Q and R (see panel b) of fig. 3.12):

$$d_{\text{CLINK}}(Q, R) = \max_{i \in Q, j \in R} d_{ij} \tag{3.10}$$

This approach does not suffer from chaining, but it has the opposite property – it identifies compact clusters that might not have well defined boundaries.

Group average method

An alternative definition of intercluster dissimilarity is offered by the unweighted pair-group average method (UPGMA, panel c) of fig. 3.12) which is the average of all dissimilarities d_{ij}:

$$d_{\text{UPGMA}}(Q, R) = \frac{1}{|Q||R|} \sum_{i \in Q} \sum_{\jmath \in R} d_{ij} \tag{3.11}$$

where i is any object in Q and j is any object in R (Sokal & Michener, 1958). This method is impervious to chaining effects and is particularly well suited for identifying ball shaped clusters (Kaufman & Rousseeuw, 1990).

3.3.2 Divisive methods

The DIvisive ANAlysis clustering method (`DIANA`) is a divisive hierarchical algorithm popularised by Kaufman & Rousseeuw (1990). We start with a single cluster and find an outlier with the highest average dissimilarity (in the Euclidean case it is the point with the highest average distance from the other points). This point is then removed from the cluster and becomes the first member of a 'splinter group'. Then

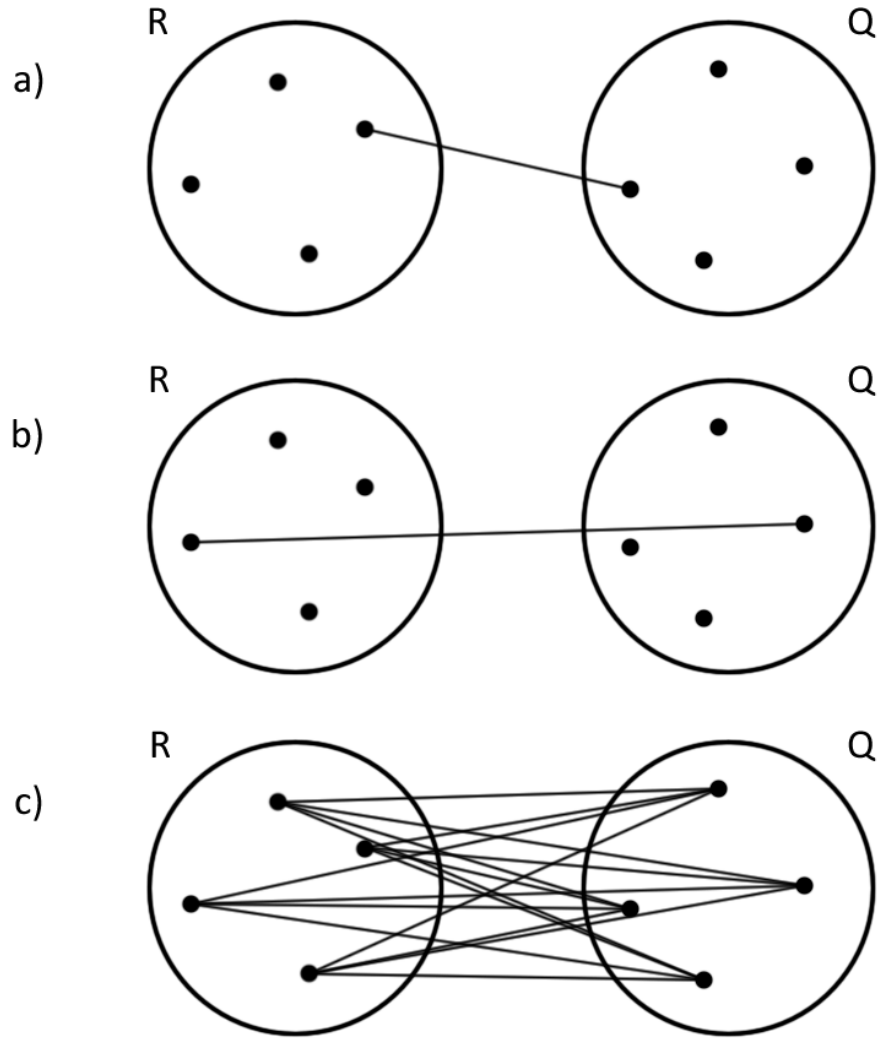


Figure 3.12: Measures of dissimilarity between sets R and Q for different linkage types. Panel a) shows single linkage clustering, panel b) shows complete linkage clustering and panel c) demonstrates the group average method.

we iterate over all the points in the cluster; each point whose dissimilarity with the splinter group is lower than its dissimilarity with the rest of the cluster is added to the splinter group. Once no new points can be added to the splinter group, it becomes a new cluster. This process is then repeated for the cluster with the highest dissimilarity radius, i.e. the largest dissimilarity between two members of the same cluster, until each point belongs to a separate cluster.

Divisive methods are more computationally expensive than the agglomerative; this particular algorithm has an exponential scaling with N. In our case, however, this is not a major problem since we are dealing with relatively small datasets. A bigger shortcoming is the fact that mutual nearest neighbours can be put in different clusters even at the top levels of this algorithm (see sec. 3.3.5).

3.3.3 Dendrograms

The results of a hierarchical clustering analysis can be conveniently presented on a dendrogram – a tree diagram. The bottom row of a dendrogram consists of individual data points, while the top row corresponds to a single cluster encompassing all the points in the distribution. The height at which two points are merged represents the intercluster dissimilarity – in this case the length of the link between the clusters.

In the agglomerative scenario we start at the bottom and join the two least dissimilar points together and join them with a horizontal line at the height corresponding to their dissimilarity. Conversely, in the divisive scenario we start at the top and divide the clusters in two at appropriate heights.

Fig. 3.13 shows the dendrograms generated for the the previously described test case consisting of three clusters. The top panel is an agglomerative dendrogram constructed using the UPGMA method; the bottom panel was created using the `DIANA` algorithm.

In the next section I shall outline a method of quantifying the information contained in a dendrogram, but even by visual inspection of the top plot we can identify three well defined areas that one could intuitively ascribe to the three clusters in the distribution. The divisive case, on the other hand, seems to contain four major components and the cuts divide the clusters into subclusters containing similar number of points.

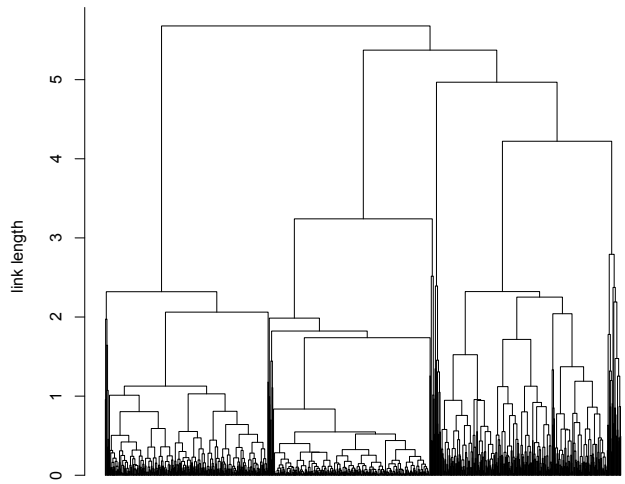
3.3.4 Number of clusters

In order to learn more about the clusters in a distribution we need to cut the dendrogram first to identify the clusters, which requires us to make a priori assumptions about their number, or what dissimilarity is important. Since the goal of our analysis is to learn about the clustering we need a way of deciding on the number of clusters based on the dendrogram. Calinski & Harabasz (1974) proposed a CH index that acts as a sui generis goodness-of-fit parameter:

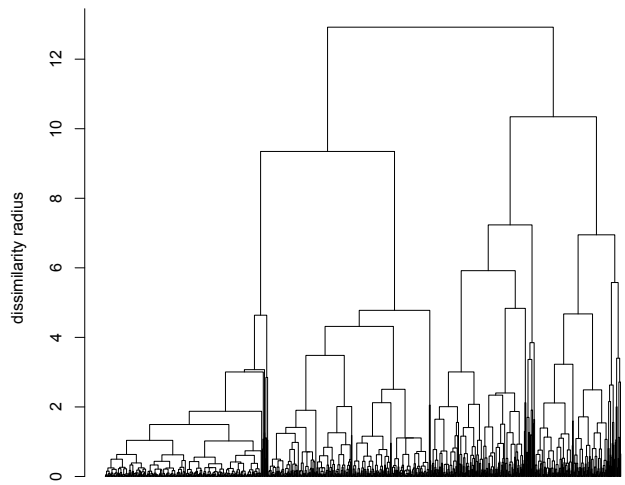
$$\mathrm{CH}(k) = \frac{\mathrm{B}(k)}{\mathrm{W}(k)} \frac{n-k}{k-1} \tag{3.12}$$

where $\mathrm{B}(k)$ is a measure of the scatter of the clusters, $\mathrm{W}(k)$ is a measure of the scatter of points within the clusters (defined by Calinski & Harabasz 1974), n is the number of points and k is the number of clusters. A high value of $\mathrm{B}(k)$ and a low value of $\mathrm{W}(k)$ mean that the clusters are well defined, therefore a high value of $\mathrm{CH}(k)$ is desirable.

Fig. 3.14 shows how CH varies with the number of clusters for the UPGMA (black symbols) and `DIANA` analysis (purple symbols) of the three cluster test case. The range of k on the plot is [2,25] since CH is not defined for a single cluster ($k = 1$). For the group average method the diagram has a well pronounced maximum for $k = 3$. The divisive case has highest value at $k = 4$ and $k = 7$. Another local maximum can



(a) UPGMA



(b) `DIANA`

Figure 3.13: Plots of dendrograms of the test case containing three clusters of different densities. The dendrogram in the top panel is the result of UPGMA analysis, while the bottom panel is the result of the `DIANA` divisive method.

be found at k=13 and it corresponds to a similar feature found in the UPGMA case.

3.3.5 Results

Fig. 3.15 shows the dendrograms with the clusters identified by the CH method imposed on them. The top panel shows the UPGMA dendrogram with $k = 3$ clusters, two bottom panels show the `DIANA` dendrogram with $k = 4$ and $k = 7$ clusters respectively. Fig. 3.16 depicts what these clusters look like in x-y plane; the colours of the points (black, red, green, blue, cyan, magenta, and yellow) represent membership to the groups identified in fig. 3.15.

Visual inspection reveals that in the UPGMA case the members of the individual sub-clusters are identified correctly, however, the background points are also classified as members of the clusters. The results of the `DIANA` algorithm are much worse – in both $k = 4$ and $k = 7$ cases the most diffuse subcluster is split into two. The division splits the cluster into two sub-clusters of similar size and there is a clear boundary between the regions. This is caused by the symmetry of the region. Since the densities on both sides of the boundary are the same, the cut occurs exactly in the middle in order to minimise the dissimilarity. The divisive methods are particularly biased by the background points that increase the dissimilarity radius. In this case, the lower density region on the right hand side of the plot is split into two lower density clusters. As a result, points that are very close to each other end up being assigned to different groups. This is further exacerbated by the fact that once two points have been assigned to two different clusters they can never be joined further down the dendrogram.

Increasing the value of k leads to some parts of the background being identified as individual 'clusters'. We could then discard them based on the low N; this approach is based on visual inspection and is therefore biased by human perception. Furthermore,

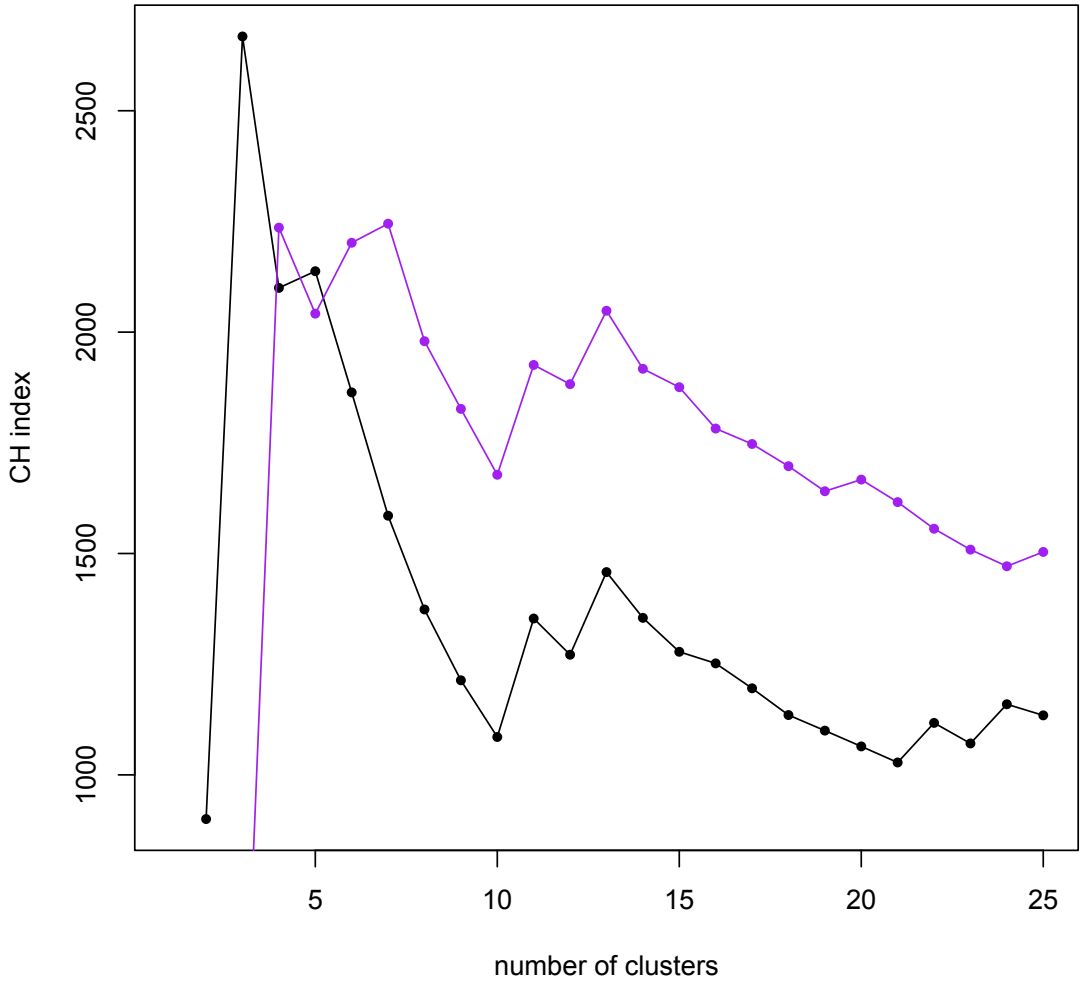


Figure 3.14: Caliński-Harabasz index (CH) as a function of the number of clusters (k). The index for the UPGMA method is shown by the black symbols. The purple symbols represent the result for the `DIANA` algorithm.

in addition to splitting the background into small groups, increasing k also results in the actual clusters being further fragmented (as demonstrated in the lower right panel of fig. 3.16), which renders this approach unacceptable, especially given the tenuous distinction between a small fragment of the cluster and a portion of the background.

3.4 Density based methods

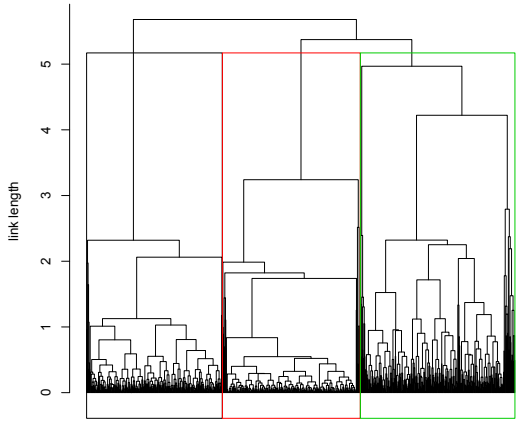
An alternative approach to identifying clusters in data is offered by the family of density-based methods. They operate on the intuitive principle that the dense regions in the distribution are clusters while the more diffuse ones are the background. In this section I outline two methods that go beyond the naïve approach and enable us to find clusters of different shapes and densities.

3.4.1 DBSCAN

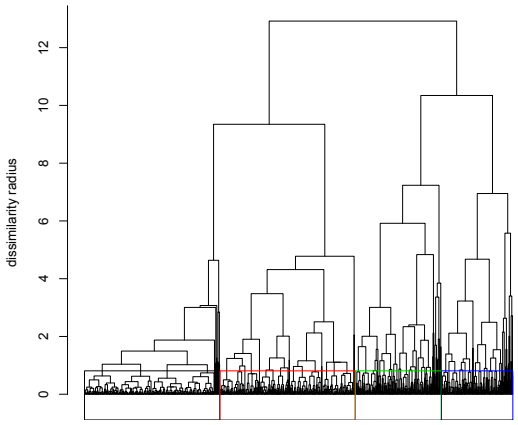
Ester et al. (1996) proposed a method called 'density-based spatial clustering of applications with noise' (DBSCAN). This algorithm uses a core radius ϵ to identify three classes of points in the data: core points, edge points and background points (outliers) by comparing the distance to the k-th nearest neighbour of each point with the radius ϵ.

A point p_1 is classified as a core point when a circle of radius ϵ drawn around it contains at least k other points. All the points within radius ϵ around a core point are referred to as *directly reachable*. Non-core points may be reachable from the core points (in this case they are known as edge points), but no points are reachable from them. The non-core points that are not reachable from any other points are classified as outliers.

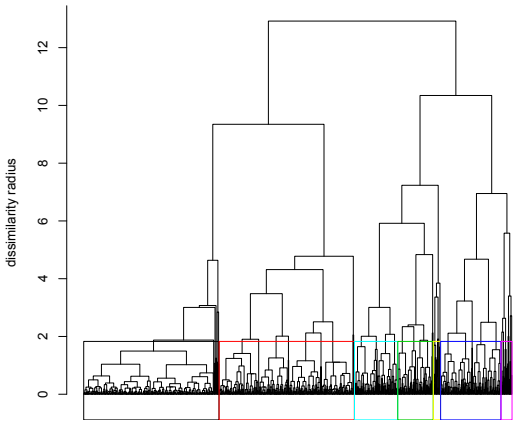
In order to identify the members of a cluster we need to start with a single



(a) UPGMA; k=3

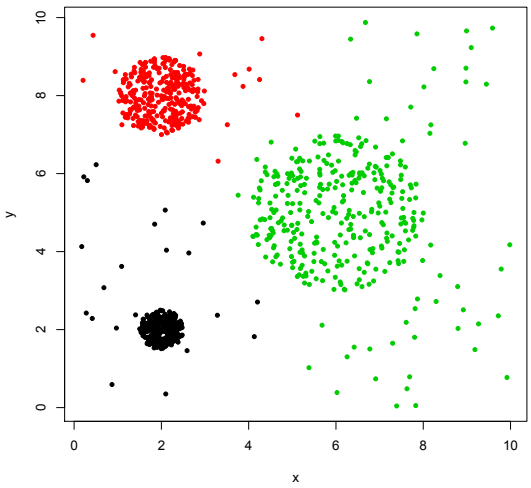


(b) `DIANA`; k=4

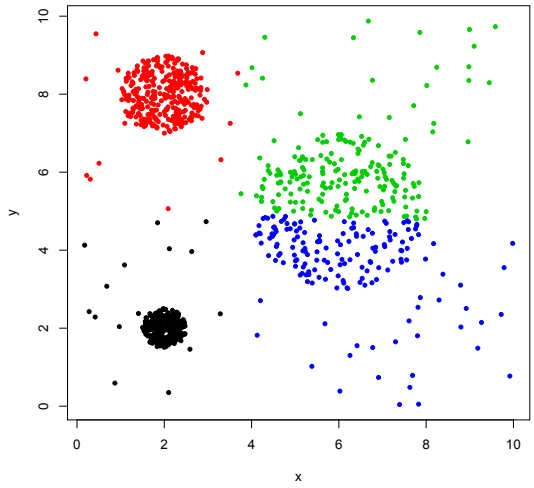


(c) `DIANA`; k=7

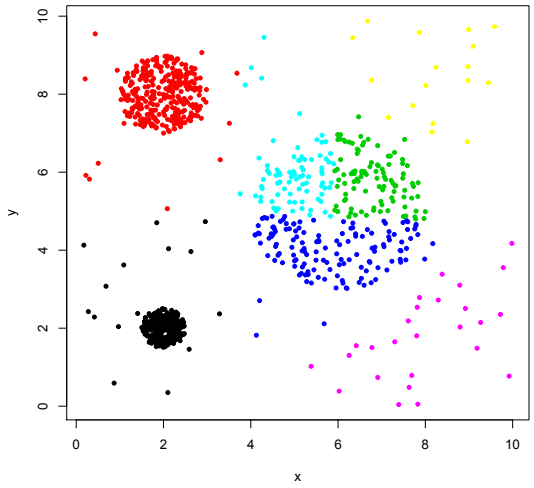
Figure 3.15: Dendrograms of the test case with three clusters. The dendrogram in the top panel was constructed using the UPGMA method, while the bottom two were generated using the `DIANA` algorithm. The coloured boxes (black, red, green, blue, cyan, magenta, and yellow) highlight the clusters identified in the data – three in panel (a), four in panel (b) and seven in panel (c).



(a) UPGMA, k=3



(b) `DIANA`, k=4



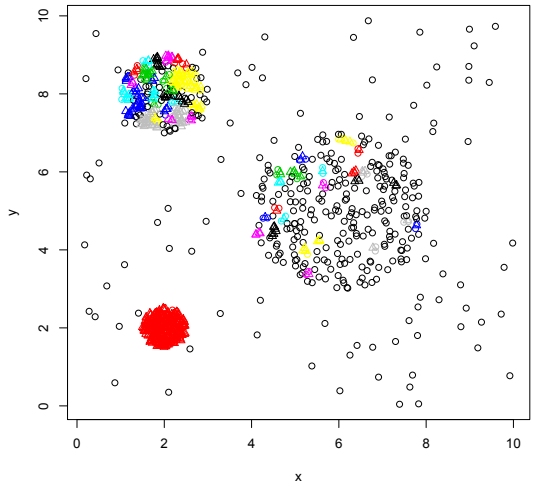
(c) `DIANA`, k=4

Figure 3.16: Plots of the test case with three clusters after applying the cuts to the dendrograms. The colours (black, red, green, cyan, magenta, and yellow) correspond to the clusters identified by the hierarchical algorithms. Panel (a) shows the result of UPGMA analysis for three clusters, panels (b) and (c) show the results of the `DIANA` algorithm for four and seven clusters respectively.

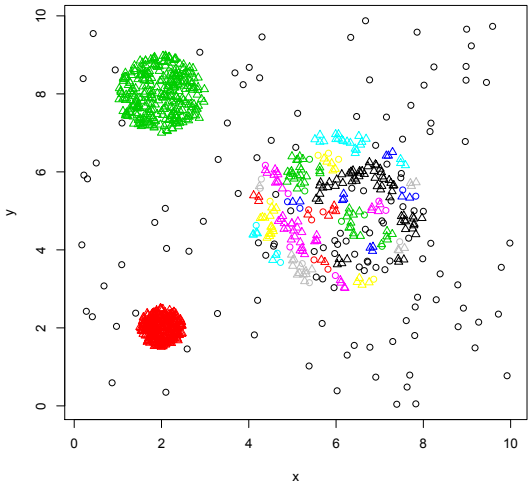
core point. All the points directly reachable from this point are added to the same cluster; we can then expand the cluster further by looking for points that are directly reachable from the newly added cluster members. This process is repeated until no more points can be added to the cluster, at which point we move to another core point that does not belong to the cluster and repeat the process. This procedure is reminiscent of the nearest neighbour clustering, but it is based on reachablility to k nearest neighbours rather than distance alone.

Fig. 3.17 shows the results of the DBSCAN analysis performed on the test case containing three clusters. The calculations were performed for a fixed value of $k = 3$ and three values of ϵ were used (0.1, 0.2, 0.3, 0.4). The triangles represent the core points, while the circles are the outliers. In all the cases the densest cluster is identified correctly (red symbols). For the lowest value of ϵ the detection of the two more diffuse clusters fails – they are split in numerous small groups and many of the cluster member are classified as background (black circles). Increasing ϵ to 0.2 results in the intermediate density cluster being identified correctly (green symbols), however, the most diffuse one is still not treated in a satisfactory manner. A further increase of ϵ to 0.3 mostly detects the members of the low density cluster correctly (blue), albeit there is still a small group of points classified as a separate cluster (cyan symbols). This problem is finally solved for $\epsilon = 0.4$; at this point, however, the value is so large that some parts of the background are mistakenly classified as clusters.

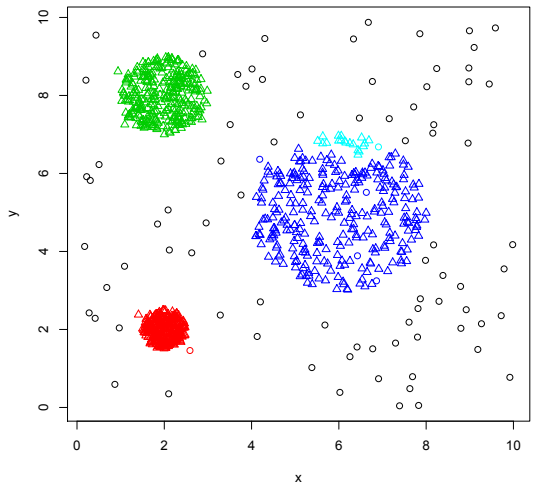
The above example is intended to exemplify two major shortcomings of the DBSCAN method: the sensitivity to input parameters and the inability to handle clusters of different densities. HDBSCAN (hierarchical DBSCAN) is modified version of the algorithm designed to fix this flaw. Unfortunately, HDBSCAN fails and detects structure in random distributions. Similarly to the MST cutting approach and other previously mentioned methods this approach requires us to set some parameters before we start the analysis. It is possible to estimate appropriate values of k and ϵ



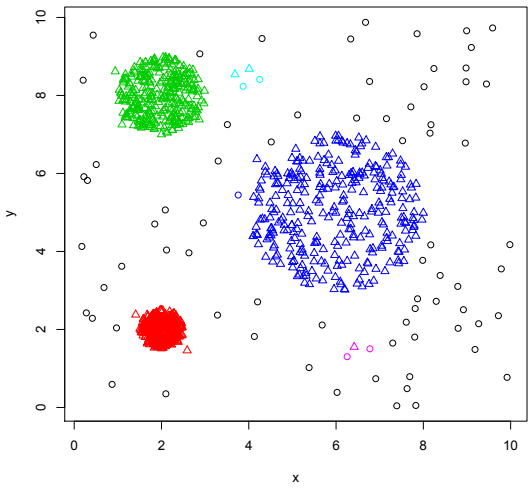
(a) $k = 3$, $\epsilon = 0.1$



(b) $k = 3$, $\epsilon = 0.2$



(c) $k = 3$, $\epsilon = 0.3$



(d) $k = 3$, $\epsilon = 0.4$

Figure 3.17: The results of the DBSCAN algorithm for different values of ϵ (0.1, 0.2, 0.3, and 0.4). The value of k is fixed at 3 for all the cases. The colours represent the membership to different clusters identified by the algorithm. Triangles are the points classified as core points, circles are the outliers.

based on the dimensions of the data and the distribution function of k-th nearest neighbour distances, however, the result is still quite sensitive to the values chosen which becomes even more apparent when the cluster has a fairly low density. If the value of ϵ is too high the algorithm will not be able to identify the cluster against the background, if it is too low the cluster will be fragmented into smaller pieces. The biggest flaw of DBSCAN that disqualifies it as a structure finding algorithm for star forming regions is the fact that it has trouble identifying clusters of different densities as demonstrated in fig. 3.17.

3.4.2 OPTICS algorithm

Ankerst et al. (1999) proposed a modification of the approach used in DBSCAN – 'ordering points to identify the clustering structure' (OPTICS) algorithm. This method still requires us to input values of k and ϵ, however, the latter is now treated as an upper limit (we can choose one as large as the region since we are dealing with relatively small datasets and the speed of the calculation is not paramount).

The implementation of the algorithm requires the introduction of the *core distance*: the distance to the k-th nearest neighbour. This value is defined only for points that have $> k$ neighbours within the radius ϵ (core points). The reachability distance of point p_2 with respect to point p_1 is defined as the distance from p_1 to p_2, if p_2 is within the radius ϵ. If p_2 is among k nearest neighbours of p_1, the reachability is instead set to the core distance. Points outside the radius of ϵ are not reachable and their reachability distance is undefined.

We start by picking an arbitrary point from a list of all points in the region. We calculate its core distance, set its reachability distance to `undefined` (since there is no point that we are reaching from) and write this information to a file. If the selected point is not a core point, we simply move to the next point on the list. If

it is a core point, we find its ϵ-neighbourhood ($N_\epsilon(p)$) i.e. the points that lie within a radius ϵ from p. We sort them in order of increasing reachability and add them to a queue ordered by reachability.

We then pull the top point of the queue (i.e. the point with the lowest reachability value) and consider its ϵ-neighbourhood. The operation of pulling (also referred to as dequeueing) removes the top entry from an ordered queue and returns its value. If the reachability from p_2 to any of the points in the queue is lower than the reachability value from any of the previously analysed points, we update the value. We then add to the queue all the points in p_2's neighbourhood that were not previously on the list and again sort the queue in order of increasing reachability. We iterate the neighbourhood queries until the queue has been emptied. If there are still unprocessed points in the region, we move to the next point on the list and repeat the process.

One of the shortcomings of DBSCAN was the fact that we had to choose a priori a single value of ϵ for the entire field. OPTICS, however, allows us to set ϵ to an arbitrarily large value, so that the radius encompasses all the points. This way they are all included in the queue at the beginning and later on we shift them appropriately, depending on their reachability distances.

Patwary et al. (2013) pointed out similarities between the OPTICS algorithm and Prim's MST algorithm. Their goal was to speed up OPTICS for large datasets by adapting it for parallel computing, however, I suggest the use of the Prim algorithm as an alternative formulation of OPTICS for arbitrarily large ϵ. When ϵ is larger than the spatial extent of the dataset, pulling the top entry from the queue and adding it to the set of the processed points is analogous to constructing a minimum spanning tree by nearest neighbour linkage (Prim's algorithm – see sec. 2.2.1 for a detailed explanation). We can simply run Prim's algorithm on the dataset – we start with a single point and at each iteration we add a point that is the closest to any point already in the tree. We compare the core distance distance of the newly added point

with the length of the edge leading to it – the larger of these values is this points reachability distance.

Once we have obtained reachability distances for all the points we can plot them against the order of the points. The resulting plot shown in fig. 3.18 (generated using ELKI data mining environment; Schubert et al. 2015), also known as *reachability plot*, represents the density based spatial clustering. The 'valleys' correspond to regions with low reachability (i.e. points that are easy to reach) – the points are close to each other so these are the cluster candidates. The points with higher values are harder to reach and represent the background.

A reachability plot contains all the information that could be obtained from a CDF of MST edge lengths, but it does not lose the information about the relative spatial distributions. The reachability plot is also similar to a dendrogram in that the points that are close on the plot are also close to each other in the x-y plane even if they are not directly connected.

The top panel represents the test case of three clusters against a random background. The depth of the valleys corresponds to the density of the clusters – the deepest valley (leftmost one) contains the points that belong to the densest clump, while the valley with the highest 'floor' contains the points from the clump with the lowest density.

The second panel shows the reachability plot of the single cluster against a random background. In this case the valley is less pronounced, especially the right hand slope is shallower. This is caused by the lower density contrast between the cluster and the background which makes it harder to distinguish the 'edge' of the cluster.

The reachability plot of the random distribution is shown in the bottom case. It contains several small dips that correspond to local overdensities in the distribution, but there is no well pronounced valley that could be interpreted as the signature of a cluster.

(a) Three clusters against a random background

(b) Single cluster against a random background

(c) Random distribution with binaries

Figure 3.18: Reachability plots of the triple cluster case (panel a), single cluster case (panel b) and a random distribution (panel c) for ϵ=10 and k=10. The spatial ordering is shown on the horizontal axis, while the vertical axis represents the reachability – the lower the value, the more reachable the point is.

This way of presenting the data retains information about the hierarchical structure of the regions and can be used to find subclusters within a cluster with deeper dips within the valleys representing regions of even higher density.

Setting ϵ to an arbitrarily large number is akin to scanning over all the values of ϵ. If we set a lower value of ϵ all the points with a reachability above ϵ would be undefined and considered to be background noise, while the points with reachability values below ϵ would be classified as cluster members. By setting ϵ to a lower value we recreate the results of DBSCAN for parameters k and ϵ'. Due to the distinction between the core distance and reachability, the bottom of the valley will have a slightly different shape than it would have in DBSCAN, but the cut would normally happen high enough for this not to change the outcome. Furthermore, the MST formulation of the OPTICS algorithm also shows that DBSCAN is analogous to MST cutting which explains why it suffers from the same problems.

OPTICS algorithm on its own gives a way of constructing the reachability plot, but it does not offer a way of identifying the clusters in the structure. Ankerst et al. (1999) proposed a ξ parameter (this approach is occasionally referred to as OPTICSxi) that gives a cutoff contrast between the valleys and the peaks. However, it is not appropriate for use in the SFR context because the clusters might have different densities and density profiles which will result in different contrasts. Sander et al. (2003) introduced a method that identifies *significant local maxima* in the reachability plot and constructs a single-linkage dendrogram based on the reachabilities, however, the applicability of this method in astronomical context remains to be tested.

3.5 Problems

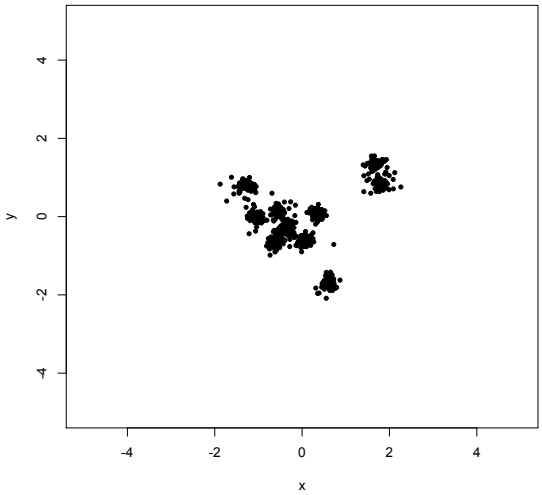
When using a structure finding algorithm we need to keep in mind the issues associated with the field of view. Observations tend to be centred around the brightest and densest parts of the star forming regions and the coverage of the outer reaches of the region might be incomplete. Conversely, some of the algorithms might not work for very large fields that cover vast amounts of background in addition to the structure.

Furthermore, if we zoomed in far enough into the clusters they would look the same as the field looked on the larger scale (save perhaps for any density profile that the cluster might have). This becomes apparent for underlying graphs that do not have a pre-determined density scale such as the $N_{\rm n}$-th nearest neighbour graph. When the background is much larger than the cluster the distant points have similar connectivity to the ones in the cluster. In the case of clique extraction method performed on such a field, the background points that are far away from the clusters will start forming their own cliques that are indistinguishable from the ones found in the cluster.

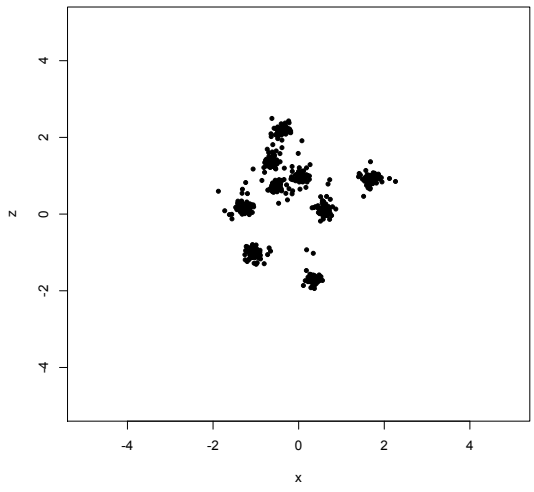
These problems are further exacerbated by the fact that we lack robust data about the third dimension. While most of the methods outlined in this chapter are compatible with multidimensional data we only have a 2D projection available. As a result sub-clusters might be superimposed and it is not possible to disentangle them without making assumptions about their density profiles. Such assumptions, however, are unlikely to hold in real life applications and could result in false detections and artefacts. What is more, projecting the same region along a different line of sight might yield different results.

I generated a three dimensional distribution consisting of 10 Plummer spheres containing 100 stars each distributed inside a larger Plummer sphere (see sec. 5.1 for details) in order to demonstrate the problems caused by projection effects. Fig. 3.19

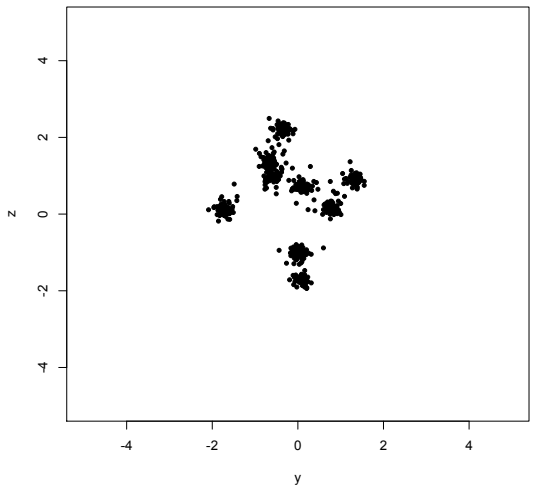
shows projections of this dataset along three orthogonal axes (z on panel (a), y on panel (b) and z on panel (c)). The field of view in all three plots is 10 pc$\times$10 pc and the barycentre can be found at (0,0). Panel (a) contains all the points in this dataset but two of the spheres are blended into one elongated object. The other two panels only contain 9 of the Plummer spheres as one of them now falls outside the field of view. Additionally, in panel (c) two of the spheres are blended along the line of sight thus making it appear as if there were only 8 subclusters in the distribution.



(a) Projection along the z-axis



(b) Projection along the y-axis



(c) Projection along the x-axis

Figure 3.19: An example region consisting of 10 Plummer spheres containing 100 stars each. The 10 pc×10 pc fields are projections of the same region along the orthogonal axes onto the planes x-y (panel a), z-x (panel b), and z-y (panel c).

Chapter 4

N-body methods

In this chapter I outline the fundamentals of the N-body simulations. I describe the basic principle of N-body integration and discuss some algorithms that can be implemented in order to improve the accuracy of the simulation and reduce the computational expense. The work presented in this thesis used the NBODY6 integrator (Aarseth, 1999) whose accuracy is comparable to that of STARLAB/KIRA (Anders et al., 2012).

4.1 Method basics

The N-body method is a way of simulating the dynamical evolution of a system of particles under the influence of forces. In this thesis it is used to analyse the dynamics of stellar clusters in which the stars interact with each other via Newtonian gravity.

In an N-body simulation each star is treated as a point mass that experiences the combined gravitational influence of all the other particles in the system. Knowing the positions of all the particles at a time t makes it possible to calculate the net force acting on every star, which in turn allows one to predict the positions and velocities of all the stars after a time increment dt. This process can then be repeated until

$t = t_{\text{final}}$, where t_{final} is the desired length of the simulation.

The net force acting on the i-th particle is equal to the sum of all the forces acting on it, and the acceleration is obtained using Newton's second law of motion:

$$\mathbf{a}_i = -\sum_{j=1, j\neq i}^{N} \frac{G m_j (\mathbf{r}_i - \mathbf{r}_j)}{|\mathbf{r}_i - \mathbf{r}_j|^3} \tag{4.1}$$

In Eqn. 4.1 $\mathbf{a}_i$ is the total acceleration experienced by the i-th particle, obtained by summing the contributions from all the other particles; $\mathbf{r}_i$ and $\mathbf{r}_j$ are the position vectors of the i-th and j-th particle respectively, while m_i and m_j are their masses; G is the gravitational constant.

4.2 Improvements

The concept behind N-body simulations is very simple indeed, however, not unlike other numerical integration problems, it suffers from inaccuracies caused by the finite length of the timestep $\mathrm{d}t$. The simple N-body method operates under the assumption that between time t_0 and $t_0 + \mathrm{d}t$ the motion of the particles is linear, which is, obviously, not the case.

Due to the sheer scale of the N-body simulations in this thesis, a brute force approach of simply choosing a very short timestep is not feasible. Simple improvements, such as second order Runge-Kutta integrator, also prove inadequate, hence the need for more sophisticated algorithms.

4.2.1 Hermite interpolation

The integration method adopted by the NBODY6 code is the fourth order Hermite interpolation (Makino & Aarseth, 1992). It uses the higher derivatives of acceleration to apply a correction to the position and velocity at the end of every timestep.

Hermite integration is a self-starting scheme i.e. it does not require any memory of the conditions at previous timesteps. We start by calculating the acceleration ($\mathbf{a}$) and *jerk* (first time derivative of acceleration; $\mathbf{a}^{(1)}$) at time t_0. We can obtain the value of the jerk if the i-th particle($\mathbf{a}_i^{(1)}$) analytically by differentiating Eqn. 4.1,

$$\mathbf{a}_i^{(1)} = -\sum_{j=1,j\neq i}^{N} \frac{Gm_j(\mathbf{v}_i - \mathbf{v}_j)}{|\mathbf{r}_i - \mathbf{r}_j|^3} + 3\sum_{j=1,j\neq i}^{N} \frac{(\mathbf{r}_i - \mathbf{r}_j)(\mathbf{v}_i - \mathbf{v}_j)\mathbf{a}_{ij}}{|\mathbf{r}_i - \mathbf{r}_j|^2} \tag{4.2}$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ are the velocities of particles i and j respectively, and $\mathbf{a}_{ij}$ is the j-th addend of $\mathbf{a}_i$. We can use the calculated values of $\mathbf{a}_i$ and $\mathbf{a}_i^{(1)}$ to express the position and velocity vectors at the end of the step (at time $t + \Delta t$) in the form of a Taylor series:

$$\begin{aligned} \mathbf{r}_i &= \mathbf{r}_0 + \mathbf{v}_0\Delta t + \frac{1}{2}\mathbf{a}_0\Delta t^2 + \frac{1}{6}\mathbf{a}_0^{(1)}\Delta t^3, \\ \mathbf{v}_i &= \mathbf{v}_0 + \mathbf{a}_0\Delta t + \mathbf{a}_0^{(1)}\Delta t^2, \end{aligned} \tag{4.3}$$

where the zero in the subscript denotes value at time t_0.

We can now calculate the values of $\mathbf{a}_1$ and $\mathbf{a}_1^{(1)}$ at the end of the timestep and also express them as a Taylor series truncated after $\mathbf{a}^{(3)}$:

$$\begin{aligned} \mathbf{a}_1 &= \mathbf{a}_0 + \mathbf{a}_0^{(1)}\Delta t + \frac{1}{2}\mathbf{a}_0^{(2)}\Delta t^2 + \frac{1}{6}\mathbf{a}_0^{(3)}\Delta t^3, \\ \mathbf{a}_1^{(1)} &= \mathbf{a}_0^{(1)} + \mathbf{a}_0^{(2)}\Delta t + \mathbf{a}_0^{(3)}\Delta t^2. \end{aligned} \tag{4.4}$$

We can now solve the above system of equations to obtain the values of $\mathbf{a}^{(2)}$ and $\mathbf{a}^{(3)}$:

$$\begin{aligned} \mathbf{a}_0^{(2)} &= 2(-3(\mathbf{a}_0 - \mathbf{a}) - 2(\mathbf{a}_0^{(1)} + \mathbf{a}^{(1)})\Delta t)\Delta t^2, \\ \mathbf{a}_0^{(3)} &= 6(2(\mathbf{a}_0 - \mathbf{a}) + (\mathbf{a}_0^{(1)} + \mathbf{a}^{(1)})\Delta t)\Delta t^3 \end{aligned} \tag{4.5}$$

These higher order derivatives can now be used to further extend the series in Eqn. 4.3 and apply a correction to the final position and velocity:

$$\begin{aligned}\mathbf{r}_i &= \mathbf{r}_0 + \mathbf{v}_0\Delta t + \frac{1}{2}\mathbf{a}_0\Delta t^2 + \frac{1}{6}\mathbf{a}_0^{(1)}\Delta t^3 + \frac{1}{24}\mathbf{a}_0^{(2)}\Delta t^4 + \frac{1}{120}\mathbf{a}_0^{(3)}\Delta t^5, \\ \mathbf{v}_i &= \mathbf{v}_0 + \mathbf{a}_0\Delta t + \mathbf{a}_0^{(1)}\Delta t^2 + \frac{1}{6}\mathbf{a}_0^{(2)}\Delta t^3 + \frac{1}{24}\mathbf{a}_0^{(3)}\Delta t^4.\end{aligned} \tag{4.6}$$

The Hermite scheme offers a fourth order polynomial fit for velocity and fifth order fit for position, hence the truncation error (τ_n) can be expressed as $\tau_n = \mathcal{O}(\Delta t^5)$. This is consistent with this integrator being a fourth order method (a method is of order p if $\tau_n = \mathcal{O}(h^{p+1})$, where h is the size of the step). Higher order Hermite methods exist but do not necessarily give a substantial advantage in terms of striking a balance between accuracy and computational cost and have not been implemented in NBODY6 (Aarseth, 2003).

Compared with the previously popular force integral method, Hermite interpolation allows for a longer timestep without compromising the accuracy (Makino, 1991). It has also the advantage of being a self starting method which means that the only initial conditions required are the positions and velocities of the particles.

4.2.2 Timestep

The length of the timestep is one of the crucial parameters that determine the accuracy of an N-body simulation. While sophisticated integration schemes such as the Hermite interpolation can improve the accuracy for a given timestep, ultimately, shortening the timestep will yield a more accurate result. The tradeoff, however, is the greater number of steps that have to be performed over the duration of the simulation and as a result an increase in the runtime. Furthermore, if the timestep is too short, the rounding errors become exacerbated.

Stellar systems exhibit a wide range of densities, hence the forces experienced by individual stars can also vary substantially (consider the differences between a tight binary near the centre of a cluster and a star travelling through the halo). This leads to a range of timescales at which the orbital parameters of stars change. In order to minimise the computational effort while preserving accuracy it makes sense to allow individual stars to have different timesteps. That way the stars on the outskirts of the system, which interact with a lower force, can have their accelerations recalculated less frequently.

The work included in this thesis used the NBODY6 code, which adopts the following timestep:

$$\Delta t_i = \left(\eta \frac{|\mathbf{a}_i||\mathbf{a}_i^{(2)}| + |\mathbf{a}_i^{(1)}|^2}{|\mathbf{a}_i^{(1)}||\mathbf{a}_i^{(3)}| + |\mathbf{a}_i^{(2)}|^2} \right)^{1/2}, \tag{4.7}$$

where η is a dimensionless factor that controls the accuracy. I used the standard value of $\eta \sim 0.2$ given by Aarseth (2003). In order to implement the individual timesteps each particle needs to have the following parameters: its individual time(t_i; initially set to zero) and timestep (Δt_i), position ($\mathbf{r}_i$) and velocity ($\mathbf{v}_i$) at time t_i, as well as $\mathbf{a}_i$ and $\mathbf{a}_i^{(1)}$ calculated at time t_i.

To perform the integration with individual timesteps we need to apply the following steps:

(i) Find the particle i for which $t_i + \Delta t_i$ is the smallest and advance the global time (t) to that value.

(ii) Predict the positions and velocities of all the other particles at that time.

(iii) Perform the Hermite integration on the particle i to calculate its position and velocity using the predicted information. Recalculate the timestep and set the value of t_i to the global time.

(iv) go back to step 1.

In a simpler integration scheme step (ii) would be equivalent to advancing the entire system by a timestep Δt_i, however, the Hermite method is an interpolation i.e. it allows us to predict the positions of the stars at any point between t_i and $t_i + \Delta t_i$ without the need to recalculate the forces.

Nonetheless, the overheads associated with the calculations are substantial, hence NBODY6 uses the block step method in order to streamline the integration. The method uses a hierarchically quantised set of timesteps, which allows groups of particles with similar Δt_i to be advanced at as a block at the same time, thus saving computational expense. Timestep of level n is given by the formula:

$$\Delta t_n = \frac{\Delta t_{\max}}{2^{n-1}}, \tag{4.8}$$

where $\Delta t_{\max}$ is the maximum timestep. The number of levels varies between systems but rarely exceeds 12 (Aarseth, 2003).

At the start of the simulation the timestep is calculated using formula given in Eqn. 4.7 and then rounded down to the nearest value given by Eqn. 4.8. Thoughout the simulation the timestep is determined by comparing the previously used timestep (Δt_p) with Δt_i from Eqn. 4.7, which results in one of three possible outcomes:

- set new timestep to $\Delta t_p/2$, if $\Delta t_i < \Delta t_p$
- set new timestep to $2\Delta t_p$, if $\Delta t_i > \Delta t_p$ and t is commensurable with $2\Delta t_p$ i.e. $t = x \cdot 2\Delta t_p$, where x is an integer
- no change, if $\Delta t_p < \Delta t_i < 2\Delta t_p$

The timestep can be halved at any point point during the integration, however, doubling is only permitted at every other step. This rule is imposed in order to keep

the blocks synchronised – all the blocks with a timestep Δt_n should be advanced at the same time in order to reap the benefits of this method. Furthermore, this ensures that all the blocks are in synchronisation at the end of the longest timestep.

The hierarchy of timesteps is represented schematically in Fig. 4.1, where n is an arbitrary level. The vertical scale represents timestep length decreasing from top to bottom while the global time increases on the horizontally from left to right. Let us consider two particles: Particle 1 starts at level $n+1$ which corresponds to a timestep $\Delta t_n/2$, Particle 2 is a level higher $n+2$ and has half the timestep. Initially, at time t_0, both particles are synchronised. As the time progresses, Particle 1 at any of its steps will always be in synchronisation with Particle 2 and all the particles occupying the levels $\geq n+1$, but Particle 2 will only be synchronised with the level $n+1$ at every other step.

At time t_b, if the above-mentioned conditions are satisfied, both particles could have their timesteps halved. Depending on the value returned by Eqn. 4.7 Particle 2 could also have its timestep doubled since $t_b + \Delta t_n/2$ is commensurable with t. Should that happen both particles would be henceforth synchronised and occupy the same level in the hierarchy. At time t_a, however, the level of Particle 2 cannot be decreased, since $t_a + \Delta t_n/2$ would not be commensurable with the global time and it could not be moved in a block with Particle 1 despite having the same timestep.

4.2.3 N-body units

In order to avoid rounding errors due to repeated multiplication by a constant and storing the exponents in memory, N-body integrators often adopt their internal unit systems in which the gravitational constant, G, is usually set to unity. In order to avoid the frequent use of large exponents associated with stellar masses, the total mass of the system is also set to 1. Furthermore, the total energy of the system is

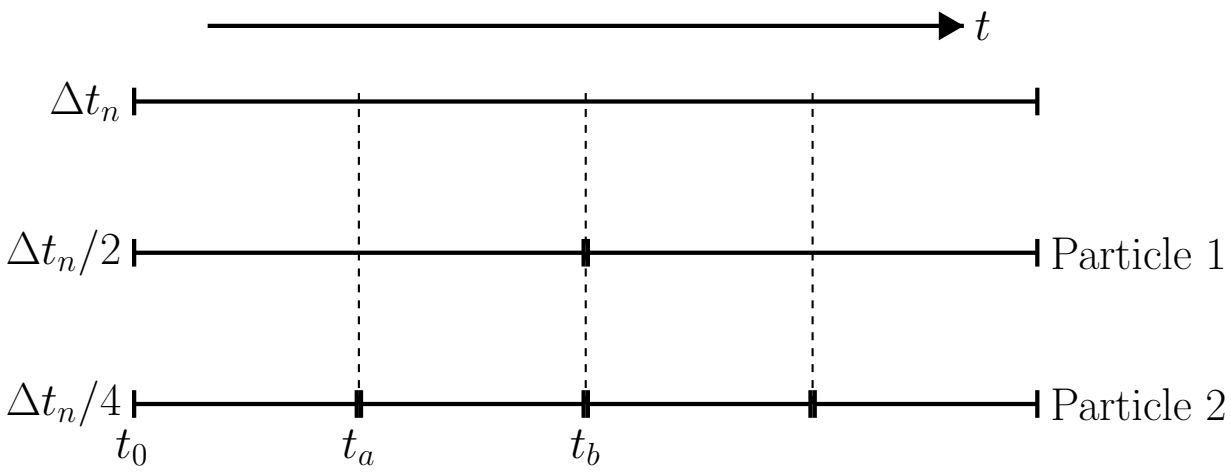


Figure 4.1: A schematic that demonstrates the hierarchical distribution of timesteps in the block step method. Horizontal scale represents the advancing time of the simulation, vertical scale corresponds to timestep length decreasing from top to bottom.

set to $-1/4$ (Heggie & Mathieu, 1986). More intuitive physical units are restored in output files which are written in the common units of solar masses, parsecs and km/s.

Chapter 5

Surviving Gas Expulsion With Substructure

In this chapter I discuss the setup of the initial conditions of the simulations whose results are included in this thesis. I explain how the initial positions and velocities are generated and how the substructure found in star forming regions is reflected in the initial setup of the simulations. I also outline how the stellar masses are sampled from the initial mass function.

It is important to keep in mind that the simulations should be treated as a numerical experiment conducted on a simplified model. The goal of a simulation is not to reproduce 'real life' with the myriad of parameters, but rather to observe some patterns of behaviour in the dynamics of the system. By beginning with a basic model and gradually increasing the complexity we can explore the parameter space and determine which parameters and physical phenomena have the most impact on the final outcome.

5.1 Plummer sphere

Stellar systems in spherical potentials are often represented using a simple density law proposed by Plummer (1911, 1915) which is usually referred to as *Plummer sphere*:

$$\rho(r) = \left(\frac{3M_{\rm tot}}{4\pi a^3}\right)\left(1+\frac{r^2}{a^2}\right)^{-5/2}, \tag{5.1}$$

where $M_{\rm tot}$ is the total mass of the system and a is the *Plummer radius* – a scale parameter that defines the spatial extent of the Plummer sphere, Spitzer (1987) gives the relationship between the half-mass radius ($r_{\rm h}$) of the cluster and the Plummer radius as $1.3a = r_{\rm h}$. This distribution is a polytrope of index n=5 and is formally infinite.

It is also useful to know the gravitational potential at radius r. The relationship between density and gravitational potential is given by the Poisson equation:

$$\nabla^2\Phi = 4\pi G\rho \tag{5.2}$$

For a spherically symmetric system (no dependency on the azimuthal angle), such as a Plummer sphere, the Poisson equation can be expressed in spherical coordinates as:

$$\frac{1}{r^2}\frac{\partial}{\partial r}(r^2\frac{\partial\Phi}{\partial r}) = 4\pi G\rho \tag{5.3}$$

Integrating eqn. 5.3 and applying the fundamental theorem of calculus results in:

$$\Phi = 4\pi G\int r^{-2}(\int \rho r^2 dr)dr \tag{5.4}$$

which after substituting eqn 5.1 for ρ reduces to:

$$\Phi = \frac{-GM}{(r^2 + a^2)^{1/2}} \tag{5.5}$$

5.1.1 Spatial distribution

Generating a set of positions and velocities based on the Plummer density profile is a non-trivial task, since stars are discrete objects while the potential is formulated for a continuous mass distribution. A prescription that allows us to randomly distribute a number of stars while keeping the potential unaltered in the limit of large N was given by Aarseth et al. (1974).

For the sake of simplicity, gravitational constant (G), Plummer radius (a), and the total system mass (M) are set to unity. Integrating eqn. 5.1 allows us to calculate the mass enclosed within the radius r:

$$M(r) = r^3(1 + r^2)^{-3/2} \tag{5.6}$$

We can now generate X_1 – a random number with a uniform distribution between 0 and 1 and equate it to $M(r)$ since $M(r) \leq 1$. Rearranging the equation allows us to obtain r corresponding to that mass:

$$r = (X_1^{-2/3} - 1)^{-1/2} \tag{5.7}$$

Once the radius is known the next step is to generate a set of suitable spatial coordinates (x, y, z). This is done by using the usual recipe for uniformly distributed points on the surface of a sphere. We generate two more random numbers: X_2 and X_3. The azimuthal coordinate, ϕ is picked from the the uniform probability distribution $\phi \in [0, 2\pi)$, this is done by simply multiplying one of the normalised random numbers by 2π:

$$\phi = 2\pi X_2 \tag{5.8}$$

The polar coordinate, θ, however cannot be generated in a similar fashion since it would result in an increased surface density close to the poles, while our goal is a uniform distribution. Instead, we can choose to pick $\cos\theta$ from the uniform distribution $\cos\theta \in [-1, 1]$:

$$\cos\theta = 1 - 2X_3 \tag{5.9}$$

Using elementary trigonometry we can now retrieve the cartesian coordinates (x, y, z):

$$\begin{aligned} z &= r(1 - 2X_3) \\ x &= r\sin\theta\cos(2\pi X_2) = (r^2 - z^2)\cos(2\pi X_2) \\ y &= r\sin\theta\sin(2\pi X_2) = (r^2 - z^2)\sin(2\pi X_2) \end{aligned} \tag{5.10}$$

5.1.2 Particle velocities

Since the Plummer sphere is initially virialised, no star can have a velocity exceeding the escape velocity at its location:

$$V_e = (-2\Phi(r))^{1/2} = \sqrt{2}(1 + r^2)^{-1/4} \tag{5.11}$$

where $\Phi(r)$ is the potential at radius r as calculated in eqn. 5.5. Let us define $q = \frac{V}{V_e}$ where V is the velocity of the star we wish to sample. The probability distribution of q is given by: $g(q) = q^2(1 - q^2)^{7/2}$ (Aarseth et al., 1974). We then generate two random numbers X_4 and X_5. Since q can take values in the range [0,1] and $q < 0.1$, if $0.1X_5 < X_4$ we set $q = X_4$. Otherwise, we draw another pair of random numbers

until these conditions are satisfied.

We then need to generate another pair of random numbers (X_6 and X_7) in order to obtain the cartesian components of velocity (v_x, v_y and v_z) in a way analogous to eqn. 5.10:

$$\begin{aligned} v_z &= V(1 - 2X_6) \\ v_x &= (V^2 - v_z^2)\cos(2\pi X_7) \\ v_y &= (V^2 - v_z^2)\sin(2\pi X_7) \end{aligned} \tag{5.12}$$

5.2 Stellar masses

When generating Plummer spheres it is essential to know the masses of the points in order to set the virial ratio of the entire system. The distribution of stellar masses in the population is known as the initial mass function (IMF). The work included in this thesis uses the L_3 parametrisation of the IMF (Maschberger, 2013). This formulation is remarkably easy to implement, in order to sample a mass from the IMF one only needs to generate a uniform random number u in the range between 0 and 1 and substitute back the equation:

$$m(u) = \mu \left([u(G(m_u) - G(m_l)) + G(m_l)]^{\frac{1}{1-\beta}} - 1 \right)^{\frac{1}{1-\alpha}}, \tag{5.13}$$

where $G(m)$ is the auxiliary function defined as $G(m) = \left(1 + \left(\frac{m}{\mu}\right)^{1-\alpha}\right)^{1-\beta}$. Parameters m_l and m_u set the lower and upper mass limits respectively, while α, β and μ determine the shape of the IMF. We used the following values recommended by Maschberger (2013): α=2.3, β=1.4, μ=0.2 $M_\odot$, m_l=0.01 $M_\odot$, m_u=150 $M_\odot$.

5.3 Physical scaling

A great advantage of the Plummer sphere lies in how easy it is to adjust its parameters. Should we choose to use a Plummer radius value other than unity we can now multiply the resulting coordinates by the chosen value of a. Now we can set a virial ratio defined as $Q = \frac{T}{|\Omega|}$, where T is kinetic energy and Ω is potential energy. This can be achieved by scaling the velocities appropriately. In order to change the virial ratio of the sphere to $Q_{\rm new}$, one needs to multiply the velocities by the factor of $(Q_{\rm new}\frac{|\Omega|}{T})^{1/2}$.

5.4 Clumpy substructure

While a Plummer sphere might be a reasonable approximation for a spherically symmetric cluster, in reality young star forming regions are not as uniform and exhibit some substructure – see Sec. 1.2 for details. In order to reflect this and introduce a degree of 'clumpiness' to our simulations we created a simple set of initial conditions consisting of Plummer spheres embedded within a larger Plummer sphere. We start by generating $N_{\rm sub}$ Plummer spheres, each of which could have a different virial ratio ($Q_{\rm sub}$), radius ($R_{\rm sub}$), and contain a different number of stars ($N_{\rm little}$). We then create a bigger Plummer sphere with radius $R_{\rm big}$, virial ratio $Q_{\rm big}$ containing $N_{\rm sub}$ points.

The next step involves replacing the points in the big sphere with the smaller Plummer spheres; we place them in such a way that the centre of mass of the small spheres coincides with the coordinates of the points in the big sphere. Their velocity vectors are simply added to the velocities of the stars in the small Plummer spheres. It is important to note that we generate the smaller spheres first in order to know the exact masses of the points in the bigger sphere since they are required to set its virial ratio.

Baumgardt & Kroupa (2007) investigated how the bound fraction depends on the initial virial ratio for a single Plummer sphere. Our setup, however, allows us to perform analogous analysis on systems with a varying degree of substructure and investigate how its presence affects the cluster survivability.

Furthermore, the virial ratio of the big Plummer sphere can be varied to produce a situation where the individual sub-clusters are supervirial (hence would be unbound if treated in isolation) yet they feel the presence of other sub-clusters moving at low relative velocities since the bigger sphere is initially in virial equilibrium

Fig. 5.1 shows an example of a clumpy initial conditions set produced with this method – a substructured region in which 10 sub-clusters form a bigger cluster. Simulating the evolution of such a system allows us to investigate the impact that the clumpiness might have on the outcome of the simulations.

5.5 Gas expulsion

Initially the natal gas from which stars form is ubiquitous in the SFR. As the star formation progresses, the feedback from supernova explosions, winds and radiation will result in gas being expelled from the region on the timescale of a few Myr. For a more detailed explanation see Sec. 1.2.2.

In our simulations we do not model the gas expulsion and assume it is instantaneous. Instead, we consider the dynamical state of the system immediately after gas expulsion. It can be described using the effective star formation efficiency (eSFE), which is defined as $\epsilon = \frac{1}{2Q}$, where Q is the virial ratio (Verschueren & David, 1989; Goodwin & Bastian, 2006; Goodwin, 2009). We define Q as the ratio of kinetic and potential energies of the system ($Q = \frac{T}{|\Omega|}$) which means that for a system in virial equilibrium Q=0.5. Systems with high virial ratios (low eSFE) have a large excess of kinetic energy and will evaporate quickly, while systems with high eSFE will contract

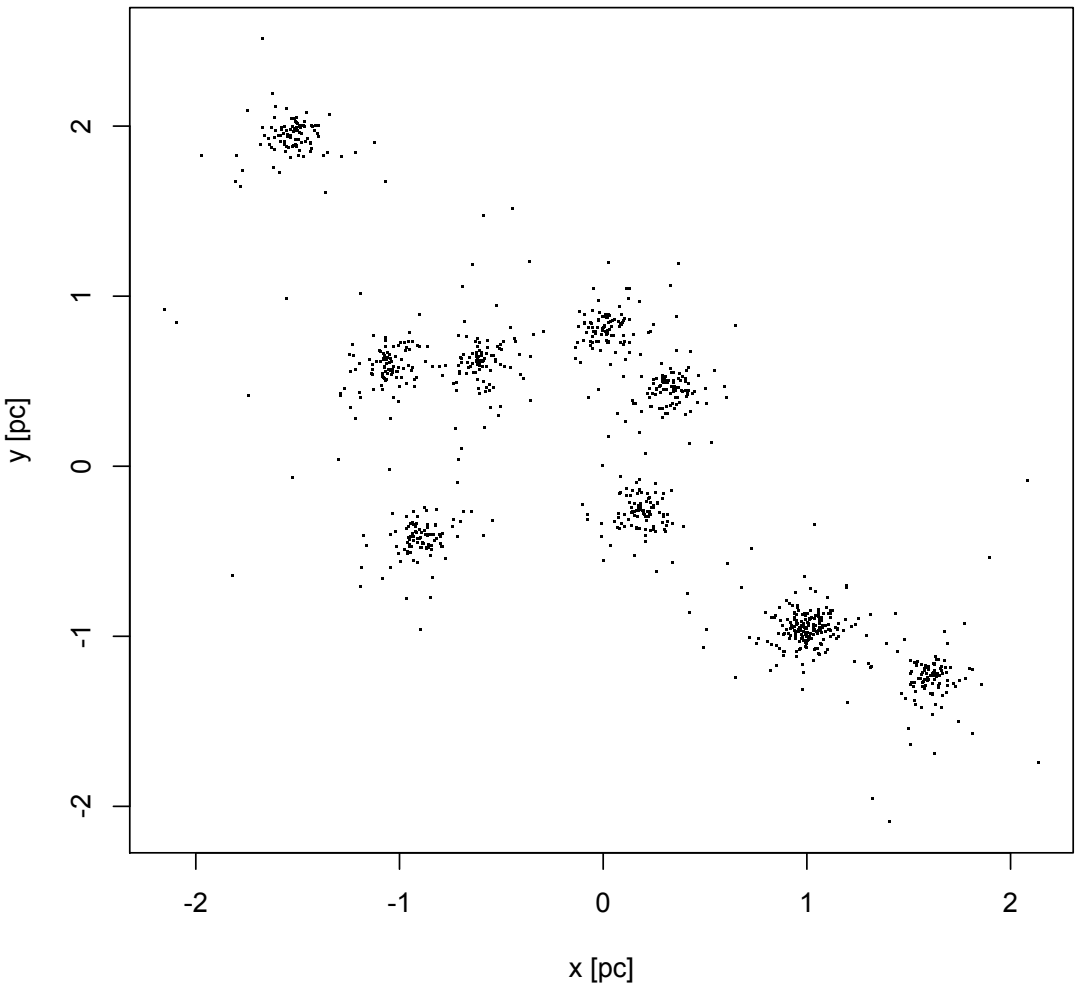


Figure 5.1: 2D projection onto the x-y plane of a set of initial conditions containing 10 sub-clusters. The big Plummer sphere has $R_{\rm big}$=1 pc and $N_{\rm sub}$=10, while each sub-cluster has a Plummer radius of 0.1 pc and contains 100 stars. The big Plummer sphere has a virial ratio $Q_{\rm big}$=0.5, while each of the sub-clusters has $Q_{\rm sub}$=4.0, however, the chosen value of virial ratio has no influence on the spatial distribution.

and remain bound. Baumgardt & Kroupa (2007) showed that for a single Plummer sphere the eSFE value of 33 per cent is critical and systems with lower eSFE cannot survive. As explained in Sec. 1.2.2, the simple relationship between eSFE and and survival holds only if prior to the gas expulsion the gas and stars are in virial equilibrium (Goodwin, 2009).

5.6 Initial conditions grid

Our systems are set up as Plummer spheres embedded in a larger Plummer sphere, as outlined in Sec. 5.4. We can vary the initial virial ratios of the sub-clusters and investigate how the bound fraction depends on the number of sub-clusters for any given virial ratio.

We set up systems containing $N_{\rm tot}$=1000 stars each. While it is possible to choose a different value of $N_{\rm little}$ for every sub-cluster, initially we chose to keep the $N_{\rm tot}$ constant for all the simulations. This results in a typical system mass of $\sim$ 300 $M_\odot$ for stellar masses sampled from the Maschberger formulation of the IMF. We divide the stars into $N_{\rm sub}$=2, 5, or 10 identical sub-clusters, each containing $N_{\rm little}$=500, 200, or 100 stars respectively. To illustrate the difference between a regular Plummer sphere and one with sub-clusters, for the virial ratios Q = 0.5, 1.0, 1.5, 2.0, and 4.0 we also set up a control run with just a single Plummer sphere which is analogous to the Baumgardt & Kroupa (2007) simulations.

Each of these sub-clusters is then put in a cluster represented by a larger Plummer sphere as prescribed in Sec. 5.4.The radius ($R_{\rm big}$) and virial ratio ($Q_{\rm big}$) of the larger sphere are kept constant throughout all the sets of input parameters and are set to: $R_{\rm big}$=1 pc and $Q_{\rm big}$=0.5.

For each $N_{\rm sub}$, however, we generate a set of initial conditions with sub-cluster *internal* virial ratio $Q_{\rm sub} = 1/2\epsilon$. We choose five different virial ratios: $Q_{\rm sub}$ = 0.5,

1.0, 1.5, 2.0, and 4.0. These correspond to effective star formation efficiencies of 100%, 50%, 33%, 25%, and 13%. According to Baumgardt & Kroupa (2007) the value of 33 per cent is critical for a single Plummer sphere and systems with lower eSFE do not survive.

In order to investigate the stochastic effects on each set of initial conditions we set up an ensemble of simulations with the same input parameters but different random number seeds.

5.6.1 Sub-cluster radii

We considered two cases – one assuming constant sub-cluster radius ($R_{\rm sub}$) and one assuming constant sub-cluster density ($\rho_{\rm sub}$). For the first case we choose $R_{\rm sub}$=0.1 pc and apply it to all the systems generated. In the latter case we fix the half-mass density of sub-clusters and adjust the radii accordingly. We choose the half-mass density of 21733 stars pc^{-3} which is the same as the density in the 5 sub-clusters case of the constant size run, which we consider 'fiducial'. The half-mass density is defined as the average density of the stars within the half-mass radius. For a Plummer sphere the half mass radius is $\sim 1.3R_{\rm plum}$ (Spitzer, 1987). Table 5.1 contains a full list of the input parameters for all the cases mentioned above.

5.7 Results

We used NBODY6 to evolve the systems for 10 Myr. At the end of the simulation we calculate the fraction of stars still bound. In order to do so we calculate the kinetic (T) and potential (Ω) energies of all the stars in the centre of mass frame of reference, and find the stars for which $T + \Omega < 0$. Then we find the centre of mass of the bound stars and move the coordinate system to coincide with this newly found centre of mass being at the origin. We also translate the velocities so that the new

Table 5.1: Initial conditions for sub-clusters of constant radius (above the line) and constant density (below the line). $N_{\rm sub}$ is the number of sub-clusters, $N_{\rm little}$ is the number of stars in a sub-cluster, $R_{\rm sub}$ is the sub-cluster half-mass radius in pc. $Q_{\rm sub}$ is the virial ratio of the sub-clusters, while $Q_{\rm big}$ is the virial-ratio of the system. $\rho_{\rm sub}$ is the half-mass number density of the sub-clusters. 10 runs with different random number seeds were simulated for each set.

$N_{\rm sub}$	$N_{\rm little}$	$R_{\rm sub}$[pc]	$Q_{\rm sub}$	$Q_{\rm big}$	$\rho_{\rm sub}$[stars pc^{-3}]
1	1000	0.1	0.5	0.5	108662
1	1000	0.1	1.0	0.5	108662
1	1000	0.1	1.5	0.5	108662
1	1000	0.1	2.0	0.5	108662
1	1000	0.1	4.0	0.5	108662
2	500	0.1	0.5	0.5	54331
2	500	0.1	1.0	0.5	54331
2	500	0.1	1.5	0.5	54331
2	500	0.1	2.0	0.5	54331
2	500	0.1	4.0	0.5	54331
5	200	0.1	0.5	0.5	21733
5	200	0.1	1.0	0.5	21733
5	200	0.1	1.5	0.5	21733
5	200	0.1	2.0	0.5	21733
5	200	0.1	4.0	0.5	21733
10	100	0.1	0.5	0.5	10866
10	100	0.1	1.0	0.5	10866
10	100	0.1	1.5	0.5	10866
10	100	0.1	2.0	0.5	10866
10	100	0.1	4.0	0.5	10866
1	1000	0.17	0.5	0.5	21733
1	1000	0.17	1.0	0.5	21733
1	1000	0.17	1.5	0.5	21733
1	1000	0.17	2.0	0.5	21733
1	1000	0.17	4.0	0.5	21733
2	500	0.14	0.5	0.5	21733
2	500	0.14	1.0	0.5	21733
2	500	0.14	1.5	0.5	21733
2	500	0.14	2.0	0.5	21733
2	500	0.14	4.0	0.5	21733
5	200	0.1	0.5	0.5	21733
5	200	0.1	1.0	0.5	21733
5	200	0.1	1.5	0.5	21733
5	200	0.1	2.0	0.5	21733
5	200	0.1	4.0	0.5	21733
10	100	0.08	0.5	0.5	21733
10	100	0.08	1.0	0.5	21733
10	100	0.08	1.5	0.5	21733
10	100	0.08	2.0	0.5	21733
10	100	0.08	4.0	0.5	21733

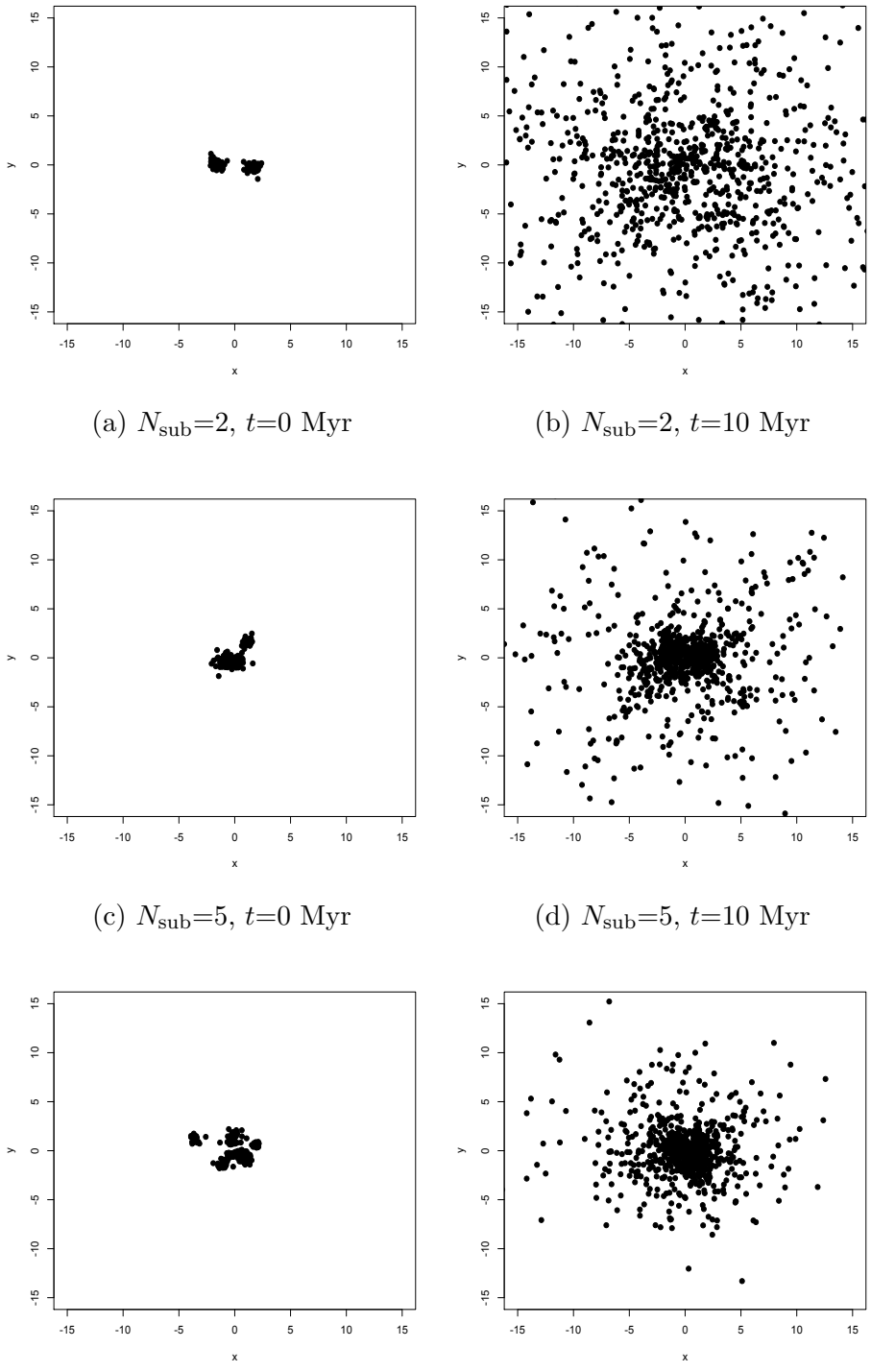
centre of mass is at rest. Now we recalculate the potential and kinetic energies and repeat the process, unless the change in $N_{\rm bound}$ from the previous step is less than 2. This is done to ensure that the bound stars are identified correctly, since a group of stars could be bound to each other but have a bulk motion with respect to the initial centre of mass.

Figure 5.2 shows the spatial distributions of stars in the clusters at the start of the simulation (left-hand column) and after 10 Myr (right-hand column) for systems with initial sub-cluster virial ratio of 1.5. While the system with $N_{\rm sub}$=2 disperses after 10 Myr, the systems with higher $N_{\rm sub} = 5$ retain about half of the stars. This fraction is even higher for systems with $N_{\rm sub}$=10.

The top panel of figure 5.3 shows how the bound fraction of stars after 10 Myr ($N_{\rm bound}/N_{\rm tot}$) changes with initial $N_{\rm sub}$ for the constant sub-cluster size case. The lines represent different values of initial sub-cluster virial ratios ($Q_{\rm sub}$), while the x-axis corresponds to increasing $N_{\rm sub}$. The error bars show the standard error in the mean calculated for ensembles of 10 simulations for each set of initial conditions.

The green line represents the initially virial system with $Q_{\rm sub} = 0.5$. After 10 Myr the system has lost some of its mass, but 80 per cent remain bound. The line is fairly flat and the bound fraction does not depend on the number of sub-clusters. A system of identical particles loses approximately 1 per cent of stars per relaxation time (in this case $\sim$ 10 Myr; from eqn. 1.17) due to evaporation (Spitzer, 1987). The presence of a mass function, however, increases the probability of ejections. Furthermore, $t_{\rm relax}$ changes as the system evolves which also affects the evaporation timescale (Bastian et al., 2008).

Bound fractions of systems with an initial sub-clump virial ratio of $Q_{\rm sub} = 1.0$ are represented by the red symbols. The change in bound fraction between a single Plummer sphere and one containing 10 sub-clusters is of the order of 10 per cent and with the exception of the $N_{\rm sub} = 1$ this case is indistinguishable from the $Q_{\rm sub} = 0.5$



(a) $N_{\rm sub}$=2, t=0 Myr

(b) $N_{\rm sub}$=2, t=10 Myr

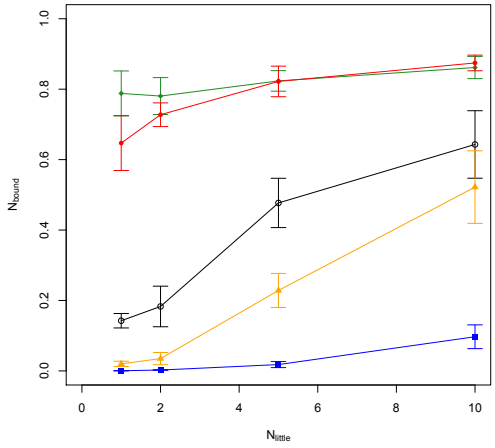
(c) $N_{\rm sub}$=5, t=0 Myr

(d) $N_{\rm sub}$=5, t=10 Myr

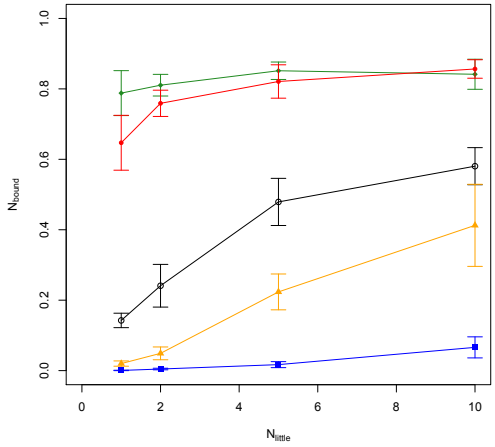
(e) $N_{\rm sub}$=10, t=0 Myr

(f) $N_{\rm sub}$=10, t=10 Myr

Figure 5.2: Spatial distributions of stars for sample sets of initial conditions with Q=1.5. Left-hand column shows systems at the start of the simulation (t=0), while the right-hand column shows the state of the same systems after 10 Myr.



(a) constant size



(b) constant density

Figure 5.3: Bound fraction ($N_{\rm bound}/N_{\rm tot}$) after 10 Myr as a function of number of sub-clusters for systems with constant sub-cluster size (top panel) and density (bottom panel). Green line represents systems with initial sub-cluster virial ratio of 0.5. Red line corresponds to $Q = 1.0$. Systems with initial virial ratio $Q = 1.5$ are represented by the black line. Orange line corresponds to $Q = 2.0$ and blue to $Q = 4.0$.

line.

The bottom line (blue) corresponds to systems with initial sub-cluster virial ratio of 4.0. It is fairly flat and stays close to zero at all times, barely reaching 10 per cent for 10 sub-clusters.

Systems with a virial ratio of $Q_{\rm sub} = 1.5$ (black), however, show a more dramatic change. In the case of a single Plummer sphere the bound fraction is less than 20 per cent, while in the case of 10 sub-clusters the bound fraction increases to 60 per cent. Similarly, initial sub-clump virial ratio of 2.0 (orange line) results in an increase from zero per cent for a single Plummer sphere to 50 per cent bound fraction for the case of 10 sub-clumps.

As expected, systems with the same $N_{\rm sub}$ but higher $Q_{\rm sub}$ have a lower virial ratio. Interestingly, however, systems with the same $Q_{\rm sub}$ but higher $N_{\rm sub}$ have higher bound fractions.

This behaviour can be explained by comparing the average velocities with the escape velocity. In virial equilibrium the kinetic energy is twice the magnitude of the potential energy of the cluster. Assuming stars of equal masses, the velocity dispersion of a Plummer sphere is:

$$\overline{v}^2 = \alpha \frac{G M_{\rm tot}}{R}, \tag{5.14}$$

where $M_{\rm tot}$ is the mass of the Plummer sphere, and $\alpha = 3\pi/32$ (Binney & Tremaine, 2008). From the formula it follows that the velocity dispersion is smaller for systems with a lower mass. Splitting the cluster mass into smaller sub-clusters will result in lower velocity dispersions: a system consisting of 2 sub-clusters, each with a mass of $M_{\rm tot}/2$ will have a velocity dispersion larger by a factor of $\sqrt{5}$ compared to a system of 10 sub-cluster, each with a mass of $M_{\rm tot}/10$.

The escape velocity of the cluster, however, depends solely on its total system

mass and radius: $v_{\rm esc} = \sqrt{2\frac{GM_{\rm tot}}{R_{\rm big}}}$. Since these values are the same in both of the above cases, substructure increases the chances of survival of a cluster by lowering the average velocity, which results in fewer stars exceeding the escape speed.

The bottom panel of figure 5.3 shows the bound fraction after 10 Myr for systems with constant sub-cluster density. The meaning of colours is analogous to the bottom panel. Systems with initial sub-cluster $Q_{\rm sub}$=0.5 are represented by the green lines. Red lines correspond to $Q_{\rm sub}$=1.0. Black ones represent $Q_{\rm sub}$=1.0 while orange and blue lines correspond to $Q_{\rm sub}$=2.0 and $Q_{\rm sub}$=4.0 respectively.

For $N_{\rm sub}$=1 there is no change as the only difference is the radius of the system. Altering the size of a Plummer sphere is just a scaling and ultimately it is the virial ratio that determines the fate of the system.

Similarly, for $N_{\rm sub}$=5 the results are identical in both cases, due to the same initial conditions. Since all the sub-cluster radii in this case were adjusted for the density to be the same as the density of the $N_{\rm sub} = 5$ constant radius case, bound fractions for systems with $N_{\rm sub} = 5$ are also exactly the same as in the left-hand panel.

For systems with $\rm N_{sub}$=2 a slight increase in the bound fraction can be observed. This behaviour is expected since, as shown in Tab. 5.1, the sub-clusters are larger ($R_{\rm sub}$=0.14 pc) than in the constant radius case ($R_{\rm sub}$=0.1 pc). Increasing their size causes a velocity dispersion decrease which results in more stars remaining bound.

On the other hand, systems with $\rm N_{sub}$=10, exhibit a smaller bound fraction since their radii decreased ($R_{\rm sub}$=0.08 pc) with respect to the previous case, which leads to a higher velocity dispersion. Larger velocities in a system whose size was not altered (the size of the whole cluster remains unchanged throughout) should lead to the system expanding more and losing more stars.

Overall, however, the differences between the plots are not significant and all the points are within their constant radius counterpart's error bars which suggests that

the stochastic effects due to the random positioning of the sub-clusters outweigh the differences due to different sub-cluster radii.

The behaviour of a single Plummer sphere was investigated by Lada et al. (1984), Goodwin (1997) as well as Baumgardt & Kroupa (2007), who showed that 33 per cent (equivalent to Q=1.5) is the critical eSFE value for instantaneous gas expulsion. For eSFE values lower than that no stars remain bound. Our results show that, if substructure is taken into account, this does not apply to the individual sub-clusters. A system comprised of Plummer spheres with Q=2 could indeed survive if the number of sub-clusters is high enough.

Star forming regions exhibit a large degree of substructure and tend to be initially subvirial with low core velocity dispersion (Kirk et al., 2007; Allison et al., 2010). The above example could represent a situation when a sub-cluster becomes locally supervirial following the gas expulsion, but is still part of a global virialised distribution.

We find that the substructure increases the chances of survival of the clusters in the post-gas-expulsion stage of their evolution. The more sub-clusters there are, the more likely it is that a significant fraction of stars is retained. Figure 5.3 also demonstrates a trend that the higher the virial ratio, the more substructure is required for survival. This happens because the average velocities of the individual stars become lower for sub-clusters with lower N_{sub} while the escape velocity of the whole cluster stays unchanged. Hence, fewer stars exceed the escape velocity and a larger fraction is retained.

5.8 Global virial ratio

In addition to trends described in Sec. 5.7, Figure 5.3 also shows that two systems with different initial sub-cluster virial ratios can result in very similar bound fractions

for example, in the constant size case, the bound fraction of a 10 sub-cluster system with a virial ratio $Q_{\rm sub}$=1.5 is the same as the bound fraction of a single Plummer sphere with $Q_{\rm sub} = 1$.

The above findings suggest that the sub-cluster virial ratio is not necessarily a good predictor of the systems fate. The survival of the cluster depends on the comparison between typical stellar velocity and the escape velocity. Therefore, we investigate how the cluster survivability is affected by the *total* initial virial ratio of the entire system ($Q_{\rm tot}$).

Fig. 5.4 shows the dependence of the bound fraction on $1/2Q_{\rm tot}$. The grey points are the results of our simulations, while the blue ones correspond to the equivalent Baumgardt & Kroupa (2007) data.

Overall, the plot shows a strong dependence of the bound fraction on the global initial virial ratio and follows the results of the Baumgardt & Kroupa (2007) simulations for single cluster. The only noticeable difference being the point corresponding to eSFE of 75 per cent which lies higher than our result. This is caused by the fact that Baumgardt & Kroupa (2007) included 10 times more stars in the simulations, hence their crossing time is greater. After 10 Myr, fewer crossing times have passed, therefore the systems are dynamically younger and have lost less mass through relaxation.

At the lower end of the eSFE scale it is possible to distinguish two populations, because both the $Q_{\rm sub}$=2.0 and $Q_{\rm sub}$=4.0 case result in systems that disperse, hence their bound fractions are scattered around 0.

On the opposite end there is very little horizontal scatter around the value of 1.0 – these are all the cases with $Q_{\rm sub}$=0.5. Putting a virialised Plummer sphere inside a bigger virialised Plummer sphere will result in a system whose $Q_{\rm tot}$ will also be close to 0.5. It is worth pointing out that one point is clearly above eSFE of 1.0. While it might seem counterintuitive, it is correct – one should not forget that *effective* star

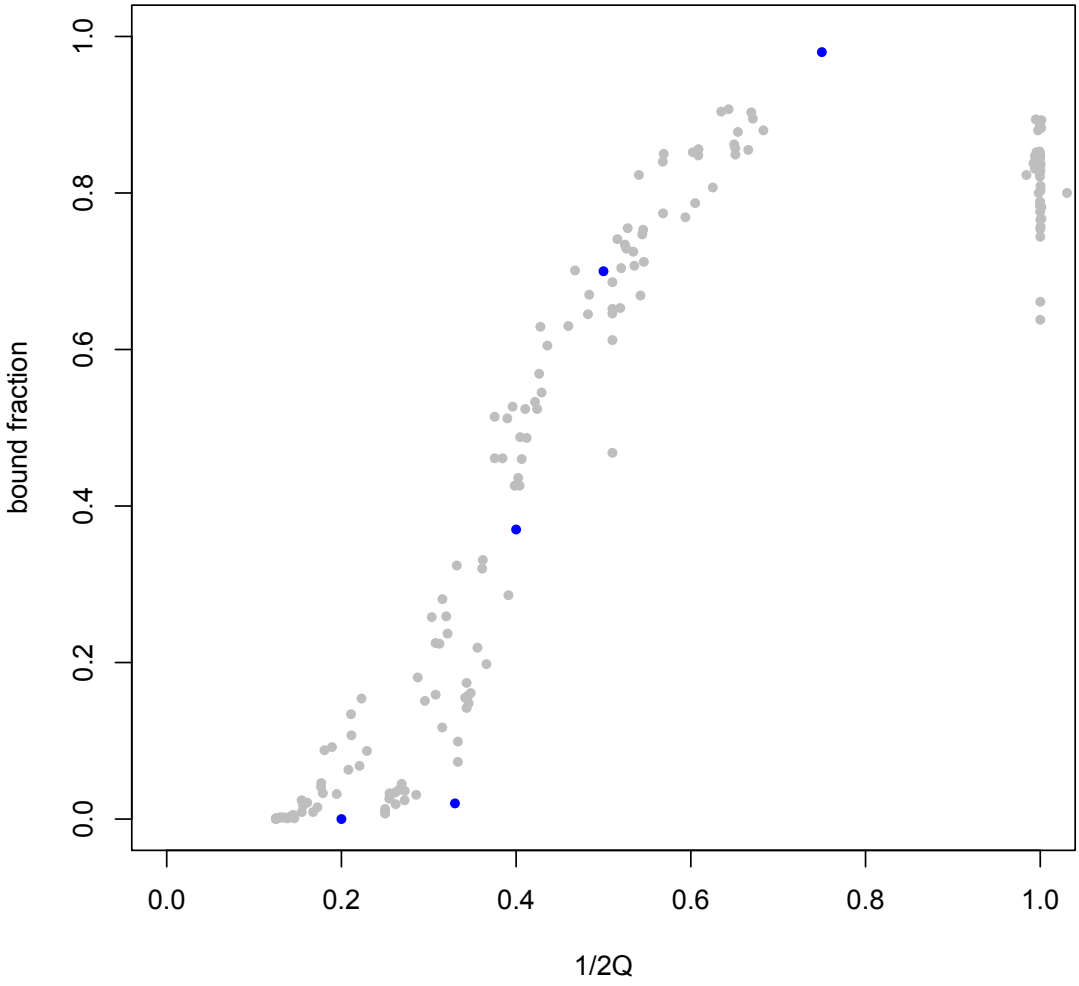


Figure 5.4: Bound fraction as a function of $1/2Q_{\rm tot}$ (eSFE). Grey points represent the results of our simulations, while blue points represent equivalent results by Baumgardt & Kroupa (2007)

formation efficiency is essentially just the inverse of virial ratio and the name is based on the assumption that initially the star-gas mix is virialised, see Sec. 1.2.2.

It might therefore seem more appropriate to plot the bound fraction against Q, however we decided to retain the eSFE convention, since the result is an analogue of the Baumgardt & Kroupa (2007) plot for individual Plummer spheres.

5.9 Messy initial conditions

We have made a few simplifying assumptions in setting up the initial conditions for the simulations. Firstly, the systems are set up as a Plummer sphere of Plummer spheres, which is not a very realistic model. While it is not unreasonable for a small, dense clump to be represented by a Plummer sphere, there is no physical reason for these sub-clusters to form a larger Plummer sphere. This setup was chosen purely due to the ease of its implementation.

Furthermore, we also make the assumption that the gas expulsion is instantaneous. Gradual gas expulsion could affect the velocity distributions but the change should not be critical.

We were trying to establish a relationship between $N_{\rm sub}$, $Q_{\rm sub}$ and the final bound fraction, hence we needed these values to be well defined for the systems. Moreover, the clusters in our simulations all have constant N and $Q_{\rm big} = 0.5$. In the runs we also keep sub-cluster size or density constant.

Now that we know that the global virial ratio is a better predictor of the fate of the system, these restrictions can be quite easily relaxed and we can generate randomised *messy* initial conditions.

I created a routine that randomly allocates values to the input parameters – following ranges: $N_{\rm sub}$ [3, 10], $Q_{\rm big}$ [0.25, 0.75], each sub-cluster can have a different $N_{\rm little}$ [100, 300] and $Q_{\rm sub}$ [0.5, 3.0], and $R_{\rm little}$ [0.75, 1.5]. Tab. 5.2 contains the

Table 5.2: An example of 3 subclusters from randomly generated set of initial conditions. $N_{\rm little}$ is the number of stars in a subcluster, $Q_{\rm sub}$ is the virial ratio of the subcluster, and $R_{\rm sub}$ is the radius of the subcluster. The virial ratio of the big Plummer sphere is $Q_{\rm big}$=0.3, the total virial ratio of the system is $Q_{\rm tot}$=0.84, the system consists of $N_{\rm tot}$=647 with a total mass of $M_{\rm tot}$=192 $M_\odot$.

$N_{\rm little}$	$Q_{\rm sub}$	$R_{\rm sub}$
221	1.92	0.94
166	1.44	1.39
260	0.82	1.26

parameters of one set of initial conditions generated this way. It contains $N_{\rm sub} = 3$ subclusters, the big Plummer sphere has a virial ratio of $Q_{\rm big} = 0.3$, its total virial ratio is $Q_{\rm tot}$=0.84 (hence $1/2Q$=0.6). The parameters of the individual subclusters (number of stars, virial ratio, and radius) are listed in the table. After evolving this system for 10 Myr its bound fraction is 0.84 which is a value that we expected based on the trend from Fig. 5.4.

This set of initial conditions is shown in red in Fig. 5.5, the *messy* initial conditions are represented by the red points, while the grey points are the results of the previous simulations.

We found that the red points follow a distribution similar as the previously discussed clusters, albeit their bound fractions tend to be slightly higher than the average for their ϵ. This result is broadly consistent with our postulate that the fate of a system is determined by the global virial ratio. The size and number of clusters, however, can also affect the result. This is the reason why at higher virial ratios it is possible to distinguish between the $Q_{\rm sub} = 4.0$ and $Q_{\rm sub} = 2.0$ populations in Fig. 5.5.

The red points having slightly higher bound fractions than the grey points suggests that the sub-clusters within one region having different sizes and different number of stars (i.e. 'messy' initial conditions) might also have an effect. This could be verified by performing more simulations of 'messy' style initial conditions and com-

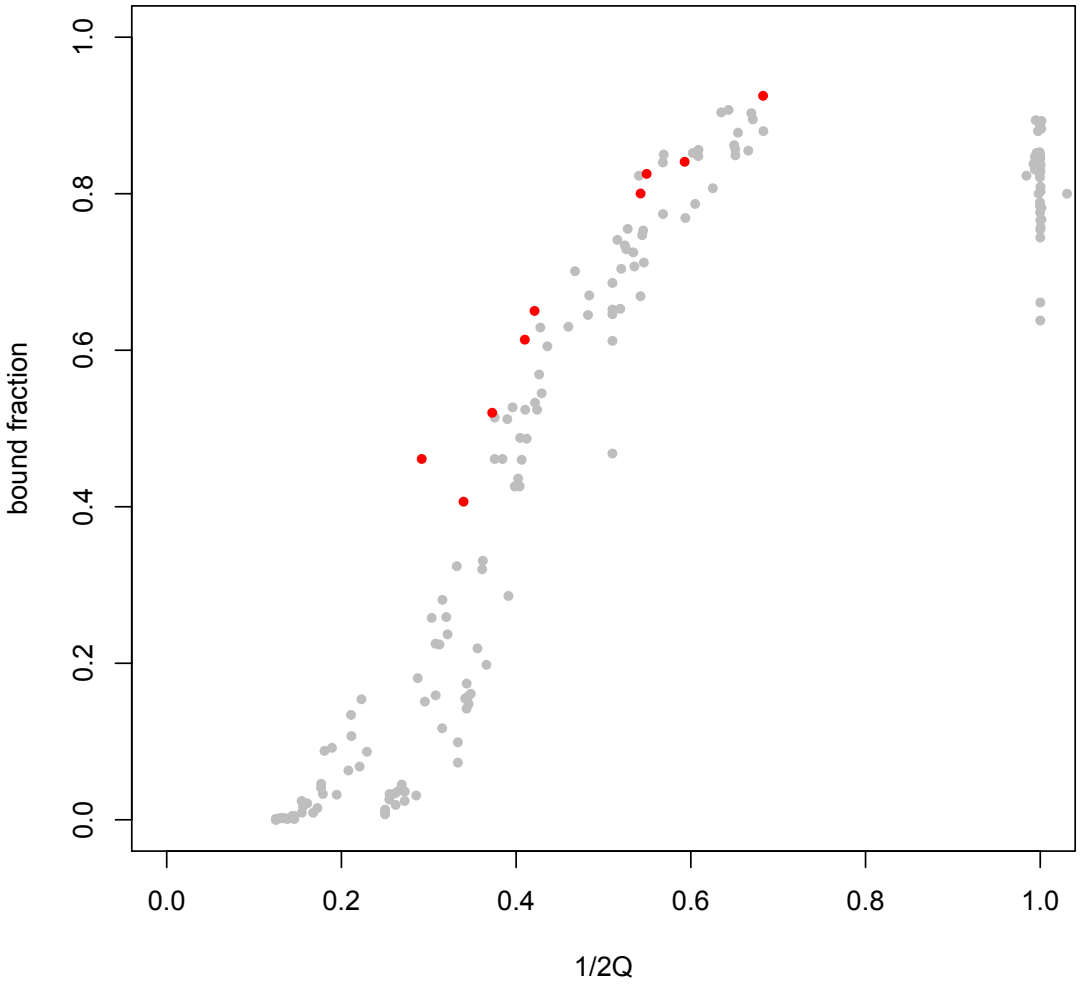


Figure 5.5: Bound fraction as a function of $1/2Q$ (eSFE). The results of simulations outlined in the earlier sections of this chapter are shown in grey; red points represent the *messy* initial conditions that were generated to by randomly picking parameters from a permitted range rather than strictly adhering to a prescribed grid.

paring whether the points from the previous initial conditions grid typically lie below them.

Chapter 6

Conclusions

In this thesis I investigated how the presence of structure affects the evolution of star forming regions. I calculated $\mathcal{Q}$ for observed regions to find out whether they are bound or unbound. I performed N-body simulations in order to investigate how substructure affects the fate of the regions after gas expulsion. I also outlined a broad range of structure finding methods used in other fields and assessed their suitability for use in the context of star forming regions.

6.1 Parameter $\mathcal{Q}$ and dynamical evolution

I explained how the $\mathcal{Q}$ parameter can be used to quantify the clumpiness of a system and how it can provide clues about the dynamical evolution of the system. I calculated $\mathcal{Q}$ of class I and class II populations in regions observed by Gutermuth et al. (2009) and used it predict if the systems are bound.

Based on the assumption that the class II sources are older, comparing the $\mathcal{Q}$ can tell us if the system is bound. Low values of $\mathcal{Q}$ are indicative of a fractal distribution, while high values mean a smoother centrally concentrated distribution. A bound system should undergo relaxation, which results in the substructure being deleted,

therefore a significant increase in $\mathcal{Q}$ implies a bound system. Conversely, if the value of $\mathcal{Q}$ stays the same, it suggests that the system is unbound and the structure has not been erased.

6.2 N-body simulations

I set up a simple model of clumpy initial conditions – several Plummer spheres, each containing a few hundred stars, distributed within a larger Plummer potential. I then varied the virial ratio of the sub-clusters and used NBODY6 to evolve the systems for 10 Myr in order to investigate how the virial ratio of the sub-clusters affects the fate of the system.

I found out, however, that it is the global virial ratio that matters – looking at an individual sub-cluster is insufficient as the region might be bound as a whole even if the individual members are supervirial. This result emphasises the need to treat the regions as a whole and demonstrates how the substructure can bias the interpretation of the observations.

6.3 Structure finding algorithms

I also explored structure finding methods frequently used in computer science and data mining and investigated if they can be used to obtain more information about star forming regions.

I have found that most of the algorithms that pick up the substructure are not satisfactory. Many are graph theory based and therefore better suited for analysing networks with explicitly defined connections between the points. However, the density based methods also performed rather poorly in situations where one region contained sub-clusters with different densities. The extent of the field of view and the

projection effects also have a detrimental effect that can bias the results. It is extremely difficult to disentangle superimposed clumps without making assumptions about the distributions or number of sub-clusters in the region. This ties in to the previously raised concerns regarding the completeness of the data and the problems caused by looking at individual sub-clusters instead of considering the region as a whole.

Some of the shortcomings could perhaps be overcome by using more advanced machine learning methods such as self organising maps, but this goes beyond the scope of this thesis.

Bibliography

Aarseth S., 1999, Publications of the Astronomical Society of the Pacific, 111, 1333

Aarseth S., 2003, Gravitational N-Body Simulations

Aarseth S., Henon M., Wielen R., 1974, Astronomy and Astrophysics, 37, 183

Adamo A., Kruijssen J. M. D., Bastian N., Silva-Villa E., Ryon J., 2015, Monthly Notices of the Royal Astronomical Society, 452, 246

Allison R., Goodwin S., Parker R., De Grijs R., Portegies Zwart S., Kouwenhoven M., 2009a, The Astrophysical Journal, 700, 99

Allison R. J., Goodwin S. P., Parker R. J., Portegies Zwart S. F., de Grijs R., 2010, Monthly Notices of the Royal Astronomical Society, 407, 1098

Allison R. J., Goodwin S. P., Parker R. J., Portegies Zwart S. F., de Grijs R., Kouwenhoven M. B. N., 2009b, Monthly Notices of the Royal Astronomical Society, 395, 1449

Anders P., Baumgardt H., Gaburov E., Portegies Zwart S., 2012, Monthly Notices of the Royal Astronomical Society, 421, 3557

André P., Belloche A., Motte F., Peretto N., 2007, Astronomy & Astrophysics, 472, 519

André P., Di Francesco J., Ward-Thompson D., Inutsuka S.-i., Pudritz R. E., Pineda J., 2014, Protostars and Planets VI, 24

Ankerst M., Breunig M. M., Kriegel H.-P., Sander J., 1999, ACM SIGMOD Record, 28, 49

Armitage P., Clarke C., Palla F., 2003, Monthly Notices of the Royal Astronomical Society, 342, 1139

Ballesteros-Paredes J., 2006, Monthly Notices of the Royal Astronomical Society, 372, 443

Ballesteros-Paredes J., Klessen R. S., Mac Low M.-M., Vazquez-Semadeni E., 2007, in Protostars and Planets V, pp. 63–80

Basri G., Brown M. E., 2006, Annual Review of Earth and Planetary Science, 34, 193

Bastian N., Covey K. R., Meyer M. R., 2010, Annual Review of Astronomy and Astrophysics, 48, 339

Bastian N., Gieles M., Ercolano B., Gutermuth R., 2009, Monthly Notices of the Royal Astronomical Society, 392, 868

Bastian N., Gieles M., Goodwin S. P., Trancho G., Smith L. J., Konstantopoulos I., Efremov Y., 2008, Monthly Notices of the Royal Astronomical Society, 389, 223

Bate M. R., 1998, The Astrophysical Journal, 508, L95

Baumgardt H., Kroupa P., 2007, Monthly Notices of the Royal Astronomical Society, 380, 1589

Bhattal A. S., Francis N., Watkins S. J., Whitworth a. P., 1998, Monthly Notices of the Royal Astronomical Society, 297, 435

Binney J., Tremaine S., 2008, Galactic Dynamics: Second Edition

Boldyrev S., 2002, The Astrophysical Journal, 569, 841

Bonnell I. A., Bate M. R., 1994, Monthly Notices of the Royal Astronomical Society, 271, 999

Bonnell I. A., Larson R. B., Zinnecker H., 2007, Protostars and Planets V, 16

Bontemps S., André P., Terebey S., Cabrit S., 1996, Astronomy and Astrophysics, 311, 858

Bressert E. et al., 2010, Monthly Notices of the Royal Astronomical Society: Letters, 409, L54

Bromm V., 2013, Reports on Progress in Physics, 76, 11

Calinski T., Harabasz J., 1974, Communications in Statistics - Simulation and Computation, 3, 1

Calvet N., Hartmann L., 1994, The Astrophysical Journal, 434, 330

Cartwright A., Whitworth A. P., 2004, Monthly Notices of the Royal Astronomical Society, 348, 589

Cartwright A., Whitworth A. P., 2009, Monthly Notices of the Royal Astronomical Society, 392, 341

Chabrier G., 2003, Publications of the Astronomical Society of the Pacific, 809, 763

Clark J. S., Negueruela I., Crowther P. A., Goodwin S. P., 2005, Astronomy and Astrophysics, 434, 949

Clarke C. J., Pringle J. E., 1991, Monthly Notices of the Royal Astronomical Society, 249, 584

Colombo D. et al., 2014, The Astrophysical Journal, 784, 3

Costa L. D. F., Rodrigues F. A., Travieso G., Villas Boas P. R., 2007, Advances in Physics, 56, 167

Crowther P. A., Schnurr O., Hirschi R., Yusof N., Parker R. J., Goodwin S. P., Kassim H. A., 2010, Monthly Notices of the Royal Astronomical Society, 408, 731

Crutcher R. M., 1999, The Astrophysical Journal, 520, 706

Csardi G., Nepusz T., 2006, InterJournal, Complex Sy, 1695

de la Fuente Marcos R., de la Fuente Marcos C., 2009, The Astrophysical Journal, 1998, 26

Dijkstra E., 1959, Numerische Mathematlk, 1, 269

Duchêne G., Kraus A., 2013, Annual Review of Astronomy and Astrophysics, 51, 269

Dunham M. M. et al., 2014, Protostars and Planets VI, 195

Elmegreen B. G., 2000, The Astrophysical Journal, 10, 277

Elmegreen B. G., 2006, "Mass loss from stars and the evolution of stellar clusters"

Erdős P., Rényi A., 1959, Publicationes Mathematicae, 6, 290

Ester M., Kriegel H.-P., Sander J., Xu X., 1996, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 226

Evans N. J., 1999, Annual Review of Astronomy and Astrophysics, 37, 311

Evans N. J. et al., 2009, The Astrophysical Journal Supplement Series, 181, 321

Gieles M., Portegies Zwart S., 2011, Monthly Notices of the Royal Astronomical Society: Letters, 410, L6

Goodman J., Hut P., 1993, The Astrophysical Journal, 403, 271

Goodwin S., 1997, Monthly Notices of the Royal Astronomical Society, 284, 785

Goodwin S., 2009, Astrophysics and Space Science, 324, 259

Goodwin S. P., 2010, Philosophical Transactions of the Royal Society, 368, 851

Goodwin S. P., Bastian N., 2006, Monthly Notices of the Royal Astronomical Society, 373, 752

Goodwin S. P., Whitworth a. P., 2004, Astronomy & Astrophysics, 4464, 12

Greene T., Wilking B., 1994, The Astrophysical Journal, 434, 614

Greene T. P., Aspin C., Reipurth B., 2008, The Astronomical Journal, 135, 1421

Gutermuth R. et al., 2008, The Astrophysical Journal, 674, 336

Gutermuth R. A., Megeath S. T., Myers P. C., Allen L. E., Pipher J. L., Fazio G. G., 2009, The Astrophysical Journal Supplement Series, 184, 18

Gutermuth R. A., Pipher J. L., Megeath S. T., Myers P. C., Allen L. E., Allen T. S., 2011, The Astrophysical Journal, 739, 17

Hartmann L., Megeath S. T., Allen L., Luhman K., Calvet N., DAlessio P., Franco-Hernandez R., Fazio G., 2005, The Astrophysical Journal, 629, 881

Heggie D., Mathieu R., 1986, The Use of Supercomputers in Stellar Dynamics

Heyer M., Krawczyk C., Duval J., Jackson J. M., 2009, The Astrophysical Journal, 699, 1092

Heyer M. H., Carpenter J. M., Snell R. L., 2001, The Astrophysical Journal, 551, 852

Hughes A. et al., 2010, Monthly Notices of the Royal Astronomical Society, 2086

Humphries M. D., Gurney K., 2008, PLoS ONE, 3

Jarník V., 1930, Práce moravské přírodovědecké společnosti, 6, 57

Kaufman L., Rousseeuw P. J., 1990, Finding Groups in Data

Kennicutt R. C., Evans N. J., 2012, Annual Review of Astronomy and Astrophysics, 50, 531

Kenyon S. J., Mercedes G., Whitney B. A., 2008, Handbook of Star Forming Regions, I, 405

King R. R., Parker R. J., Patience J., Goodwin S. P., 2012, Monthly Notices of the Royal Astronomical Society, 421, 2025

Kirk H., Johnstone D., Tafalla M., 2007, The Astrophysical Journal, 668, 1042

Kolmogorov A., 1941, Doklady Akademiia Nauk SSSR

Kroupa P., 2002, Science, 295, 82

Lada C., 1987, Star Forming Regions

Lada C., Margulis M., Dearborn D., 1984, The Astrophysical Journal, 285, 141

Lada C. J., Lada E. A., 2003, Annual Review of Astronomy and Astrophysics, 41, 57

Larson R. B., 1969, Monthly Notices of the Royal Astronomical Society, 145, 271

Larson R. B., 1981, Monthly Notices of the Royal Astronomical Society, 194, 809

Larson R. B., 2000, in Star formation from the small to the large scale. ESLAB symposium, p. 13

Longmore S. N. et al., 2014, Protostars and Planets VI, 24

Low C., Lynden-Bell D., 1976, Monthly Notices of the Royal Astronomical Society, 176, 367

Lynden-Bell D., 1967, Monthly Notices of the Royal Astronomical Society

Magnani L., Blitz L., Mundy L., 1985, The Astrophysical Journal, 295, 402

Makino J., 1991, The Astrophysical Journal, 369, 200

Makino J., Aarseth S., 1992, Publications of the Astronomical Society of Japan, 44, 141

Maschberger T., 2013, Monthly Notices of the Royal Astronomical Society, 429, 1725

Masunaga H., Inutsuka S.-I., 2000, The Astrophysical Journal, 531, 350

Masunaga H., Miyama S., Inutsuka S.-I., 1998, the Astrophysical Journal, 495, 346

McKee C. F., Ostriker E. C., 2007, Annual Review of Astronomy and Astrophysics, 45, 565

Men'shchikov A. et al., 2010, Astronomy and Astrophysics, 518, L103

Oka T., Hasegawa T., Sato F., 2001, The Astrophysical Journal, 562, 348

Opsahl T., Colizza V., Panzarasa P., Ramasco J. J., 2008, Physical Review Letters, 101, 168702

Palla F., 2012, AIP Conference Proceedings, 1480, 22

Parker R. J., Dale J. E., Ercolano B., 2015, Monthly Notices of the Royal Astronomical Society, 446, 4278

Parker R. J., Goodwin S. P., 2015, Monthly Notices of the Royal Astronomical Society, 449, 3381

Parker R. J., Goodwin S. P., Kroupa P., Kouwenhoven M. B. N., 2009, Monthly Notices of the Royal Astronomical Society, 397, 1577

Parker R. J., Wright N. J., Goodwin S. P., Meyer M. R., 2014, Monthly Notices of the Royal Astronomical Society, 438, 620

Patwary M. A., Palsetia D., Agrawal A., Liao W.-k., Manne F., Choudhary A., 2013, Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis on - SC '13, 1

Plummer H., 1911, Monthly Notices of the Royal Astronomical Society, LXXI, 460

Plummer H., 1915, Monthly Notices of the Royal Astronomical Society, LXXVI, 107

Preibisch T., Kim Y., 2005, The Astrophysical Journal, 160, 401

Prim R., 1957, Bell System Technical Journal, 36, 1389

R Core Team, 2012, R: A Language and Environment for Statistical Computing

Rawiraswattana K., 2012, PhD thesis, University of Sheffield

Rees M. J., 1976, Monthly Notices of the Royal Astronomical Society, 176, 483

Salpeter E., 1955, The Astrophysical Journal, 121, 161

Sander J., Qin X., Lu Z., Niu N., Kovarsky A., 2003, Advances in Knowledge Discovery and Data Mining. 7th Pacific-Asia Conference, PAKDD 2003, 75

Schubert E., Koos A., Emrich T., Züfle A., Schmid K. A., Zimek A., 2015, PVLDB

Shu F., Adams F., Lizano S., 1987, Annual review of astronomy, 25, 23

Smith R., Fellhauer M., Goodwin S., Assmann P., 2011, Monthly Notices of the Royal Astronomical Society, 414, 3036

Soderblom D. R., Hillenbrand L. A., Jeffries R. D., Mamajek E. E., Naylor T., 2014, Protostars and Planets VI

Sokal R., Michener C., 1958, University of Kansas Science Bulletin, 2, 1409

Spitzer L., 1987, Dynamical evolution of globular clusters

Spitzer L. J., 1969, The Astrophysical Journal, 158, L139

Verschueren W., David M., 1989, Astronomy and Astrophysics, 219, 105

Ward-Duong K. et al., 2014, Proceedings of the International Astronomical Union, 8, 74

Ward-Thompson D., André P., Crutcher T., Johnstone D., Onishi T., C W., 2007, in Protostars and Planets V, pp. 33–46

Watts D. J., Strogatz S. H., 1998, Nature, 393, 440

Whitworth A. P., 2000, Astronomy & Geophysics, 41, 18

Whitworth A. P., Stamatellos D., 2006, Astronomy & Astrophysics, 829, 817

Williams J. P., Blitz L., McKee C. F., 2000, in Protostars and Planets IV, pp. 97–120

Wright N. J., Parker R. J., Goodwin S. P., Drake J. J., 2014, Monthly Notices of the Royal Astronomical Society, 438, 639

Zhou S., Mondragon R., 2005, IEEE Communications Letters, 8, 180