Active and Merging Galaxies

Caroline Scott

Astrophysics Group
Department of Physics
Imperial College London

Thesis submitted for the Degree of Doctor of Philosophy to
Imperial College London
· 2015 ·
Galaxy close pairs are studied to investigate the effects of gravitational interactions on star formation and black hole accretion processes in merger progenitors. We derive star formation rates from near-ultraviolet luminosities; this is a new method for studying mergers and provides unique insight into recent star formation rates. A range of progenitor masses are considered, as well as the separation between merging galaxies and the environment they inhabit. Star formation enhancements in major versus minor close pairs are also considered. Pairs are extracted from the SDSS by identifying galaxies with small angular separation and small recessional velocity difference. Optical photometry in five filters is available for these galaxies. The pairs sample is cross-matched with near-ultraviolet flux measurements from GALEX and specific star formation rates are derived. We study the fraction of active galaxies as a function of separation in close pairs and seek observational evidence for merger activity triggering black hole accretion. Optical emission lines are used to identify progenitors harbouring active galactic nuclei, and the ratio of active galaxies in close pairs is compared to that of non-mergers.

The variable properties of a sample of quasi-stellar objects (QSOs) are analysed. We present optimal QSO classification algorithms that exploit time series variability features calculated from Pan-STARRS light curves. This groundbreaking work crosses boundaries between astrophysics, statistics and machine learning. Spectroscopically confirmed QSOs and stars are used to train Support Vector Machine and Random Forest algorithms. We compare and evaluate the outcome of these models then apply them to Pan-STARRS light curves over nine medium deep fields, each covering 7°-squared and located uniformly across the sky, to predict likely QSO candidates. We present a host of new variability features to characterise and provide measures of QSO variability.
Contents

Abstract 2
List of Tables 7
List of Figures 8
Publications and Conference Appearances 14
Declaration and Copyright 15
Foreword 17

1 Introduction 19
  1.1 The Standard Paradigm for Structure Formation 20
    1.1.1 Monolithic Collapse 24
    1.1.2 Hierarchical Evolution of Galaxies through Mergers 25
    1.1.3 Downsizing: An Issue with the Hierarchical Formation Model? 28
  1.2 The Hubble Sequence and Hubble Types 29
    1.2.1 Elliptical 29
    1.2.2 Spiral 32
    1.2.3 Lenticular/SO 33
    1.2.4 Irregular 34
  1.3 The Role of Active Galaxies in Evolutionary Models 34
    1.3.1 Active Galactic Nuclei 34
    1.3.2 AGN Variability 36
    1.3.3 AGN Feedback and the Regulation of Star Formation 39
    1.3.4 Linking Active and Merging Galaxies in Evolutionary Models 42
  1.4 Studying Galaxy Mergers 44
    1.4.1 Initial Studies 44
    1.4.2 Star Formation Enhancement 45
4 Star Formation and AGN Activity in Close Pairs 104
   4.1 Recent Star Formation in Close Pairs 106
      4.1.1 SF Enhancement as a Function of Separation and Mass 106
      4.1.2 SF Enhancement as a Function of Separation and Environment 107
   4.2 Major and Minor Mergers 110
      4.2.1 SF Enhancement as a Function of Separation and Mass 110
      4.2.2 Impact of Mass Ratio on SF Enhancement in Minor Mergers 113
      4.2.3 SF Enhancement as a Function of Separation and Environment 115
   4.3 AGN Activity in Close Pairs 118
      4.3.1 Emission-Line Analysis for Close Pairs 118
      4.3.2 Emission-Line Analysis for Major and Minor Mergers 119
   4.4 Summary and Discussion 122

5 Time Series Methods of QSO Classification 127
   5.1 Machine Learning Classification Models 129
      5.1.1 Support Vector Machines 130
      5.1.2 Decision Trees, Ensemble Learning and Random Forest Classifiers 131
      5.1.3 Machine Learning in Astrophysics 132
   5.2 Data Preparation 134
      5.2.1 Recalibration of Magnitudes 134
      5.2.2 Preparing the Light Curves 135
   5.3 Features 136
      5.3.1 Autocorrelation and Slotted Autocorrelation 136
      5.3.2 List of Features 139
   5.4 Invariance Between Pan-STARRS Fields 153
   5.5 Summary 157

6 Random Forest and SVM QSO Classification Models 158
   6.1 Modelling 160
      6.1.1 Creating the Training Set 160
      6.1.2 Random Forest Model 161
      6.1.3 SVM Model 163
   6.2 Identifying the Best Features 165
   6.3 Random Forest Results 173
      6.3.1 RF Blind Test 177
      6.3.2 RF Predicted QSOs 178
   6.4 SVM Results 185
6.4.1 SVM Blind Test .................................................. 188
6.4.2 SVM Predicted QSOs .......................................... 188
6.5 QSOs Predicted by both RF and SVM ......................... 194
6.6 Summary and Discussion ....................................... 198

7 Conclusions .............................................................. 201
  7.1 Summary and Discussion ...................................... 201
  7.2 Outlook ............................................................. 205

Acknowledgements ....................................................... 207

References ................................................................. 210
List of Tables

2.1 GALEX Baseline Mission Surveys (up to GR4, completed in Fall 2007) .................................................. 57
4.1 Median SSFR ($yr^{-1}$) derived from NUV luminosity for each stellar mass and separation bin. .............................. 107
4.2 Median SSFR ($yr^{-1}$) derived from NUV luminosity for each environment and separation bin. .............................. 109
4.3 Summary of the SSFR enhancements from the lowest separation bin to the highest separation bin (i.e. close pairs versus the control sample) from the various analyses in this chapter. ........................................ 123
6.1 Cross Validation results for Stars (S) and QSOs (Q) for RF Models I, II and III. There were 1,468 QSOs and 3,643 stars in the training set. ........................................ 162
6.2 Cross Validation results for Stars (S) and QSOs (Q) for SVM Models I, II and III. There were 1,468 QSOs and 3,643 stars in the training set. ........................................ 164
6.3 Analysis of Top Features: RF Cross Validation results for combinations of the top features in $g$-band only. ........................................ 174
6.4 Predictions from RF Model I, RF Model II and RF Model III. ........................................ 175
6.5 RF blind test evaluation table: precision and recall rates are shown for Stars and QSOs predicted from RF Models I, II and III. ........................................ 178
6.6 Predictions from SVM Model I, SVM Model II, and SVM Model III. ........................................ 186
6.7 SVM blind test evaluation table: precision and recall rates are shown for Stars and QSOs predicted from SVM Models I, II and III. ........................................ 188
List of Figures

1.1 Centaurus A .................................................. 27
1.2 The Hubble tuning fork illustrates Edwin Hubble’s system of morphological classification; showing elliptical, SO, spiral and barred spiral galaxies. 30
1.3 Top left: M87 elliptical galaxy. Top right: Lenticular galaxy NGC 5866. Bottom left: M101 spiral galaxy, also known as the Pinwheel galaxy. Bottom right: NGC 6822, also known as Barnard’s Galaxy. 31
1.4 Very Large Array (4.9GHz) image of the radio emission from the quasar 3C 175 at $z = 0.77$ (Bridle et al., 1994). 35
1.5 Illustration of the AGN Unification model. 36
1.6 Top (excerpted from Fan (2012)): Evolution of the density of luminous quasars. Bottom (excerpted from Bouwens et al. (2011) and Fan (2012)): Rest-frame continuum UV luminosity density at $z \sim 10$, and star formation rate density derived from the extinction-corrected luminosity density. 41
1.7 Excerpted from Larson & Tinsley (1978): The two-colour plots for morphologically normal and peculiar galaxies with latitudes $|b| > 20^\circ$. 45
1.8 Excerpted from Bridge et al. (2007): Number of mergers per galaxy ($L_{IR} \geq 5 \times 10^{10}$) as a function of redshift. 51

2.1 Excerpt from Morrissey et al. (2007), Table 1. 56
2.2 BOSS spectroscopy for two QSO TDSS targets. Left: $z = 3.53$, right: $z = 5.01$ (excerpted from a TDSS presentation by Paul Green, 2013). 59
2.3 Excerpted from Schawinski et al. (2007a): Volume-limited UV color-magnitude relation. 64
2.4 Excerpted from Yi (2008), originally used by Yi, Demarque & Oemler (1998): The composite spectrum of the giant elliptical galaxy NGC 4552 shows a classic example of the UV upturn. 65
2.5 Excerpted from Shimasaku et al. (2001): Correlation of concentration index with visually classified Hubble type, $T$, for 426 galaxies with $R_{50} \geq 2''$. 67

2.7 Excerpted from Kauffmann et al. (2003a): BPT plot with emission-line flux ratio [OIII]/Hβ versus the ratio [NII]/Hα for 55,757 galaxies where all four lines are detected with S/N > 3.

3.1 Top: NUV- r colour-magnitude plot. Bottom left: Optical colour-magnitude plot for galaxies from the optical close pairs sample after our K-correction and extinction correction. Bottom right: NUV- r distribution.

3.2 Comparison between our mass estimates with MPA masses for both close and wide pairs samples.

3.3 Top: NUV- r against stellar mass plot (left) for the wide pairs (black) and close pairs (red) samples. Bottom: Stellar mass distribution for the wide pairs (black histogram) and close pairs (red histogram) samples.

3.4 Halo mass distribution for the close and wide pairs samples.

3.5 BPT plots: $\log_{10}(\text{[NII]}/\text{H}\alpha)$ is plotted against $\log_{10}(\text{[OIII]}/\text{H}\beta)$ for the close pairs (top) and wide pairs (bottom). Objects are classified as Star-forming, Transition, LINER or Seyfert depending on the region they inhabit on the BPT diagram.

3.6 Emission-line luminosity-derived SFRs from the MPA SDSS DR7 catalogue (black) are compared with NUV photometry-derived SFRs for our wide pairs sample (green) and close pairs sample (red).

3.7 NUV- r and $\log_{10}$ SFR plot before (top) and after (bottom) internal extinction correction.

3.8 SDSS galaxy: ra, dec = (179.714, 42.7187), z = 0.0019. The red squares show peaks identified by the pair-extraction code; since there are two peaks our procedure incorrectly concludes that there are two neighbouring galaxies.

3.9 Examples from the SDSS online image viewer of galaxy mergers from our close pairs sample.

3.10 Examples from the SDSS online image viewer of galaxy mergers from our close pairs sample.

3.11 Examples from the SDSS online image viewer of galaxy mergers from our close pairs sample.
3.12 Optical colour-magnitude contour plots showing SDSS galaxies from the close pair catalogue separated by visually classified morphological type. We overplot the visually classified bulge-dominated spirals with green contours in the right plot.  

3.13 Redshift distribution (top) and optical colour-magnitude plot (bottom) for major and minor close pairs samples.  


3.15 Emission-line luminosity-derived SFRs from the MPA SDSS DR7 catalogue are compared with NUV photometry-derived SFRs for our wide pairs sample for the majors sample (left) and minors sample (right).  

4.1 Top: Median NUV-$r$ colours for close and wide pairs binned by separation for both low stellar mass galaxies and high stellar mass galaxies. Bottom: SSFR difference between the separation bin in question and the widest separation bin, for both low and high mass galaxies.  

4.2 Top: Median NUV-$r$ colours for close and wide pairs binned by separation for pairs in field, group and cluster environments. Bottom: SSFR difference between the separation bin in question and the widest separation bin, for each environment.  

4.3 Median NUV-$r$ and $\Delta$SSFR values for each environment/separation bin are plotted for low stellar mass galaxies and high stellar mass galaxies.  

4.4 Median NUV-$r$ and $\Delta$SSFR values are plotted for the major mergers sample (with pair mass ratio $>1/3$) -top, and and minor mergers sample (with pair mass ratio $<1/3$) -bottom.  

4.5 Median NUV-$r$ and $\Delta$SSFR values are plotted for the minor mergers sample split into primary progenitors and secondary progenitors.  

4.6 Median NUV-$r$ colour for minor merger primaries (red) and secondaries (blue) binned by mass ratio (only close pairs are considered here).  

4.7 Median NUV-$r$ colour for minor merger primaries (dotted) and secondaries (dashed) binned by mass ratio.  

4.8 Median NUV-$r$ and $\Delta$SSFR values are plotted for the major mergers sample (with pair mass ratio $>1/3$) -top, and and minor mergers sample (with pair mass ratio $<1/3$) -bottom, for pairs in field (black), group (blue) and cluster (red) environments.
4.9 Top: primaries and secondaries for minor mergers in field environments. Bottom: primaries and secondaries for minor mergers in group/cluster environments.

4.10 Fraction of Transition, LINER and Seyfert galaxies in four separation bins. Top Left: All masses and environments. Top Right: Restricted to high mass galaxies. Bottom left: High mass galaxies in field environments. Bottom right: High mass galaxies in group environments.

4.11 Fraction of Transition, LINER and Seyfert galaxies in four separation bins. Major merger sample (top) and minor merger sample (bottom).

4.12 Fraction of Transition, LINER and Seyfert galaxies in four separation bins for major mergers in field and group/cluster environments.

5.1 Left: Support vectors are the data points located nearest to the hyperplane. Right: Non-linearly separable data points can often be mapped to a higher dimensional space where they can be separated by a linear hyperplane.

5.2 Decision tree to classify whether one should or should not play outdoor sports depending on the weather. In this example the weather conditions are the feature instances.

5.3 Light curve showing magnitude as a function of modified Julian date (MJD) for the object with ra = 35.18956671 and dec = -4.85610289 in all five optical bands (g, r, i, z, y).

5.4 \(r\)-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed QSOs.

5.5 \(r\)-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed QSOs. The top object is a narrow absorption line QSO.

5.6 \(r\)-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed narrow absorption line QSOs.

5.7 \(r\)-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed A-type stars.

5.8 \(r\)-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed B-type stars.

5.9 \(r\)-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed F-type stars.
5.10 $r$-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed G-type stars.

5.11 $r$-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed K-type stars.

5.12 $r$-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed M-type stars.

5.13 Invariance of modified $g$-band features between MD01 and MD09 for Eta (top), Cusum (middle) and Stetson $K$ (bottom).

5.14 Invariance of modified $g$-band features between MD01 and MD09 for Stetson $K$ (SAC) (top), Eta(SAC) (middle) and $\sum SAC > e^{-1}$ (bottom).

6.1 Colour-magnitude and colour-colour plots for the full sample with known stars and QSOs over-plotted.

6.2 Bar plot of the absolute value of SVM kernel model weights for $g$-band features. The highest SVM kernel weight is indicative of the success of the feature during the classification modelling.

6.3 Normalised counts of various $g$-band feature distributions for Stars (blue) and QSOs (red) in the training set. The top six features and bottom two features from the SVM kernel weight bar plot are shown.

6.4 Normalised count of the $g$-band Skewness distribution for Stars (blue) and QSOs (red) in the training set with higher resolution binning.

6.5 Various $r$-band features are plotted for Stars and QSOs in the training set.

6.6 More two dimensional feature plots. Various $r$-band features are plotted for Stars (blue) and QSOs (red) in the training set.

6.7 More two dimensional feature plots. Various $r$-band features are plotted for Stars (blue) and QSOs (red) in the training set.

6.8 Normalised histogram of RF Model I outputted confidences for QSOs (left) and stars (right) for predicted objects from the known (spectroscopic) sample (bold) and predicted objects from the unknown sample (dotted).

6.9 Histogram and table showing RF Model I outputted confidences for QSO predictions from the full unknown sample.

6.10 Blind test results. Normalised histogram of RF Model I outputted confidence for the QSOs (left) and stars (right) for spectroscopically confirmed and falsely predicted objects.

6.11 QSO predictions summary for each RF model.
6.12 The top two plots show $g$-band magnitude and colour-magnitude plots for QSOs predicted by RF Model III but not RF Models I or II. Bottom: Normalised histogram of confidences for QSO predictions. .......................... 181

6.13 RF $u - g$ and $g - r$ colour-colour plots. .......................... 182

6.14 RF $g$-band magnitude and $u - g$ colour-magnitude plots. .......................... 183

6.15 These plots show QSOs predicted by RF Model I but not RF Model III. .......................... 184

6.16 Normalised histogram of SVM Model I outputted confidences for QSOs (left) and stars (right) for predicted objects from the known and unknown samples. .......................................................... 187

6.17 Histogram and table showing SVM Model I outputted confidences for QSO predictions from the full unknown sample. .......................................................... 187

6.18 Blind test results. Normalised histogram of SVM Model I outputted confidence for the QSOs (left) and stars (right) for spectroscopically confirmed objects and falsely predicted objects. .......................................................... 189

6.19 QSO predictions summary for each SVM model. .......................................................... 190

6.20 The top two plots show $g$-band magnitude and colour-magnitude plots for QSOs predicted by SVM Model III but not SVM Models I or II. Bottom: Normalised histogram of confidences for QSO predictions. .......................................................... 191

6.21 SVM $u - g$ and $g - r$ colour-colour plots. .......................................................... 192

6.22 SVM $g$-band magnitude and $u - g$ colour-magnitude plots. .......................................................... 193

6.23 These plots show QSOs predicted by SVM Model I but not SVM Model III. 195

6.24 RF Model I confidences (top) and SVM Model I (bottom) for various RF and SVM Confidence Q cutoffs. .......................................................... 196

6.25 Colour-colour (top) and colour-magnitude (bottom) plots for predicted QSOs with RF Model I and SVM Model I Q confidences $\geq 0.7$ in green and Q confidences $\geq 0.9$ in red. .......................................................... 197
Publications and Conference Appearances

- “Star Formation and AGN Activity in Interacting Galaxies: A Near-UV Perspective”,

- “Star Formation and AGN Activity in Major and Minor Mergers”,

- “QSO Selection Model Using Light Curve Variability Features from Pan-STARRS”,

- “Mauna Kea Observatories, Hawaii”,

- “Identifying Quasars from their Variability Features”,
  Caroline Scott, Post-Graduate Research Physics Symposium, Imperial College London, U.K., 2013, presentation

- “Star Formation and AGN Activity in Interacting Galaxies: A Near-UV Perspective”,
  Caroline Scott, Interacting Galaxies and Binary Quasars Meeting, International Centre for Theoretical Physics, Trieste, Italy, 2012, presentation

- “Star Formation and AGN Activity in Massive Interacting Galaxies”,
  Caroline Scott, International Astronomical Union General Assembly XXVIII, Beijing, China, 2012, presentation
• “IAU Meeting, China”,  
  Article written by Caroline Scott for Astronomy Wise Magazine, October 2012,  

• “Let’s Talk... Caroline Scott Interview”,  
  blogspot.com/2012/07/astronomy-wise-e-zine-is-out-now-july.html

• “Studying Close Pairs in the Near-UV”,  
  Caroline Scott & Sugata Kaviraj, National Astronomy Meeting, Manchester,  
  UK, 2012, poster

Declaration and Copyright

This thesis is my own work, except where explicitly indicated in the text.

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Caroline Scott
19 February 2015
- To all family, friends and colleagues who have been a continuous source of support and inspiration during the process of writing this thesis -

“Willst du ins Unendliche schreiten, Geh nur im Endlichen nach allen Seiten.”
(Do you want to stride into the infinite? Then explore the finite in all directions.)
- Johann Wolfgang von Goethe

Sunset from Mauna Kea Observatories, Hawaii, during an observing trip at the James Clerk Maxwell Telescope in January 2014.
Foreword

The last century has born witness to some of the greatest revelations about our physical surroundings in history. Technological innovations have allowed scientists to probe microscopic to macroscopic scales in greater detail than ever before. Discoveries in particle physics are leading to a more detailed understanding of the building blocks of our natural world. Currently, advancing research into the Higgs-particle endeavours to uncover the nature of mass and the force of gravity. Astronomical explorations are allowing us to translate these fundamental laws to describe physical conditions and processes in other planets, stars and galaxies. Even the nature of the Universe itself can be explored using data from a number of modern telescopes that have been specifically designed for cosmological studies. Such studies seek to explain the formation of large-scale structures that now harbour galaxies, to uncover the ratio of ordinary to unknown/exotic matter, and to probe the nature of the mysterious force known as dark energy that drives the accelerated expansion of the Universe.

When two or more galaxies merge we are granted a rare glimpse into the act of evolution within the Universe. The strength of gravitational forces present during a merger can induce changes in the morphology, star formation rate, central black hole mass and internal galactic processes of progenitors. In the early Universe, the galaxy population would have been very different to how it appears now, almost 14 billion years later. We endeavour to observe galaxies at different stages in this merging act and to understand the role that mergers play in galaxy evolution. We also aim to study the internal mechanisms of active galaxies, to understand their role in galaxy evolution and to determine whether interactions with other galaxies may trigger the process of black hole accretion in active galaxies. The aim of this thesis is to contribute to these areas of research by providing new insight and analyses of active and merging galaxies. To facilitate this, we use data from a range of new-generation astronomical instruments.

The research presented in this thesis was conducted between Imperial College London (my home institution), the Harvard-Smithsonian Center for Astrophysics
(where I was granted a Smithsonian Astrophysical Observatory Predoctoral Fellowship between April 2012 and April 2014) and Harvard’s Institute for Applied Computational Science (where I was granted a Research Fellowship between December 2012 and April 2014). At Imperial College, I was supervised by Sugata Kaviraj and Stephen Warren; we worked within the field of galaxy mergers (this work is detailed in Chapters 3 and 4). At Harvard I was supervised by Pavlos Protopapas and Paul Green; we applied contemporary computer science methods to classify active galaxies (this work is detailed in Chapters 5 and 6).

This thesis begins by introducing the reader to underlying topics and essential terminology in the field of galaxy evolution. We then introduce the tools and techniques used to obtain and prepare data for our research in Chapter 2. Original research is presented in Chapters 3 to 6. To conclude, we summarise and discuss this work in Chapter 7.
Chapter 1

Introduction

Overview

In this introductory chapter we present a review of the standard paradigm of structure formation in the framework of the ΛCDM cosmological model. We discuss the prominent features of this model and give examples of its successes and failures when tested with observational data. We describe the two main competing models that have dominated the field of galaxy evolution in recent decades; monolithic collapse and hierarchical formation through galaxy mergers. Both models are evaluated and we explain why the hierarchical formation model is more convincing within a ΛCDM framework and in better agreement with observations. As new studies are presented using increasingly reliable observational data and powerful computer simulations, many refinements have been made to the hierarchical model of galaxy evolution. We summarise and evaluate the results of many relevant studies here. A precise parametrisation of the merger rate with redshift can provide meaningful constraints for modelling hierarchical evolution; various observational studies that have estimated the merger rate are reviewed.

We familiarise the reader with nomenclature in this field of research by introducing terms such as dark matter and dark energy. Various galaxy types are described that were initially categorised morphologically by Edwin Hubble; these include elliptical, spiral, lenticular and irregular galaxies. Active galaxies and potential mechanisms of black hole accretion in active galaxies are introduced. Observations of photometric variability in active galaxies are discussed, and theoretical propositions that this variability could arise from accretion disk instabilities are evaluated.
1.1 The Standard Paradigm for Structure Formation

Constituents of the Universe

The standard paradigm states that the Universe was initially radiation dominated; i.e. the radiation density was higher than the matter density and Cosmological constant density. However, expansion over time led to cooling and the radiation density decreased rapidly ($\propto a^{-4}$; where $a(t)$ is the scale factor\textsuperscript{1}) and the Universe shifted into its current matter-dominated era\textsuperscript{2}. This matter is present in two fundamental forms; baryonic and non-baryonic. The word baryon is derived from the Greek word ‘heavy’, and baryonic matter typically describes heavy particles; protons, neutrons, and all atoms and ordinary matter constructed from these particles.

The term ‘dark matter’ was initially used by Fritz Zwicky (1933) when he applied the Virial Theorem to the Coma galaxy cluster and found that the expected total galaxy mass was much higher than that implied by their observed luminosity. Jan Oort (1932) had already observed that stellar motions within the Milky Way implied that the galactic plane mass should be higher than that inferred from observed matter. Nearly fifty years later, dark matter was proposed to explain rotation curve measurements of stars within spiral galaxies that were inconsistent with standard Newtonian dynamics; observable matter alone is not sufficient to explain the surprisingly high velocity of distant objects in orbit when confined to standard Newtonian gravity (Rubin, Ford & Thonnard, 1980). The true nature of dark matter is currently unknown, but it is hypothesised to consist of mostly exotic and yet-to-be-discovered non-baryonic matter that does not absorb or emit detectable amounts of electromagnetic radiation. Baryonic matter could potentially constitute some of the dark matter fraction in the form of massive compact halo objects that emit undetectable levels of electromagnetic radiation; gravitational microlensing surveys have been utilised to search for evidence of this form of matter (Paczynski, 1986; Griest, 1991; Ricotti & Gould, 2009; Iocco et al., 2011).

\textsuperscript{1}The scale factor describes expansion in a homogeneous and isotropic Universe. The expansion rate, or \textit{Hubble parameter}, $\dot{a}/a$ is an expression frequently encountered in Cosmology.

\textsuperscript{2}This shift to a matter dominated Universe occurred as the matter density decreases less rapidly than the radiation density ($\rho_{\text{radiation}} \propto a^{-4}$ whereas $\rho_{\text{matter}} \propto a^{-3}$); physically this reflects a loss in photon energy due to redshifting, leading to particles becoming non-relativistic and constituting ordinary matter.
The Standard Paradigm for Structure Formation

ΛCDM

The ΛCDM model is currently the most widely-accepted cosmological model stemming from the Big Bang theory (e.g. Liddle, 2003; Rebolo et al., 2004; Navarro et al., 2010; Komatsu et al., 2011; Dunkley et al., 2011; Hinshaw et al., 2013; Planck Collaboration, 2013). It balances simplicity with impressive accuracy at matching many observed properties of the cosmos such as accelerated expansion and the large-scale structure distribution. This model describes a homogeneous and isotropic Universe which features a cold dark matter component and a cosmological constant term, Λ (as predicted from Einstein’s equations of general relativity). The cosmological constant has negative pressure and is believed to describe the accelerating expansion whereby galaxies are generally found to be moving away from each other and the Universe appears to be expanding in all directions. Cosmological expansion was predicted theoretically by Georges Lemaître in 1927 as a consequence of general relativity, and was shortly after observed by Edwin Hubble by measuring the recessional velocities of nearby objects (Lemaître, 1927; Hubble, 1929). The hypothetical force driving accelerated expansion has been termed ‘dark energy’, and although the nature of this force is not yet understood, dark energy is routinely included in most cosmological models. It is currently postulated that the Universe comprises ∼4.9% baryonic matter, ∼26.8% dark matter and ∼68.3% dark energy (Planck Collaboration et al., 2013b).

Evaluating ΛCDM

Over the last decade, observational data from instruments such as the Wilkinson Microwave Anisotropy Probe (WMAP) and Planck have helped to corroborate various aspects of the ΛCDM model through studies of anisotropies in the cosmic microwave background (CMB; see explanation below). Detailed supernovae measurements indicate an accelerated expansion in the Universe consistent with ΛCDM (Paal, Horvath & Lukacs, 1992; Riess et al., 1998; Knop et al., 2003; Wang & Tegmark, 2005; Astier et al., 2006; Blake et al., 2011). However, many gaps exist within this standard model. There are a number of observations for which ΛCDM does not offer explanations by itself, and extensions are frequently proposed to explain observed properties that ΛCDM does not predict directly.

For example, the very early Universe is thought to have undergone a period of momentous inflation which led to the flat, homogeneous and isotropic properties that we observe. Inflation offers an explanation for an issue that stems from ΛCDM known as the ‘horizon problem’, where isolated regions of the Universe that have no
apparent causal connection seem to have evolved as if causally connected with regards to consistent temperature and curvature. Inflation is thought to have carried quantum fluctuations to outside of the Hubble radius \( R = c/H \), for Hubble parameter, \( H \), and speed of light, \( c \), leading to isotropy. Various models of inflation have been proposed, and this is currently an active area of research in cosmology. Also additional to \( \Lambda \)CDM, baryogenesis models seek to explain how the matter fraction came to outweigh that of the anti-matter fraction in the early Universe (since the Big Bang is predicted to have produced equal matter/anti-matter fractions), allowing pockets of matter to form and the observed structure of the cosmos to emerge.

The main contender to \( \Lambda \)CDM is Modified Newtonian Dynamics (MOND), proposed by Mordehai Milgrom (Milgrom, 1983). This seeks to explain the puzzle of rotation curve measurements by modifying aspects of Newtonian gravity at small accelerations, such that objects at large radius in gravitational orbits have a larger velocity. The dark matter versus MOND debate is ongoing and is an active area of research for cosmologists and particle physicists (Bertone, Hooper & Silk, 2005; Kroupa et al., 2010; Famaey & McGaugh, 2012; Milgrom, 2013); MOND is beyond the scope of this thesis and we assume a \( \Lambda \)CDM cosmological model throughout.

**Analysing the Seeds of Structure Formation**

Analysis of the CMB is particularly useful for studying the evolution of large-scale structure in the Universe, as the relics of structure over-densities and under-densities are encoded in light that has been travelling since the era of decoupling. Prior to this time, the Universe was hot, dense and opaque due to the absorption of electromagnetic radiation by hydrogen plasma. Recombination describes the transition period where cooling due to expansion allowed radiation and plasma to cool enough for neutral hydrogen and helium to form, and thus for structure to emerge. This is estimated to have occurred \( \sim 400,000 \) years after the big bang. Recombination led to the decoupling of photons that were no longer scattered by electrons in the hot dense plasma of protons and electrons, and the light from this decoupling period is still reaching us from all directions. This light has now been redshifted from the microwave to the radio band.

The CMB power spectrum has been mapped in detail by COBE (Fixsen et al., 1996; Jaffe et al., 2001), WMAP (Spergel et al., 2003; Bonaldi et al., 2007) and most recently Planck (Planck Collaboration et al., 2013a). It is observed to have a black body spectrum with incredible precision. However, tiny fluctuations quantified in the power spectrum reveal slight under-densities and over-densities at the
time of decoupling. $Λ$CDM predicts that primordial density fluctuations should be isotropic and Gaussian distributed, and therefore much about the validity of this model can be learned from studying CMB anisotropies (Eriksen et al., 2004; Emir Gümrukçuoglu, Contaldi & Peloso, 2007; Planck Collaboration et al., 2013c).

With the latest Planck data, anisotropies in the CMB are measured with incredible precision. These observations must be satisfied by all credible cosmological models, enabling tight constraints to be placed on inflationary and structure formation models.

Primordial density fluctuations are predicted to have been amplified during inflation shortly after the Big Bang and to have undergone continuous contraction/expansion thereafter into regions of dark matter over-densities/under-densities. The primordial density fluctuations can be traced by minute temperature fluctuations, and are estimated to have varied by an order of $10^{-5}$ at recombination. These density perturbations are thought to have continually evolved under their own gravitational instability, causing regions with deep potential wells to be inhabited by concentrated baryonic matter; this evolution could then have led to structure formation. By ‘structure’, we mean condensed matter on all scales from galaxies to galaxy clusters and superclusters (also known as ‘galaxy filaments’; these are the largest structures presently known) as well as the extended void regions between.

‘Jeans instability’ occurs when the force of gravitational collapse is stronger than the outward pressure from gas or radiation. Since dark matter does not interact with radiation, the outward radiation pressure does not oppose dark matter from flowing freely into gravitationally dense regions, leading to gravitational clustering and forming a cosmic web of dark matter halos. Once this network of structure was established galaxies would then have formed hierarchically by the merging of dark matter halos (White & Rees, 1978; White & Frenk, 1991; Kauffmann, White & Guiderdoni, 1993; Parkinson, Cole & Helly, 2008). It is important to understand the structure of dark matter clustering, as this determines the spatial distribution of galaxies and therefore affects how galaxies interact and evolve. The distribution of galaxies has been mapped by surveys including the 2dF Galaxy Redshift Survey (Magliocchetti & Porciani, 2003; Cole et al., 2005; Ribeiro et al., 2009). The Millennium Simulation (Boylan-Kolchin et al., 2009) was a three-dimensional N-body simulation using more than $10^{10}$ particles to trace dark matter structure evolution to its present state. It provided a mostly convincing re-enactment of structure formation (e.g. Fakhouri, Ma & Boylan-Kolchin, 2010). However, it predicted more small-scale dark matter sub-halos than there is observational evidence for; we might expect to see objects such as dwarf galaxies and globular clusters inhabiting these
regions.

Protogalactic clouds of concentrated baryonic matter in the central regions of
dark matter halos would then form and cool to below the virial temperature, finally
collapsing under their own gravity to form the very first stars (Silk, 1977; Abel,
Bryan & Norman, 2002; Bromm et al., 2009), which are known as ‘population III’
stars. Through a series of mergers, the first galaxies could then begin to grow
and evolve morphologically. Some galaxies have been observed using Hubble Ultra
Deep Field Imaging dating back to only 600 million years after the Big Bang (e.g.
Bouwens et al., 2010), allowing us to place constraints on the era when the first
galaxies formed. It will be shown in Section 1.2 that many morphological types of
galaxies now exist.

1.1.1 Monolithic Collapse

The monolithic collapse model was proposed by Eggen, Lynden-Bell & Sandage
(1962). It is contradicted by many recent observations and is generally neglected in
favour of the hierarchical formation model (see Section 1.1.2), however, we illustrate
it here due to its historical significance. Much work in the past has been devoted to
modelling and simulating galaxy formation from a monolithic collapse perspective,
and thus it plays a significant role in how research in the field of galaxy evolution
has progressed.

The monolithic collapse model suggests that galaxies are formed by the collapse
of massive gas clouds under their own gravity in a short and efficient burst. In this
model, early-type galaxies formed at high redshift with little structural evolution
since. These are often referred to as ‘red-and-dead’ galaxies; all stars are thought to
have formed in a single starburst and then evolved passively from then on (Eggen,
suggest that these initial starbursts must last less than 1 Gyr to reproduce the
stellar populations observed in elliptical galaxies. Indeed, Kriek et al. (2008) found
that \( \sim 45\% \) of \( K \)-bright massive galaxies already have evolved stellar populations by
\( z \sim 2.3 \) with little ongoing star formation. This model offers a natural explanation
for observations of cosmic downsizing (see Section 1.1.3). It is also consistent with
the small scatter and lack of evolution with redshift observed in the fundamental
plane for early-type galaxies (see Section 1.2.1).

It was shown by Kauffmann, Charlot & White (1996) using the colours of 125
local galaxies (from the Canada-France Redshift Survey) that only \( \sim 1/3 \) of local el-
liptical and SO galaxies could have been assembled and contained passively evolving
stellar populations by $z = 1$. Kauffmann et al. concluded that other processes must play a role in their evolution, such as morphological disturbances and recent star formation caused by mergers. Van Dokkum et al. (2008) observed that the quiescent early type galaxies in the sample used by Kriek et al. (2008) are very compact (with median effective radius $r_e = 0.9$ kpc) compared to galaxies of similar mass in the nearby Universe (which have sizes $\sim 5$ kpc) and that fully assembled early-type galaxies only comprise at most $\sim 10\%$ of $K$-selected quiescent galaxies at $z \sim 2.3$. This shows that considerable evolution must have taken place after $z \sim 2.3$ (through dry mergers or other processes) and provides strong evidence against the monolithic collapse model.

Monolithic collapse also fails to satisfy other observational properties; the rotation of the protogalactic cloud would imply that stars would move in elliptical orbits in the same direction (a trend which is not observed), and we would expect globular clusters to have formed at the same time within a narrow time frame (however, we observe a broad range in globular cluster ages). If monolithic collapse were to be considered as a viable model, elements of hierarchical evolution would still have to be considered to account for these observational discrepancies.

### 1.1.2 Hierarchical Evolution of Galaxies through Mergers

The hierarchical model of galaxy formation attributes the evolution of galaxies to the process of repeated mergers between smaller galaxies (e.g. Toomre & Toomre, 1972; Fall & Efstathiou, 1980; Kauffmann, White & Guiderdoni, 1993; Cole et al., 2000; Steinmetz & Navarro, 2002; Jiang et al., 2008; Brook et al., 2012; Shankar et al., 2013; Wetzel et al., 2013). Because of this sequential evolution from smaller to larger galaxies, the standard paradigm for galaxy evolution is known as a ‘bottom-up’ model.

Most galaxies with substantial spheroidal components are thought to harbour supermassive black holes at their centre; these are expected to be linked with properties of the host galaxy (Kormendy & Richstone, 1995; Magorrian et al., 1998; Di Matteo, Springel & Hernquist, 2005; Ishibashi & Fabian, 2014). When a merger takes place it is thought that the surrounding dark matter halos merge and the gas cools and condenses, forming a rotating disk at the halo’s centre in which star formation begins to take place (Somerville & Primack, 1999; Kauffmann & Haehnelt, 2000; Hatton et al., 2003).

As the galaxies merge some of this gas is funneled into the black hole nuclei of the progenitor galaxies and the rest is thought to be used up in starbursts. A
tight correlation can be seen between supermassive black hole mass and velocity dispersion in the galactic bulge, adding confidence to the idea that there is a strong link between spheroid formation and black hole growth (Richstone et al., 1998; Cattaneo, Haehnelt & Rees, 1999; Kauffmann & Haehnelt, 2000; Monaco, Salucci & Danese, 2000; Cavaliere & Vittorini, 2000; Gebhardt et al., 2000; Haehnelt & Kauffmann, 2000; Hu, 2008). During the final stages of a merger, the supermassive black holes orbit and finally merge (Begelman, Blandford & Rees, 1980; Debuhr, Quataert & Ma, 2011). The small percentage of the remaining gas is accreted onto the new black hole (this is thought to occur over a timescale of order $10^7$ years) with the rest of the gas being transformed into stars (e.g. Kauffmann & Haehnelt, 2000). General relativity predicts that gravitational waves are emitted during the coalescence of super massive black holes, and so final stage mergers provide a potential observational test for general relativity.

Disk galaxies are thought to be formed through the acquisition of angular momentum via tidal torques in interacting dark matter halos (Silk, 2003; Dekel, Sari & Ceverino, 2009; Agertz, Teyssier & Moore, 2011). Gravitational contraction leads to rotational support as angular momentum is conserved, and baryonic cooling within dense regions of the resulting disk leads to star formation. These low mass, spiral-shaped galaxies are then predicted to undergo a succession of mergers that result in higher mass galaxies, eventually forming massive elliptical-shaped galaxies (Toomre, 1978; Schweizer, 1982; Wright et al., 1990; Bournaud, Jog & Combes, 2007). After conducting three-dimensional simulations between gravitationally interacting bodies (each containing $\sim 10^4$ particles), Barnes (1988) found that the observed morphological parameters and intrinsic properties of galaxies could be reproduced by this model of merging disk galaxies. In this work, Barnes modified merger simulations originally conducted by Toomre & Toomre (1972) to include dark matter halos in the merger scenario. Even without the inclusion of dark matter halos, and with much more basic simulations (between only two bodies), Toomre and Toomre came to the same conclusion: galaxy mergers explain many of the properties that we see in neighbouring galaxies, such as tidal tails and bridges.

Kauffmann & Charlot (1998) present a model of elliptical galaxy formation in which the majority of stars are thought to have formed in disk galaxies that then go through a series of mergers to form ellipticals. In this work they assume that supernovae explosions allow the transfer of metals between the stars, the cold gas and the hot gas halo components (i.e. it is a non-closed box model). The semi-analytic models they adopted imply that the inter-cluster medium between elliptical galaxies in cluster environments will have seen very little evolution since $z < 1$. This
1.1 The Standard Paradigm for Structure Formation

is because more than 80% of metals belonging to galaxies with circular velocities
\( < 250 \text{km s}^{-1} \) will have been ejected through supernova explosions for \( z > 1 \). In
this model, bright elliptical galaxies are thought to have formed from the merging
of massive disk galaxies, whereas faint ellipticals form from lower mass mergers.
Since mergers inherit most of the independent stellar populations of the progenitor
galaxies as well as new stars that are formed throughout the merger, modelling the
resulting stellar population of a merger is a difficult task.

A local example of a galaxy merger is Centaurus A (see Figure 1.1); the fifth
brightest galaxy observed from Earth. Although the subject of debate, this appears
to be an elliptical galaxy in the process of accreting a spiral galaxy (Tubbs, 1980).
Active star formation and black hole accretion are seen in Centaurus A. Our own
galaxy, the Milky Way, and the spiral galaxy Andromeda (which is located approxi-
mately 2.5 million light years from Earth) are currently approaching each other and
are expected to merge within 5 billion years (Cox & Loeb, 2008).

Modelling Hierarchical Growth

Despite impressive advances in computing, N-body simulations are still unable to
fully reproduce the dynamical evolution of galaxies since many physical processes
are not fully understood. Accurate star formation and dust models are required,
as well as models describing the accretion of gas and dust onto central black holes.
Semi-analytic processes adopting Monte Carlo simulations can be used to simulate
1.1 The Standard Paradigm for Structure Formation

hierarchical structure growth (see Section 1.3.4) by assuming cosmological density perturbation theory and using parameters such as radiative cooling, star formation, supernova feedback, the stellar initial mass function, metallicity, dust extinction, stellar winds, and merging rates of galaxies (Baugh et al., 1998; Somerville & Primack, 1999; Kauffmann & Haehnelt, 2000; Hatton et al., 2003; Dubois et al., 2012, etc.). When using Monte Carlo methods, processes must be simplified; spherical symmetry and certain flow properties must be assumed (e.g. Cole et al., 2000). As well as permitting cosmological modelling for large numbers of galaxies, semi-analytic simulations are very useful for testing feasible models within galaxies; such as modelling the evolution between an active galactic nucleus, black hole growth, bulge formation and star formation (see Section 1.3.3).

N-body simulations offer an alternative to Monte Carlo simulations; however, such methods are computationally heavy, resolution limited, and extremely time consuming (e.g. Heggie & Hut, 2003). Steinmetz & Navarro (2002) used N-body methods to simulate the formation and evolution of a galaxy population and concluded that hierarchical growth is likely to play a significant role in galaxy evolution.

1.1.3 Downsizing: An Issue with the Hierarchical Formation Model?

Using a nearly complete sample of 393 Keck spectroscopically observed galaxies, Cowie et al. (1996) found that the most massive and luminous galaxies appear to have already formed and ceased star formation at high redshift; whereas low mass galaxies have prolonged star formation that is still ongoing. They called this trend ‘downsizing’. Bower, Lucey & Ellis (1992) found that \( \leq 10\% \) of the stellar population in early-type and SO galaxies was formed in starbursts in the last 5 Gyr. Ellis et al. (1997) built upon the work of Bower et al. and showed that the bulk of the stellar population in dominant spheroidal galaxies in clusters would have formed before \( z \approx 3 \). Terlevich, López & Terlevich (2007) used a sample of local HII galaxies, which typically have low mass and intense star formation, to investigate downsizing within low mass galaxies. They found that the lowest mass systems are generally younger with lower metallicity than the more massive ones and hence downsizing can even be seen within low mass distributions of galaxies.

This presents an apparent contradiction to hierarchical formation since this model predicts that the formation of massive galaxies should have taken place after that of lower mass galaxies, thus we might expect to see more recent star formation in massive galaxies. However, Neistein, van den Bosch & Dekel (2006) commented
that the downsizing effect is not necessarily contradictory to a hierarchical model of halo clustering if cooling and baryonic feedback effects are included in merger models. Their simulations show that with certain parameterisations of these processes, efficient star formation can be extended in low mass galaxies, but quenched in high mass galaxies. Faber et al. (2007) show that downsizing is likely to be the result of various processes that quench star formation in blue galaxies, causing them to migrate into the red sequence; this would offer an explanation for the observed increase in red sequence galaxies since $z \sim 1$. Examples of quenching processes include the massive halo quenching model (Dekel & Birnboim, 2006; Cattaneo et al., 2006), satellite quenching (Faber et al., 2007) and active galactic nuclei feedback (see Section 1.3.3).

Stringer et al. (2009) used the Millennium Simulation as a basis for mock observations of halo clustering, then used semi-analytic methods to simulate the evolution of a population of galaxies. Utilising the radio-mode feedback available in GALFORM (see Section 1.3), they were able to reproduce the effects of cosmic downsizing within a hierarchical scenario. However, their model led to the over-excessive quenching of star formation for intermediate mass galaxies, and failed to reproduce the observed colour distribution of galaxies for their full redshift range ($0.4 < z < 1.4$).

Observations of cosmic downsizing show that star formation in massive galaxies at early times must have been more efficient than it is now. To be fully accepted, hierarchical formation models must be further developed in order to provide an explanation for these effects.

### 1.2 The Hubble Sequence and Hubble Types

In *The Realm of the Nebulae*, (1936), Edwin Hubble proposed a morphological classification scheme; this is illustrated in the well-known Hubble *tuning fork* in Figure 1.2. The Hubble classification sequence uses the morphological properties of galaxies to classify them as either elliptical, spiral, lenticular or irregular. We now look at the properties of each Hubble type.

#### 1.2.1 Elliptical

Defined by their ellipsoidal shape (see Figure 1.3, top left), these galaxies have approximately elliptical isophotes\(^3\). Elliptical galaxies have the property $0.3 \lesssim \epsilon \leq 1$.

\(^3\)Isophotes are lines of constant surface brightness. It has been estimated that only one third of elliptical galaxies have perfectly elliptical isophotes; another third are thought to have box-like
where $\epsilon = \frac{b}{a}$ and $\frac{b}{a}$ is the ratio of semi-minor to semi-major axes. Elliptical galaxies are assigned a Hubble type according to $10 \times (1 - \epsilon)$, thus the Hubble type for elliptical galaxies ranges from $E0$ (with circular isophotes) to $E7$ (with axis ratio 0.3). An axis ratio $\lesssim 0.3$ has never been observed; this is believed to be due to the Firehose instability. The Firehose instability arises because for $\epsilon \lesssim 0.3$ (an approximate axis ratio 1:3) the system is susceptible to bending in the direction perpendicular to the elongated axis which results in the elongated axis ratio becoming shorter; i.e. the galaxy becomes rounder (Hernquist, Heyl & Spergel, 1993; Jessop, Duncan & Levison, 1997). Elliptical galaxies generally have a de Vaucouleurs brightness profile given as follows:

$$I_E(r) = I_e e^{-7.67((r/r_e)^{1/4} - 1)}.$$  

(1.1)

Elliptical galaxies are the most massive galaxy types and have a high escape velocity. Gas is needed to fuel star formation and elliptical galaxies have a smaller gas content than other galaxy types. Using interferometric $^{12}$CO(1-0) observations, Davis et al. (2013) finds that, although elliptical galaxies show less molecular gas than spiral galaxies, the gas content is dependent on the environment of the elliptical, with ellipticals in higher density environments showing less gas and implying a different path of evolution from field ellipticals. When supernovae explode, remaining interstellar gas can be heated and ejected from the galaxy by resulting galactic winds (Mathews & Baker, 1971; Larson, 1974; Loewenstein, 2013). This effect is

\footnote{The Firehose instability is also thought to aid the formation of bulges in barred spiral galaxies by \textit{puffing up} the elongated bar component into a bulge.}
1.2 The Hubble Sequence and Hubble Types

Figure 1.3: Top left: M87 elliptical galaxy. Situated near the centre of the Virgo cluster, M87 boasts a high number of globular clusters (∼10,000); Canada-France-Hawaii Telescope, J.C. Cuillandre, Coleum. Top right: Lenticular galaxy NGC 5866; imaged by HST and released in the original NASA press release. Globular clusters inhabit the outer halo, each containing nearly ∼1,000,000 stars. Bottom left: M101 spiral galaxy, also known as the Pinwheel galaxy; composite image from 51 HST exposures and ground based images and released by NASA. It is situated in the Ursa Major constellation (25MLyrs from Earth) and spans nearly twice the diameter of the Milky Way. Bottom right: NGC 6822, also known as Barnard’s Galaxy is an irregular galaxy member of our Local Group in the constellation of Sagittarius. It is a small galaxy and has low surface brightness yet it contains relatively bright HII regions which have sparked a lot of interest in studying it.

http://apod.nasa.gov/apod/ap040616.html
http://www.spacetelescope.org/images/opo0624a/
http://hubblesite.org/newscenter/archive/releases/2006/10/image/a
dampened somewhat by radiative cooling but still has dramatic effects for lowering
gas and metallicity levels, especially in less massive galaxies where the gravitational
potential well is shallower.

Observations in optical wavebands indicate that little star formation takes place
in ellipticals (Baum, 1959; Sandage & Visvanathan, 1978; Bower, Lucey & Ellis,
1992). The ‘Fundamental Plane’ is a three-dimensional space with axes of velocity
dispersion, effective radius and effective surface brightness. Elliptical galaxies only
span (approximately) a two-dimensional subspace of the Fundamental Plane (Djor-
govski & Davis, 1987) and show little scatter (Jorgensen, Franx & Kjaergaard, 1996;
Cappellari et al., 2006; Graham, 2013); a lack of evolution with redshift is shown in
this scatter (Bower, Lucey & Ellis, 1992), implying that the bulk of star formation
took place by $z = 2$ (Peebles, 2002). This lack of observable recent star formation
led to massive elliptical galaxies being described as ‘early-type’ or ‘red and dead’.
However, more sensitive indicators of star formation show that a significant number
of ellipticals do in fact show ongoing star formation (e.g. Kaviraj et al., 2007; Ford
& Bregman, 2013). Kaviraj et al. (2007) used NUV-$r$ colours from $\sim$2100 early-
type galaxies from SDSS DR3 (crossmatched with GALEX measurements for NUV
magnitudes) and found that at least 30% show evidence of recent star formation to
a 95% confidence level.

For quite some time, observational evidence has implied that early-types have a
cold gas content which could be used to fuel star formation. Cold gas ($\leq$100K) has
been identified in early-types using HI 21cm emission line measurements (Knapp,
Turner & Cunniffe, 1985; Wardle & Knapp, 1986; Morganti et al., 2006; Serra et al.,
2012) and CO emission line measurements (Wiklind & Rydbeck, 1986; Phillips et al.,
1987; Knapp & Rupen, 1996; Young, 2005; Young et al., 2011). Since dust radiates
in the far-IR, the IRAS satellite has granted a new perspective by which to measure
interstellar dust and to map dust regions that could potentially fuel star formation.
Knapp et al. (1989) used an IRAS sample of $\sim$1150 early type galaxies and found a
significant interstellar dust content in a large fraction of early type galaxies. Row-
lands et al. (2012) compare the dust properties using submillimetre measurements
of elliptical and spiral galaxies and find that submillimetre-selected ellipticals can
show as much dust as typical spirals.

### 1.2.2 Spiral

Hubble classified spiral galaxies as either standard spirals (S) or barred spirals (SB);
he acknowledged that these classifications are not disjoint and a small proportion
of galaxies lie between between S and SB. Spiral galaxies are rotating disk galaxies characterised by a central bulge and spiral arms extending from the centre, or from the edge of the bar in the case of an SB galaxy (see Figure 1.3, bottom left).

The central bulge is generally a concentrated region of older stars (Graham, 2013). The spiral arms are thought to be caused by spiral gravitational density waves which are independent of the rotational motion of the stars and gas within the galaxy; they are visible as a result of orbiting matter being compressed in these regions when passing through (Lin & Shu, 1964; Kim & Kim, 2014). The increased density in spiral arms causes gas clouds to approach their Jeans limit; if this limit is surpassed the gas clouds will collapse to create starbursts. As a result spiral arms are often inhabited by young, OB-type stars which makes them brighter and bluer.

Spirals tend to have younger stellar populations than elliptical galaxies. When plotted on optical colour-magnitude diagrams spirals are usually found in the so-called blue cloud, with the bluer colours indicating that more star formation is taking place, whereas ellipticals tend to inhabit the red region (Strateva et al., 2001; Bell et al., 2003). This bimodal colour distribution provides a simple, yet not always reliable, way to distinguish between spiral and elliptical galaxies. Hubble’s classification system was modified by de Vaucouleurs (1959) with particular emphasis on categorising spiral galaxies to include more detailed features such as diffuse/broken spiral arms with a lack of bulge component (Sd, SBd), and highly irregular appearance with a lack of bulge component (Sm, SBm). In Hubble’s classification system these galaxies were grouped together as Irr galaxies. The light profile of the bulge in spirals is generally described by a de Vaucouleurs profile, with that of the disk following an exponential brightness profile given by

\[ I_S(r) = I_s e^{-(r/r_s)} \]  

1.2.3 Lenticular/SO

SO galaxies lie between the morphological classifications of elliptical and spiral galaxies and are also known as lenticular galaxies because of their lentil-like shape (see Figure 1.3, top right). They are disk galaxies with bulges and usually with unclear spiral arms, yet they generally have low star formation rates like elliptical galaxies because they have been stripped of most of their interstellar gas (Cappellari et al., 2006). Hubble later modified the SO classification to distinguish between non-barred lenticular, SO, galaxies and barred, SBO, lenticular galaxies. Hubble died before publishing some modifications to his classification system, however Allan Sandage, the successor to Hubble at the Mt. Wilson and Palomar Observatories, collected
Hubble’s notes and continued his work (de Vaucouleurs, 1959).

### 1.2.4 Irregular

Galaxies which did not fit into any of the three above morphological classifications were categorised by Hubble as *irregular* (Hunter & Elmegreen, 2006). These galaxies, which lack rotational symmetry and a dominating nucleus, account for \( \sim 2\text{-}3\% \) of the population and are usually star forming; often at a rate similar to that of spiral galaxies (Hunter, 1997). They are morphologically peculiar galaxies, for example the Magellanic Clouds (irregular dwarf galaxies), Messier 82 (although this was first classified as irregular from optical observations, spiral arms have since been detected in the near-Infrared) and NGC 6822 (see Figure 1.3, bottom right).

### 1.3 The Role of Active Galaxies in Evolutionary Models

#### 1.3.1 Active Galactic Nuclei

Active galactic nuclei (AGN) activity occurs in the central region of massive active galaxies, outputting radiation that traverses most of the electromagnetic spectrum (Lynden-Bell, 1969). This substantial emission is thought to be triggered by the accretion of gas and dust onto a central supermassive black hole. An accretion disk forms from the in-falling interstellar material and generates a range of extreme physical processes surrounding the galaxy nucleus (Rees, 1984; Lin & Papaloizou, 1996; Ulrich, Maraschi & Urry, 1997; Mirabel & Rodríguez, 1999; Kembhavi & Narlikar, 1999; Sadowski et al., 2013; Suzuki & Inutsuka, 2014). During the accretion process, gravitational potential energy from in-falling material is converted into kinetic energy which can very efficiently be transformed into heat and radiated away due to friction within the accretion disk. The origin of AGN activity is not yet understood, but it could potentially be triggered by tidal interactions or mergers with other galaxies; this is a question that we address in Chapter 4.

Photometric and variability properties are distinguished from non-active galaxies, where the majority of emitted light comes from stellar or nebular activity. Quasars (quasi-stellar radio sources) and their radio-quiet companions QSOs (quasi-stellar objects) are the most luminous types of AGN. They have broad emission lines and an extremely high luminosity emanating from the compact galactic nucleus, allowing them to be detected at high redshift as point-sources (Villata et al., 2006;
1.3 The Role of Active Galaxies in Evolutionary Models

Mortlock et al., 2011). The accretion disk is generally surrounded by a dusty torus and optically thick plasma that can obscure our view of the active nucleus. In some radio-loud quasars, two radio lobes can be seen extending roughly symmetrically from quasars (de Vries, Becker & White, 2006); these are often connected by outflowing jets of relativistic particles that are thought to provide a path for energy transfer between the compact central core and radio lobes (see Figure 1.4). Together, these radio components have been observed to span up to $\sim 1$ Mpc. Optical jets were first observed extending from the active elliptical galaxy M87 (located in the Virgo cluster) by Curtis (1918) (Mirabel & Rodríguez, 1999).

![Figure 1.4: Very Large Array (4.9GHz) image of the radio emission from the quasar 3C 175 at $z = 0.77$ (Bridle et al., 1994). The radio emission spans nearly 200 kpc. Notice that the radio jets extend from a compact source, the galaxy center, located at $(\Delta \alpha, \Delta \delta) = (0,0)$.](image)

The lower redshift, lower luminosity analogs of Quasars and QSOs are the Seyfert Type I galaxies. Still characterised as active nuclei with high luminosity and broad emission lines, Seyferts tend to have host galaxies resolvable by optical telescopes. Type II QSOs and Seyfert Type II galaxies are similar but show only narrow emission lines. Other types of AGN have been identified; such as BL Lacs (named because the galaxy BL Lacertae is a prototypical example), radio galaxies (radio-bright elliptical galaxies), and blazars (compact quasars where the relativistic jets are directed at the observer). Orientation-based unification schemes have attempted to group different types of AGN into one standard model where we attribute various properties (such as luminosity, emission-width etc.) to the angle...
from which we are viewing a single type of AGN (see Figure 1.5) (Antonucci, 1993; Urry & Padovani, 1995). In this scenario, the main difference between Seyfert types is that a Seyfert I has a direct view of the central AGN, whereas the nucleus is obscured by the optically thick, dusty torus in the Seyfert II case. Material surrounding the torus can scatter optical and X-ray light emitted from the nucleus, as well as polarised broad emission lines, into the line of sight of a Seyfert II observer; this provides evidence that the Seyfert I is hidden within the torus and bolsters confidence in the unification model.

### 1.3.2 AGN Variability

The colossal activity taking place in active galaxies often leads to photometric variability detectable with repeated observations of these objects over time. Measuring this variability provides a very useful perspective for studying the inner processes of
active galaxies. Time series variability is observed over a wide range of wavelengths and time-scales from hours to years (Heckman, 1976; Hook et al., 1994; Hawkins, 2002; MacLeod et al., 2012, references therein), with more luminous objects showing less variability (Vanden Berk et al., 2004) and at longer timescales (Hawkins, 1993; Giveon et al., 1999). More than 90% of quasar light curves in Stripe 82 show variability at the 0.03 mag level on timescales of a few years (Sesar et al., 2007). There are many potential sources of variability, mainly due to accretion disk instabilities or processes taking place in jet regions (Rees, 1984; Kawaguchi et al., 1998; Trèvese, Kron & Bunone, 2001; Pereyra et al., 2006). Other potential sources include supernovae bursts (e.g. Terlevich et al., 1992) and microlensing events by compact objects such as stars situated along the line of sight (e.g. Hawkins, 1993). Continuum variability can be observed from gamma to radio wavelengths, with most studies so far having been conducted in the optical band.

Different types of AGN tend to show different variability properties. Quasars tend to vary more at shorter wavelengths (de Vries, Becker & White, 2003). Since emission from blazars is dominated by the jets, they usually show strong (i.e. >1 mag) flux variations on a range of time-scales, from days to months, and over a broad range of frequencies, from radio to gamma. BL Lac and OVV galaxies often show large-amplitude, short-timescale (i.e. days) variability that could be caused by relativistic beaming effects (e.g. Bregman et al., 1990; Fan & Lin, 2000; Vagnetti, Trevese & Nesci, 2003). In UV-optical bands, Seyfert I and quasars generally show less variability (<0.5 mag), on larger time-scales of more than a few months; although large variations (>1 mag) on time-scales of days have been detected using X-rays in some Seyfert galaxies.

A structure function analysis measures the power distribution over a range of timescales to describe the temporal structure of variations (Simonetti, Cordes & Heeschen, 1985; Hughes, Aller & Aller, 1992; de Vries, Becker & White, 2003; Vanden Berk et al., 2004; de Vries et al., 2005). The structure function is defined as

\[
S(\tau) = \left\{ \frac{1}{N(\tau)} \sum_{i<j} [m(i) - m(j)]^2 \right\}^{1/2}
\]

for time intervals \( \tau = t_j - t_i \), for all exposures \( i < j \). For a total of \( N \) exposures, de Vries et al. (2005) group all \( n(n-1)/2 \) possible time-lag permutations into bins containing at least 200 measurements. The structure function outputted for each bin is the root mean square of the magnitude variations. Hughes, Aller & Aller (1992) found that the total structure functions of QSOs and BL Lacs show similar
power-law slopes; they conclude from this that there are likely the same physical processes causing the variability in these two classes of active galaxies and they attribute this to shocks in the jet regions. They find that $\gtrsim 85\%$ of sources that vary on timescales $> 10$ years are QSOs. Schmidt et al. (2010) model the light curve structure function using a power-law for a subset of spectroscopically confirmed quasars and use this to classify likely quasar candidates with single-band multi-epoch photometry from the Sloan Digital Sky Survey (SDSS) Stripe-82 survey. The SDSS is a multi-filter photometric and spectroscopic survey and will be introduced in more detail in Section 2.1.1. Stripe-82 is a 300 square degrees equatorial field that has been repeatedly imaged by the SDSS; there is also a wealth of data available for objects in this region from many other photometric and spectroscopic surveys. The autocorrelation function is similar to the structure function and can also be used to identify quasars from their intrinsic variability; it is described and used to characterise variability for classification purposes in Section 5.3.1.

The fluctuation power-density spectrum of optical variability in QSO light curves can be reasonably well described by a positive power-law with slope $\sim 2$ (Giveon et al., 1999; Collier & Peterson, 2001). Andrae, Kim & Bailer-Jones (2013) showed that QSO light curves are generally stochastic in nature, as opposed to being describable by simple deterministic models. QSO optical variability can be approximated by a damped random walk model (e.g. Kelly, Bechtold & Siemiginowska, 2009; Kozłowski et al., 2010; MacLeod et al., 2010; Zu et al., 2013; Andrae, Kim & Bailer-Jones, 2013), where the light curve is characterised as a stochastic process with exponential covariance function $S(\Delta t) = \sigma^2 \exp(-|\Delta t/\tau|)$ for amplitude, $\sigma$, and characteristic timescale, $\tau$. Parameters $\sigma$ and $\tau$ are expected to correlate with quasar properties such as rest-frame wavelength, luminosity and black hole mass (Kelly, Bechtold & Siemiginowska, 2009; Kozłowski et al., 2010; MacLeod et al., 2010; Zu et al., 2013). For typical quasars, damped random walk modelling of optical variability shows a return to the mean on a timescale of $\sim 200$ days with variability amplitude $\sim 10\%-20\%$ (Kelly, Bechtold & Siemiginowska, 2009; Kozłowski et al., 2010; MacLeod et al., 2010; Dexter & Agol, 2011). QSO classification from intrinsic variability features is an active field of research, since it enables large samples of likely QSOs to be selected from photometric surveys instead of relying on spectra (which is time consuming). QSO classification by colour is a useful photometric alternative (e.g. Sesar et al., 2007); however, as we will show later in this thesis, QSO colours may overlap with those of stars and so this method alone is unreliable. The main difficulty in variability studies is differentiating QSOs from variable stars, which are morphologically similar (i.e. point-like
1.3 The Role of Active Galaxies in Evolutionary Models

sources). However, variable stars are also of great interest; for example Cepheid variables are used as standard candles in Hubble constant measurements to determine cosmological distance scales (e.g. Freedman et al., 2001; Riess et al., 2011), RR Lyrae stars are utilised in globular cluster studies (e.g. Carretta et al., 2000; Catelan, 2009) and galaxy structure studies (e.g. Oort & Plaut, 1975; Vivas et al., 2001) and Mira variables are used to estimate globular cluster distances (e.g. Feast et al., 1989; Knapp et al., 2003). Ideally variability classification studies would not only be able to identify QSOs, but also able to determine stellar types for large samples (e.g. Pichara & Protopapas, 2013; Kim et al., 2014). Large samples of verified QSOs will allow researchers to investigate accretion mechanisms in detail, to explore the relationship of AGN variability with black hole mass, and to understand how various parameters scale with variability, e.g. investigating how variability correlates with intrinsic luminosity and redshift. We present our research into QSO classification using variability features in Chapters 5 and 6.

1.3.3 AGN Feedback and the Regulation of Star Formation

Downsizing tells us that massive galaxies often cease star formation at high redshift, and low mass galaxies continue star forming for a longer duration. One possible reason for this could be the quenching of star formation in massive galaxies due to AGN feedback (Schawinski et al., 2007b; Fabian, 2012; Cicone et al., 2014; Barai et al., 2014). Silk & Rees (1998) found that feedback processes from radiation, winds and jets from the AGN could drive enriched material from the central region out to the intergalactic medium and so AGN processes could regulate star formation in galaxies. In active galaxies, gas in the galaxy halo is shock-heated and AGN feedback can keep this gas hot, thus reducing radiative cooling and preventing star formation from taking place (Dekel & Birnboim, 2006). Feedback can also stifle black hole accretion, thus regulating the size of the central black hole and regulating the proportion between black hole mass and host galaxy mass (Springel, Di Matteo & Hernquist, 2005; Di Matteo, Springel & Hernquist, 2005; Ciotti & Ostriker, 2007; Fabian, 2012; Dubois et al., 2012; Ishibashi & Fabian, 2014).

Schawinski et al. (2007a) use 16,000 early-type galaxies within redshift range $0.05 < z < 0.1$ from the SDSS and find that AGN feedback can quench star formation in early-type galaxies and predict that after a transition period of $\sim 1$ Gyr star forming active galaxies would become quiescent and settle on the red sequence.

Kauffmann et al. (2003a) use a sample of 22,623 narrow-line AGN galaxies from the SDSS with Petrosian $r$-band magnitudes $14.5 < r < 17.7$ and find a link
between AGN activity and concurrent star formation in the host galaxy. They find that galaxies hosting low-luminosity AGN tend to have older stellar populations, like those found in early-type galaxies, and that galaxies hosting high-luminosity AGN tend to have younger stellar populations. The [OIII] emission line is found at $\lambda 5007$ and can be used to measure AGN activity (see Section 2.4). It can also indicate the presence of massive stars; however in metal-rich, star-forming galaxies massive stars only have a small influence compared with AGN activity. Kauffmann et al. conclude that AGN galaxies with strong [OIII] emission lines tend to have younger stellar populations.

Concurrent Peaks in the AGN Accretion Density and Star Formation Rate Density

The observed QSO luminosity density as a function of redshift (see Figure 1.6, top) peaks at $z = 2 - 3$ and shows a decline in the QSO population over a period of $\sim 2$ billion years (Rees, 1990; Bouwens et al., 2011; Fan, 2012). This appears to scale with observations for global star formation rate (SFR) evolution with cosmological time (see Figure 1.6, bottom). Boyle & Terlevich (1998) found a striking relation; for $z < 4$,

$$40 \times \text{QSO luminosity density (at 2800Å)} \approx \text{SFR luminosity density.} \quad (1.4)$$

This relation demonstrates a link between star formation with cosmic time and QSO activity. Silk & Rees (1998) observed a time delay of $\delta z \approx 1$ between the epoch of maximum quasar activity ($z = 2 - 3$) and the epoch of maximum SFR ($z = 1 - 2$) in the Universe. Bouwens et al. (2011) describe a $z \approx 10$ galaxy candidate then show that the star formation rate density was much smaller (by $\sim 10\%$) at this time than at $z \approx 8$. They propose that the 100-200 Myr period before $z \approx 10$ was a crucial phase for galaxy assembly, and that there was a phase of rapid galaxy build-up leading to increases in the luminosity and volume density between $z = 8 - 10$. Future observations from the James Webb Space Telescope (due to launch in 2018) in the infra-red waveband will hopefully reveal the secrets of galaxy assembly at this era ($z \approx 15$).

Research has found increasingly fascinating links between various aspects of galaxy evolution; the evolution of individual active galaxies through black hole growth and star formation in the host galaxy is closely linked with AGN feedback mechanisms, and the evolution in the overall quasar number density to its peak at $z \sim 2$ is closely connected with the Universal SFR density (Cavaliere, Perri
1.3 The Role of Active Galaxies in Evolutionary Models

Figure 1.6: Top (excerpted from Fan (2012)): Evolution of the density of luminous quasars based on the SDSS and 2dF surveys. Next generation near-IR surveys will extend these measurements to \( z > 7 \), and determine whether there is a sharp cutoff as we approach the epoch of the first billion \( M_\odot \) black hole formation. Bottom (excerpted from Bouwens et al. (2011) and Fan (2012)): The rest-frame continuum UV luminosity density (right axis) at \( z \sim 10 \), and the star formation rate density (left axis) derived from the extinction-corrected luminosity density. The comoving star formation density in the Universe has a broad peak at \( z = 2 - 5 \) and appears to be declining towards higher redshift at the end of the reionization epoch.
1.3 The Role of Active Galaxies in Evolutionary Models

& Vittorini, 1997; Efstathiou & Rees, 1988; Rees, 1990; Kauffmann & Haehnelt, 2000; Bouwens et al., 2011; Fan, 2012). Next we present one further link of vital importance: the observed relationship between merger activity and AGN activity.

1.3.4 Linking Active and Merging Galaxies in Evolutionary Models

Quasars allow valuable constraints to be placed on models of galaxy evolution. For example, for quasars to be detected out to $z \gtrsim 6$ a portion of active galaxies must have already been formed and viable models must predict their existence by this period. High redshift quasars help to constrain the epoch of reionization; this is the phase where neutral hydrogen gas in the intergalactic medium became ionized and this transition was expected to have peaked at $z \sim 10$ and ended at $z \sim 6 – 7$ (Dunkley et al., 2009; Fan, 2012). The detection of high redshift galaxies is complicated by the absorption of flux by intervening neutral hydrogen. Currently the highest redshift quasar ever observed has $z = 7.085$ (Mortlock et al., 2011). Surveys such as the UK Infra-red Telescope Deep Sky Survey (Lawrence et al., 2007; Mortlock et al., 2011), the Sloan Digital Sky Survey (Fan et al., 2006; Jiang et al., 2009), the Hubble Ultra Deep Field (Bouwens et al., 2011), and the Canada-France High-z Quasar Survey (Willott et al., 2010) aim to identify more high redshift quasars ($z \gtrsim 6$); so far approximately 50 have been identified at this redshift.

Observations of quasars imply that they have a relatively short lifetime and suggest that quasars could be a necessary phase in certain types of galaxies for a short period of time; this period of quasar activity could be a vital stage in a galaxy’s evolution (Rees, 1990; Marconi et al., 2004). Saikia & Jamrozy (2009) suggest that AGN activity could be episodic; recurring on a currently unknown timescale. Schmitt, Storchi-Bergmann & Cid Fernandes (1998) compare the stellar populations of elliptical galaxies with twenty Seyfert II galaxies and find the Seyfert IIs to have a lower population of old metal rich stars (where $z \geq z_{\text{Sun}}$ and age $\sim 10$ Gyr) and a higher population of stars aged $\sim 100$ Myr. The higher proportion of metal rich stars in elliptical galaxies could indicate that AGN activity has taken place in the past. If quasar activity is merely a phase in a galaxy’s evolution, one would wonder what type of physical event could ignite such activity.

Studies of radio galaxies and quasars find that a significant number of radio galaxies are currently interacting or show signs of recent merger activity (e.g. Smith et al., 1986; Hutchings, 1987), suggesting a link between galaxy mergers and AGN activity. Hutchings (1987) note that AGN activity is generally observed in the larger
of two interacting galaxies and suggest that in minor mergers the smaller galaxy can serve to fuel nuclear activity in the massive (often elliptical) galaxy. Alonso et al. (2007) report that the AGN fraction is larger for pairs that show strong evidence for recent interactions, although only by $\sim 10\%$ increase compared with non-interacting galaxies.

Observations show a distinct drop in the number of quasars from $z < 2$, as well as a lack of identified high redshift quasars at $z \gtrsim 6$ (see figure 1.6). If AGN activity is indeed linked with merger activity, then the decline in the quasar population would be a natural consequence in a hierarchical evolution model since the merger rate is predicted to decline at lower redshifts. We saw earlier that calculations of the merger rate show such a decline (see Section 1.4.5). Kauffmann & Haehnelt (2000) assume a $\Lambda$CDM cosmology with a hierarchical formation model and suggest that the evolution in the quasar population can be attributed to:

- less frequent mergers between galaxies
- a decline in the amount of cold gas in galaxy nuclei which is needed to fuel AGN activity
- longer time scales for the process of gas accretion onto the central black hole.

In a hierarchical scenario, the increase in SFR resulting from mergers at higher redshifts follows naturally since there is a higher gas fraction available for higher redshift galaxies, and the correlation between black hole growth and the build up of stellar mass has long been implied (Sanders & Mirabel, 1996; Hopkins et al., 2005; Debuhr, Quataert & Ma, 2011). Sanders et al. (1988) propose an evolutionary connection between Ultra Luminous Infra-Red Galaxies (ULIRGs) and quasars, and suggest that ULIRGs are formed during major mergers or forceful interactions between gas-rich spirals. Kauffmann & Haehnelt (2000) link the assembly history of supermassive black holes (which they assume have evolved through a series of smaller black hole mergers) with starbursts caused by merger activity (traced from the cold gas content by damped Lyman alpha systems) and QSO evolution.

Monaco, Salucci & Danese (2000), Cavaliere & Vittorini (2000) and Fabian, Celotti & Erlund (2006) discuss a connection between QSO and bulge formation in galaxy evolution. A tight correlation is observed between black hole mass and bulge formation in the host galaxy, where only a small scatter is observed in the correlation between black hole mass and bulge velocity dispersion (Haehnelt & Kauffmann, 2000; Ferrarese & Merritt, 2000; Gebhardt et al., 2000; Umemura, 2001; Marconi & Hunt, 2003; Daddi et al., 2007). Kauffmann & Haehnelt (2000) suggest that
correlations between these systems could be explained by one model of hierarchical evolution if:

- major mergers are the primary cause of black hole formation and growth
- cold gas fractions available to fuel black hole growth increase with redshift.

1.4 Studying Galaxy Mergers

Galaxy mergers are fundamental to galaxy evolution; they drive star formation, black hole activity and morphological transformations. Because of their importance, researchers strive towards a detailed and accurate understanding of how mergers trigger these processes across a broad redshift range. Here we provide an overview of the extensive research that has been conducted in the field of galaxy mergers; this is by no means comprehensive, but serves to outline the chronological progression of research. These previous studies provide a basis for current research into interacting galaxies and are relevant to the work that we present in Chapters 3 and 4. Methods and instrumentation have varied immensely depending on the equipment available at the era of each study. In recent times, technological advances have led to significant improvements in the quality and scale of work that can be conducted in this field.

1.4.1 Initial Studies

Erik Holmberg was one of the first to study galaxy pairs. His dissertation, *A Study of Double and Multiple Galaxies* (1937), described the clustering properties of galaxies. In his first major paper on galaxy interactions he used light bulbs to represent stellar systems, with each galaxy represented by 37 light bulbs (Holmberg, 1941). The bulb intensity was interpreted as proportional to mass in order to determine to what extent the energy loss in close-passing galaxies will result in a merger. In this creative, pre-computer simulation experiment, he found an increase in attraction between the galaxies with a peak in attraction after the galaxies had passed-by.

In 1966 a catalogue of 338 morphologically peculiar galaxies was formed by Dr. Halton C. Arp; named the *Atlas of Peculiar Galaxies* (Arp, 1966). This catalogue was compiled from 1961 to 1966 using plate files from the Palomar and Mount Wilson Observatories, and provided a robust sample for research into galaxies that have been disrupted by interactions. Using galaxies from the Arp catalogue, Larson & Tinsley (1978) showed that morphologically peculiar galaxies show much more
scatter than normal galaxies in a $U-B$, $B-V$ colour-colour plot (see Figure 1.7); suggesting that more star formation is taking place in the peculiar galaxies. Further, they showed that nearly all of this scatter is from galaxies showing signs of recent tidal interactions, i.e. close pair galaxies.

**Figure 1.7:** Excerpted from Larson & Tinsley (1978): The two-colour plots for morphologically normal and peculiar galaxies with latitudes $|b| > 20^\circ$. Panel (a) shows all Hubble Atlas galaxies with colours in the RC2 (Second Reference Catalogue of Bright Galaxies) that are not in sample (b). Panel (b) shows Arp Atlas galaxies with colours in the RC2, plus a few with colours from other sources listed in the text; the two open circles are Type I Seyfert galaxies. The curve in both plots is an eye-estimated mean line through the Hubble Atlas sample. The average mean errors of the RC2 colours for each sample are indicated.

### 1.4.2 Star Formation Enhancement

Galaxy mergers are now understood to cause the instability needed for gas clouds to collapse and form starbursts (e.g. Kauffmann & Haehnelt, 2000; Tissera et al., 2002; Cox et al., 2006; Di Matteo et al., 2007; Tonnesen & Cen, 2012). Observations of bluer optical colours in close pair systems became more frequent as larger scale surveys were introduced (e.g. Patton et al., 1997, 2005, using the first and second Canadian Network of Observational Cosmology catalogues (CNOC1 and CNOC2)).
Perhaps intuitively, close pair systems where galaxies are in the early stages of merging (i.e. at relatively large separation and only beginning to interact gravitationally) generally show less resulting star formation than further advanced mergers (Larson & Tinsley, 1978; Barton, Geller & Kenyon, 2000). Therefore, it is interesting to study the star formation in close pairs as a function of the separation between galaxies; this is a central aspect of our work in Chapter 4.

Wong et al. (2011) used UV photometry to look at NUV-\(r\) and FUV-\(r\) colours for intermediate redshift close pairs (\(0.25 \leq z \leq 0.75\)) drawn from the Prism Multi-Object Survey (PRIMUS). They find an \(\sim 15 - 20\%\) increase in SSFR for close pairs with projected separation \(\leq 50h^{-1}\)kpc, and an \(\sim 25 - 30\%\) increase in SSFR for close pairs with projected separation \(\leq 30h^{-1}\) kpc. Using redshift slices \(z = 0.4, 0.8, 1.5, 2.2\) from the HiZELS narrow-band H\(\alpha\) survey (Geach et al., 2008; Sobral et al., 2013), Stott et al. (2013) find a correlation between SSFR and merger fraction at these redshifts, implying that galaxies with enhanced SSFR are progressively more likely to be mergers.

Ultra Luminous Infra-Red Galaxies (ULIRGs) are expected to be experiencing, or to have recently experienced strong interactions (e.g. Sanders et al., 1988; Dasyra et al., 2006; Lin, Hashimoto & Foucaud, 2013), or multiple strong interactions (Borne et al., 2000). Kennicutt et al. (1987) used a complete sample of close pair spiral and irregular galaxies, as well as a sample from the Arp catalogue of peculiar galaxies, with H\(\alpha\) emission line and IRAS far-IR measurements to investigate star formation induced by interactions. Both samples generally show enhanced H\(\alpha\) and far-IR emission when compared with a control sample, indicating higher star formation levels in interacting galaxies. However, a smaller fraction of galaxies from the close pairs sample are found to show higher than average levels of star formation than that of the Arp sample (in which all objects were classed as starburst galaxies). Kennicutt speculates that this is because the Arp sample is naturally biased to galaxies that have experienced stronger interactional effects (enough to render the galaxies morphologically peculiar), but also claims there is a bias towards unusually high surface brightness, actively star forming galaxies in the Arp catalogue. This bias explained why Larson & Tinsley (1978), Joseph et al. (1984), and Lonsdale, Persson & Matthews (1984) had found a significant rise in star formation for interacting galaxies (when using morphologically peculiar samples).
1.4.3 Major and Minor Mergers

Mergers with galaxies of similar mass are commonly known as major mergers and those with a small mass ratio (usually with mass ratio less than 1:3) are referred to as minor mergers. Darg et al. (2010b) estimated that the fraction of volume-limited ($M_r < -20.55$) major mergers in the local Universe is $1-3 \times C\%$, where $C \sim 1.5$ is a correction factor to account for spectroscopic incompleteness. Using simulations, Cox et al. (2008) find that merger induced star formation is a strong function of the mass ratio of progenitors; where progenitors with similar mass generate stronger tidal forces and produce more intense star formation bursts.

Ellison et al. (2008) took a sample of 1716 galaxies selected from the SDSS with mass ratio $0.1 < M_1/M_2 < 10$; where either $M_1$ or $M_2$ could be the greater mass. For close pair galaxies (with $< 30 - 40 h^{-1}_{70}$ kpc projected separation and rest-frame velocity difference $\Delta V < 500$ km$^{-1}$) they found an enhancement in star formation of up to 70% compared with a control sample of 40,095 galaxies that had the same mass distribution. This enhancement in star formation is greatest for mergers where the progenitors have mass ratio $0.5 < M_1/M_2 < 2$.

Patton et al. (2005) find that, as well as close pair galaxies triggering star formation, we observe greater asymmetry in gravitationally interacting galaxies as they are stretched and pulled by each others gravitational influence. Patton et al. (CNOC2) estimate that $\sim 40\%$ of close pair galaxies are asymmetric and that $\sim 25\%$ are strongly asymmetric. For the close pairs that do not show asymmetry effects, Patton et al. speculate that they may be too early in the merging process and thus currently at a relatively high projected separation. They also comment that some galaxies (particularly early-types) are less likely to show asymmetry effects, and that the orbital and rotational properties can have an effect on the asymmetries observed. Patton et al. also find evidence of a higher bulge fraction for the bluest pair galaxies when compared with the bluest isolated galaxies in their sample; this could suggest that central starbursts are triggered in paired systems.

Using H$\alpha$ as a diagnostic for star formation, Woods & Geller (2007) found evidence that the lower mass progenitor in a minor merger will experience the most star formation. They used a large SDSS (DR5) sample of spectroscopically confirmed close pairs, and relative magnitudes were used to estimate mass ratio. However, they highlight an important issue when studying minor mergers; it is quite likely that the minor pairs sample will be contaminated by false classified objects that have been incorrectly deblended (estimated $\sim 15\%$ of the sample). Minor mergers are thought to play a strong role in galaxy evolution. Kaviraj (2014) estimates that $\sim 40\%$ of star formation in local spirals is triggered by minor mergers. In hierarchical models,
massive galaxies are thought to have formed from repeated merging with other galaxies (e.g. Tasca et al., 2014), with simulations predicting an estimated order of magnitude more minor mergers than major mergers (Hernquist & Mihos, 1995).

Gas-rich major mergers are predicted to supply gas to the central regions of progenitors and to potentially trigger AGN activity (Heckman et al., 1986; Sanders et al., 1988; Springel, Di Matteo & Hernquist, 2005; Hopkins et al., 2006); observational evidence of recent merger activity in galaxies with AGN activity is increasingly being presented to support this model (e.g. Canalizo & Stockton, 2001; Bennert et al., 2008; Urrutia, Lacy & Becker, 2008; Ramos Almeida et al., 2011; Villar-Martín et al., 2012; Bessiere et al., 2014). Although it is still unknown at which stage of a major merger AGN activity is most likely to be triggered and how this timescale correlates with merger-induced starbursts. Some studies have suggested a delay between merger-induced starbursts and merger-triggered AGN activity (e.g. Wild, Heckman & Charlot, 2010; Canalizo & Stockton, 2013). Wild, Heckman & Charlot (2010) find that the average black hole accretion rate rises sharply $\sim 250$ Myr after starburst activity. Bessiere et al. (2014) describe quasar-like activity in J002531-104022 which is thought to have been triggered in a major, gas rich merger; by modeling the ages and reddenings of the stellar populations, they find evidence that quasar activity and star formation were triggered quasi-simultaneously, contrary to having the delays predicted by simulations. Dasyra et al. (2006) found that the majority of ULIRGs are triggered by major mergers (with average mass ratio 1.5:1) and suggest that minor mergers (with mass ratio $>3:1$) generally do not drive enough gas into the galaxy center to cause ULIRG activity.

In Chapter 4 we study the median star formation differences in major versus minor close pair systems with different mass and environment properties. We also look for evidence of AGN being triggered in major merger systems in field environments.

### 1.4.4 Environmental Effects on Interacting Pairs

Galaxies in dense environments, such as clusters, often have lower gas fractions as a result of tidal fields and ram pressure stripping (Byrd & Valtonen, 1990); as a consequence, less gas is thought to be available to fuel star formation during a merger. Accordingly, close pairs in galaxy clusters and groups have been found to show comparatively less star formation than close pairs in field environments (Kauffmann et al., 2004; Ellison et al., 2010; Alonso et al., 2012).

At higher redshifts, clusters are found to have galaxy members which tend to be bluer than those at low redshifts. This is known as the Butcher-Oemler effect.
(Butcher & Oemler, 1978; Butcher & Oemler, 1984; Margoniner et al., 2001) and is often attributed to an evolution in cluster galaxy populations where younger galaxies in clusters have younger (and hence bluer) stellar populations. As these stellar populations age and a higher metallicity level is achieved, the galaxies are expected to redden.

Due to large velocity dispersions resulting from gravitational interactions in clusters, mergers tend to occur in lower density field and group environments. The group environment population is dominated by SO galaxies (Wilman et al., 2009) and these likely formed from major mergers (Schweizer, 1993). McGee et al. (2009) suggest that a significant fraction of galaxies in cluster environments have been accreted from galaxy groups, where most of their evolution took place through mergers. Although mergers in clusters are less likely, close pair interactions are more common (e.g. Lin et al., 2010). Galaxies in this environment will be moving much faster due to stronger gravitational effects and are less likely to get locked into a merger scenario.

Using SDSS DR4 data, Ellison et al. (2010) found a decrease in star formation in local dense environments when using asymmetry and optical colours to indicate interaction-induced star formation. They concluded that the higher levels of star formation detected in close pairs in low density environments is a result of the higher gas fraction available to fuel star formation. Ellison et al. found that, even though mergers in high density environments show little star formation, they show strong morphological asymmetries (particularly towards the cluster centre) which suggest that forceful gravitational interactions are taking place. Alonso et al. (2012) find that galaxy pairs tend to be located closer towards the group centre, and that disturbed pairs are more likely to contain the brightest galaxy in a group.

Schawinski et al. (2007a) find that UV-bright galaxies are most likely to be found in the field. They saw 25% less star formation in environments denser than the field; however, this trend ends after intermediate densities and they find similar measurements for group and cluster environments. Schawinski et al. find a stronger correlation between recent star formation and environment for more massive galaxies, and they hypothesise that the star formation that we see in dense environments could be a result of starbursts in inter-cluster gas after it has been expelled from the galaxies.
1.4.5 Merger Rate

A detailed understanding of the evolution of the merger rate over different epochs paves the way for tighter cosmological constraints and more reliable evolutionary models. More mergers are expected to have taken place when the Universe was younger and more condensed; this is corroborated by observations of a larger star formation density at higher redshifts than in the local Universe (e.g. Madau et al., 1996; Fan, 2012). We might also expect star formation to be greater in higher redshift galaxies since they are thought to be more gas rich (e.g. Kauffmann & Haehnelt, 2000).

Zepf & Koo (1989) estimated that the merger rate increases with redshift as $(1 + z)^{4\pm2.5}$ by using faint galaxy pairs as analogues of moderate redshift pairs. Their result assumes that faint galaxy pairs adequately represent moderately high redshift pairs. Burkey et al. (1994) suggested a merger rate of $(1 + z)^{2.5\pm0.5}$ when using galaxy close pairs imaged by the HST Wide-Field Camera, and Patton et al. (1997) suggested an estimate of $(1 + z)^{2.8\pm0.9}$ after accounting for various selection effects that arise when identifying mergers.

The merger rate appears to be dependent on the assumptions made, although a reasonable approximation can be made by combining the results of various studies that have employed different methods. Figure 1.8 shows an excerpt from Bridge et al. (2007) that summarises merger rates derived using different estimation methods; including the CAS (concentration, asymmetry, and clumpiness) quantitative classification system (Conselice, 1997).

Bell et al. (2006) used the projected correlation function of a large sample of galaxies from COMBO-17 to measure the expected fraction of galaxies in close pairs to explore the merger rate. Robaina et al. (2010) continued this work by Bell et al. by studying the evolution in the two-point correlation function for massive galaxies ($M > 5 \times 10^{10} M_\odot$) from the COSMOS and COMBO-17 surveys as a function of redshift. They estimate that the close pair fraction of massive galaxies (with three-dimensional separation $< 30$ kpc) evolves as $F(z) = (0.0131 \pm 0.0019) \times (1 + z)^{1.21 \pm 0.25}$. From this result, they deduce that galaxies with $M > 10^{11} M_\odot$ have participated in an average of 0.5 massive mergers (where both progenitors have $M > 5 \times 10^{10} M_\odot$) since $z = 0.6$. 
1.5 Summary

In this chapter we introduced various concepts that will be relevant throughout the thesis and summarised how some of these ideas have developed over decades of research into astrophysics and cosmology. The constituents of the Universe were introduced; baryonic matter (e.g. protons and neutrons) and non-baryonic matter (e.g. potentially dark matter). The search to determine the composition of dark matter is ongoing, but because its existence is known it must be accounted for in models of galaxy formation and evolution.

The ΛCDM cosmological framework was presented and evaluated by considering evidence from CMB and supernovae studies. Extending this framework to assume that a momentous cosmic inflation occurred shortly after the Big Bang, the proposed history for structure formation was described; beginning with quantum fluctuations before inflation that gradually expanded into the over-dense and under-dense regions which galaxies and voids now inhabit. The first galaxies were thought to have formed...
due to the gravitational collapse of concentrated baryonic matter in dense regions of dark matter halos, and to have evolved hierarchically from then on by merging with other galaxies. A competing model for galaxy evolution, Monolithic Collapse, was also presented since this was a popular model in the 1960s; we described how this contradicts recent observations and why this model is generally neglected in favour of the hierarchical evolution model. The ΛCDM cosmological framework with hierarchical galaxy evolution is assumed throughout the thesis.

Various types of galaxies were introduced; elliptical, spiral, lenticular and irregular. We then introduced active galaxies and discussed the role of AGN in evolutionary models. Studies on AGN feedback and its effect on star formation within the host galaxy were reviewed, and we discussed evolutionary models that link galaxy mergers with AGN activity, star formation and black hole growth. AGN variability, ways of quantifying this variability, and the implications that variability can have on processes within the host galaxy were presented. In Chapters 5 and 6 we utilise various measures of AGN variability to distinguish AGN from stars in a large-scale photometric survey.

Galaxy mergers were introduced and previous research in this field was presented. Interactions between galaxies are known to trigger star formation episodes, and the amount of observed star formation is expected to depend on the properties of interacting galaxies and the environment that they inhabit. We will study the effects of merger activity between galaxies as a function of separation, mass and environment density in Chapters 3 and 4.
Chapter 2

Tools and Techniques

Overview: Optical and UV Sky Surveys

We now introduce the surveys and instruments from which we gathered our data and describe various research techniques that will be employed in future chapters. We use data from a number of telescopes and surveys including the Sloan Digital Sky Survey, Galaxy Evolution Explorer, Panoramic Survey Telescope and Rapid Response System, and the Time Domain Spectroscopic Survey. The first section of this chapter provides information, technical specifications and science goals for these surveys and instruments, and details limitations that should be taken into account when analysing data from each.

Different techniques are introduced for measuring star formation in galaxies using continuum measurements and nebular recombination lines; the advantages and disadvantages of each method are evaluated. Emphasis is placed on UV photometric measurements of star formation, and we explain why this method was chosen for our close pairs study.

We look at various methods for classifying galaxy samples. Morphology-based methods are described; using concentration, asymmetry and symmetry features. Ways to distinguish between galaxy populations using spectra and colour are explored. Colour-magnitude diagrams are introduced, and their usefulness at distributing galaxy samples into star forming versus non-star forming regions is discussed. BPT diagrams are described and we discuss how these can be used to classify galaxies according to their sources of ionization; such an analysis permits some types of active galaxies to be distinguished from Starburst and Transient galaxies. Galaxies can also be classified according to the density of their environment, and we discuss methods which determine whether a galaxy exists as part of a field, group or cluster environment.
2.1 Surveys and Telescopes

2.1.1 SDSS

The Sloan Digital Sky Survey (SDSS) is one of the largest surveys ever conducted and utilises a 2.5m telescope at Apache Point Observatory, New Mexico. A multi-object fibre spectrograph observes over an area of $\sim10,000\ \text{deg}^2$ (about one quarter of the celestial sphere). Photometric data is imaged using five optical filters $u, g, r, i, z$ over a 3,000-11,000Å range (with effective wavelengths 3550Å, 4770Å, 6230Å, 7620Å and 9130Å respectively) using a large format mosaic CCD camera (Fukugita et al., 1996; Gunn et al., 1998). The atmospheric UV cut-off wavelength is 3,000Å and the silicon sensitivity limit of the CCDs is 11,000Å.

The telescope scans the sky along great circles (in sidereal time) and is easily switched from photometric to spectroscopic mode by replacing the imaging camera with a fibre plug plate which sends incoming light to the spectrographs (York et al., 2000). Two digital fibre-fed spectrographs are used, each containing up to 640 fibres that can observe separate objects in a $1.49\degree$ radius circular tile (Blanton et al., 2003). Each fibre has a 0.2mm diameter, which translates to $3''$ on the sky (York et al., 2000). Due to physical limitations when placing two fibres close together, there is a minimum fibre separation of $55''$; spectra for objects within $55''$ can only be obtained if they are observed by overlapping tiles. These so-called fibre collisions present an incompleteness issue when studying galaxy close pairs and only $\sim30\%$ of the $10,000\ \text{deg}^2$ area covered by the SDSS is observed by overlapping tiles (Darg et al., 2010a). This spectroscopic incompleteness can be quantified by using the photometric data available (Patton & Atfield, 2008). A tiling algorithm developed by Blanton et al. (2003) helps to reduce fibre collision incompleteness by strategically allocating fibres to desired targets (instead of covering the sky with uniformly spaced tiles) to optimise the area of the sky covered by overlapping tiles.

For the primary sample, also referred to as the main galaxy sample, the Petrosian system was introduced to avoid potential biases when dealing with galaxy photometry. This is because galaxies, unlike stars, do not necessarily have similar radial brightness profiles nor well-defined boundaries (Petrosian, 1976). The SDSS uses a modified version of the Petrosian system to measure a constant proportion of the total detected light, independent of the object’s distance (Deng et al., 2006). The Petrosian ratio, $R_p$, at radius $r$ from the galaxy’s centre is defined as the ratio

\[ R_p = \frac{\text{Total Detected Light}}{\text{Petrosian Radius}} \]

\[ 50 \text{ of these 640 fibres are used for calibration purposes.} \]
of local surface brightness (in an annulus at \( r \)) to mean surface brightness within \( r \);

\[
R_P(r) = \frac{\int_{0.8r}^{1.25r} 2\pi r'dr' I(r')/\left[\pi\left(1.25^2 - 0.8^2\right)r^2\right]}{\int_0^r 2\pi r'^{4}I(r')/(\pi r^2)}
\]

(2.1)

where \( I(r) \) is the azimuthally averaged surface brightness profile.

The Petrosian radius, \( r_p \), is defined such that \( R_p = 0.2 \); i.e. it is the largest local radius such that the surface-brightness averaged in an annulus of radius \( r \) is at least 20\% of the mean surface brightness interior to this annulus. The Petrosian magnitude is the total flux within a circular aperture of diameter twice the Petrosian radius, i.e.

\[
F_P = \int_0^{2r_p} 2\pi r'dr' I(r')
\]

(2.2)

This is a large enough aperture size to enclose nearly all of the flux for most spiral galaxies and \( \sim 80\% \) of the flux from elliptical galaxies (Kauffmann et al., 2003a). The primary spectroscopic sample for the SDSS only targets galaxies with \( r \)-band apparent Petrosian magnitude \( r < 17.77 \); this cut-off corresponds to \( M_r = -20.55 \) at \( z = 0.1 \) (Schawinski et al., 2007a). The median redshift of the main sample is 0.1, and few galaxies are observed with \( z > 0.25 \) (Deng et al., 2006).

In the SDSS frames pipeline, overlapping objects which have initially been detected as one parent galaxy are deblended by separating the various sub-peaks into children components. This process takes place across all five optical bands, and after deblending has taken place the properties of the individual children are measured such that the sum of the optical flux of the children is equal to that of the parent.

### 2.1.2 GALEX

GALEX (Galaxy Evolution Explorer) is a NASA space based all-sky survey observing at ultraviolet wavelengths with 4-6" resolution and \( \sim 50\text{cm}^2 \) effective area (Morrissey et al., 2007). It was launched on April 28th 2003 into a circular Earth-orbit with a 98.6 minute period. GALEX was scheduled to run for 29 months, however this has been extended and it is still presently functioning; it has a projected orbit lifetime of at least 25 years and is expected to be fully functional until 2015. It does not require the use of consumables and is self-sufficient. GALEX images in 1°2 diameter circular fields at 1770-2730Å (NUV) and 1350-1780Å (FUV) simultaneously, using a modified Ritchey Chrétien telescope. It has a spectroscopic observing mode, though we only use photometric measurements in this work.

GALEX is faced with various constraints because of the high sensitivity of its detectors (which is a necessary requirement when observing in the UV). It is
incapable of observing when the detectors are facing the Sun, Earth, Moon or bright planets; this limits the feasible observing time for most individual targets to \( \sim 10\% \) of the year. During the daytime period of its orbit, GALEX is oriented so that its solar panels face the sun (with its detectors facing in the opposite direction so as to avoid damage) and its batteries are recharged. Observations are only performed during the night period and the telescope is oriented towards its target. Observations are flux limited since stars or regions which emit strongly in the UV can saturate and damage the detectors. The FUV limit is 5000 cts/s, \( m_{AB} = 9.5 \), \( F_\lambda = 7 \times 10^{-12} \) erg cm\(^{-2}\) s\(^{-1}\) Å\(^{-1}\) and the NUV limit is 30,000 cts/s, \( m_{AB} = 8.9 \), \( F_\lambda = 6 \times 10^{-12} \) erg cm\(^{-2}\) s\(^{-1}\) Å\(^{-1}\).

### Table 1: Summary of Measured Performance Parameters for GALEX

<table>
<thead>
<tr>
<th>Item</th>
<th>FUV Band</th>
<th>NUV Band</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandwidth(^{a})</td>
<td>1344–1786 Å</td>
<td>1771–2831 Å</td>
</tr>
<tr>
<td>Effective wavelength (( \lambda_{\text{eff}} ))(^{b})</td>
<td>1538.6 Å</td>
<td>2315.7 Å</td>
</tr>
<tr>
<td>Mean effective area(^{c})</td>
<td>19.6 cm(^2)</td>
<td>33.6 cm(^2)</td>
</tr>
<tr>
<td>Peak effective area(^{d})</td>
<td>36.8 cm(^2) at 1480 Å</td>
<td>61.7 cm(^2) at 2200 Å</td>
</tr>
<tr>
<td>Astronomical (( 1^\circ, R \leq 0.6^\circ ))</td>
<td>0.057(^{f})</td>
<td>0.069(^{f})</td>
</tr>
<tr>
<td>Field of view(^{g})</td>
<td>1.27(^{g})</td>
<td>1.25(^{g})</td>
</tr>
<tr>
<td>Photometry (( p ))</td>
<td>( \pm 0.05 m_{AB} )</td>
<td>( \pm 0.03 m_{AB} )</td>
</tr>
<tr>
<td>Zero point (( m(0,0) ))</td>
<td>18.82</td>
<td>20.08</td>
</tr>
<tr>
<td>Image resolution (FWHM)</td>
<td>4.2(^{i})</td>
<td>5.3(^{i})</td>
</tr>
<tr>
<td>Spectral resolution (( /\Delta \lambda ))</td>
<td>200</td>
<td>118</td>
</tr>
<tr>
<td>Spectral dispersion(^{j})</td>
<td>1.64 Å arcsec(^{-1})</td>
<td>4.04 Å arcsec(^{-1})</td>
</tr>
<tr>
<td>Detector background (typical):</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>78 counts s(^{-1})</td>
<td>193 counts s(^{-1})</td>
</tr>
<tr>
<td>Diffuse</td>
<td>0.66 counts s(^{-1}) cm(^{-2})</td>
<td>1.82 counts s(^{-1}) cm(^{-2})</td>
</tr>
<tr>
<td>Hotspots</td>
<td>47 counts s(^{-1})</td>
<td>107 counts s(^{-1})</td>
</tr>
<tr>
<td>Sky background (typical)(^{l})</td>
<td>10,000 counts s(^{-1})</td>
<td>10,000 counts s(^{-1})</td>
</tr>
<tr>
<td>Limiting magnitude (5(^{m}))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIS (100 s)</td>
<td>19.9</td>
<td>20.8</td>
</tr>
<tr>
<td>MIS (1500 s)</td>
<td>22.6</td>
<td>22.7</td>
</tr>
<tr>
<td>DIS (50000 s)</td>
<td>24.8</td>
<td>24.4</td>
</tr>
<tr>
<td>Linearity:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global (10% roll off)</td>
<td>18,000 counts s(^{-1})</td>
<td></td>
</tr>
<tr>
<td>Global (50% roll off)</td>
<td>91,000 counts s(^{-1})</td>
<td></td>
</tr>
<tr>
<td>Local (10% roll off)</td>
<td>114 counts s(^{-1})</td>
<td>303 counts s(^{-1})</td>
</tr>
<tr>
<td>Pipeline image format</td>
<td>3840 x 3840 elements with 1.5(^{p}) pixels</td>
<td></td>
</tr>
</tbody>
</table>

\(^{a}\) Includes wavelengths with effective area at least 10\% of the peak value.
\(^{b}\) Wavelength-weighted: \( \lambda_{\text{eff}} = \int \lambda F_\lambda d\lambda / \int F_\lambda d\lambda \).
\(^{c}\) These correspond to 246 and 601 photons s\(^{-1}\) cm\(^{-2}\) s\(^{-1}\) Å\(^{-1}\), respectively.
\(^{d}\) These are worst-case values for point sources.

**Figure 2.1:** Excerpt from Morrissey et al. (2007), Table 1.

The primary aim is to study star formation and evolution in galaxies. GALEX allows us to collect photometric and spectroscopic data on hundreds of thousands of stars and galaxies for targets aged up to \( \sim 10 \) billion years old. It enables us to estimate levels of recent star formation and constrain star formation histories for galaxies.

In Chapter 4, we use the GR4/GR5 database (from data releases 4 and 5) which
### 2.1 Surveys and Telescopes

<table>
<thead>
<tr>
<th>Survey</th>
<th>Exposure Time (seconds)</th>
<th>Sky Coverage (deg²)</th>
<th>Depth (m&lt;sub&gt;AB&lt;/sub&gt;)</th>
<th>GR4 (no. tiles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-sky Imaging (AIS)</td>
<td>100</td>
<td>26,000</td>
<td>20.5</td>
<td>28,000**</td>
</tr>
<tr>
<td>Medium Imaging (MIS)</td>
<td>1,500</td>
<td>1,000</td>
<td>23.5</td>
<td>1615</td>
</tr>
<tr>
<td>Deep Imaging (DIS)</td>
<td>30,000</td>
<td>80</td>
<td>25.0</td>
<td>193</td>
</tr>
<tr>
<td>Nearby Galaxy (NGS)</td>
<td>1,500</td>
<td>300</td>
<td>28*</td>
<td>433</td>
</tr>
</tbody>
</table>

**Table 2.1:** GALEX Baseline Mission Surveys (up to GR4, completed in Fall 2007). Exposure time, sky coverage, depth and the number of tiles released in GR4 are shown for each survey. This information was sourced from the GALEX webpage. *surface density (mag/sq arcsec) **projected.

combines data from the following imaging surveys: All-sky Imaging Survey (AIS), Deep Imaging Survey (DIS), Medium Imaging Survey (MIS), and Nearby Galaxy Survey (NGS) (Martin et al., 2005; Morrissey et al., 2005). For exposure times, sky coverage, depth and the number of tiles released see Table 2.1. AIS covers around 3/4 of the sky and aims to provide an all-sky survey with similar depth to the Palomar Observatory Sky Survey II and the SDSS. Regions in the vicinity of the Galactic plane and the Magellanic clouds were avoided due to the sensitivity of detectors. As a result, imaging in the surrounding areas can be patchy, but otherwise fields are mostly adjacent. MIS is positioned to have a significant overlap with the SDSS; there is a 7325 deg² overlap between SDSS DR7 and AIS GR5 and 1103 deg² with the MIS GR5 (Bianchi, 2011).

#### 2.1.3 Pan-STARRS

The Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) is an in-construction wide-field project that aims to build four 1.8m telescopes, each with a 7 square degrees field of view, giving the ability to scan the visible sky<sup>2</sup> in less than one week (Hodapp et al., 2004; Kaiser et al., 2010). Pan-STARRS aims to monitor the trajectory of Near Earth Objects (NEOs), such as asteroids and comets, to identify any that are potentially threatening. Pan-STARRS provides wider coverage and deeper imaging (by ~3 mags) than the previous leading sky survey for NEOs; the Catalina Sky Survey (Larson et al., 2006). Most of the funding is provided by the United States Air Force.

The ‘PS4 telescope’ refers collectively to the four planned individual telescopes. The first telescope, Prototype Telescope 1 (PS1), has been operational since 13th May 2010 on Mount Haleakala in Maui, Hawaii, under the direction of the PS1

<sup>2</sup>~3/4 of the entire sky is visible from this location in Hawaii.
2.1 Surveys and Telescopes

Science Consortium (currently a U.S.A., U.K., Germany, Taiwan collaboration). It was built in the space where NASA’s Satellite Laser Ranging system was located; this was decommissioned in 2004. Many of the design features that will be implemented on the full PS4 telescope are being tested by PS1, such as the optical and camera technology. The imager is equipped with a 1.4 Gigapixel mosaic focalplane CCD camera at the back of the telescope\(^3\), with 10\(\mu\)m pixels that subtend 0.258 arcsec (Tonry et al., 2008; Onaka et al., 2008; Stubbs et al., 2010; Tonry et al., 2012).

The automated image processing pipeline takes only 15 hours to process data from one night of observations (the expected data to be processed is approximately one trillion pixels per night, or \(\sim\)50 Terrabytes per month). The broad-band imaging ranges from 400-1000nm with optical filters \(g, r, i, z, y\). The \(g, r\) and \(i\) filters are very close to the corresponding filter wavelengths in the SDSS. First light for PS2 was expected in early 2013, but is currently delayed due to funding cuts made by the United States Congress in 2011 (although an anonymous $3 million donation and a $5 million donation from NASA has allowed the project to continue).

As well as variable stars such as Cepheids and RR Lyraes, Pan-STARRS can also identify eclipsing binaries, supernovae and micro-lensing events. Pan-STARRS allows galaxies with AGN activity, such as QSOs, to be identified by both colour and variability features. In this work we focus on low to intermediate redshift objects (\(z < 5\)) from the Medium Deep and \(3\pi\) surveys (which are limited to \(i\)-band \(\text{mag}_{\text{AB}} < 25\)).

The \(3\pi\) survey mode will offer \(3\pi\) steradian coverage of the sky, the entire visible sky from Hawaii (north of -30° declination), in each band every five nights with 12 epochs in each filter (Magnier, 2007; Magnier et al., 2013). It has exposure times of <60 seconds and median redshift \(z \sim 0.7\). The \(3\pi\) survey accounts for 56% of Pan-STARRS observing time. The survey will take 3.5 years, and will re-image each field 4 times per band per year. The Medium Deep (MD) survey is a deeper survey than \(3\pi\) with more frequent exposures, a longer exposure time, and 8 exposures per filter (these are dithered and then stacked each night). It covers ten uniformly distributed 7°-squared fields (with some regions overlapping Stripe 82), and accounts for 25% of observing time. Each exposure takes a single ‘snapshot’ of the 7°-squared field (Rest, 2011).

\(^3\)The camera comprises an array of 64×64 CCD devices spread over 40×40 centimetres, each with \(\sim\)600×600 pixels.
2.1.4 TDSS

In Chapters 5 and 6 we utilise spectra from the BOSS spectrograph on the SDSS telescope; each spectrum was measured for either the SDSS or the Time Domain Spectroscopic Survey (TDSS). TDSS is a collaboration between SDSS & Pan-STARRS led by Paul Green (Harvard-Smithsonian Center for Astrophysics) and Scott Anderson (University of Washington). TDSS is in the process of building a candidate list of 100,000 variable objects using Pan-STARRS and SDSS photometry. Colours and other statistical features that are useful for extracting variable objects are considered during this selection process. Single-epoch spectroscopic follow-up of these candidates from the Baryon Oscillation Spectroscopic Survey (BOSS) allows these objects to be analysed in more detail. This project was approved by SDSS-III as a BOSS Ancillary project for PS1 Medium Deep Fields 1 and 3. During SDSS-IV, variables selected from the PS1 $3\pi$ survey will be targeted as part of the eBOSS survey at a surface density of about 10 deg$^{-2}$, yielding anywhere from 75,000 to 100,000 follow-up spectra from the upgraded BOSS spectrograph.

BOSS (Bolton et al., 2012; Smee et al., 2012; Dawson et al., 2013) is one of the leading surveys from SDSS-III (Eisenstein et al., 2011), the third phase of the SDSS project. Its primary goal is to map the spatial distribution of massive galaxies and quasars to accurately measure the length scale imprinted on large-scale mass structures by baryon acoustic oscillations (BAO), and thus constrain dark energy models. The SDSS multi-object fibre spectrographs were upgraded in 2009 and are in use for BOSS. Its successor in SDSS-IV is eBOSS, which will target a much larger number of QSOs, as well as luminous red galaxies and emission-line galaxies across a wider redshift range than BOSS.

![Figure 2.2: BOSS spectroscopy for two QSO TDSS targets. Left: $z = 3.53$, right: $z = 5.01$ (excerpted from a TDSS presentation by Paul Green, 2013).](image)

TDSS variable objects include quasars and different types of variable stars. In
the pilot survey results in 2012, out of 371 TDSS targets, 45.5% (169/371) were confirmed to be QSOs (the highest redshift object is shown in Figure 2.2 (right) and has $z = 5.01$) and 54.3% (201/371) were confirmed to be stars. Of these 201 stars, 5 were A-type, 17 were F-type, 27 were G-type, 59 were K-type and 91 were M-type. The pilot survey used the Pan-STARRS Medium Deep Survey, as opposed to 3$\pi$ light curves. We utilise this pilot survey spectra in Chapters 5 and 6.

### 2.2 Star Formation Diagnostics

The analysis of star formation rates (SFRs) provides insight into the past and present activity of galaxies, giving a measure by which to analyse and classify galaxies. We can measure the SFR density for galaxies as far as $z > 10$ (Bouwens et al., 2010) and can analyse how SFR densities have evolved with cosmological time (e.g. Madau et al., 1996; Steidel et al., 1996; Hopkins, 2004; Bouwens et al., 2010; Fan, 2012). Evolution in SFR density has implications about the evolution of Hubble types and can be used to gain a broader understanding of how the galaxies we see in the local Universe have developed.

Various methods exist by which to approximate the SFR of galaxies, each following a different set of assumptions and with different limitations. These diagnostics fit into the two categories of integrated continuum measurements (in various wavelength regions) and nebular recombination lines. Amongst these diagnostics include Balmer lines, forbidden lines such as [OII] and [OIII], PAH features in the mid-IR, UV-continuum measurements and IR-continuum measurements. Other diagnostics are available but we focus on these because they are the main methods used in galaxy evolution studies.

Given that the most common system of classifying galaxies is by morphology, it is not surprising that much work has been devoted to finding correlations between morphology and star formation history (e.g. Tinsley, 1968; Searle, Sargent & Bagnuolo, 1973; Struck & Smith, 2003; Williams et al., 2011). Tinsley (1968) was one of the first to observe a close link between morphological characteristics and the star formation history of a galaxy.

Researchers benefit from having access to a wide range of available instruments and surveys which employ different technology, make different assumptions, and have independent physical limitations for measurements. We briefly introduce some of the methods available and then move forward to describe UV measurements in detail. We then discuss why UV photometry was chosen to calculate SFRs for our close pairs sample.
2.2 Star Formation Diagnostics

2.2.1 Nebular Recombination and Ionization Lines

Galaxy spectra can be used to indicate the presence of star formation since various emission-lines are produced during star forming processes; e.g. hydrogen recombination emission-lines indicate that HII star forming regions are present within a galaxy. Many ionizing processes which result from the UV flux radiated from young stars can be traced using nebular recombination and ionization lines.

- **Balmer lines**: $H\alpha$ is the spectral emission-line (at 6562.8Å) produced when an electron in a hydrogen atom transitions from the third quantised energy level to the second. It indicates that hydrogen ionization is taking place. When hydrogen is ionized, the proton and electron soon recombine. At recombination the electron can start at any energy level and will cascade to the first energy level, and for the n=3 to n=2 transitions we see $H\alpha$ emission.

Hydrogen recombination occurs when HII star forming regions are present within the galaxy. $H\alpha$ luminosity scales directly with the hydrogen-ionizing radiation from massive stars ($>10^{10}M_\odot$), and so it is a very effective tracer of star formation (Moustakas, Kennicutt & Tremonti, 2006; Domínguez et al., 2013). $H\alpha$ is observed in the red in the optical band, however, it can only be measured using optical CCDs for $z \lesssim 0.4$ (Aragón-Salamanca et al., 2003). Once it is redshifted into the near-infrared (NIR) it can be observed for $0.7 \lesssim z \lesssim 2.5$ sources. $H\beta$ emits at 4861.4Å when electrons transition from the fourth to the second energy level. $H\beta$ and higher order Balmer emission-lines are potential alternatives, however these lines are relatively weak and can be influenced by stellar absorption more than $H\alpha$ (Kennicutt, 1998).

- **[OII] and [OIII]**: The forbidden emission-lines [OII] and [OIII] trace ionising photons (Kennicutt, 1998; Moustakas, Kennicutt & Tremonti, 2006; Domínguez et al., 2013). The [OII] forbidden line doublet occurs at 3726-3729Å and the doubly ionised oxygen, [OIII], forbidden emission-line doublet occurs at 4959Å and 5007Å (in the optical waveband). These are a result of UV radiation from young stars photoionizing heavier elements (in this case neutral oxygen).

[OII] can be measured up to $z \sim 1.4$ in the optical and $z \sim 5.4$ in the NIR, making it an attractive alternative to $H\alpha$ for high redshift galaxies (Gallagher, Hunter & Bushouse, 1989; Gallego et al., 2002). However, the accuracy with which it measures SFRs is strongly dependent on the metallicity in the surrounding environment, and although the equivalent widths are well correlated
2.2 Star Formation Diagnostics

with H\textalpha, the flux is on average half that of H\textalpha.

- **PAH features:** Polycyclic aromatic hydrocarbon (PAH) dust molecules contain $\approx$50 atoms and absorb FUV photons that are produced in star forming regions; re-emitting at the following infrared wavelengths: 3.3, 6.2, 7.7, 8.6 and 11.3\(\mu m\) (Leger & Puget, 1984; Allamandola, Tielens & Barker, 1985; Popescu et al., 2011; Meidt et al., 2012). Thus we can use these emission-lines to quantify FUV flux, and as a result approximate the SFR of the galaxy. Peeters, Spoon & Tielens (2004) evaluate the effectiveness of this method to measure star formation and find that PAHs trace B stars well but are unable to trace massive (O-type) star formation well. PAH emission is dependent on the galaxy metallicity and the distribution of HII and dust regions (Calzetti et al., 2007).

2.2.2 Continuum Measurements

- **Infrared Continuum:** Young stars emit UV flux and ideally we could measure recent star formation by looking directly at the UV emission. However, star forming regions are necessarily dusty regions, and the UV radiation from young stars is absorbed by this dust and re-emitted in the FIR at $\approx$10-300\(\mu m\) (Kennicutt, 1998; Buat et al., 2010; Schisano et al., 2014). Therefore, by measuring FIR luminosity we can gain valuable information about the star formation which is taking place beneath the dust obscuration in optically thick regions.

Rowan-Robinson & Crawford (1989) used a sample of 227 galaxies from the IRAS Point Source Catalogue with measured flux in the four IRAS bands (12, 25, 60, 100\(\mu m\)) to model the IR spectra (10-100\(\mu m\)). They split the typical dust regions in galaxies into the following three components to accurately model IR re-emission;

- Cool disk component: Models the IR re-emission from interstellar dust that is illuminated by the stellar population
- Warm starburst component: Models the IR re-emission from optically thick dust clouds that surround young stellar populations in central starbursts
- Hot Seyfert component: Models the IR re-emission from the narrow-line region in Seyfert galaxies.
Assumptions must be placed on extinction levels, galaxy morphology and the initial mass function (IMF). The SFR is sensitive to these parameters and can change dramatically if, for example, a different IMF is assumed. Buat & Xu (1996) studied the effects of extinction on derived SFRs and found a large variation in optical depth for galaxies in their sample that causes a dispersion of factor 2 (at the 1σ level) when converting FIR flux to SFR. Because light is re-emitted in the FIR after dust absorption, infrared selected galaxy samples suffer from a selection-bias and often contain a large population of dust obscured objects (Casey et al., 2012).

- **Ultraviolet continuum:** UV continuum measurements allow us to detect particularly young stellar populations. It is especially useful for galaxies with weak Hα lines and where AGN emission has contaminated Hα measurements. Ground-based surveys are difficult because of substantial UV-absorption by the Earth’s atmosphere, and so detailed UV studies must utilise space-based telescopes such as GALEX. The International Ultraviolet Explorer (IUE; Kondo, 1987) measured spectra in the UV (1200-3000Å) and paved the way for modelling star formation from a UV perspective; however, it only measured spectroscopy. GALEX measures both photometry and spectroscopy; its spectral range (1350-2800Å) lies conveniently between the Lyα forest and longer wavelength spectral features caused by older stellar populations.

For distant galaxies (1 ≲ z ≲ 5), the FUV continuum produced by hot (eg. O, B and A) stars is redshifted into optical wavelengths (Kennicutt, 1998; Steidel et al., 1996) and can be detected from ground-based telescopes. The mid-UV (MUV) filter in GALEX can detect young stellar populations with as small as ∼1% mass fraction, providing an effective tracer for recent star formation (Schawinski et al., 2007a).

The main benefits of the near-UV (NUV) waveband were unveiled when it was discovered to be much more sensitive to recent star formation (up to ∼1 Gyr) than optical filters; this was first observed in studies of early-type galaxies. Yi et al. (2005) used GALEX data to construct an NUV-optical colour-magnitude relation (CMR, see Section 2.3.2) for early-type galaxies with z ≤ 0.25. Early-type galaxies are classically thought to be non-star forming, however, the sensitivity of the NUV to young stellar populations enabled Yi et al. to find evidence for recent star formation (≲ 1Gyr) in ∼15% of local early-type galaxies (with z < 0.13)⁴. Schawinski et al. (2007a) found

⁴Consistent with this result, Osterbrock (1960) found that 15 ± 5% of moderate-to-giant ellip-
2.2 Star Formation Diagnostics

Figure 2.3: Excerpted from Schawinski et al. (2007a): Volume-limited UV color-magnitude relation. Ellipticals are green, lenticulars are blue, and red dots denote galaxies that were rejected during a visual inspection as late-types. The red circles show those galaxies that are host to a strong AGN (from BPT analysis). The dashed line indicates the $\text{NUV}-r = 5.4$ cutoff for recent star formation. The fraction of UV blue galaxies that are not genuine early-type galaxies is significant: both late-types and AGN candidates are significantly bluer. The error bars in the top left show typical $1\sigma$ errors, although the reddest galaxies may have slightly larger errors as they tend to be very faint in the NUV.

that a much higher population of early-type galaxies show evidence in the UV of recent star formation. They used a sample of 839 early-type galaxies with $r$-band magnitudes from the SDSS and found that $29\pm3\%$ of elliptical galaxies and $39\pm5\%$ of lenticular galaxies show recent star formation (see Figure 2.3).

The monolithic formation model for early-type galaxies predicts that the entire population of stars in early-type galaxies forms at high redshift and evolves passively thereafter. However, UV measurements of star formation have provided convincing evidence against this model. The NUV waveband is less influenced by old stars that are UV bright than FUV measurements.

The main disadvantage of using the UV to test for star formation is dust obscuration; since dust is always present in starburst regions. Many attempts...
have been made to optimise methods to correct for dust attenuation for UV continuum emission (Calzetti, Kinney & Storchi-Bergmann, 1994; Kennicutt, 1998; Hopkins et al., 2001; Salim et al., 2007). The advantages are that the UV is highly sensitive to young stars and that it is an integrated quantity, so galaxies can be detected in the UV even when they are undetected in $H\alpha$.

In Chapters 3 and 4, SFRs are calculated from NUV luminosities derived from GALEX fluxes. Its sensitivity to recent starbursts makes it ideal to measure star formation that has been recently triggered during a merger. Also the GALEX fields overlap well with SDSS, so NUV data is readily available for our SDSS close pairs sample. Using $r$-band photometry from SDSS, we plot NUV-$r$ CMRs like Yi et al. (2005) and Schawinski et al. (2007a) to learn about the star formation distribution of our sample.

**UV Upturn**

Figure 2.4 shows the composite spectrum of giant elliptical galaxy NGC 4552 and illustrates a bump in the spectrum known as the UV upturn. When present, the UV upturn lies between the Lyman limit and 2500Å. Much work was devoted to finding out whether this bump is caused by hot young stars, or if old low mass stars could produce such UV emissions (e.g. Code & Welch, 1979; Greggio & Renzini, 1990; Brown et al., 1997; Greggio & Renzini, 1999; Yi & Yoon, 2004; Yi et al., 2005; Yi, 2008).

![Figure 2.4](https://example.com/figure2.4.png)

**Figure 2.4:** Excerpted from Yi (2008), originally used by Yi, Demarque & Oemler (1998): The composite spectrum of the giant elliptical galaxy NGC 4552 shows a classic example of the UV upturn. The mosaic spectrum is originated from HUT (FUV), IUE (NUV), and ground-based telescope (optical).
It is now generally accepted that old core HB stars are the source of the UV upturn; but these are easily swamped by even a tiny amount of star formation, therefore the impact from the UV upturn on our work in later chapters is negligible.

2.3 Classifying Galaxy Populations

When analysing large samples it is necessary to classify and group galaxies with similar properties together. Galaxy classes are loosely defined and often refer to multiple properties; including morphology (e.g. spiral, elliptical, bulge fraction, direction of rotating spiral arms etc.), star formation rate, stellar-kinematic behaviour, composition/metallicity etc. We are often limited by the means by which we classify galaxies; for example, if we study them purely by morphology we may miss details of other important properties such as metallicity and star formation rate. Hubble types are intuitive and are the most common classification types; however, they are subjective to the observer, and are not applicable for \( z > 1 \) when it is harder to see detailed features such as spiral arms.

2.3.1 Automated Classification

**Brightness Profile Fitting** Brightness profile fitting allows us to estimate morphological parameters of merger progenitors without having to visually inspect the galaxy. In brightness profile fitting, a linear combination of the elliptical nature of a galaxy’s brightness and its spiral nature is used to infer whether it is an early type or late type galaxy. The SDSS fracdev parameter gives the best fit (in terms of the fraction of light fit) to a de Vaucouleurs profile in each of the 5 optical bands. This gives the extent of the elliptical nature of the galaxy; a pure de Vaucouleurs profile fit would have fracdev value 1 and objects with fracdev 0.7-1 are often classified as elliptical. For the lowest fracdev values, \( \sim 0-0.5 \), the profile has the maximum deviation from an elliptical and the galaxy is classified as spiral.

An interesting correlation has been found between Hubble type and concentration parameter \( C \), defined as the ratio \( R_{90}/R_{50} \) where \( R_{90} \) and \( R_{50} \) are the 90% and 50% \( r \)-band Petrosian radii respectively (Shimasaku et al., 2001; Strateva et al., 2001). Elliptical galaxies generally have \( C \sim 5.5 \) and spirals generally have \( C \sim 2.3 \). Strateva et al. (2001) suggest that morphological classifications can be made by taking \( C > 2.6 \) to be early type and \( C < 2.6 \) to be late type galaxies. Shimasaku et al. (2001) differ slightly and take \( C = 3 \) to be the critical value. Figure 2.5 shows the correlation that Shimasaku et al. found between concentration and Hubble type.
They estimate that 15-20% of galaxies would be misclassified as either early-type or late-type using their system of automated classification compared with visual classification. This high number of misclassifications is a clear disadvantage in using automated methods to conduct morphological classifications. However, advantages are that automated methods take less time than visual inspection and can classify much larger samples; and they also yield an unbiased classification.

**Figure 2.5:** Excerpted from Shimasaku et al. (2001): Correlation of concentration index with visually classified Hubble type, $T$, for 426 galaxies with $R50 \geq 2''$. Here $C$ is the inverted concentration index $R50/R90$ measured in the $r'$ band. $T$ correlates to morphological type as follows: $T < 0.5$ -Early-type; $0.5 \leq T < 1.5$ -SO; $1.5 \leq T < 2.5$ -Sa; $2.5 \leq T < 3.5$ -Sb; $3.5 \leq T < 4.5$ -Sc; $4.5 \leq T < 5.5$ -Sdm and $5.5 \leq T$ -Im.

**CAS** Conseilice (2003) comments that most of the current classification methods were not based enough on the most important physical features. Instead of classifying galaxies by shape or colour, he suggests that we classify them by the following three features;

- concentration of stellar light (C)
- asymmetric distribution (A)
2.3 Classifying Galaxy Populations

- clumpiness (S)

The concentration index, $C$, is similar to that used by Shimasaku et al. (2001) and Strateva et al. (2001). It is defined such that

$$C = 5 \times \log\left(\frac{r_{80\%}}{r_{20\%}}\right)$$

(2.3)

for Petrosian radii $r_{80}$ and $r_{20}$ within 1.5 times the Petrosian inverted radius at $r$.

To calculate the Asymmetry index, the rotated object image is subtracted from the original object image and the residuals are then compared with the flux from the galaxy before rotation. The clumpiness index, $S$, is the ratio of light contained in high-frequency structures to the total light in the galaxy; this index is usually $\sim 0$ for ellipticals. Conselice (2003) explains that the resulting three-dimensional CAS volume can then be used to approximate the past formation history and the merger status of the galaxy.

The CAS classification system is a useful tool, particularly for intermediate redshift classifications since morphological features like spiral arms are harder to determine visually for higher redshift, and often lower resolution, images (the Hubble classification method is not applicable for $z > 1$). CAS is model-independent, convenient for high-redshift classification and also permits measurements on a continuous scale; whereas Hubble classifications only allow galaxies to be binned into discrete categories (spiral, barred spiral, elliptical etc.).

**Automated Merger Classification** Not only can the CAS system be used to classify individual galaxies, it can also provide a useful model-independent method by which to identify major mergers. The Asymmetry parameter from CAS identifies galaxies with a strongly asymmetric distribution of (rest-frame optical) stellar light; i.e. galaxies that are likely to have recently merged.

Conselice (2006) proposes that galaxies should ideally be classified by (i) mass, (ii) star formation and (iii) merger activity; since these three properties are responsible for the physical state of a galaxy and unveil evolutionary traits. These can be measured using CAS. A merger index, $I$, is used to identify interacting galaxies; it is defined as the ratio of the HI line width at 20% of maximum level and at 50% of maximum level in the 21cm line profile. During interactions HI gas is disturbed, leading to a shallow rise or wings in the HI profile (as opposed to a relatively narrow, or unaltered profile for non-interacting galaxies) (Conselice, Bershady & Gallagher, 2000). Thus the index $I = W_{20}/W_{50}$ increases for interacting galaxies; with $I > 1.5$ interpreted as a recently interacting galaxy. This method is limited
2.3 Classifying Galaxy Populations

to galaxies with HI, and is therefore mostly applicable to spirals. Conselice (2006) reports that the merger index, $I$, is a more robust merger identifier than the asymmetry index. This is due to its sensitivity to interacting galaxies right through to the merging stage, whereas the asymmetry index is tailored to identify galaxies that have already merged.

Another automated method for identifying mergers is to determine galaxies which are located close together on the celestial sphere, and are also located close together in the line-of-sight direction; i.e. those which have a small angular separation and small recessional velocity difference. Spectra must be available for this method of classification. Patton et al. (2002) used a projected separation of $20h^{-1}$kpc (which translates into $\sim$30kpc) and a difference in recessional velocities of 500km s$^{-1}$ (which translates to $\Delta z \sim 0.0017$ at low redshifts). At very low redshifts, peculiar velocities can dominate over the cosmic expansion velocity, so the recessional velocity constraint must be set appropriately to account for this effect and minimise contamination from false pairs.

Patton et al. justify their constraints theoretically and observationally, since at least half of their sample of Southern Sky Redshift Survey 2 (da Costa et al., 1998) galaxy pairs with this criteria showed morphological signs of interactions. We follow Patton et al.’s constraints to extract our close pairs sample in Section 3.1.1.

2.3.2 Classification by Colour

We can use colours from broadband photometry to classify galaxies. In Chapters 3 and 4 we use NUV-$r$ colours as an indicator of recent star formation taking place in galaxy mergers. Optical colours can be used to classify QSOs (e.g. Sesar et al., 2007) and in Chapter 6 we test photometric features for QSO classification against classification models that use only optical colours.

Late-type spiral galaxies are generally optically bluer than early-type ellipticals; this is because they are predominantly star forming and early-types are thought to lack the molecular gas needed to fuel star formation (Baum, 1959; Sandage & Visvanathan, 1978; Bower, Lucey & Ellis, 1992). Larson (1974) attributed this discovery to the mass-metallicity relation, where more massive galaxies with higher escape velocities retain interstellar metals because they have a deeper gravitational potential well and are observed as redder because of their high metallicity fraction. Age is also found to be a factor in this reddening (Kaviraj et al., 2005).

Strateva et al. (2001) identified a strong correlation between colour and morphological type with galaxies imaged by the SDSS when they found a bimodal $u-r$
colour distribution where early type (E, S0, Sa) galaxies tend to be found in the red region and late type (Sb, Sc, Irr) galaxies tend to be found in the blue region; this bimodal distribution is independent of magnitude. Bell et al. (2004) found a similar bimodal colour distribution up to $z \sim 1$ using data from the 0.78 deg$^2$ survey COMBO-17; they suggest that this bimodality may be used to define optically red galaxies as early-type and blue galaxies as late-type for any redshift.

**Colour-Magnitude Relation:** The colour-magnitude relation (CMR) allows us to study star formation activity and to constrain models of galaxy evolution. In Figure 2.6 we can see a bimodal density distribution in the colour of SDSS galaxies. We see separate distributions for redder and bluer galaxies (in $u-r$) with a peak for low luminosity, blue galaxies and a peak for high luminosity, red galaxies. Notice the dotted lines in the right plot showing galaxies of similar mass; the critical mass $M_{\text{crit}} = 3 \times 10^{10}M_{\odot}$ provides a crude line of separation, where most galaxies with $M_* > M_{\text{crit}}$ are early-type and galaxies with $M_* < M_{\text{crit}}$ are late-type (Dekel & Birnboim, 2006). The sparsely populated region between the two peaks is sometimes known as the *green valley*.

![Figure 2.6](image-url): Excerpted from Baldry et al. (2004): Colour-magnitude distributions. (a): Observed bimodal distribution, corrected for incompleteness. The contours are on a logarithmic scale in number density, doubling every two levels. The dashed lines represent the colour-magnitude relations of the red and blue sequences. (b): Deconvolved and parametrised distributions. The solid contours represent the red distribution and the dashed contours represent the blue distribution. The dotted lines represent galaxies that have similar stellar masses, near the midpoints of the transitions.
2.3 Classifying Galaxy Populations

Bower, Lucey & Ellis (1992) compared the optical CMR for the Virgo and Coma galaxy clusters and found that they have similar distributions. They found very small scatter in colour-colour and colour-velocity dispersion diagrams for early-type galaxies in the Virgo and Coma clusters and inferred that the stellar populations within these clusters were equivalent. They suggested that it is likely that the optical CMR has a universal form for all clusters and as a result can be used to determine their relative distances to $\sim 20\%$ accuracy; an idea first proposed by Sandage (1972).

The sensitivity of UV filters has allowed CMRs to be extended to include the UV waveband. In the last decade, NUV-optical CMRs have been employed (e.g. Yi et al., 2005; Kaviraj et al., 2007; Schawinski et al., 2007a) and detections of recent star formation have been permitted in some early-type galaxies where optical CMRs would have classified them as not star forming. NUV-$r$ CMRs are particularly useful for classifying galaxies with recent star formation.

Classification by colour has the advantage that photometry is provided in multiple filters for a large number of local galaxies and it is easily accessible (e.g. SDSS, GALEX and Pan-STARRS). However, the main disadvantage is that observations are sensitive to dust obscuration, especially in star forming regions where there are dust clouds, and these need to be adequately corrected for.

NUV-$r$ provides a quick and simple way to identify recently star forming galaxies in our large sample. However, it must be recognised that studies using colour alone are limited; we also use NUV luminosity-derived specific star formation rates for a quantitative measure of recent star formation.

2.3.3 Classification by Spectra

In *A Spectral Classification of Galaxies*, Morgan & Mayall (1957) classified 47 local bright galaxies by estimating their spectral types using spectra from the Mount Wilson, Palomar and Lick Observatories. They discovered that the spectral types dominant in galaxies are correlated to some of their morphological parameters. Stellar systems with spectra dominated by A-type stars tend to have a minor central concentration of light; whereas spectra dominated by F-type stars implies a larger central concentration and K-type dominated spectra systems have the highest central concentration. Since the work of Morgan & Mayall, much work has been devoted to the classification of galaxies by spectra (e.g. Dressler & Gunn, 1992; Sodré & Cuevas, 1994; Connolly & Szalay, 1999; Yip et al., 2004; Ascasibar & Sánchez Almeida, 2011; Karampelas et al., 2012).
2.3 Classifying Galaxy Populations

Galaxy spectra are a combination of the spectra of billions of stars within the galaxy, and therefore the combined spectral energy distribution tends to be quite flat. In contrast, certain spectral features can be very prominent. The 4000\AA\ break is due to the absorption of high energy radiation by metals in the stellar atmospheres causing an increase in opacity just below 4000\AA; it is the accumulation of a large number of spectral lines in a narrow wavelength region (Bruzual, 1983; Balogh et al., 1999; Yu et al., 2013). Impressive correlations have been shown between the 4000\AA\ spectral index and morphological type (Hamilton, 1985; Sodré & Cuevas, 1994) and it has been used to place constraints on the mean stellar ages of galaxies and the stellar mass fraction formed in recent bursts (Kauffmann et al., 2003b; Kriek et al., 2011). In Section 2.4 we show that intensity ratios of pairs of strong spectral emission-lines (such as the Balmer lines H$_\alpha$ and H$_\beta$, and the forbidden lines [NII] and [OIII]) can be used to separate starburst galaxies from AGN.

Kinematic Classification: The advent of integral-field spectroscopy has provided a new avenue by which to classify galaxies, and in particular early-type galaxies, according to their stellar-kinematic behaviour. The SAURON and ATLAS$^{3D}$ surveys (Emsellem et al., 2007, 2011) have led to a paradigm change whereby early-type galaxies are now best divided into fast- and slowly-rotating stellar systems. All classical lenticular galaxies fall in the fast-rotating class, as do a good fraction of elliptical galaxies. On the other hand, truly slowly-rotating systems are relatively rare, making up only 13\% of the early-type galaxy population and being generally fairly massive. Cappellari et al. (2013) note how the structural and kinematic properties of fast-rotators could connect to those of spirals, which suggests an evolutionary link between them. For practical reasons, our work will not utilise kinematic classification.

Spectroscopic classification techniques are advantageous since galaxies can be classified out to higher redshifts than we can visually classify, and spectra are insensitive to the effects of dust attenuation or obscuration (unlike classification by colour). However, the main disadvantage to spectroscopic techniques is that building spectroscopic samples is expensive and time consuming. As well as proving useful for galaxy classification, spectral features allow us to classify stellar types. In Chapter 6 we use spectra where they are available to distinguish QSOs from stars and galaxies in our Pan-STARRS sample.
2.3 Classifying Galaxy Populations

2.3.4 Visual Classification

Ideally, all galaxies would have their images classified visually to determine morphology, merger status and other interesting features such as spiral arms or a bulge. In Section 3.3 we visually inspect our close pairs sample and discard non-merger contaminants. It is also useful for spectra to be checked visually, and in Chapters 5 and 6 we rely on visually classified spectra for Pan-STARRS objects to determine whether they are galaxies, QSOs or stars (and their stellar type).

Visual classification allows exceptional cases to be identified and flagged in ways that computer software cannot always achieve. The SDSS pipeline often incorrectly processes separate extended individual sources into multiple objects, or mistakes bulge dominated spirals as having elliptical morphology, and checks must be done by eye to ensure an uncontaminated sample. However, visual inspection is a very time consuming process and, given the amount of galaxies imaged by the SDSS (more than fifty million galaxies), it is usually inefficient for larger surveys.

The Galaxy Zoo project has been very successful at utilising public interest by inviting members of the public to classify merging galaxies online (Lintott et al., 2008; Bamford et al., 2009; Lintott et al., 2011). This project has recruited over 250,000 people so far to visually classify $\sim 10^6$ galaxies from the SDSS into mergers, non-mergers and morphological types. Since its launch in July 2007, Galaxy Zoo has evolved and now allows more detailed properties of merging galaxies to be classified; it now considers features such as galaxy smoothness, roundness and the presence of disks. Once these galaxies have been classified numerous times by various Galaxy Zoo volunteers, a weighted-merger-vote fraction, $f_m$, is assigned to the galaxy system as a measure of confidence that a merger is taking place.

$$f_m = \frac{W n_m}{n_{e,s,b,m}},$$

where $W$ is a weighting factor that describes the classification accuracy of the volunteers who have participated, $n_m$ is the number of merger classifications, and $n_{e,s,b,m}$ is the total number of classifications for the object; the options were elliptical (e), spiral (s), star/bad image (b) or merger (m).

Darg et al. (2010b) extracted a population of 3003 galaxies pairs with $f_m > 0.4$ and spectroscopic redshift $0.005 < z < 0.1$ (with absolute magnitude $M_r < -20.55$) and found that the spiral-to-elliptical ratio of galaxies in mergers is approximately a factor of 2 higher than the spiral-to-elliptical ratio for the global population. Darg et al. suggest that the reason that a higher spiral-to-elliptical ratio is observed in mergers is because of the longer time-scales of detectability of spirals in mergers.
compared to ellipticals in mergers. Another recent discovery resulting from Galaxy Zoo classifications is that spiral galaxies do not rotate in any preferential direction. Jimenez et al. (2010) find that neighbouring spirals with similar star formation histories often spin in the same direction if most of their stars were formed more than 10 Gyr ago. They suggest that this is because these galaxies were formed in the same dark matter halo filament at roughly the same time.

Visual classification is without a doubt the most effective method of classification for low redshift galaxies, and thanks to Galaxy Zoo it is now a relatively quick process (taking on average 45 million classifications over an 8 month period; i.e. \( \sim 8,000 \) classifications per hour). No artificial intelligence algorithm to date has been as effective at recognising patterns such as those required to classify galaxy mergers and types as the human brain; human judgement and reasoning is crucial in determining exceptional cases and subtle features such as dim spiral arms. However, accurate morphological classifications are only considered to be possible for galaxies of average size and brightness up to \( z \sim 0.5 \) since the effects of seeing and cosmological surface brightness dimming (this becomes an issue at \( \sim (1 + z)^4 \)) limit our observations; thus we are limited to classifying low redshift galaxies (Sodrè & Cuevas, 1994). As we move to higher redshift galaxies only the most massive bright galaxies can be seen clearly enough for a visual classification to be made and so classifications become biased.

2.4 BPT Analysis

In 1981, Baldwin, Phillips & Terlevich (BPT) showed that certain combinations of emission-line spectra can be used to categorise galaxies according to their sources of ionisation. Plotting ratios of emission-lines allows us to determine the predominant mechanism of excitation in galaxies, e.g. HII regions (that are photoionised by O and B stars), photoionisation by power-law continuum source, or photoionisation by shock-wave heating. These are known as BPT plots (Baldwin, Phillips & Terlevich, 1981; Kewley et al., 2001; Kauffmann et al., 2003a; Kewley et al., 2006; Trichas et al., 2010; Kalfountzou et al., 2011). The current form of these diagnostic diagrams uses revised line ratios which were chosen by Veilleux & Osterbrock (1987) to fully exploit internal differences between classes. The lines in each ratio were also chosen to be very close in wavelength so as to minimise the effects of the reddening correction and errors in the flux calibration.

A BPT analysis allows us to determine the predominant mechanism of excitation in our sample galaxies. Intensity ratios of pairs of strong emission-lines are
used to separate starburst galaxies from AGN. The regions in BPT plots are defined by modelling gas that has been photoionised by a central AGN, young stars, or shock-wave heating, where the hardness of the ionising continuum can distinguish photoionisation by OB-stars from other excitation sources. For example, low values of $\text{[OIII]}/\text{H}$/$\beta$ correspond to photoionisation from HII regions and high values correspond to photoionisation from Seyfert nuclei.

Figure 2.7: Excerpted from Kauffmann et al. (2003a): BPT plot with emission-line flux ratio $\text{[OIII]}$/H$\beta$ versus the ratio $\text{[NII]}$/H$\alpha$ for 55,757 galaxies where all four lines are detected with S/N $> 3$. The dotted curve shows the demarcation between starburst galaxies and AGN as defined by Kewley et al. (2001). The dashed curve shows Kauffmann’s revised demarcation.

An example of a BPT plot is given in Figure 2.7, where log([NII]/H$\alpha$) is plotted against log([OIII]/H$\beta$) (excerpted from Kauffmann et al. (2003a)). Emission-line ratios are used to separate AGN and non-AGN galaxies. Kewley et al. (2001) defined a line using the following parametrisation that separates likely AGN and starburst galaxies;
This line gives the dotted curve which we see in Figure 2.7. Their parametrisation was modified by Kauffmann et al. (2003a) and in Figure 2.7 they use the following expression (shown by the dashed line) to identify likely AGN galaxies:

$$\log([\text{OIII}]/H\beta) = \frac{0.61}{\log([\text{NII}]/H\alpha) - 0.47} + 1.19. \quad (2.5)$$

Kauffmann et al. (2003a) used a sample of $0.02 < z < 0.3$ galaxies from the SDSS with signal-to-noise ratio greater than 3 for H$\alpha$, H$\beta$, [NII] and [OIII] emission-lines to test AGN host properties. Type 1 AGN galaxies were eliminated and $\sim 40\%$ of the 55,747 remaining galaxies are believed to be Type 2 AGN (N.B. with Type 1 AGN we view the broad line emitting region of the AGN directly whereas with our perspective of a Type 2 AGN, we see the obscured dusty torus region (see Section 1.3.1)). Kauffmann et al. found that out of this sample of emission-line galaxies at least 80% with $M_* > 10^{11}M_\odot$ are (Type 2) AGN and the AGN fraction for $M_* < 10^{10}M_\odot$ rapidly decreases with decreasing mass. Different star formation trends can be seen for high mass galaxies than for low mass galaxies, and given the high proportion of high mass galaxies that have AGN activity, it is possible that this difference can be attributed to processes such as AGN feedback that may quench star formation. It was shown by Sarzi et al. (2010) that fast shocks are unlikely to be a primary source of ionization in early-type galaxies.

The main advantage of using emission-line ratios is we can avoid dependency issues with stellar age, star formation history and dust attenuation that must be addressed when working with individual emission-lines (e.g. Brinchmann et al., 2004).

## 2.5 Galaxy Environments

Research into the spatial distribution of galaxies shows that galaxies are often observed to cluster together (e.g. Magliocchetti & Porciani, 2003; Cole et al., 2005; Ribeiro et al., 2009, using the 2dF Galaxy Redshift Survey). In a hierarchical ΛCDM model, clustering follows naturally since galaxies would have formed in dense dark matter halo regions within the cosmic web and would further cluster together during gravitational interactions. Galaxy environments describe the number of galaxies gravitationally clustered together. In field environments galaxies exist independently.

There are two popular methods used to classify galaxies as field galaxies or
2.5 Galaxy Environments

group/cluster members in large scale surveys such as the SDSS and 2dFGRS; methods which count galaxies (e.g. Schawinski et al., 2007a), and group-finding methods such as the friends-of-friends algorithm (Huchra & Geller, 1982) and the Yang et al. (2005) halo-based galaxy group finder. The friends-of-friends group-finding algorithm focuses on one galaxy and checks the 3-dimensional space around it (using angular separation and recessional velocity to determine the line-of-sight separation) to identify nearby galaxies. If a nearby galaxy is found, the companion galaxy is recorded and then the algorithm scans for other galaxies in the vicinity of the companion galaxy and records these together as a group; otherwise, the initial galaxy is listed as isolated.

Yang et al.’s halo-based method is an algorithm that assumes a ΛCDM model and consists of the following 5 steps;

(i) The friends-of-friends algorithm is used with a small span around each anchor galaxy in order to find the centre of potential groups. Since central galaxies are the brightest in groups, galaxies are identified as the potential group centres if they are the brightest in a cylinder of radius $1h^{-1}\text{Mpc}$ and velocity depth $\pm 500\text{km s}^{-1}$.

(ii) The total luminosity, $L_{\text{total}}$, of potential groups is estimated using the galaxy luminosity function from Norberg et al. (2002).

(iii) Using $L_{\text{total}}$ and a model for the group mass-to-light ratio, properties such as the halo mass, halo radius, virial radius and virial velocity are derived (NB the computed halo mass is sensitive to the mass-to-light ratio assumed).

(iv) For each galaxy, a loop is run over all groups and the distance between the galaxy and each potential group centre is computed. The galaxy is assigned a probability of belonging to each dark matter halo; if the galaxy could potentially be assigned to two groups then it is defaulted to the group that it has the highest probability of belonging to. If all galaxies in two groups can be identified with a single group then the two groups are merged into one.

(v) Once members have been assigned to each group, the centre is recomputed and the algorithm skips back to step (ii). This process continues until the same outcome is achieved upon each run through.

The halo mass value can be used as a parameter to describe the local environment of galaxies. Field pairs refer to an isolated close pairs system as opposed to individual galaxies.
2.6 Summary

In this chapter we introduced SDSS, GALEX and Pan-STARRS; these are the instruments from which our data was gathered. We described relevant technical specifications and limitations for each telescope. Some of these limitations have implications for our research. Fibre collisions in the SDSS lead to only $\sim 30\%$ of galaxies, and hence only $\sim 30\%$ of close pairs being detected; we will discuss this incompleteness as pertaining to our sample in Chapters 3 and 4. SDSS only targets galaxies with $r$-band Petrosian magnitude $r < 17.77$, with the main galaxy sample having median redshift 0.1, and so our close pairs sample is a low redshift sample.

GALEX faces various constraints because it has very sensitive detectors and it must avoid bright sources; this leads to an observing time of only $\sim 10\%$ of the year for most targets. In Chapter 3 we crossmatch our SDSS close pairs sample with GALEX data; a significant number of SDSS pairs do not have GALEX NUV or FUV measurements.

Pan-STARRS has the ability to repeatedly image and store data for an enormous number of objects in the local sky. We use the numerous exposures recorded for each object to form light curves in Chapter 5 and we look in detail at the statistical properties of these light curves in Chapter 6. We also use spectra where available from the TDSS pilot survey to accurately determine whether these Pan-STARRS objects are stars, galaxies or AGN.

We presented various methods of measuring star formation within galaxies and discussed the advantages and disadvantages for each method. In Chapters 3 and 4 we calculate star formation rates from NUV luminosities for the first time in a close pairs study. This provides a different perspective from which to study star formation in close pairs since most previous studies have used optical or IR continuum measurements or emission-lines. We also use NUV-{$r$} colours as an indicator of recent star formation in our sample.

We looked at various ways to classify galaxies; it is important to be able to adequately categorise objects in large samples. Galaxies can be classified using brightness profile fitting into Hubble types such as spiral and elliptical, and automated methods such as the CAS system can identify mergers. However, automated methods can be unreliable. The most reliable classification method is visual inspection, and the Galaxy Zoo project was very successful at doing this for a large number of SDSS main galaxy sample objects. Spectral and colour properties are also used to classify galaxies. We will use an automated method to extract close pairs, and then conduct a visual classification to check our sample.
Tools such as BPT analyses and the galaxy environment classification algorithm from Yang et al. (2005) can be used to categorise galaxies according to their sources of photoionization and by the environment density in which they are located. We employ each of these methods in the next chapter.
Chapter 3

The Properties of Close Pairs

This chapter is based on the first half of the following paper:


Overview

In this chapter and the next, our research into close pairs of galaxies is presented: recent star formation is examined in close pair galaxies at various stages of the merger process. In Chapter 1, we saw that mergers are fundamental to the standard hierarchical paradigm of galaxy formation and evolution. Models predict that they produce intense star formation episodes (Mihos & Hernquist, 1996; Cox et al., 2008; Bournaud et al., 2011), contributing to the build-up of stellar mass and black holes (Sanders & Mirabel, 1996; Hopkins et al., 2005; Debuhr, Quataert & Ma, 2011), and alter the morphological mix of the Universe (Toomre & Toomre, 1972; Kauffmann, White & Guiderdoni, 1993; Mihos, 1995; Kauffmann & Haehnelt, 2000). Past studies from the previous generation of astrophysical instruments found that interactions could enhance star formation in close pair galaxies (Larson & Tinsley, 1978; Joseph et al., 1984; Lonsdale, Persson & Matthews, 1984) and potentially trigger AGN activity (Smith et al., 1986; Hutchings, 1987; Sanders et al., 1988). Since the emergence of large-scale galaxy surveys, such as the SDSS and GALEX (see Section 2.1), robust statistical analyses can now be conducted on galaxy-galaxy interactions, resulting in extensive studies of mergers over recent years.

Substantial observational data now serves to complement results from simulations, which claim that close pair interactions and mergers cause the instability needed for gas clouds to collapse and produce starbursts (e.g. Kauffmann &
3.1 Sample Description

Haehnelt, 2000; Tissera et al., 2002; Cox et al., 2006; Di Matteo et al., 2007; Tonnesen & Cen, 2012). Spectroscopic and photometric evidence from large scale surveys such as IRAS (Kennicutt et al., 1987) and the SDSS (Ellison et al., 2008), and high resolution instruments like the HST (Patton et al., 2005) have bolstered such claims. Morphological asymmetry effects are also a trademark signature of close pair systems; with most of the galaxies in the Arp catalogue (Atlas of Peculiar Galaxies, Arp, 1966) showing signs of recent tidal interactions (Larson & Tinsley, 1978) and ∼40% of close pair systems expected to show asymmetry effects (Patton et al., 2005).

Interaction-induced effects are expected to depend on properties of the progenitor galaxies and external pair-properties of close pair systems. Such properties may include the environment in which the merger is taking place, pair separation, the type of galaxies merging (e.g. morphology, mass), and central galactic processes within the progenitors; such as feedback from AGN activity. In Section 1.4 we reviewed previous studies which investigated the effects of these properties as a function of star formation in close pairs. We now build on this previous work by using a new measure of star formation that has not been used before in close pairs studies; we use NUV-luminosity derived specific star formation rates. This provides a new perspective from which to study galaxy close pairs.

We begin this chapter by introducing our close pairs sample; we describe how our close pairs were extracted from the SDSS database and then cross-matched with GALEX data. We then show how star formation rates were derived and AGN were classified, and how various properties such as mass and environment were quantified. Then in Chapter 4 we use this sample and the properties we have derived to investigate star formation and AGN activity in close pairs with various mass and environment parameters.

3.1 Sample Description

3.1.1 Extracting the Close Pairs

Our close pairs catalogue is extracted from the SDSS Data Release 7 (DR7) main galaxy sample (see Section 2.1.1; Fukugita et al. (1996); Gunn et al. (1998); York et al. (2000)) by using an automated procedure to seek galaxies with small angular separation and a small recessional velocity difference (i.e. low separation in the line-of-sight direction). We follow Patton et al. (2002) who suggest a projected separation of $20h^{-1}\text{kpc}$ (which translates into $\sim30\text{kpc}$) and a difference in recessional velocities...
of 500km s\(^{-1}\) (which translates to \(\Delta z \sim 0.0017\) at low redshifts). Automated methods of close pairs extraction were discussed in Section 2.3.1.

Since the sample is extracted from a low redshift survey and magnitude-limited survey, we do not impose a maximum redshift constraint. In our emission-line analysis we only use galaxies with [NII] and H\(\alpha\) emission-line signal-to-noise ratio >3, so this spectroscopic constraint will limit our sample to lower redshifts. Spectra for objects within 55\(^{\prime\prime}\) can only be obtained when observed in overlapping tiles and so fibre collisions lead to incompleteness within our sample. Incompleteness due to fibre collisions, magnitude limits, large peculiar velocities, minor mergers (low mass galaxies are often below the SDSS spectroscopic limit of \(r < 17.77\)) etc. result in only \(\sim 30\%\) of all close pairs being detected (Darg et al., 2010b; Strauss et al., 2002; Blanton et al., 2003). Since the pairs that are detected are drawn randomly from a homogeneous sample, our catalogue constitutes a representative (yet incomplete) sample of low redshift close pairs.

We cross match our close pairs catalogue with the GALEX GR4/GR5 database (see Section 2.1.2) to get NUV (1770-2730\(^{\AA}\)) and FUV (1350-1780\(^{\AA}\)) measurements for our galaxies. All systems where more than one SDSS object (including the primary object) is within 5\(^{\prime\prime}\) of a GALEX object (5\(^{\prime\prime}\) being the GALEX resolution) are removed. Simard et al. (2011) find that photometry for close pairs from the standard SDSS pipeline is sometimes poor for pairs with projected separation \(\lesssim 20\text{kpc}\), and in these cases the deblending process can be unreliable: setting the 5\(^{\prime\prime}\) GALEX constraint also serves to minimise this SDSS deblending issue for low separation close pairs.

Some of our SDSS close pairs did not have counterpart NUV or FUV measurements from GALEX so the close pairs were split into two samples; one for optical-NUV studies (using the crossmatched SDSS/GALEX sample) and one for optical studies using purely SDSS data (this was used in our AGN analysis in Section 4.3 since it is a bigger sample and only spectra was required, not NUV photometry). The close pairs sample consists of 6668 galaxies with SDSS data and 2902 galaxies with both SDSS and NUV photometry. The median redshift is \(z \sim 0.07\). A \textit{wide pairs} sample with projected separation 30-150kpc and \(\Delta z \sim 0.0017\) was also extracted to be used as a control sample with which to compare the close pairs. The wide pairs sample consists of 34,294 galaxies with SDSS data and 19,202 galaxies with both SDSS and NUV data.

At very low redshifts peculiar velocities can dominate over the cosmic expansion velocity, and deblending issues become more significant, so we impose a minimum redshift constraint \(z \geq 0.01\) to reduce contamination. The only way to be sure that
the close pairs identified by the automated procedure are actually interacting is to
re-classify visually, which we discuss in Section 3.3.

3.1.2 K-correction and Extinction Correction

KCORRECT V4.2 (Blanton & Roweis, 2007) was used to calculate K-corrections for
the galactic-extinction-corrected SDSS and GALEX apparent magnitudes. The code
fits restricted spectral energy distribution models, from templates created by the
high resolution stellar population synthesis methods of Bruzual & Charlot (2003),
to the optical and UV photometry for our sample. Results are therefore dependent
upon the accuracy of the fit to the underlying spectral-energy distribution and as-
sume that the templates are adequately representative of the true spectral energy
distribution. Templates fitted to GALEX and SDSS photometry using this method
were tested by Blanton & Roweis (2007) and found to be effective for these data
sets.

The NUV-band is extremely sensitive to interstellar reddening, more so than
optical and IR photometry (Meurer, Heckman & Calzetti, 1999; Pannella et al.,
2009). We correct our flux measurements for interstellar extinction using the Balmer
decrement from the SDSS spectrum. Then absolute magnitudes were derived as
follows;

\[ M(\lambda) = m(\lambda) - 5\log(D/1\text{pc}) - 5 - \text{k-correction}(\lambda) - A_{\text{ISM}}, \]

for each photometric band, \( \lambda \), where \( \lambda = \text{NUV, FUV, } u, g, r, i, z \) and \( m \) is the
apparent magnitude. \( A_{\text{ISM}} \) is the intrinsic extinction in the observed galaxy; we use
the interstellar extinction calculated from the Balmer decrement in Section 3.1.5.
\( 5(\log_{10}(D/1\text{pc}) + 1) \) is the distance modulus; this transforms the magnitudes of all
galaxies to their expected (absolute) magnitude at a distance of 10 parsecs.

Figure 3.1 (bottom left) shows a \( u - r \) colour-magnitude diagram using our
absolute magnitudes after calculating the K-correction and extinction correction
terms in Equation 3.1 for our optical close pairs sample. Figure 3.1 (top) shows an
NUV-r colour-magnitude plot for the K-corrected close and wide pairs samples. The
horizontal dotted lines are median lines for these two distributions. Notice that the
close pairs sample median is significantly bluer in NUV-r than the wide pairs sample
median. From previous close pairs research (see Section 1.4.2), we expect close pairs
to exhibit more recent star formation, and hence the bluer NUV-r distribution is
not a surprise as this indicates that more star formation is taking place.

The coloured blocks at the left of the horizontal median lines in Figure 3.1 (top)
and at the bottom of the vertical median lines in Figure 3.1 (bottom right) are from a process called *bootstrapping*. Bootstrapping takes data points from the original sample and keeps redistributing them randomly to generate new populations from the original data points; it then finds the mean output from each of these new populations. By repeatedly resampling the original data and then taking the mean output, we gain insight into how an overall population might look. This is useful.

**Figure 3.1:** Top: NUV-r colour-magnitude plot for the close pairs sample (red points) and the wide pairs sample (black points). Bottom left: Optical colour-magnitude plot for galaxies from the optical close pairs sample after our K-correction and extinction correction. Bottom right: NUV-r distribution for the close pairs sample (red histogram) and wide pairs sample (black histogram). Coloured blocks on median lines show bootstrapping results.
for finding statistics for small samples. When plotted as in Figure 3.1, the blocks tell us whether or not our results are statistically significant; if the blocks do not overlap then the distributions are thought to be significantly different, as is the case in the plot. We will frequently use bootstrapping blocks on median lines.

### 3.1.3 Calculating Galaxy Masses

Stellar masses were approximated using our absolute magnitudes and the following formula from Wang et al. (2006) (based on Bell et al. 2003):

\[
\log \left( \frac{M_*}{M_\odot} \right) = -0.4[M(r) - 4.67] - 0.306 + 1.097[M(g) - M(r)] + 0.15, \tag{3.2}
\]

where 4.67 is the \(r\)-band solar absolute magnitude from the SDSS. This assumes a Salpeter (1955) stellar IMF with \(dN/dM \propto M^{-2.35}\) and \(0.1M_\odot < M < 100M_\odot\). The median uncertainty in stellar mass values is \(\sim 0.1\) dex, with the maximum uncertainty expected to be \(\sim 0.3\) dex.

![Figure 3.2: Comparison between our mass estimates with MPA masses for both close and wide pairs samples. The dashed line black shows equality between the axes and is shown for reference.](image)

The Max Planck Institute for Astrophysics and John Hopkins University released an online catalogue of stellar masses for SDSS objects; we will refer to this as the MPA catalogue\(^1\). Their stellar masses are obtained from photometric fits

\(^1\)www.mpa-garching.mpg.de/SDSS/DR7/
which estimate the mass inside the spectroscopic fibre (Brinchmann et al., 2004). There is an aperture bias since extended objects are likely to have regions that are not fully imaged spectroscopically; we will come back to this point in Section 3.2.2. They then scale these fits to get estimated total stellar masses. In Figure 3.2 we compare our mass estimates with MPA mass estimates. Our mass distributions for both samples are higher than the MPA mass estimates. This is likely because we calculate our masses purely from photometry, a measure which covers the full galaxy for extended sources, whereas MPA are limited by finite spectroscopic fibres and have to extrapolate to estimate full stellar masses.

Figure 3.3: Top: NUV-r against stellar mass plot (left) for the wide pairs (black) and close pairs (red) samples. Bottom: Stellar mass distribution for the wide pairs (black histogram) and close pairs (red histogram) samples. Dotted lines represent median values, and coloured blocks on median lines show bootstrapping results.

Figure 3.3 shows the stellar mass distribution; the close pairs sample has a
median stellar mass of $10^{10.8}M_\odot$ and the wide pairs sample has a slightly lower median stellar mass of $10^{10.6}M_\odot$. Bundy et al. (2009) and Darg et al. (2010a) find a similar bias towards higher stellar masses (by $\sim 0.2$ dex) for close pairs. Simard et al. (2011) show that photometry for close pairs from the standard SDSS pipeline is sometimes poor for pairs with projected separation $\lesssim 20$ kpc and Patton et al. (2011) suggest that this has led previous authors to incorrectly perceive an extremely red population in close pairs samples. This reddening effect would explain our bias towards higher stellar masses for the close pairs, however, since the difference lies within the 0.3 dex stellar mass error we justify their use.

### 3.1.4 Environment

The close and wide pairs samples were crossmatched with an environment catalogue derived from a halo-based group finder (Yang et al., 2005, 2007). This method primarily uses the friends-of-friends algorithm to find groups (Huchra & Geller, 1982), then approximates the group centre using its brightest member. An initial mass is calculated using the mass-to-light ratio, then the total luminosity of potential groups is estimated using the luminosity function from Norberg et al. (2002). All galaxies are then assigned a probability of belonging to each group’s dark matter halo. The algorithm iteratively assigns each galaxy to its most probable group (merging smaller groups that can be identified as one), updates the assigned group centre, and re-calculates the total luminosity. The dark matter halo mass is then calculated from this finalised characteristic luminosity, and is used to approximate environment density (see Section 2.5).

We crossmatched the close pairs sample with the DR7 version of the environment catalogue from F. C. van den Bosch. This approximates the halo mass in which the galaxy is located using the Yang et al. (2007) method. The halo mass value can be used as a parameter to describe the local environment of the galaxy. The catalogue has the derived dark matter halo mass from the characteristic luminosity of groups from the SDSS main galaxy sample with $0.01 < z < 0.2$ and SDSS completeness parameter $> 0.7$. There are 369,447 entries in the catalogue, and these also include galaxies with ALTO spectra from 2dFGRS. The $\log_{10}$ based halo masses are given in units of $M_\odot$ as discreet positive integers ranging from 10 to 15. We assume that a halo mass value from $10^{10}$ to $10^{13}M_\odot$ describes a local field environment, $10^{13}$ to $10^{14}M_\odot$ a group environment, and $10^{14}$ to $10^{15}M_\odot$ a cluster environment (Kaviraj et al., 2009). Figure 3.4 shows the distribution of the halo mass parameter for the close and wide pairs samples.
3.1 Sample Description

3.1.5 Emission-Line Analysis

A BPT analysis (Baldwin, Phillips & Terlevich, 1981; Kewley et al., 2001; Kauffmann et al., 2003a; Kewley et al., 2006; Trichas et al., 2010; Kalfountzou et al., 2011) allows us to determine the predominant mechanism of excitation in our sample galaxies (see Section 2.4). We use intensity ratios of pairs of strong emission-lines from the OSSY catalogue (Oh et al., 2011); this is a database of publicly available absorption and emission-line measurements for SDSS DR7 galaxies. The emission-line process utilises the Gas and Absorption Line Fitting code (GANDALF; Sarzi et al. (2006)). GANDALF fits stellar population and emission-line templates to the galaxy spectrum simultaneously to separate the stellar continuum and absorption lines from the ionized gas emission.

We plot log([NII]/Hα) against log([OIII]/Hβ) and classify each galaxy as Star-forming, Transition, LINER or Seyfert on the BPT diagram as defined by Schawinski et al. (2007b). See Figure 3.5 for the BPT plots for the close and wide pairs samples, showing in different colours the Star-forming, Transition, LINER and Seyfert regions. Transition objects are defined as those located in the region between the two classification curves of Kewley et al. (2001) and Kauffmann et al. (2003a); shown in Equations 2.5 and 2.6. Star-forming objects are located further to the left of the
Figure 3.5: BPT plots: $\log_{10}([\text{NII}] / \text{H}\alpha)$ is plotted against $\log_{10}([\text{OIII}] / \text{H}\beta)$ for the close pairs (top) and wide pairs (bottom). Objects are classified as Starforming, Transition, LINER or Seyfert depending on the region they inhabit on the BPT diagram.

Kauffmann et al. (2003a) curve on the BPT diagram. Objects to the right of the Kewley et al. (2001) classification curve are separated into LINERs and Seyferts according to the following empirical line from Schawinski et al. (2007a):

$$\log([\text{OIII}] / \text{H}\beta) = 1.05 \log([\text{NII}] / \text{H}\alpha) + 0.45.$$  

(3.3)

Only galaxies with emission-line signal-to-noise ratio $>3$ are used are shown in
3.2 Deriving Star Formation Rates

This is the first time that luminosity-derived specific star formation rates (SSFRs) have been used to study close pairs. We derive SFRs directly from NUV magnitudes, and this gives a measure with which to meaningfully quantify the recent star formation rate (rSFR) in close pairs. First we justify our approach, then we describe how the SSFRs were derived from GALEX fluxes.

3.2.1 Justification for NUV-derived SFRs

Since recently formed stars ($\lesssim 1$ Gyr) are responsible for most of a galaxy’s UV luminosity, rSFR correlates strongly with UV luminosity (Iglesias-Páramo et al., 2006). This requires the assumption of a constant rate of star formation. Salim et al. (2005) find that star formation, i.e. the ratio of current to past-averaged star formation rates, $b$, is approximately constant for low mass galaxies but decreases for higher mass galaxies. They also find a tight correlation between $b$ and NUV-$r$, suggesting that NUV-$r$ colour alone is sufficient to estimate the star formation history of galaxies. Salim et al. (2005) show that GALEX is sensitive to star formation levels as low as $\sim 10^{-3} M_\odot yr^{-1}$.

The conversion between UV flux and SFR can be derived from synthesis models (e.g. Kennicutt, 1998; Iglesias-Páramo et al., 2006). Results normally vary due to the choice of stellar libraries and initial assumptions. It is often appropriate to assume constant SFR over a period much longer than the considered stellar population, i.e. $>8$ Myr (Kennicutt, 1998). Results also vary depending on the assumed IMF. Using a Salpeter (1955) IMF results in an approximately constant luminosity over the range 1500-2800Å.

GALEX’s NUV-band (effective wavelength: 2271Å) allows us to study star formation with approximately an order of magnitude more sensitivity to star formation than optical filters (Yi et al., 2005); making the NUV well suited for studying star formation that has been recently ignited during merging (Yi et al., 2005; Kaviraj et al., 2007; Donas et al., 2007; Bianchi, 2011). However, UV photometry provides a measure of emission from young stars for the full extent of a galaxy. One caveat is the NUV-band is extremely sensitive to interstellar reddening, more-so than optical and IR photometry (Meurer, Heckman & Calzetti, 1999; Pannella et al., 2009). We

the BPT plots. Galaxies with weaker emission-lines are classified as Quiescent. We consider Seyfert galaxies as Type II AGN.
3.2 Deriving Star Formation Rates

correct our GALEX flux measurements for interstellar extinction using the Balmer decrement from the SDSS spectrum.

Spectra collected using fixed apertures results in an aperture bias which needs to be corrected for; these corrections are inherently uncertain. Galaxies can have extremely varied colour profiles, and spectra from only one area can be collected using fixed apertures, leading to different portions being sampled for different galaxies. As a result, SFRs are often based on measurements from a particular region of the galaxy and do not reflect star formation that may take place outside. For example, at the median survey redshift, SDSS spectra only sample approximately one third of the total galaxy light (Brinchmann et al., 2004). Star formation estimates from limited spectroscopic regions may be corrected by utilising photometric images which cover broader areas (e.g. Brinchmann et al., 2004; Salim et al., 2007); the specific star formation rate likelihood distribution for a given set of photometric colours is calculated. A key advantage of deriving SFRs from UV photometry instead of spectroscopy is that a total flux for the full galaxy is measured and as a result SFRs are more likely to be reliable.

3.2.2 Deriving SSFRs

NUV fluxes were first corrected for internal reddening in the observed galaxy via the Balmer decrement, measured using GANDALF (Sarzi et al. (2006); see Section 3.1.5). GANDALF calculates an internal E(B-V) from the emission-lines in the standard way, via the Balmer decrement assuming Case B recombination. This internal E(B-V) likely traces the E(B-V) in the star-forming regions and is used to derive intrinsic NUV fluxes and star formation rates. The dust law from Calzetti et al. (2000) (see also Kaviraj et al. (2007)) allows us to estimate reddening for a given value of E(B-V).

Galactic extinction (i.e. extinction caused by our own galaxy) estimates for each object are taken from the maps of Schlegel, Finkbeiner & Davis (1998). To correct for internal and galactic extinction, we multiply the E(B-V)$_{\text{internal}}$ and E(B-V)$_{\text{galactic}}$ components by 2.751 (according to Schlegel, Finkbeiner & Davis (1998)) to correct our SDSS r-band magnitudes and by 8.2 (according to Calzetti et al. (2000); Kaviraj et al. (2007)) to correct our GALEX NUV magnitudes. We calculate the NUV luminosity of our galaxies as follows:

\[
L_{\text{NUV}} = 10^{-0.4(M_{\text{NUV}}-M_{\odot \text{NUV}})}L_{\odot \text{NUV}},
\]

where $M_{\odot \text{NUV}}$ is the apparent magnitude of the Sun in GALEX’s NUV filter. The
3.2 Deriving Star Formation Rates

$M_{NUV}$ values are our NUV-band K-corrected and extinction corrected absolute magnitudes. These luminosities were then converted into units of bolometric solar luminosity and then our (log$_{10}$) star formation rates were calculated using the following expression from Iglesias-Páramo et al. (2006):

$$\log SFR_{NUV}(M_\odot yr^{-1}) = \log L_{NUV}(L_\odot) - 9.33.$$  \hspace{1cm} (3.5)

Using our stellar mass estimates from Equation 3.2, we divide the NUV photometry derived SFRs by stellar mass for each galaxy to get specific star formation rates (SSFRs).

Figure 3.6: Emission-line luminosity-derived SFRs from the MPA SDSS DR7 catalogue (black) are compared with NUV photometry-derived SFRs for our wide pairs sample (green) and close pairs sample (red). Vertical lines show sample median values. The SFR distributions are normalised.

Figure 3.6 shows the normalised SFR distribution for the close pairs (red line) and wide pairs (black line) samples. As we would expect, our close pairs sample has a higher SFR distribution than the wide pairs sample since the galaxies are interacting more. We compare our results with SDSS SFRs from the MPA-JHU DR7 catalogue (Max Planck Institute for Astrophysics and John Hopkins University)$^2$. Plotted for comparison are the SDSS DR7 SFRs available from the online catalogue (black line) provided by the Max Planck Institute for Astrophysics (MPA). These were derived from emission-line measurements (Gaussian fitted to continuum-subtracted spectra) based on the method of Brinchmann et al. (2004) for non-Starburst galaxies (i.e. AGN, Composite, low S/N SF, low S/N LINER and Unclassifiable), and the model

---

$^2$www.mpa-garching.mpg.de/SDSS/DR7/
3.2 Deriving Star Formation Rates

of Charlot & Longhetti (2001) was used for Starburst galaxies.

The full MPA SDSS DR7 SFR sample contains 775,559 galaxies, but we only compare distributions for objects in our close and wide pairs samples. We would expect this distribution to follow that of our wide pairs sample more than our close pairs sample since the wide pairs act as a control sample. The SFRs derived from emission-lines are generally lower than our NUV photometry-derived SFRs; this is likely an impact of aperture bias. We believe that the MPA SFRs, as well as their mass estimates, have been under-estimated because their measurements are aperture limited.

![Figure 3.7: NUV-r and log_{10} SFR plot before (top) and after (bottom) internal extinction correction.](image-url)
In Figure 3.7 we plot NUV-$r$ colours against $\log_{10}$ SFR for the close and wide pairs samples before our internal extinction correction (top) and after (bottom). Notice that both the close and wide pairs samples are shifted to bluer NUV-$r$ colours.

### 3.3 Visual Classification of Close Pairs

Our catalogue is likely to be contaminated with a subset of objects that are not actually merging and we need to know the size of this population. Should either (or both) galaxies have a large peculiar velocity, the automated procedure may wrongly classify them as close enough to merge. Soares (2007) compares projected separation with spatial separation for a Monte Carlo simulated random sample of gravitationally-bound pairs and finds that over 50% with projected separation $\leq 50\text{kpc}$ actually have spatial separation $> 50\text{kpc}$. Thus, a sample defined as close pairs using projected separation as a proxy for 3-dimensional separation (e.g. Patton et al., 2002; Scott & Kaviraj, 2014) must not be confused with a true mergers sample; we accept that projected separation is merely an approximation to spatial separation, and that the physical separation is likely to be higher.

We cannot assume that the SDSS will have obtained spectra for both close pair galaxies. The SDSS will only target photometric children with $r < 17.77$ and it will often be the case that merging galaxies do not both satisfy this criterion. This happens particularly in minor merger scenarios where the smaller galaxy is less bright, or when galaxies are in an advanced stage of merging so that only a single peak is detected instead of two nearby galaxy cores (Darg et al., 2010b).

Another concern with the automated deblending process is that single, large, bright galaxies can be mistakenly deblended into children that are wrongly classified as close pairs (Strauss et al., 2002). In the SDSS frames pipeline, overlapping objects that have initially been detected as one parent galaxy are deblended by separating the various sub-peaks into children components. This process takes place across all five optical bands, and after deblending has taken place these individual children are assigned surface brightness profiles such that the sum of the optical flux of the children is equal to that of the parent (Stoughton et al., 2002; Strauss et al., 2002). This process of deblending is often unreliable and frequently mistakingly classifies single galaxies into two separate galaxy close pairs. Figure 3.8 shows an example of this; the automated procedure has identified two peaks within the same galaxy and wrongly deduced that there is a merger taking place.

By imposing redshift and magnitude limits the contamination from non-mergers can be reduced. However, the only way to be sure that the close pairs identified
3.3 Visual Classification of Close Pairs

by the automated procedure are genuine close pairs is to re-classify visually. We classified the close pairs sample using their accompanying images from the SDSS SkyServer. In Figures 3.9-3.11 we show examples drawn directly from the SDSS online image-viewer of galaxies in the close pairs sample. Figure 3.9 (top left) shows a galaxy which is being deformed as it is pulled in by a massive elliptical galaxy, and in the center-left image we see a pair interacting spiral galaxies. The bottom left image shows an example of multiple galaxies interacting, and the bottom-right image shows a visually striking example of two interacting spiral galaxies.

14% of the close pairs sample were found to be incorrectly classified as close pairs and 0.66% of the sample were deemed unclassifiable morphologically. This is consistent with the ~15% of a minor pairs sample extracted using a similar automated procedure (0.027 < z < 0.17) found to be false pairs upon visual inspection by Woods & Geller (2007). We extrapolate that ~14% of the wide pairs sample will be contaminated by spurious objects.

Figure 3.12 shows $u - r$, $M(r)$ colour-magnitude plots split by visually classified morphological type. As we would expect, and as found by Strateva et al. (2001), elliptical galaxies lie in the red region of the bimodal colour distribution and spiral
3.3 Visual Classification of Close Pairs

Figure 3.9: Examples from the SDSS online image viewer of galaxy mergers from our close pairs sample.
Figure 3.10: Examples from the SDSS online image viewer of galaxy mergers from our close pairs sample.
3.3 Visual Classification of Close Pairs

Figure 3.11: Examples from the SDSS online image viewer of galaxy mergers from our close pairs sample.
We overplot the visually classified bulge-dominated spirals with green contours in the right plot.

Figure 3.12: Optical colour-magnitude contour plots showing SDSS galaxies from the close pair catalogue separated by visually classified morphological type. We overplot the visually classified bulge-dominated spirals with green contours in the right plot.

galaxies lie in the blue region. In Figure 3.12 (right), the bulge dominated spirals lie on the so-called *green valley* between these two distributions. The wide pairs sample is too large for visual inspection, and so we used the $r$-band fracdev parameter from the SDSS pipeline to morphologically classify the wide pairs.

Close pairs with fracdev$_r$ from 0.7 to 1 are classified as elliptical and objects with fracdev$_r$ from 0 to 0.5 are classified as spiral. We compare these with our visual classifications to test the reliability of the SDSS fracdev$_r$ parameter. 84% of the visually classified ellipticals had fracdev$_r$ from 0.7 to 1 (with 57% having fracdev$_r$ equal to 1) and 16% of the ellipticals had fracdev$_r < 0.7$ and so have been incorrectly classified by fracdev as non-elliptical. Of the galaxies that were visually classified as spirals; 68% had fracdev$_r$ from 0 to 0.5, with 32% misclassified as non-spirals by fracdev$_r$.

We had originally hoped to use the fracdev$_r$ determined morphologies for the wide pairs sample to test how SFR varies with morphology in our pairs sample. However, we decided that ~32% misclassification for spiral galaxies and a 16% misclassification rate for ellipticals was too unreliable to conduct a robust morphological analysis.

### 3.4 Major and Minor Pairs

We split our close pairs and wide pairs samples into major and minor merger systems; pairs with mass ratio $<1/3$ were categorised as major mergers (we introduced major and minor mergers in Section 1.4.3). We have 4431 galaxies in the major mergers
sample (2405 with NUV measurements) and 4483 galaxies in the minor mergers sample (2700 with NUV measurements). It is likely that many minor mergers have not been detected since the SDSS will not obtain spectra for photometric children with $r < 17.77$; this may be the case for smaller, less bright galaxies in a minor merger. Figure 3.13 (top) shows the redshift distribution for our major and minor close pairs samples. Since the low mass progenitor in a minor merger may not be detected at higher redshift, the major mergers have a significantly higher redshift distribution with median $z = 0.07$, whereas the minor mergers redshift distribution has median $z = 0.04$.

In Figure 3.13 (bottom) and Figure 3.14 (top left and top right) we see that the major mergers have a redder distribution in optical $(u - r)$ and NUV$-r$ colours. This is to be expected since the major mergers sample has a higher mass distribution than the minor mergers sample (see Figure 3.14, bottom left). Major mergers also show a preference for higher density environments (see Figure 3.14, bottom right). This is a consequence of the morphology-density relation (Dressler, 1980; Goto et al., 2003; Holden et al., 2007; Cappellari et al., 2011), where massive, early-type galaxies tend to inhabit larger dark matter halos in the local Universe.

The wide pairs sample is also used as a control sample with which to compare the major and minor mergers samples. The wide pairs sample was split according to the mass ratio of pairs, where pairs with mass ratio $< 1/3$ were categorised as a major mergers control sample (there were 26,463 such galaxies, and 17,704 with NUV measurements) and the rest were classified as a minor mergers control sample (there were 27,942 such galaxies, and 20,434 with NUV measurements).

Figure 3.15 shows how our NUV luminosity-derived SFRs compare with the MPA-JHU DR7 catalogue for the majors sample (left) and the minors sample (right). The comparison is only shown for wide pairs, where $30 < r_p < 150$ kpc, since these comprise our control sample. The SFRs derived from emission-lines are generally lower than our NUV photometry-derived SFRs (this is likely because the MPA catalogue is aperture limited) but overall the two methods show reasonable agreement.
Figure 3.13: Redshift distribution (top) and optical colour-magnitude plot (bottom) for major and minor close pairs samples.
3.4 Major and Minor Pairs

Figure 3.14: Top (left and right): NUV−r distribution for major and minor mergers. Bottom left: Stellar mass distribution for major and minor mergers. Bottom right: Halo mass distribution for major and minor mergers.

Figure 3.15: Emission-line luminosity-derived SFRs from the MPA SDSS DR7 catalogue (black) are compared with NUV photometry-derived SFRs for our wide pairs (30 < r_p < 150kpc) sample (green) for the majors sample (left) and minors sample (right). Vertical lines show sample median values. The SFR distributions are normalised.
3.5 Summary

We began this chapter by introducing the close pair and wide pair data sets and described how these were extracted from the SDSS DR7 and cross-matched with the GALEX GR4/GR5 database for NUV measurements. We explained how the optical and NUV magnitudes were K-corrected and then extinction corrected for both galactic and internal extinction. Galaxy masses and environment densities were calculated for each close pair, and a BPT analysis allowed us to classify each object according to the predominant mechanism of excitation by using SDSS spectra. The median uncertainty in stellar mass is $\sim 0.1$ dex, with the maximum uncertainty expected to be $\sim 0.3$ dex. Only galaxies with [NII], [OIII], H$\alpha$ and H$\beta$ emission-line signal-to-noise ratio greater than 3 are considered in the BPT analysis.

For the first time in close pair studies, we use NUV luminosity-derived SSFRs. We highlighted the advantages of using the NUV waveband as opposed to optical photometry and spectroscopy, and explained how our SSFRs were calculated from NUV fluxes by using a formula from Iglesias-Páramo et al. (2006).

Close pairs and wide pairs with mass ratio $< 1/3$ were categorised as major mergers, and the remaining pairs were classified as minor mergers. It is likely that many minor mergers have not been detected since the SDSS will not obtain spectra for photometric children with $r < 17.77$; this may be the case for smaller, less bright galaxies in a minor merger.

Now that we have familiarised the reader with the advantages and limitations of our sample, we move on to study how the NUV-derived SSFR in close pair galaxies evolves as mergers progress. We consider various mass and environment parameters during the course of this study, and also test for observational evidence of AGN activity being triggered in close pairs.
Chapter 4

Star Formation and AGN Activity in Close Pairs

This chapter is an elaboration of the results presented in the following two papers:


Overview

In the previous chapter we described how the close and wide pairs samples were extracted from the SDSS by assuming a projected separation of <30kpc for the close pairs, 30-150kpc for the wide pairs, and a recessional velocity difference <500km s\(^{-1}\). We described the process by which these samples were cross-matched to get GALEX NUV magnitudes and how the properties for our study (such as mass, environment, BPT classification etc.) were calculated. Mass ratios were also used to categorise major and minor mergers. NUV fluxes from GALEX were used to derive specific star formation rates (SSFRs). We now use our close pairs sample to study the key factors affecting star formation (SF) and AGN activity triggered during galaxy interactions.

Galaxy close pairs are studied to investigate the effects of gravitational interactions on SF and black hole accretion processes in merger progenitors. Properties such as mass and environment are shown to impact SF in mergers as a function of separation (i.e. as the galaxy pairs draw closer together). We first study the SF enhancement as a function of mass and separation for the whole sample in Section
4.1.1. We then study the SF enhancement as a function of environment density and separation in Section 4.1.2. Then, by using our major and minor merger samples, we investigate how SF enhancements in close pairs are affected by the mass ratio between the interacting galaxies in Section 4.2.

Theoretical models predict that merger activity can trigger black hole accretion and we seek observational evidence for this. In Section 4.3, we study the fraction of Starburst, Transition, LINER and Seyfert galaxies in separation bins from 0-150kpc and investigate changes in these fractions as galaxies draw closer to merging. The evolution in these fractions with decreasing separation is studied as a function of galaxy mass and environment density and is also explored in major and minor mergers.

Efforts to investigate the direct effects of properties such as mass, environment, nuclear activity etc. on merger-induced SF are complicated by correlations which exist between them. Examples include:

- A morphology-density relation where spirals tend to be found in low density environments and ellipticals in high density environments (Dressler, 1980; Goto et al., 2003; Holden et al., 2007; Cappellari et al., 2011)
- A colour-morphology relation where massive optically-red galaxies tend to be elliptical and less massive optically-blue galaxies tend to be spiral (Strateva et al., 2001; Bell et al., 2004)
- A mass-AGN relation where massive galaxies are found to be more likely to host AGN activity (Kauffmann et al., 2003a). Kauffmann et al. (2003a) found that at least 80% of galaxies in an SDSS sample of 0.02 < z < 0.3 galaxies with $M_\ast > 10^{11} M_\odot$ are classified as type 2 AGN by a BPT analysis, and the AGN fraction with $M_\ast < 10^{10} M_\odot$ rapidly decreases with mass.

For this reason, we subdivide our sample by various properties where possible; e.g. during the BPT analysis we split our Seyfert sample according to low and high mass galaxies.

In recent years, close pairs have been studied extensively in optical bands; mostly using SDSS data (Kauffmann et al., 2004; Alonso et al., 2007; Ellison et al., 2010). The near-UV (NUV) band allows us to study SF with much more sensitivity than optical filters to recent star formation (rSF), up to $\sim$1Gyr (e.g. Yi et al., 2005; Kaviraj et al., 2007; Schawinski et al., 2007a), and detections of rSF have been permitted in some early-type galaxies where optical colours would have classified them as non-starforming. This work offers a similar investigation to previous research into
close pairs, but from an optical-UV perspective. Our results are summarised and discussed in the context of a hierarchical model of galaxy evolution and a summary of the SSFR enhancements found in the various analyses explored in this chapter is given in Table 4.3.

4.1 Recent Star Formation in Close Pairs

4.1.1 SF Enhancement as a Function of Separation and Mass

![Figure 4.1](image_url)

**Figure 4.1:** Top: Median NUV-\(r\) colours for close and wide pairs binned by separation (0-15kpc, 15-30kpc, 30-60kpc and 90-130kpc) for both low stellar mass galaxies (\(10^8-10^{11}M_\odot\), shown in blue) and high stellar mass galaxies (\(10^{11}-10^{13}M_\odot\), shown in red). Bottom: SSFR difference between the separation bin in question and the widest separation bin, for both low and high mass galaxies. The vertical dotted line at 30kpc is a reminder of the border between the close pair and wide pair samples. The fractional \(\Delta\)SSFR error is \(\sim 10\%\) for the low mass sample and \(\sim 30\%\) for the high mass sample.

In Figure 4.1 we split the close pairs and wide pairs samples into low stellar mass (\(10^8-10^{11}M_\odot\), shown in blue) and high stellar mass (\(10^{11}-10^{13}M_\odot\), shown in red) galaxies. Notice that the low mass sample has a bluer NUV-\(r\) distribution which indicates higher levels of rSF; we would expect this since low mass galaxies tend to
4.1 Recent Star Formation in Close Pairs

show more SF than high mass galaxies (e.g. Karim et al., 2011), a consequence of cosmic downsizing (see Section 1.1.3).

To investigate the SF enhancement in advancing mergers, both low mass and high mass galaxies are binned further by projected separation and we analyse median NUV-r and SSFR trends in these bins as separation decreases; the widest separation bin is treated as a control sample, representative of the general population of non-close pair galaxies. For both stellar mass bins, the median NUV-r colours become bluer as projected separation decreases, indicating that rSF is enhanced as pairs draw closer together. This decrease in NUV-r occurs at approximately the same rate for the lower and the higher stellar mass bins.

<table>
<thead>
<tr>
<th>Mass (Log_{10} M_\odot)</th>
<th>Projected Separation (kpc)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-15</td>
</tr>
<tr>
<td>8-11</td>
<td>7.62×10^{-11}</td>
</tr>
<tr>
<td>11-13</td>
<td>2.21×10^{-12}</td>
</tr>
</tbody>
</table>

Table 4.1: Median SSFR (yr^{-1}) derived from NUV luminosity for each stellar mass and separation bin.

In Table 4.1 we show the median SSFR for the galaxies in each mass/separation bin, and at the bottom of Figure 4.1 we plot the difference in SSFR (\Delta SSFR) between the separation bin in question and the widest separation bin; this quantity is by definition zero for the widest separation bin. The \Delta SSFR error bars are negligible so instead of plotting these we provide a fractional error; this error is \sim 10% for the low mass sample and \sim 30% for the high mass sample.

We find a difference of 6.1×10^{-11}yr^{-1} in SSFR from the widest (90-130kpc) to the smallest separation bin (0-15kpc) for low stellar mass galaxies, and a difference of 1.1×10^{-12}yr^{-1} for high stellar mass galaxies. This indicates a factor of 5.1 ± 0.7 increase in SSFR for low stellar mass close pair galaxies and a factor of 2.0 ± 0.9 increase in SSFR for high stellar mass close pairs compared to the general galaxy population. These increases, along with the results from other analyses in this chapter, are summarised in Table 4.3 at the end of this chapter.

4.1.2 SF Enhancement as a Function of Separation and Environment

We now split the sample by environment and separation instead of stellar mass and separation. The three environment bins are field (with halo mass 10^{10} to 10^{13}M_\odot),
4.1 Recent Star Formation in Close Pairs

Group (with halo mass $10^{13}$ to $10^{14}M_\odot$) and cluster (with halo mass $10^{14}$ to $10^{15}M_\odot$). Notice in Figure 4.2 that the NUV-r colour distribution shifts to bluer colours from the high density cluster environment to the low density field environment; we would expect this since more SF tends to take place in galaxies inhabiting lower density environments (e.g. Kauffmann et al., 2004).

![Figure 4.2](image)

**Figure 4.2:** Top: Median NUV-r colours for close and wide pairs binned by separation (0-15kpc, 15-30kpc, 30-60kpc and 90-130kpc) for pairs in field (black), group (blue) and cluster (red) environments. Bottom: SSFR difference between the separation bin in question and the widest separation bin, for each environment. The fractional $\Delta$SSFR error is $\sim20\%$.

SSFRs for each bin are shown in Table 4.2. We find a difference of $1.4 \times 10^{-11} \text{yr}^{-1}$ (i.e. a factor of $1.8 \pm 0.5$ increase) in SSFR from the widest to the smallest separation bin for pairs in field environments, and no significant increase for pairs in group and cluster environments. Since stellar mass and environment are correlated, such that increasingly massive galaxies tend to be found in higher density environments (a consequence of the morphology-density relation), we attempt to break the degeneracy by splitting our sample by stellar mass and environment.

Figure 4.3 shows the environment/separation analysis, but now restricted to low mass galaxies ($10^8$-$10^{11}M_\odot$ -top) and higher mass galaxies ($10^{11}$-$10^{13}M_\odot$ -bottom). In the low mass analysis, NUV-r colours in field and group environments become noticeably bluer with decreasing projected separation; with no clear trend for cluster environments. We find an average rise in SSFR of $4.4 \times 10^{-11} \text{yr}^{-1}$ (a factor of $2.4 \pm 0.7$
4.1 Recent Star Formation in Close Pairs

<table>
<thead>
<tr>
<th>Environment</th>
<th>Projected Separation (kpc)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-15</td>
</tr>
<tr>
<td>Field</td>
<td>$2.98 \times 10^{-11}$</td>
</tr>
<tr>
<td>Group</td>
<td>$2.62 \times 10^{-12}$</td>
</tr>
<tr>
<td>Cluster</td>
<td>$1.76 \times 10^{-12}$</td>
</tr>
</tbody>
</table>

**Table 4.2:** Median SSFR (yr$^{-1}$) derived from NUV luminosity for each environment and separation bin.

![Figure 4.3: Median NUV-$r$ and ΔSSFR values for each environment/separation bin are plotted for low stellar mass galaxies (10$^8$-10$^{11}$M$_\odot$ -top) and high stellar mass galaxies (10$^{11}$-10$^{13}$M$_\odot$ -bottom).](image-url)
4.2 Major and Minor Mergers

We now look to our major and minor mergers samples. The major mergers sample comprises pairs with mass ratio $>1/3$ and the minor mergers sample comprises pairs with mass ratio $<1/3$. The major/minor mergers terminology extends to our full close and wide pairs samples, including projected separations up to 150kpc; at wide separation the pairs will not actually be merging, but we refer to the full separation range as major/minor mergers for simplicity. As with Section 4.1, we take the widest separation bin as a control sample to compare the close pairs with, and we test pairs in different mass and environment bins to see how these properties impact rSF in major and minor pairs.

4.2.1 SF Enhancement as a Function of Separation and Mass

Figure 4.4 (top) shows the galaxies in the major merger sample, and Figure 4.4 (bottom) shows the galaxies in the minor merger sample; we further split these samples by mass. For both the major and minor samples, the NUV-$r$ colour distributions for lower mass galaxies are bluer in NUV-$r$ as we would expect from Section 4.1.1 and due to cosmic downsizing.

There is a slightly bluer NUV-$r$ distribution for high mass galaxies in major mergers and also a slightly redder NUV-$r$ distribution for low mass galaxies in major mergers. This is likely because of the narrower mass range in the major mergers sample (see Figure 3.14 (bottom left)) leading to a slightly more heterogeneous
4.2 Major and Minor Mergers

NUV-$r$ distribution.

![Graph showing median NUV-$r$ and ΔSSFR values for major and minor mergers](image)

**Figure 4.4:** Median NUV-$r$ and ΔSSFR values are plotted for the major mergers sample (with pair mass ratio >1/3) -top, and and minor mergers sample (with pair mass ratio <1/3) -bottom.

For the major mergers sample, we find a difference of $5.3 \times 10^{-11} \text{yr}^{-1}$ in SSFR from the widest (90-130kpc) to the smallest separation bin (0-15kpc) for low stellar mass galaxies (an average factor of 4.6 ± 0.7 increase in SSFR), and a difference of $1.3 \times 10^{-12} \text{yr}^{-1}$ for high stellar mass galaxies (an average factor of 1.8 ± 0.7 increase in SSFR). For the minor mergers sample, we find a difference of $7.1 \times 10^{-11} \text{yr}^{-1}$ in SSFR from the widest (90-130kpc) to the smallest separation bin (0-15kpc) for low stellar
mass galaxies (an average factor of 3.9 ± 0.5 increase in SSFR), and a difference of 2.2×10^{-12}yr^{-1} for high stellar mass galaxies (an average factor of 3.0 ± 1.3 increase in SSFR).

Figure 4.5: Median NUV-r and ∆SSFR values are plotted for the minor mergers sample split into primary progenitors and secondary progenitors.

Galaxies in the major mergers sample have similar mass by definition, so we might expect similar SFRs for both galaxies in a major merger system. However, in a minor merger we might expect to see a stronger impact on the low mass progenitor (as Woods & Geller (2007) reported; see Section 1.4.3), since it is interacting with a significantly more massive progenitor that will have a strong gravitational influence. From now on, we refer to the higher mass galaxy in a minor merger as the ‘primary’ progenitor and the lower mass galaxy as the ‘secondary’ progenitor.

We split the minor mergers sample according to the primary and secondary progenitors in each system (see Figure 4.5). Notice that the median NUV-r colour for the smallest separation bin, 0-15kpc, becomes sharply bluer for the primary progenitors as well as the secondary progenitors, indicating a significant enhancement in rSF for both the primary and secondary progenitors in low separation minor merger pairs.

We find a difference of 2.1×10^{-11}yr^{-1} in SSFR from the widest (90-130kpc) to the smallest separation bin (0-15kpc) for primary galaxies (an average factor
4.2 Major and Minor Mergers

of 13.5 ± 3.8 increase in SSFR), and a difference of 11.5×10^{-11}\text{yr}^{-1} for secondary galaxies (an average factor of 4.9±1.4 increase in SSFR); i.e. we see a higher relative increase in SSFR in primary galaxies than in secondary galaxies in minor mergers. This is a surprising result since Woods & Geller (2007) found SSFR enhancements in minor merger secondaries, but no evidence for an enhancement in SSFR in primaries. We expect that this is because the level of SF in primaries is relatively low and since Woods and Geller used H\alpha as an SSFR diagnostic small changes have not been detected, whereas the NUV is particularly sensitive to measuring changes in rSFR.

4.2.2 Impact of Mass Ratio on SF Enhancement in Minor Mergers

We now investigate the impact of mass ratio between progenitors on SF enhancement in minor mergers. First the minor mergers sample is split according to the mass ratio between merger progenitors, then we look at how the median NUV-\text{r} colour changes for different mass ratios compared to a control sample of non-close pairs with the same mass ratio.

In Figure 4.6, the minor merger primaries and secondaries are binned by mass ratio (in mass ratio intervals of 0.05), and the median NUV-\text{r} colour is shown for each bin. Median NUV-\text{r} colours for control sample primaries and secondaries are shown in black for comparison and only objects in the widest separation bin 80 < r_p < 150kpc are used. Overall the secondaries show bluer NUV-\text{r} colours in both the close pair and control samples; these samples will naturally have a lower mass distribution and this trend is due to cosmic downsizing. Both primaries and secondaries in close pairs show bluer NUV-\text{r} colours than their control samples, indicating that enhanced rSFR is taking place as shown in Figure 4.5, and further, showing that this enhancement occurs for all minor merger mass ratios.

Notice that there is a trend towards bluer NUV-\text{r} colours in both primaries and secondaries where the mass ratio is smaller (i.e. the secondary is interacting with a primary that is significantly more massive). This trend exists in the secondaries control sample but not in the primaries control sample. We would expect to see this behaviour in non-close pairs because the mass difference will in general be most pronounced when the mass ratio is smallest, therefore the primaries at lowest mass ratio will be very massive galaxies (and hence usually show redder NUV-\text{r} colours) and the secondaries at lowest mass ratio will be very small galaxies (and hence usually show bluer NUV-\text{r} colours). It is very interesting that the primaries in minor mergers indicate a rise in rSF when the mass ratio is smallest contrary to
what is shown in the primaries control sample. This could imply that in minor mergers where the secondary is relatively small, gas is more efficiently funneled into the primary progenitor leading to a large SF enhancement; this is a result that we have not seen reported in any other literature, and may rely on the sensitivity of the NUV waveband to rSF to be detected.

**Figure 4.6:** Median NUV-r colour for minor merger primaries (red) and secondaries (blue) binned by mass ratio (only close pairs with \( r_p < 30 \)kpc are considered here). Control samples for progenitors with minor merger mass ratios and \( 80 < r_p < 150 \)kpc are shown in black for primaries and secondaries.

In Figure 4.7 we explore this trend further by splitting by separation. Figure 4.7 (top left) shows minor merger primary and secondary progenitors with projected separation \( 0 < r_p < 15 \)kpc and Figure 4.7 (top right) shows minor merger primary and secondary progenitors with projected separation \( 15 < r_p < 30 \)kpc as a function of mass ratio. These are compared with wide pair minor merger primary and secondary progenitors in the bottom two plots. The primary and secondary progenitor samples at the closest separation show significant decreases in median NUV-r colour at the lowest mass ratio. Notice that the higher separation primaries in the bottom plot are on average redder at low mass ratio; the \( 15 < r_p < 30 \)kpc separation sample indicates an increase in rSF in primaries when the mass ratio is larger, and an increase in rSF for secondaries when the mass ratio is lower.
4.2 Major and Minor Mergers

Figure 4.7: Median NUV-r colour for minor merger primaries (dotted) and secondaries (dashed) binned by mass ratio. Top left: minor merger galaxies with $0 < r_p < 15\text{kpc}$. Top right: minor merger galaxies with $15 < r_p < 30\text{kpc}$. Bottom left: minor merger galaxies with $30 < r_p < 80\text{kpc}$. Bottom right: minor merger galaxies with $80 < r_p < 150\text{kpc}$. The bottom plots are in black to differentiate the wide pairs from the close pairs sample.

4.2.3 SF Enhancement as a Function of Separation and Environment

We now split our major and minor mergers samples by environment (see Figure 4.8). Major merger pairs generally show more SF in field environments than major merger pairs in higher density environments. This is only slightly enhanced at low separation when compared with the wide separation control sample; showing some signs of enhancement in SSFR for major mergers in field environments (and only a slight enhancement for major mergers in higher density environments). However, there is significantly more SF taking place in minor mergers in field environments at low separation when compared with the control sample, indicating that major mergers in field environments lead to an enhancement in SSFR. There is a slight
indication of an enhancement in rSF for minor mergers in group environments but no evidence of SF enhancements in cluster environments.

Figure 4.8: Median NUV-r and ∆SSFR values are plotted for the major mergers sample (with pair mass ratio >1/3) -top, and and minor mergers sample (with pair mass ratio <1/3) -bottom, for pairs in field (black), group (blue) and cluster (red) environments.

Figure 4.9 shows the minor mergers sample split into primary (red) and secondary (blue) progenitors for both field (top) and group/cluster environments (bottom). Note that the ∆SSFR axes are now in units of $10^{-12}\,\text{yr}^{-1}$. In field environments, both primary and secondary progenitors show evidence of SSFR enhancement...
in minor merger pairs at low separation. However, there is no significant evidence for SSFR enhancement in either primary or secondary progenitors in higher density environments.

We report a difference of $2.2 \times 10^{-11}$ yr$^{-1}$ in SSFR from the widest (90-130kpc) to the smallest separation bin (0-15kpc) for minor merger primary galaxies in field environments (an average factor of $12.8 \pm 3.6$ increase in SSFR), and a difference of $2.4 \times 10^{-11}$ yr$^{-1}$ for minor merger secondary galaxies in field environments (an average

---

**Figure 4.9:** Top: primaries and secondaries for minor mergers in field environments. Bottom: primaries and secondaries for minor mergers in group/cluster environments.
factor of $1.7 \pm 0.5$ increase in SSFR). Again, we see a larger relative increase in SSFR in the primary progenitor sample than in the secondary progenitor sample, although overall the secondaries are on average showing more SF.

4.3 AGN Activity in Close Pairs

We wish to study how AGN activity evolves as a function of separation in close pairs. The BPT catalogue categorises the close and wide pairs samples into Starburst, Transition, Quiescent, LINER and Seyfert classifications. We use the BPT classification scheme from Schawinski et al. (2007b), for SDSS DR7 emission lines [NII], Hα, [OIII] and Hβ drawn from the OSSY catalogue (Oh et al., 2011). Seyferts are considered as Type II AGN. Sarzi et al. (2010) propose that very few objects with LINER emission are truly ionised by a central AGN.

We analyse how the Transition, LINER and Seyfert fractions change as pairs advance to the lowest separation bin. Our aim is to see how the AGN fraction changes; i.e. if AGN activity is somehow ignited in some close pairs as the merging process advances. The pairs sample is split into four projected separation bins (0-15kpc, 15-30kpc, 30-80kpc and 80-150kpc).

4.3.1 Emission-Line Analysis for Close Pairs

We first conduct our emission-line analysis on the close pairs sample. Figure 4.10 (top left) shows the fraction of galaxies in each separation and BPT classification bin. Although we have restricted our analysis by excluding Starburst and Quiescent galaxies in this figure, we include these in the total sample when calculating fractions. The error bars shown are standard Poisson number count errors. As pair separation decreases, we see a steady rise in the Transition class from the widest separation bin to the smallest separation bin of $3.1\% \pm 0.5\%$. We see a slight, yet non-significant, increase in the Seyfert fraction of $0.1\% \pm 0.4\%$, and no significant evolution in the LINER class. Since galaxies with stellar mass below $\sim 10^{10}M_\odot$ are unlikely to host AGN (Kauffmann et al., 2003a), we now restrict the analysis to higher stellar mass galaxies ($M \geq 10^{10}M_\odot$); see the top right plot in Figure 4.10. This plot shows a similar distribution to the full stellar mass sample, but now each category accounts for a higher fraction. Here, we see a more defined increase in the Transition class ($4.3\% \pm 0.7\%$) and an increase in the Seyfert fraction of $1.2\% \pm 0.5\%$.

We further split the sample into high mass pairs in the field (Figure 4.10, bottom left) and pairs in group/cluster environments (bottom right). A larger Tran-
4.3 AGN Activity in Close Pairs

Figure 4.10: Fraction of Transition, LINER and Seyfert galaxies in each of the following four separation bins: 0-15kpc, 15-30kpc, 30-80kpc and 80-150kpc. Top Left: All masses and environments. Top Right: Restricted to high mass (M$_{\geq 10^{10} M_\odot}$) since AGN activity is unlikely in lower mass galaxies. Bottom left: High mass galaxies (M$_{\geq 10^{10} M_\odot}$) in field environments. Bottom right: High mass galaxies (M$_{\geq 10^{10} M_\odot}$) in group environments.

A transition fraction is noticed in field environments, where we see a dramatic increase of 4.6% ± 1.0%, compared with an increase of 2.8% ± 1.0% in higher density environments. The Seyfert fraction shows small increases, of 0.9% ± 0.7% in the field and 1.2% ± 0.7% in higher density environments; a slightly higher Seyfert fraction over all separations (from wide pairs to close pairs) is noticed in field environments, with a more pronounced increase from the wide pairs control sample to close pair, interacting galaxies, in higher density environments. Little significant evolution is seen in the LINER fraction.

4.3.2 Emission-Line Analysis for Major and Minor Mergers

We now conduct our emission-line analysis on the major and minor mergers samples. Figure 4.11 shows the fraction of major merger (top) and minor merger (bottom)
4.3 AGN Activity in Close Pairs

galaxies in each separation and BPT classification bin. We see a steady rise in the Transition fraction for both samples; an increase of $3.4\% \pm 0.8\%$ for the major mergers and an increase of $3.1\% \pm 0.7\%$ for the minor mergers. These similar trends show a general increase in the Transition class for close pairs, regardless of whether they are in major or minor mergers. The Seyfert fractions are also similar; an increase of $0.9\% \pm 0.5\%$ is seen for the major mergers and an increase of $0.6\% \pm 0.5\%$ for the minor mergers. These trends are not distinct between the major and minor samples.

Figure 4.11: Fraction of Transition, LINER and Seyfert galaxies in each of the following four separation bins: 0-15kpc, 15-30kpc, 30-80kpc and 80-150kpc. Major merger sample (top) and minor merger sample (bottom).
Since gas-rich major mergers are predicted to trigger AGN activity (see Section 1.4.3), we might expect major mergers in field environments (where there is more gas available) to show stronger signs of AGN activity, so we further split the major mergers into field and group/cluster environments. In Figure 4.12 (top) we show a BPT analysis for major mergers in field environments. We see increases in the fraction of Transition, LINER and Seyfert objects; a rise of 3.7% ± 1.2% is seen in the Transition class, a rise of 1.2% ± 0.8% in Seyferts and a rise of 1.5% ± 0.7% in LINERs.
4.4 Summary and Discussion

The sample size is small and the error bars are large for the analysis of major mergers in group and cluster environments, however, we still see some evolution in the BPT classes between wide and close pairs in Figure 4.12 (bottom). The Transition fraction rises by $2.3\% \pm 1.2\%$. The observed rise in the Seyfert fraction is not statistically significant; $0.7\% \pm 0.8\%$. There is an interesting evolution in the LINER fraction where it rises by $1.2\% \pm 0.7\%$ by the 15-30kpc separation bin, and then there is a non-statistically significant dip in the LINER fraction at the 0-15kpc separation bin.

4.4 Summary and Discussion

We have studied close pair galaxies using NUV-$r$ colours and NUV-derived SSFRs. This serves to add to previous work on close pairs in which optical colours and emission lines are used as indicators of SF. Whereas emission-line measurements are limited by finite fibre size and rely on corrections, NUV photometry provides a broader measure of young SF for an entire galaxy. Our sample consisted of SDSS optical spectroscopy and photometry, and GALEX photometry for low redshift close pair systems, and we also extracted a sample of wide pair galaxies to be used as a control sample representative of the general population.

We found a factor of $5.1 \pm 0.7$ increase ($6.1 \times 10^{-11} \text{yr}^{-1}$) in SSFR for low stellar mass close pair galaxies and a factor of $2.0 \pm 0.9$ increase ($1.1 \times 10^{-12} \text{yr}^{-1}$) in SSFR for high stellar mass close pairs compared to the general galaxy population. We found a difference of $1.4 \times 10^{-11} \text{yr}^{-1}$ (i.e. a factor of $1.8 \pm 0.5$ increase) in SSFR from the widest to the smallest separation bins for pairs in field environments, and no significant increase for pairs in group and cluster environments. In the low mass sample we found an average rise in SSFR of $4.4 \times 10^{-11} \text{yr}^{-1}$ (a factor of $2.4 \pm 0.7$ increase) for pairs in the field and an average rise of $1.2 \times 10^{-11} \text{yr}^{-1}$ (a factor of $3.3 \pm 0.9$ increase) for pairs in groups. For high mass pairs we saw an average rise of $2.2 \times 10^{-12} \text{yr}^{-1}$ (a factor of $2.5 \pm 0.7$ increase) in SSFR in field environments. These results are summarised in Table 4.3.

Our results are consistent with Wong et al. (2011), who used NUV-$r$ and FUV-$r$ colours from GALEX as a proxy for SSFR for intermediate redshift close pairs ($0.25 \leq z \leq 0.75$) drawn from PRIMUS. Note that their PRIMUS sample was much smaller than our SDSS sample and at $z > 0.3$ GALEX does not detect red sequence galaxies well, so their sample was biased towards star forming galaxies. They found an $\sim 15 - 20\%$ increase in SSFR for close pairs with projected separation $\leq 50h^{-1}$ kpc, and an $\sim 25 - 30\%$ increase in SSFR for close pairs with projected separation
## 4.4 Summary and Discussion

<table>
<thead>
<tr>
<th>Analysis</th>
<th>SSFR Enhancement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mass</strong></td>
<td></td>
</tr>
<tr>
<td>Low Mass</td>
<td>$5.1 \pm 0.7$</td>
</tr>
<tr>
<td>High Mass</td>
<td>$2.0 \pm 0.9$</td>
</tr>
<tr>
<td><strong>Environment</strong></td>
<td></td>
</tr>
<tr>
<td>Field</td>
<td>$1.8 \pm 0.5$</td>
</tr>
<tr>
<td>Group</td>
<td>$1.2 \pm 0.3$</td>
</tr>
<tr>
<td>Cluster</td>
<td>$1.0 \pm 0.3$</td>
</tr>
<tr>
<td><strong>Environment (Low Mass)</strong></td>
<td></td>
</tr>
<tr>
<td>Field</td>
<td>$2.4 \pm 0.7$</td>
</tr>
<tr>
<td>Group</td>
<td>$3.3 \pm 0.9$</td>
</tr>
<tr>
<td>Cluster</td>
<td>$1.6 \pm 0.4$</td>
</tr>
<tr>
<td><strong>Environment (High Mass)</strong></td>
<td></td>
</tr>
<tr>
<td>Field</td>
<td>$2.5 \pm 0.7$</td>
</tr>
<tr>
<td>Group</td>
<td>$1.6 \pm 0.5$</td>
</tr>
<tr>
<td>Cluster</td>
<td>$1.1 \pm 0.3$</td>
</tr>
<tr>
<td><strong>Major Mergers</strong></td>
<td></td>
</tr>
<tr>
<td>Low Mass</td>
<td>$4.6 \pm 0.7$</td>
</tr>
<tr>
<td>High Mass</td>
<td>$1.8 \pm 0.7$</td>
</tr>
<tr>
<td><strong>Minor Mergers</strong></td>
<td></td>
</tr>
<tr>
<td>Low Mass</td>
<td>$3.9 \pm 0.5$</td>
</tr>
<tr>
<td>High Mass</td>
<td>$3.0 \pm 1.3$</td>
</tr>
<tr>
<td><strong>Minor Mergers</strong></td>
<td></td>
</tr>
<tr>
<td>Primaries</td>
<td>$13.5 \pm 3.8$</td>
</tr>
<tr>
<td>Secondaries</td>
<td>$4.9 \pm 1.4$</td>
</tr>
<tr>
<td><strong>Major Mergers</strong></td>
<td></td>
</tr>
<tr>
<td>Field</td>
<td>$1.4 \pm 0.4$</td>
</tr>
<tr>
<td>Group</td>
<td>$0.9 \pm 0.3$</td>
</tr>
<tr>
<td>Cluster</td>
<td>$1.1 \pm 0.3$</td>
</tr>
<tr>
<td><strong>Minor Mergers</strong></td>
<td></td>
</tr>
<tr>
<td>Field</td>
<td>$2.6 \pm 0.7$</td>
</tr>
<tr>
<td>Group</td>
<td>$1.3 \pm 0.4$</td>
</tr>
<tr>
<td>Cluster</td>
<td>$0.9 \pm 0.3$</td>
</tr>
<tr>
<td><strong>Minor Mergers in Field Environments</strong></td>
<td></td>
</tr>
<tr>
<td>Primaries</td>
<td>$12.8 \pm 3.6$</td>
</tr>
<tr>
<td>Secondaries</td>
<td>$1.7 \pm 0.5$</td>
</tr>
</tbody>
</table>

**Table 4.3:** Summary of the SSFR enhancements from the lowest separation bin to the highest separation bin (i.e. close pairs versus the control sample) from the various analyses in this chapter.
4.4 Summary and Discussion

≤30h\(^{-1}\) kpc. PRIMUS spectra are lower resolution than SDSS spectra (redshifts are accurate to within \(\sigma_z/(1+z)\) with \(\lesssim3\%\) catastrophic outliers), and the coverage is less (9.1 deg\(^2\) of the sky, using multiple independent fields). Our SSFRs were calculated directly from the NUV luminosity, whereas Wong et al. used colours to estimate SSFR; yet despite different methods for estimating SSFR and our statistics being more reliable, the implications of enhancement in close pairs are consistent. A combination of these two studies shows evidence for SSFR enhancement from local to intermediate redshift close pairs.

Our results are also consistent with previous work on close pairs using optical colours and emission lines as tracers of SF. Karim et al. (2011) found that low mass galaxies tend to show the most SF after calculating SFRs from stacked 1.4 GHz radio continuum emission. Due to the phenomenon of cosmic downsizing we expect to see more SF in low mass galaxies in the local Universe (Cowie et al., 1996; Terlevich, López & Terlevich, 2007; Faber et al., 2007). We showed a bluer NUV-\(r\) colour distribution for galaxies in low density environments when compared to those in higher density environments; this is true for our close pairs and wide pairs so reflects the general galaxy population (i.e. not only close pairs). Our results are in-line with those of Kauffmann et al. (2004), who also found higher levels of (optical emission-line-derived) specific SF in galaxies (not necessarily close pairs) in lower density environments, and Schawinski et al. (2007a) who found that UV-bright galaxies are most likely to be found in the field. We showed a bluer NUV-\(r\) colour distribution for lower mass galaxies in both the close and wide pairs samples; indicating from our NUV perspective that low mass galaxies in general show more rSF than massive galaxies.

Using asymmetry effects and optical colours as tracers, Alonso et al. (2005) and Ellison et al. (2010) also found an enhancement in SF for close pairs in low density environments. This is likely to be due to the higher gas fraction available in low density environments to fuel SF, since tidal fields and ram pressure stripping can reduce gas fractions in higher density environments.

Upon splitting our minor mergers sample into primary and secondary progenitors, we find a difference of \(2.1\times10^{-11}\) yr\(^{-1}\) in SSFR from the widest (90-130kpc) to the smallest separation bin (0-15kpc) for primary galaxies (an average factor of 13.5 ± 3.8 increase in SSFR), and a difference of \(11.5\times10^{-11}\) yr\(^{-1}\) for secondary galaxies (an average factor of 4.9 ± 1.4 increase in SSFR). Woods & Geller (2007) found that the lower mass progenitor in a minor merger experiences the most SF, and we did indeed see that SF is enhanced in the secondary sample through our NUV-\(r\) colour analysis and the \(\Delta\)SSFR analysis. However, we also saw indications
of a rise in SF in the higher mass galaxy (the primary) in a minor merger at small separation from both our NUV-\(r\) colours and the \(\Delta\)SSFRs in Figure 4.5. We could not find any other literature showing a general and significant impact on rSF in the primary progenitor of minor mergers; it may be that these effects require the sensitivity of the NUV-waveband to become apparent. Woods & Geller (2007) used H\(\alpha\) measurements and it is likely that the relatively low SSFR occurring in primaries is undetectable using this SF diagnostic.

We also showed a trend where there is a significant decrease in median NUV-\(r\) colour for both primary and secondary progenitors where the mass ratio is smallest and the close pairs have low separation. This is surprising, yet statistically significant, evidence that both the primary and secondary progenitors in minor mergers on average show enhancements in rSF, particularly for very low mass ratios (where the primary is at least three times more massive than the secondary) and at the lowest separation (i.e. as the close pairs are reaching an advanced state of merging). Perhaps in a minor merger where the masses are very different, the smaller progenitor stirs up gas in the primary more efficiently than if they are of similar masses. We were unable to find evidence of a similar result in any other literature, however, this trend is statistically significant in our research and it is likely that it has only now been detected due to the sensitivity of the NUV waveband to rSF.

We then split our minor mergers sample by environment. We reported a difference of \(2.2\times10^{-11}\text{yr}^{-1}\) in SSFR from the widest (90-130kpc) to the smallest separation bin (0-15kpc) for minor merger primary galaxies in field environments (an average factor of \(12.8\pm3.6\) increase in SSFR), and a difference of \(2.4\times10^{-11}\text{yr}^{-1}\) for minor merger secondary galaxies in field environments (an average factor of \(1.7\pm0.5\) increase in SSFR). Again, we saw a larger rate of increase in SSFR in the primary progenitor sample as pairs progress from the widest to smallest separation than we saw in the secondary progenitor sample; although overall the secondaries on average showed more SF at all separations.

We found only a slightly significant increase in the Seyfert fraction as close pair interactions progressed to the nearly coalesced stage. However, we did see a stronger rise in the Transition fraction of \(3.1\%\pm0.5\%\). This rise in the Transition fraction was more pronounced, \(4.3\%\pm0.7\%\), for high mass (\(\sim10^{10}\text{M}_\odot\)) close pairs. Kauffmann et al. (2003a) showed that galaxies with stellar mass below \(\sim10^{10}\text{M}_\odot\) are unlikely to host AGN; we found a slight increase of \(1.2\%\pm0.5\%\) in the Seyfert fraction when restricted to this mass range. A larger relative increase in the Seyfert fraction, of \(1.2\%\pm0.7\%\), was seen in group and cluster environments, compared to \(0.9\%\pm0.7\%\) in field environments; however, both of these observed rises in the
4.4 Summary and Discussion

Seyfert fraction are very slight. Rises in the Transition fraction were particularly strong in field environments, with a $4.6\% \pm 1.0\%$ rise, compared with a $2.8\% \pm 1.0\%$ rise in group and cluster environments. Little significant evolution was found in the LINER fraction for any of the mass and environment restrictions.

We found little distinction between evolutions in the BPT diagram for minor or major mergers; the Seyfert fraction rose slightly in both (with a $0.9\% \pm 0.5\%$ rise for major mergers and $0.6\% \pm 0.5\%$ for minors mergers) as did the Transition fraction (with a $3.4\% \pm 0.8\%$ rise for major mergers and $3.1\% \pm 0.7\%$ for minors mergers). The rise in the Transition fraction was particularly strong for major mergers in field as opposed to group/cluster environments. We found no significantly stronger rise in the Seyfert fraction for major mergers than for minor mergers, whereas theoretical models often link AGN activity with major mergers.

We saw strong evidence that merging can cause a change in emission-line processes, leading to an evolution in a galaxy’s location in the BPT diagram. However, based on our BPT analysis, where Seyferts are the only category to definitely harbour AGN activity, we saw little evidence that mergers are triggering much AGN activity during the close pairs stage of merging. Very small increases were seen in the Seyfert fraction in pairs at very low separation, paralleled with an increase in the Transition fraction; this suggests that AGN activity may increase but it may be overwhelmed by SF in low separation close pairs. Kocevski et al. (2012) used Chandra X-ray selected AGN and found no evidence that mergers trigger AGN activity using a higher redshift sample; they showed that at $z \sim 2$, moderate-luminosity AGN are no more likely to be involved in an ongoing merger or interaction relative to non-active galaxies of similar mass. There was no significant evidence of increased AGN activity in major mergers over minor mergers in our sample. We propose that, if AGN activity is ignited in some interacting massive galaxies as theoretically predicted (e.g. Sanders et al., 1988; Springel, Di Matteo & Hernquist, 2005; Hopkins et al., 2006), this process may lead to another class of AGN activity, or take place at the post-merger stage once the merging black holes have coalesced (see Carpineti et al., 2012).
Chapter 5

Time Series Methods of QSO Classification

This chapter and the next chapter are based on the following paper,

“QSO Selection Models Using Light Curve Variability Features from Pan-STARRS”, Caroline Scott, Pavlos Protopapas, Paul Green, Dae-Won Kim, Eric Morganson, 2014, in progress

Overview

Over the next two chapters we present our Quasi-Stellar Object (QSO) classification algorithms: these utilise cutting-edge machine learning techniques to classify QSOs from state-of-the-art photometric light curves from Pan-STARRS. QSOs, and their role in galaxy evolution, were introduced in Section 1.3. They are thought to be caused by the accretion of gas and dust onto a galaxy’s central supermassive black hole, triggering a range of extreme physical processes and emitting over most of the electromagnetic spectrum (Rees, 1984; Lin & Papaloizou, 1996; Ulrich, Maraschi & Urry, 1997; Mirabel & Rodríguez, 1999; Kembhavi & Narlikar, 1999). With the rise of wide field time-resolved surveys, e.g. Pan-STARRS (Kaiser, 2004) and the Large Synoptic Survey Telescope (Ivezic et al., 2008), much effort has been devoted to finding an efficient method to confidently identify QSOs and distinguish them from a large number of potential contaminants, such as variable stars. Spectroscopic confirmation of QSOs is expensive and selection from photometric colours can be unreliable. As an efficient alternative, a number of QSO classification algorithms have been proposed that focus on identifying QSOs by the variable nature of their light curves (e.g. van den Bergh, Herbst & Pritchet, 1973; Hawkins, 1983,
QSO flux variability occurs over a wide range of wavelengths and timescales (Hook et al., 1994; Hawkins, 2002; MacLeod et al., 2012, references therein), and is often attributed to accretion disk instability (Rees, 1984; Kawaguchi et al., 1998; Trèvese, Kron & Bunone, 2001). Once QSOs can be confidently extracted from large scale surveys, substantial QSO samples will allow us to further investigate variability mechanisms in detail; for example, to explore the relationship between AGN variability and black hole mass. Samples at a range of redshifts could also aid cosmological research such as probing the epoch of reionization.

Machine learning algorithms can be trained to classify QSOs by exploiting their unique photometric properties to distinguish them from other objects. Statistical features from light curves can be used to characterise variability trends associated with QSOs; the features from known QSOs can be used as attributes to train classification models to identify other likely QSOs from a large sample of light curves. Some of the features we will use to train classification models are utilised in general time series analysis, i.e. they are not unique to QSO classification, and other features were created specifically for this research and are shown to successfully characterise QSO variability behaviour.

In Section 5.1.2 we introduce one of our classification methods, random forest, which is particularly new to research in astrophysics. We also use support vector machine methods and compare their performance with random forest methods to find the best classification model. In Section 5.3.2 we present a host of new variability features. One of these new features, Stetson K(SAC), is shown to be our most successful feature for QSO classification from time series methods. Another new feature, Eta(SAC), is shown to be our third most successful feature.

Autocorrelation features have previously been shown to be particularly useful for QSO classification (e.g. Thomson & Schild, 1997; Schild, Lovegrove & Protopapas, 2009; Kim et al., 2011; Pichara et al., 2012). Our team have used support vector machine methods (Kim et al., 2011) and random forest classification methods (Pichara et al., 2012) to classify QSOs in MACHOS and EROS-2 data sets. Previous methods were not applicable to Pan-STARRS light curves because the most effective classification features were based on autocorrelation results which were sampling-dependent. Since light curves from different Pan-STARRS medium deep fields (MDFs) have different sampling patterns, these features were not consistent between fields and were unsuccessful for classification in more than one field.
5.1 Machine Learning Classification Models

We modify some features to use slotted autocorrelation instead of standard autocorrelation; this accounts for differences in sampling patterns and provides more robust features for QSO classification for our Pan-STARRS light curves. We will show that the success of our slotted autocorrelation in producing features that are invariant between sampling patterns is key to the application of our classification methods to other time-resolved photometric surveys; such as the upcoming Large Synoptic Survey Telescope (LSST)\(^1\). Slotted autocorrelation makes our models more robust than previous classification methods which were limited to autocorrelation methods. We also modify some features that were not based on autocorrelation yet still found to be dependent on the sampling pattern, to make them invariant between fields.

Since colours are a more conventional method for identifying QSOs in optical photometric surveys (e.g. Sesar et al., 2007), we compare our models and our prediction rates using our variability features with a colour-only analysis. We will show that our variability features can significantly improve performance when compared to a colours-only analysis, and can identify a population of QSOs that a colour analysis alone does not.

The photometric data that we use was not publicly available at the time of conducting this work, and I am grateful for the access provided to me by the Harvard-Smithsonian Centre for Astrophysics during my Predoctoral Fellowship there. This chapter focuses on introducing relevant concepts from computer science, details our data preparation steps, and introduces variability features that will be used to train the models. We review previous research in this field and detail our contributions. In Chapter 6, we present our classification models and evaluate the most successful features for QSO selection from their time series. Some of the most successful features for QSO selection using time series analysis were created during this research and provide new and useful measures to characterise QSO variability. This work crosses boundaries between astrophysics, statistics and machine learning; we first introduce the reader to the concept of machine learning and describe various methods of classification that utilise machine learning models.

5.1 Machine Learning Classification Models

In machine learning, supervised classification models are created by analysing properties of a training set of objects of a known type, or class, with the aim of creating an optimally defined model and applying it to a set of unknown objects to deter-

\(^1\)The LSST is due to conduct a ten year survey, which is planned to commence in January 2022 and will provide even higher resolution light curves than Pan-STARRS.
mine their class. Since the training set comprises a set of known types, the modelling process is called \textit{supervised learning}. A set of quantifiable observations that can be measured for each object is used to determine the object class; these are known as \textit{features} in machine learning but are often referred to as \textit{explanatory variables} in statistical studies. The observations of these features are commonly called \textit{instances} and the output categories are called \textit{outcomes}. Examples of supervised learning models include decision trees (Quinlan, 1993; Yuan & Shaw, 1995), naive Bayes (Duda & Hart, 1973; Zhang, 2004), neural networks (Rumelhart, Hinton & Williams, 1986; Rumelhart, McClelland & PDP Research Group, 1986; Bishop, 1995), and support vector machines.

\section*{5.1.1 Support Vector Machines}

From the research field of machine learning, support vector machine (SVM) methods allow optimised classification models to be formed from a training subset where the true class is known (Boser, Guyon & Vapnik, 1992; Cristianini & Shawe-Taylor, 2000; Bennett & Campbell, 2000; Hsu, Chang & Lin, 2003; Zhang & Zhao, 2004). SVM was traditionally used for binary classification. If training points belong to a class (e.g. in our case QSOs) they are considered positive (otherwise negative; such as foreground stars in our case), and the SVM algorithm constructs an optimally defined hyperplane to separate positive and negative points; thus optimally classifying the objects.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{svm_diagram.png}
\caption{Left: Support vectors are the data points located nearest to the hyperplane (http://www.cac.science.ru.nl/people/ustun/). Right: Non-linearly separable data points can often be mapped to a higher dimensional space where they can be separated by a linear hyperplane (http://www.sbaban.org).}
\end{figure}

Data points located nearest to the hyperplane are the most challenging to classify and are referred to as the support vectors; they directly affect where the op-
5.1 Machine Learning Classification Models

timally separating hyperplane is placed (see Figure 5.1, left). The hyperplane is optimally defined by weighting its vector coordinates so as to maximise the margin between positive and negative support vectors; this margin is known as the functional margin. When support vectors are moved, the hyperplane is shifted (this is not true if other vectors are moved), so the weights assigned by the SVM algorithm, and thus the resulting boundary, are fully determined by the support vectors. Often a data set is not linearly separable by a hyperplane. In this case, positive-class data points that lie in the negative-class region are penalised and a transformation can shift these points so that all data points are then linearly separable.

Often non-linearly separable data points can be mapped to a higher dimensional space where they can be separated by a linear hyperplane (see Figure 5.1, right). However, it is rare that a hyperplane can be defined to completely separate points near the functional margin and various types of SVM have been developed to minimise errors (e.g. soft margin SVM; Cortes & Vapnik, 1995). Once the hyperplane has been defined, broader data sets where the true class is unknown can be plotted in the higher dimensional feature space and a probability can be assigned to their class by the region they occupy in this space.

5.1.2 Decision Trees, Ensemble Learning and Random Forest Classifiers

Decision trees, also known as classification trees in data mining, map feature instances to a classification outcome using a tree structure. Each node represents an input feature, branches stemming from nodes represent conjunctions of feature values, and the leaves represent the class to be outputted (see Figure 5.2). Decision trees are supervised learning methods that infer and learn from the training sample decision rules to be applied at each node (based on the known class of each training set object), creating a trained algorithm that can then be applied to a broader sample.

Ensemble machine learning methods, such as random forests, generate an ensemble of individual classifiers that are diverse yet appropriately representative of the original sample. Random forests (Breiman, 2001) take the output classifications from an ensemble of decision trees and output the most popular result, i.e. the mode output from all decision trees, as the final class. Each decision tree uses a random vector training set that is bootstrapped from the original training set. There are various ways to generate these random vector training sets. Earlier methods utilise bagging techniques developed by Breiman (1996) to perform random selec-
5.1 Machine Learning Classification Models

![Decision Tree Diagram](image)

**Figure 5.2:** Decision tree to classify whether one should or should not play outdoor sports depending on the weather. In this example the weather conditions are the feature instances.

Based on an example from:
http://learnitdaily.com/six-ways-to-address-collinearity-in-regression-models/

...tion (without replacement) from the original training set. More advanced methods for generating random training sets were developed by Dietterich (2000), Breiman (1999) and Ho (1998) to produce optimally accurate and diverse ensembles of classifiers.

5.1.3 Machine Learning in Astrophysics

Machine learning classification methods have previously been used in astrophysics with features derived from time-resolved photometry. Woźniak et al. (2004) used support vector machines to distinguish Mira variables from other types of red variables. Wachman et al. (2009) use cross-correlation and SVM methods to classify Cepheid, RR Lyrae and eclipsing binary stars from the OGLE II periodic variable star catalogue. Debosscher et al. (2007) compare the computational speed and robustness of various automated supervised classification methods for variable stars and are able to adequately separate monoperiodic variables, some of their multi-periodic variables, and eclipsing binaries. Richards et al. (2011) use a random forest variable-star classifier and report a 24% improvement over the best classifier from Debosscher et al. (2007), which was a Bayesian model method of averaging of artificial neural networks. They attribute this success to more accurate periodic-feature estimates and the random forest classifier being more flexible and better suited for multi-class scenarios.

Zhang & Zhao (2004) compare a class of supervised neural networks known as learning vector quantisation (LVQ; see Bazell & Peng (1998)), another class of
supervised neural networks called single-layer perceptron (SLP) and SVM classification methods for AGN selection. They find that when few features are considered, LVQ and SLP perform better; however, when more features are available SVM performs best. Kim et al. (2011) developed a QSO-selection algorithm with the aim of extracting QSOs from a sample of all known variable sources. The algorithm utilises a Support Vector Machine (SVM) and uses 58 known QSOs, 1,629 known variable stars, 4,288 known non-variable stars and also considers microlensing events for model-training purposes. The light curve data set was from the MACHO survey (Alcock et al., 2000), which ran for just over 7 years from 1992 and monitored millions of stars in the LMC and SMC regions with the aim of detecting potential microlensing events.

The most problematic stars to distinguish from QSOs using selection algorithms are Be stars, since they have similar variability properties to QSOs (e.g. Eyer, 2002). Some recent publications of QSO selection algorithms have shown impressive results (e.g. Schmidt et al., 2010; MacLeod et al., 2010; Butler & Bloom, 2011), but these studies were restricted to samples that avoided problematic stars (i.e. those with long fluctuation timescales similar to QSOs). MacLeod et al. (2010) used a damp random walk model to describe correlations between variability and luminosity, black hole mass, and rest-frame wavelength in a sample of ~9000 spectroscopically confirmed quasars in Stripe 82 (Sesar et al., 2007). Butler & Bloom (2011) and MacLeod et al. (2011) used damp random walk models to parametrise the ensemble QSO structure function for known QSOs in Stripe 82, then used their models to predict other QSOs in Stripe 82; in both of these studies, known QSOs were predicted correctly by these models with precision > 90%.

Certain regions of the sky can prove to be particularly problematic for classification models due to issues such as stellar crowding, excessive recent star formation and dust extinction. The Large Magellanic Cloud is such a region, although QSO algorithms specifically for data sets from this area have been created (e.g. Geha et al., 2003; Kim et al., 2011; Pichara et al., 2012). Kim et al. (2011) introduced a method that can be applied to all variable sources in the MACHO data base and is found to correctly identify ~80% of QSOs with ~25% false positive rate\(^2\). Pichara et al. (2012) then improved this quasar identification algorithm by creating a boosted version of a random forest classifier and applied the trained model to MACHOS and EROS-2 objects; their model has ~90% precision and ~86% recall.

We take a similar approach to Kim et al. (2011) and Pichara et al. (2012), now

\(^2\)We will refer to the percentage of correctly identified objects as the model *precision* and the true positive rate as the model *recall*. 
using state-of-the-art Pan-STARRS data. We process light curves directly from the data base over nine medium deep fields (MDFs), each covering a $7^\circ$-squared region of the sky. Exposures are taken using five filters; $g, r, i, z, y$. Zero points are recalibrated for optimised magnitude measurements (Schlafly et al., 2012) and outlier exposures are removed. QSO luminosity function estimates imply that there are $\sim 500$ QSOs in each MDF with $g < 22$ (Fine et al., 2012; Croom et al., 2009), so over nine MDFs we would expect to find 4500 QSOs.

5.2 Data Preparation

5.2.1 Recalibration of Magnitudes

To optimise photometry from the Pan-STARRS pipeline, one has to correct for factors that may contribute to, or contaminate, flux from the source. Such factors include galactic and extragalactic extinction, observational constraints (such as weather, atmospheric seeing and airmass conditions during exposures), and systematic constraints such as detector sensitivity. Systematic effects, i.e. from the telescope and detector, can be more easily understood and accounted for, whilst atmospheric impacts on photometry arise from a continuously changing dynamical system and are harder to identify and correct for.

We found the zero point calibrations directly from the Pan-STARRS pipeline to be unreliable and so adopted zero points from Schlafly et al. (2012) to recalibrate the photometry. They provide photometric calibrations for the first 18 months of Pan-STARRS1 photometry with estimated precision $<0.001$ mag ($g, r, i$-band) and $\sim 0.001$ mag ($z, y$-band). They filtered out $\sim 1/5$th of detections as they were found to be unreliable. Schlafly et al. followed the method used by Padmanabhan et al. (2008) to recalibrate SDSS photometry, to build what they have called the ‘übercalibrated’ data set. They use repeated observations of a set of sources and, insisting on the ideal scenario where the flux should be consistent at all times, solve for other atmospheric and systematic factors.

Morganson et al. (2014) crossmatch Pan-STARRS objects with SDSS DR9 point source objects (classified by the SDSS as having a point-source morphology). This limits the sample to point sources, and removes much of the contamination found in Pan-STARRS data sets (e.g. asteroids, extended sources and galaxies). With the exception of MD02 (since this region does not have SDSS coverage), the sample has sources in the other 9 Pan-STARRS MDFs; MD01 to MD10. We initially were working on a multi-class model to classify QSOs, stars and galaxies from Pan-STARRS...
light curves; however since the morphology constraint removed most galaxies, we then moved to making a binary classification model (QSOs versus stars). There remains a small percentage of galaxies with non-extended morphology that are now viewed as contamination (see Section 6.1.1); these are removed from the training set (where we can determine that they are galaxies from their spectra), but we estimate that 0.73% of our full sample are galaxies.

5.2.2 Preparing the Light Curves

![Figure 5.3: Light curve showing magnitude as a function of modified Julian date (MJD) for the object with ra = 35.18956671 and dec = -4.85610289 in all five optical bands (g, r, i, z, y).](image)

Some exposures may be unreliable due to imaging problems (such as saturated pixels, bad weather conditions etc.), though many of these exposures are flagged out by Schlafly et al. (2012). We attempt to remove potentially remaining outliers by excluding exposures that fall greater than 3σ deviant from the median magnitude for each light curve; this process is done for each filter g, r, i, z, y. We then remove exposures where the magnitude error is greater than 3 times the median magnitude error for each light curve (per filter). Figure 5.3 shows a light curve example from our sample with the median, μ, standard deviation, σ, and χ² shown per filter. Since our sample contains a subset of intrinsically variable objects, our constraints
were set liberally and empirically tested to optimise the removal of outliers without
discarding legitimate data points from intrinsically variable objects. We only use
light curves with at least 25 remaining exposures in each filter after the magnitude
cuts, giving a sample of 253,106 objects.

5.3 Features

We derive various statistical features that can characterise variability behaviour;
such as light curve periodicity, amplitude and autocorrelation features. These vari-
ability measurements can be used to separate QSOs from non-variable objects and
variable stars. It is well known that QSOs can be reasonably well distinguished by
their colours (e.g. Sesar et al., 2007), however, we aim to classify QSOs based on
variability features to offer an alternative means of classification and to be used in
conjunction with colour-based methods to boost performance. In Chapter 6, we use
a combination of these variability features and colours to train an optimal classifi-
cation model, and compare with models trained using only the variability features
and only colours.

5.3.1 Autocorrelation and Slotted Autocorrelation

Some of the features are based on an autocorrelation function, which takes the
cross-correlation of each light curve with itself and describes the similarity between
exposures in the light curve as a function of the time separation between them (in
this case discreet time lags, $\tau$).

$$AC(\tau) = \frac{1}{(N - \tau)\sigma^2} \sum_{i=1}^{N-\tau} (m_i - \bar{m})(m_{i+\tau} - \bar{m}),$$ \hspace{1cm} (5.1)

where $N$ is the number of exposures in the light curve, $\tau = 1, 2, ..., N - 1$ is the time
lag, $\sigma^2$ is the variance, $m$ is the magnitude, $\bar{m}$ the mean magnitude, and $i$ is the
exposure index. $AC(\tau)$ gives the autocorrelation function amplitude at timescale, $\tau$.
The autocorrelation function describes the timescales of variability, i.e. timescales
where magnitudes are well correlated.

Autocorrelation has been used before to quantify quasar variability (e.g. Thom-
son & Schild, 1997; Schild, Lovegrove & Protopapas, 2009; Kim et al., 2011; Pichara
et al., 2012). Autocorrelation values $\sim 0$ imply that there is little variability taking
place between exposure magnitudes at that timescale. Even for a periodic light
curve (for example a pure sinusoidal curve) the sampling may be sparse relative to
the period and thus show little autocorrelation. However, a QSO that has little periodicity but first exhibits brighter magnitudes for a long time, then switches to exhibit fainter magnitudes for a long time, will show strong autocorrelation because the trend of previous points helps to predict upcoming points. See Kim et al. (2011) for a comparison between autocorrelation outputs for non-variables, various types of variable stars, and QSOs.

Autocorrelation outputs are in the range [-1,1], with the first output being (by definition) the correlation of the light curve with itself; thus the first output is perfectly correlated and has value 1. The lags, i.e. the series \( \tau \), are created by taking a fixed interval and multiplying this by integers \( k = 0, \ldots, N - 1 \); this fixed interval has to be chosen appropriately in order for the autocorrelation to give us meaningful results. Essentially, we are averaging the cross-products \( <m_i, m_j> \) for all magnitudes \( m_i \) and \( m_j \) that are separated by the given time lag, \( \tau \). If this average is high then we infer that the samples separated by \( \tau \) are very similar and this implies periodicity at this timescale. However, with uneven sampling we are unable to ensure a large number of exposures that are separated exactly by \( \tau \). Instead of using a constant, \( \tau \), for the lag we can use intervals, i.e. slots, and use the exposures that lie within these slots.

**Slotted Autocorrelation**

Within each MDF the sampling rate of exposures is mostly consistent between objects since each Pan-STARRS image takes a snapshot of the whole MDF; although data points can be missing where exposures have been flagged out because of photometric problems or bad atmospheric conditions. Therefore, even though the sampling is ‘patchy’, the time series features that we calculate give a consistent statistical measure of intrinsic variability for sources within individual fields. This means that if we were only working in one MDF, the standard autocorrelation formula would be helpful to distinguish variables from non-variables (e.g. Kim et al., 2011; Pichara et al., 2012, with MACHOS and EROS-2 light curves).

However, between MDFs the sampling pattern is different, and so features that rely on the autocorrelation no longer give a consistent way to characterise variability. We work around the sampling unevenness by using slots instead of discrete time lags (Mayo, 1974; Tummers & Passchier, 1996).

Each lag \( k \) is defined by a slot \( [(k - 0.5) \times \text{slot\_size}, (k + 0.5) \times \text{slot\_size}] \), and in the slotted correlation we average the cross-products that fall into each slot. We set the slot size empirically to twenty days to satisfy the following constraints:
(i) If the slot size is set too high, resolution will be sacrificed in the autocorrelation function

(ii) If the slot size is set too low, the risk increases of some slots being sparsely filled (leading to no good average and/or empty slots).

We set the maximum lag to be the full time span of the light curve, and define the number of lags by dividing the length of the signal (i.e. the full timespan of the light curve) by our chosen slot size. The slots are defined by multiplying the slot size by a series of integers \( k = 1, \ldots, N \). All magnitudes are scaled by subtracting the mean and dividing by the variance. For all exposures, \( i, j \), each exposure time difference, \( t_j - t_i \), is allocated to one of these slots; the corresponding scaled magnitudes are then multiplied, \( m_i \ast m_j \). These magnitude products are summed for each slot, \( k \Delta \tau \), and then finally divided through by the number of cross-products in the slot.

For each slot, \( k \Delta \tau \), where \( k = 1, \ldots, N \), the slotted autocorrelation function is defined as follows:

\[
SAC(k \Delta \tau) = \frac{1}{N} \sum_{i,j} (m_i \ast m_j),
\]

where \( N \) is the number of cross-products within a slot, \( \Delta \tau \) is the slot width (set to 20 days), and the sum is over all cross-products \( m_i \ast m_j \) for all epochs \( i, j \) with a time difference in the interval \((k - 0.5)\Delta \tau < t_i - t_j < (k + 0.5)\Delta \tau\).

**SAC Examples for Various Light Curves:** Figures 5.4-5.12 show examples of \( r \)-band light curves from our sample and the corresponding SAC output (as a function of the time lag). Examples of QSOs and various stellar types are shown. For the first time lag the SAC output is always 1, since this is the correlation of the light curve with itself. The SAC error bars are defined by \( 1/\sqrt{N} \) where \( N \) is the number of cross-products within the slot. Error bars for each object are centered around the line \( SAC = 0 \) as opposed to around the data points themselves, and these errors are joined by the smoothed line; SAC points which lie outside of this error-region (shown in pink) are considered non-zero with significance. Non-zero SAC points indicate light curve variability on the shown slotted timescale.

Observe that for the QSOs in Figures 5.4 and 5.5 the SAC output is very erratic; indeed we expect QSOs to show strong variability at various timescales. Interestingly, out of 18 spectroscopically confirmed narrow absorption line (NAL) QSOs, we found 3 that show very little variability from the SAC output. These are shown in Figure 5.6 and have spectroscopic redshifts \( z=2.48 \) (top), \( z=2.27 \) (middle) and \( z=2.58 \) (bottom). The 15 other NAL QSOs show normal variability (one is
shown in 5.5 (top); this has $z=2.48$ and was flagged as a NAL QSO at high velocity in the spectroscopic catalogue).

Figure 5.7 shows the SAC output for some A-type stars; observe that the bottom object displays variability when compared to the other two examples. We found most examples of variable stars when looking at K-types (Figure 5.11; middle and bottom) and M-types (Figure 5.12; top and bottom).

### 5.3.2 List of Features

Here we describe the features used for our supervised learning models. Some of these features were proposed by Kim et al. (2011); the features based on their autocorrelation output were updated to use our slotted autocorrelation output (Eqn. 5.2). New features which were created during this work are also described below. These features are all calculated per filter ($g, r, i, z, y$) for each light curve.

#### Non-SAC Indices

(i) $Eta$: For a normal distribution with variance, $\sigma^2$, and mean square successive difference

$$
\delta^2 = \frac{1}{(N - 1)} \sum_{i=1}^{N-1} (m_{i+1} - m_i)^2,
$$

the ratio $\eta^*$ is defined as

$$
\eta^* = \frac{\delta^2}{\sigma^2}.
$$

For further information, see the definition in von Neumann (1941) and the discussion in Press (1969). The statistic $\eta^*$ is used to quantify the extent to which magnitudes $m_0, ..., m_N$ are independent (see Kim et al. (2011) for an analysis of how $\eta^*$ varies for various sources).

We found $\eta^*$ to be dependent on the sampling pattern of the light curves, which in turn depends on which MDF the object lies in. We modified this feature to make it invariant between MDFs as follows:

$$
Eta = \frac{1}{(N - 1)} \sum_{i=1}^{N-1} \frac{(m_{i+1} - m_i)^2}{(t_{i+1} - t_i)},
$$

(5.5)
Figure 5.4: $r$-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed QSOs.
Figure 5.5: $r$-band light curves and the corresponding slootted autocorrelation plotted against time lag (in days) for spectroscopically confirmed QSOs. The top object is a narrow absorption line QSO.
Figure 5.6: $r$-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed narrow absorption line QSOs.
Figure 5.7: \( \tau \)-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed A-type stars.
Figure 5.8: \(r\)-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed B-type stars.
Figure 5.9: r-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed F-type stars.
Figure 5.10: \(r\)-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed G-type stars.
Figure 5.11: $r$-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed K-type stars.
Figure 5.12: r-band light curves and the corresponding slotted autocorrelation plotted against time lag (in days) for spectroscopically confirmed M-type stars.
for magnitudes $m_0, ..., m_N$ and exposure times $t_0, ..., t_N$. See Figure 5.13 to compare the distribution of $Eta$ for MD01 and MD09.

(ii) **Casum**: the range of the cumulative sum, $S$, of each light curve (Ellaway, 1978; Kim et al., 2011) is defined as $\text{max}(S)-\text{min}(S)$, where

$$S_t = \frac{1}{\sqrt{N\sigma}} \sum_{i=1}^{t} (m_i - \bar{m}), \tag{5.6}$$

for $t = 1, 2, ..., N$.

We found this feature to vary between MDFs and modified it by dividing by the modified Julian Date (mjd) range: $\text{mjd\_range} = \text{mjd(\text{where(max(S))})} - \text{mjd(\text{where(min(S))})}$.

$$\text{Cusum} = \frac{\text{max}(S)-\text{min}(S)}{\text{mjd\_range}} \tag{5.7}$$

(iii) **Con**: This index is based on the number of consecutive magnitudes that lie past a boundary region outside of which magnitude fluctuations are thought to signify variability (Wozniak, 2000; Shin, Sekora & Byun, 2009). Wozniak (2000) defined variable objects in part by measuring the amount of consecutive data points that lie above or below a $3\sigma$ boundary line from the median flux. In our case, $\text{Con}$ calculates the number of magnitude data points, with a minimum of three consecutive points, that lie outside of a $2\sigma$ boundary region from the median flux. This number is normalised by $N-2$, where $N$ is the number of data points. $\text{Con}$ is naturally larger for variable objects.

(iv) **Anderson-Darling**: The Anderson-Darling statistic is used to assess whether a data set is drawn from a specified probability distribution (Anderson & Darling, 1952); e.g. it can test whether data can be well described by a normal distribution. If a data set is indeed drawn from a given probability distribution then the data can be mapped to a Uniform distribution; therefore, upon testing, the distance from uniformity should be small. The data are first ordered, $\{Y_1 < ... < Y_n\}$ (in our case these are magnitudes per filter), then the Anderson-Darling statistic $A^2$ is calculated as follows:

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \left[ \ln(\Phi(Y_i)) + \ln(1 - \Phi(Y_{n+1-i})) \right] \times C, \tag{5.8}$$

where $\Phi$ is the cumulative distribution function of the assumed underlying
probability distribution, $n$ is the number of data points and $C = (1 + \frac{4}{n} - \frac{25}{n^2})$. The factor $C$ is a modification to the original Anderson-Darling statistic to test for normality in our case, where both the mean and variance are unknown. This statistic can then be compared with critical values for the distribution it is being tested against\(^3\). We use the statistic in Equation 5.8 as a feature to test for variability.

(v) $\text{Std}_\text{Err}$: This is the ratio between the standard deviation of the magnitudes and the photometric uncertainty:

$$\text{Std}_\text{Err} = \text{std}(\text{mag})/\text{median}(\text{mag\_error}),$$

(5.9)

where std(mag) is the standard deviation of magnitudes and mag\_error is the magnitude error.

(vi) $\text{Skewness}$: This determines if the distribution is skewed, i.e. not centred around 0 (e.g. a Normal distribution would have skewness value 0).

(vii) $\text{Kurtosis}$: Kurtosis is the fourth central moment divided by the square of the variance.

(viii) $\text{Stetson\_K}$: This variability index was derived to characterise how magnitudes are distributed between the maximum and minimum values (Stetson, 1996). The Stetson\_K value can distinguish pure sinusoidal from Gaussian distributions, and can separate non-variable sources from variables.

$$\text{Stetson\_K} = \frac{1}{\sqrt{N - 1}} \frac{\sum_{i=1}^{N} |\delta(i)|}{\sqrt{\sum_{i=1}^{N} \delta(i)^2}}$$

(5.10)

where $N$ is the number of exposures in the light curve and $\delta(i) = (m_i - \bar{m})/\epsilon(m_i)$ for exposure index $i$. $\epsilon(m_i)$ is the magnitude error for exposure $i$.

(ix) $\text{Quartile Range}$: We define this as the first magnitude quartile (i.e. the median of the bottom 25% of sorted magnitudes) subtracted from the third magnitude quartile (i.e. the median of the top 25% of sorted magnitudes).

$$Q_{\lambda,3} - Q_{\lambda,1}$$

(5.11)

\(^{3}\text{See} \url{http://en.wikipedia.org/wiki/AndersonDarling_test} \text{for more information on the Anderson-Darling test.}\)
where $Q_{\lambda,1}$, $Q_{\lambda,3}$ are respectively the first and third magnitude quartiles.

(x) \textit{SNR}: This feature calculates a form of signal-to-noise measure as follows:

$$\frac{Q_{\lambda,3} - Q_{\lambda,1}}{\epsilon_\lambda}$$ \hfill (5.12)

where $Q_{\lambda,1}$, $Q_{\lambda,3}$ are respectively the first and third magnitude quartiles and $\epsilon_\lambda$ is the median magnitude error in each filter, $\lambda$.

(xi) \textit{Standard Deviation}: The standard deviation, $\sigma$, was calculated as follows:

$$\sigma = \sqrt{\frac{1}{N_\lambda - 1} \sum_{i=1}^{N_\lambda} (m_{i,\lambda} - \bar{m}_\lambda)^2}$$ \hfill (5.13)

for each magnitude, $m$, with mean magnitude, $\bar{m}$, and total number of exposures, $N$, in each filter $\lambda$.

(xii) $\sigma/m$: This is a simple variability index and is defined as the ratio of the standard deviation, $\sigma$, to the mean magnitude, $m$. If a light curve displays strong variability, $\sigma/m$ is usually large.

(xiii) \textit{Chi}^2: Based on the standard $\chi^2$ statistic:

$$Chi^2 = \frac{1}{N_\lambda - 1} \sum_{i=1}^{N_\lambda} \frac{(m_{i,\lambda} - \bar{m}_\lambda)^2}{\epsilon_{i,\lambda}^2}$$ \hfill (5.14)

for each magnitude, $m$, with mean magnitude, $\bar{m}$, median magnitude error, $\epsilon$, and total number of exposures, $N$, in each filter $\lambda$.

(xiv) \textit{Median magnitude} and \textit{Median magnitude error}: We calculate the median magnitude and median magnitude error over all magnitude points in each filter $g, r, i, z, y$ for each light curve. Data points with relatively large magnitude errors were previously removed as described in the light curve preparation stages; the median of remaining magnitude errors for each light curve are used as a feature for model training.

(xv) \textit{Colours}: We calculate ten photometric colours: $g-r$, $g-i$, $g-z$, $g-y$, $r-i$, $r-z$, $r-y$, $i-z$, $i-y$, $z-y$ as follows:

$$\text{colour}_{\lambda-\lambda'} = m_\lambda - m_{\lambda'}$$ \hfill (5.15)
for each filter $\lambda, \lambda'$ in $g, r, i, z, y$. We calculated the median colour $\lambda - \lambda'$ as median$(m_\lambda)$-median$(m_{\lambda'})$; i.e. when we refer to the colour $g - r$ we describe the median colour.

(xvi) $N$: The number of light curve data points used to calculate the features.

**Note:** Since colours are already a standard tool for quasar identification, we use colours in addition to our variability features to test whether a features-only model works better than a model trained using only colours, or to see if colours in addition to our features improve performance.

**SAC Indices**

The slotted autocorrelation (SAC) outputs an array of SAC, SAC error, and time lag, $\tau$, values and we use these to create SAC indices. Kim et al. (2011) define upper and lower boundary lines at $\pm 4\sigma$ from the average AC value. These boundary lines were defined empirically after comparing AC outputs for variable and non-variable objects, and accordingly we define new boundary lines for our SAC outputs (the first two features described below are based on three new boundary line definitions).

- $\sum SAC < e^{-1}$ and $\sum SAC > e^{-1}$: the sum of SAC points where SAC absolute values lie respectively below and above the boundary line defined by the natural exponential to the power of -1; $e^{-1}$.
- $N(SAC < e^{-1.2})$ and $N(SAC > e^{-1.2})$: the number of SAC points where SAC absolute values lie respectively below and above the boundary line defined by the natural exponential to the power of -1.2; $e^{-1.2}$.
- Median(Cusum(SAC)): The median of the cumulative sum of all SAC outputs.
- Median(Cusum_20(SAC)): The median of the cumulative sum of the SAC output for the first twenty time lags. This encodes useful information about the SAC at small timescales.
- Sigma(SAC): Standard deviation of all SAC outputs.
- Eta(SAC): This is analogous to Eta and measures the extent to which SAC outputs for neighbouring time lags are independent.
\[ Eta(SAC) = \frac{1}{(N-1)} \sum_{i=1}^{N-1} \frac{SAC_{\tau+1} - SAC_{\tau}}{(\tau_{i+1} - \tau_i)}, \] (5.16)

for SAC values \( SAC_{\tau(0)}, ..., SAC_{\tau(N)} \) and exposure times \( \tau_0, ..., \tau_N \).

- **Stetson\_K(SAC)**: This is analogous to Stetson\_K, where we now characterise how SAC outputs are distributed between the maximum and minimum values. This quantifies how much the SAC distribution deviates from 0 and is therefore a useful measure; we can see in Figure 5.4 that QSO SAC values show strong deviation from 0 whereas stars generally do not.

\[
\delta^2 = \frac{1}{(N-1)\sigma^2} \sum_{i=1}^{N-1} (SAC_{\tau+1} - SAC_{\tau})^2, \] (5.17)

\[
Stetson\_K(SAC) = \frac{1}{\sqrt{N-1}} \frac{\sum_{\tau=1}^{N} |\delta(\tau)|}{\sqrt{\sum_{\tau=1}^{N} \delta(\tau)^2}} \] (5.18)

where \( N \) is the number of lags and \( \delta(\tau) = (SAC_{\tau} - \bar{SAC})/\epsilon(SAC_{\tau}) \) for lag \( \tau \). \( \epsilon(SAC_{\tau}) \) is the SAC error for lag \( \tau \).

### 5.4 Invariance Between Pan-STARRS Fields

In order to create an optimised QSO classification method that is fully applicable between Pan-STARRS MDFs and to future light curve data sets (such as from the next generation LSST) it is important that our features are invariant between light curve sampling patterns. Previous research has relied on data with the same sampling patterns to give consistent features; such as MACHOS light curves (Kim et al., 2011) and EROS-2 light curves (Pichara et al., 2012). However, the various Pan-STARRS MDFs have unique sampling patterns and this leads to different distributions for some variability features between MDFs. This significantly reduces the ability of machine learning models to classify objects between MDFs and limits robust modelling to individual Pan-STARRS MDFs.

Using the original features used by Kim et al. (2011) and Pichara et al. (2012), we found that QSO models that were optimised in MD01 and then applied to other MDFs failed to work well. We modified Cusum and Eta as described in Section 5.3.2 in order to make these features invariant regardless of the light curve sampling
5.4 Invariance Between Pan-STARRS Fields

pattern. The autocorrelation features were particularly affected by differences in the sampling patterns; through slotted autocorrelation we have addressed this issue. Kim et al. (2011) originally used features called \( N_{\text{above}} \) and \( N_{\text{below}} \), which counted the number of points situated above and below boundary lines defined empirically at \( \pm 4\sigma \) from the average AC value. We defined a host of new features using our slotted autocorrelation which were designed and tested to optimally characterise variability. These new SAC features are no longer dependent on the sampling pattern; they are described in the SAC Indices list in Section 5.3.2.

Figure 5.13 illustrates the invariance of some of our non-SAC indices (Eta, Cusum and Stetson\_K) between MD01 and MD09. There are small observable differences between the Cusum and Stetson\_K distributions in Figure 5.13; MD09 has more than twice as many light curves as MD01 and we feel that these distributions agree well. Figure 5.14 illustrates the invariance of some of our SAC indices (Stetson\_K (SAC) (top), Eta(SAC) (middle) and \( \sum SAC(i) > e^{-1} SAC(i) \) (bottom)) between MD01 and MD09. These improved invariant features lead to a significant improvement in our QSO classification models; this will be discussed in the next chapter.
Figure 5.13: Invariance of modified $g$-band features between MD01 and MD09 for Eta (top), Cusum (middle) and Stetson_K (bottom).
5.4 Invariance Between Pan-STARRS Fields

Figure 5.14: Invariance of modified $g$-band features between MD01 and MD09 for Stetson $K$ (SAC) (top), Eta(SAC) (middle) and $\sum SAC > e^{-1}$ (bottom).
5.5 Summary

In this chapter we introduced machine learning classification methods such as support vector machines and random forest classifiers. These methods will be employed in the next chapter to build QSO classification algorithms. We reviewed recent literature where modern computer science methods have been used in astrophysics and particularly for QSO classification.

Our light curve data set from Pan-STARRS was introduced. This data base spans nine medium deep fields, each covering a $7^\circ$-squared region of the sky with exposures in five optical filters; $g, r, i, z, y$. Only objects determined by the SDSS as having a point-source morphology are used, which removes much of the contamination usually found in Pan-STARRS data sets (e.g. asteroids, extended sources and galaxies). MD02 is not used in our work, since this region does not have SDSS coverage. Exposures that were likely to be unreliable due to imaging problems were removed, and after this process only light curves with at least 25 remaining exposures in each filter were used; this resulted in a sample of 253,106 objects.

We then introduced various features and calculated their value for each light curve. These features characterise variable behaviour by measuring light curve properties such as periodicity, amplitude and autocorrelation features, and can be utilised to separate QSOs from non-variable objects and variable stars. Some of the features were previously used for QSO classification models by Kim et al. (2011) (with MA-CHOS light curves) and Pichara et al. (2012) (with EROS-2 light curves); however, we found that a number of the features failed to provide consistent measures for light curves in different Pan-STARRS medium deep fields due to differences in the sampling patterns between fields. We modified these features to account for sampling differences so as to make them invariant between fields. Cusum and Eta were modified, and all autocorrelation features were updated to use slotted autocorrelation instead. We contributed a host of new features, including Stetson $K$ (SAC), which we will show in the next chapter to be our strongest feature for QSO selection. In the next chapter we propose classification algorithms which utilise our variability features to optimally classify QSOs in our photometric sample.
Chapter 6

Random Forest and SVM QSO Classification Models

Overview

Likely QSO candidates are often initially selected from their photometric colours before being proposed for follow-up spectroscopy. Most SDSS spectroscopically measured QSOs were first colour-selected (Richards et al., 2002; Schneider et al., 2007). The Time Domain Spectroscopic Survey (TDSS) is in the process of building a candidate list of 100,000 variable objects using Pan-STARRS and SDSS photometry (see Section 2.1.4). TDSS pilot survey QSO candidates were initially colour-selected by our team (led by Paul Green at the Harvard-Smithsonian Center for Astrophysics). However, as we will show in this chapter, QSO selection by colours can be unreliable. There was an overlap between members of the TDSS collaboration and members of the Time Series Center at Harvard’s Institute of Applied Computational Science (led by Pavlos Protopapas); as a result, contemporary time series methods were considered as an alternative means of QSO classification for Pan-STARRS objects.

We aim to apply time series data analysis methods to reliably identify QSO candidates for the TDSS by utilising photometric variability features instead of relying on colour-selected candidates. Previous work on QSO classification using time series methods was discussed in Section 5.1.3, and in Section 5.4 we highlighted that previous methods are not directly applicable to Pan-STARRS data because of sampling differences between fields. We create new, optimised models to account for these sampling differences. The selection of likely QSO candidates from Pan-STARRS for future TDSS follow-up spectra is the main motivation behind this work, although the models we will present are applicable to other time-resolved surveys (such as the future LSST mission).
We create Random Forest (RF) and Support Vector Machine (SVM) algorithms to classify Quasi-Stellar Objects (QSOs) using time series data from the time-resolved, optical, photometric survey Pan-STARRS. We train our models using a set of spectroscopically confirmed QSOs and stars, and make use of various light curve features which characterise variability trends associated with QSOs. These features were presented in Section 5.3 and measure time series properties such as periodicity, amplitude and statistical distributions. The most successful features in our training process are identified and discussed. Many of the features are new and have been derived or modified specifically for this work; these features provide useful statistical measures that can be used to quantify and characterise QSO light-curve variability for future studies. Our features can be added to the vocabulary of useful time series measures in the field of computer science; the potential applications lie beyond astrophysics (e.g. the analysis of time series data in fields such as economics, neuroscience and climate modelling).

We train the RF and SVM models with our features in 5 optical filters ($g, r, i, z, y$) and with a set of optical colours, then compare models trained using both variability features and colours with models trained using only features and only colours. The method of model testing by cross-validation is described and a 10-fold cross-validation test is run on our models. This utilises objects from the training set to approximate what the prediction and recall rates may be when the models are applied to a full sample of unknown objects.

We conduct a more robust test by leaving a portion of our spectroscopic sample aside during the training process, running our models on the full sample, and then comparing the predicted types with known types. This provides a ‘blind test’ to evaluate the success of the model since these objects have not been involved in the model training; the blind test sample is as close to a representation of the full sample as we can get with our limited spectroscopic catalogue. Results of our tests show that our most successful models were trained using variability features in addition to photometric colours. We show that models trained using only colours predict a large and unrealistic number of QSOs, and that by including our time series features in the modelling process the number of predicted QSOs is significantly reduced to a more realistic amount that still has a high precision and recall rate for known QSOs. Models trained using our variability features also predict a population of likely QSOs that colours alone do not.
6.1 Modelling

6.1.1 Creating the Training Set

We use a training set of objects that were visually classified as QSOs or stars from spectra measured by the SDSS-III’s Baryon Oscillation Spectroscopic Survey (BOSS). 371 objects in our spectroscopic catalogue were measured for the TDSS pilot survey, which is currently building a candidate list of 100,000 variable objects using the BOSS spectrograph, and the rest are from the SDSS catalogue. QSOs from the SDSS catalogue were mostly colour-selected (Richards et al., 2002; Schneider et al., 2005). Some of the non-colour-selected spectra were obtained because, for example, they were the closest optical object to a ROSAT (X-ray) or FIRST (20cm radio) source, or they may have been targeted as a potentially interesting class of star such as a Cepheid variable or white dwarf (Schneider et al., 2007). The TDSS pilot survey candidates were colour-selected by our team (led by Paul Green at CfA).

Only visual classifications with a high confidence level are used. Of our 253,106 light curves; 2,048 were visually classified as QSOs, 5,116 as stars, 19 as galaxies (0.25% of the confidence 1 training set) and the rest remain unclassified. Most galaxies are removed by the SDSS point source morphology constraint when building the light curve data set. The 19 visually identified galaxies are contaminants and are removed so as to not affect the model training, but this highlights that a small percentage of the unknown sample will be galaxies and will incorrectly be predicted to be either QSOs or stars. After removing outliers we are left with a set of 7,164 known stars and QSOs. All features in both the full sample and the training set are normalised by subtracting the mean and dividing by the standard deviation for that feature. Figure 6.1 shows colour-magnitude and colour-colour distributions for our full sample of 253,106 light curves using magnitudes from SDSS DR8 (since Pan-STARRS does not have u-band filters), and shows where the known QSOs (red) and stars (blue) lie on these distributions.

For evaluation purposes, we use a subset of approximately five sevenths of the known QSOs (1,468) and five sevenths of the known stars (3,643) to train our model. We then apply our models to the full set of light curves and use the remaining two sevenths of known QSOs and stars to conduct a ‘blind test’ to determine if the correct type has been predicted for spectroscopically known objects that were not included in the training.

Our training set is not fully representative of our full Pan-STARRS data set. The ratio of QSOs to stars in the training set is \( \sim 0.4 \), which is much higher than
the realistic ratio of QSOs to stars in our full light curve sample. We expect contamination from galaxies classified as point source objects by the SDSS; these are expected to comprise <1% of the sample. Most objects in the training set were selected for spectroscopic follow-up based on their colours. In a complete QSO sample, we would expect the region occupied by QSOs in Figure 6.1 (left) to extend to fainter $g$-band magnitudes; this implies that our training set is biased towards brighter objects than the full sample of unknowns (shown in black) which could lead to unreliable classifications for fainter unknowns. At $z \sim 2.7$ SDSS colours for QSOs overlap with the colour region inhabited by A-type stars and blue horizontal branch stars, and it is difficult to distinguish QSOs in the redshift range $2.2 < z < 3.5$ from stars, especially since they vastly outnumber QSOs (Fan, 1999; Ross et al., 2012). Since our QSOs were mostly colour-selected from the SDSS, they are unlikely to adequately represent this population of QSOs in the training set.

### 6.1.2 Random Forest Model

We conduct our modelling using RapidMiner, an open-source software platform that provides an integrated environment for machine learning. We use the Weka:W- RandomForest module with 10 decision trees to train a model from our training set. Three models are trained using the following attributes for the training set:

- RF Model I: all features and colours; including features for all filters $g, r, i, z, y$
6.1 Modelling

- RF Model II: \( g \)-band only features
- RF Model III: colours only

We train RF Model II with \( g \)-band features only so that colours are not indirectly used in the modelling (i.e. so that there is not a combination of the median magnitude features for different filters). For comparison purposes, we train RF Model III with colours only since separation by colour is an established method for QSO identification (e.g. Richards et al., 2002; Sesar et al., 2007; Schneider et al., 2007).

Cross-Validation

A cross-validation was then run to test the performance of the learning operator. The cross-validation randomly chooses 90\% of the training set to re-train the model, then tests this model on the remaining 10\%. Since the true type for the 10\% is known, this allows a comparison to test the accuracy of the model predictions. Both class precision and class recall are outputted. This process is repeated ten times and the average outputted success-rate is recorded. See Table 6.1 for the cross-validation results for all three RF models. Notice that the cross-validation results for RF Model I shows the highest accuracy recall and precision rate for QSOs (with 96.9\% and 98.6\% respectively); this is where we use a combination of features and colours as opposed to \( g \)-band features only or colours only.

<table>
<thead>
<tr>
<th>RF Cross Validation Table</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True S</strong></td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td><strong>RF Model I</strong></td>
</tr>
<tr>
<td>Predicted S</td>
</tr>
<tr>
<td>Predicted Q</td>
</tr>
<tr>
<td><strong>Class Recall</strong></td>
</tr>
<tr>
<td><strong>RF Model II</strong></td>
</tr>
<tr>
<td>Predicted S</td>
</tr>
<tr>
<td>Predicted Q</td>
</tr>
<tr>
<td><strong>Class Recall</strong></td>
</tr>
<tr>
<td><strong>RF Model III</strong></td>
</tr>
<tr>
<td>Predicted S</td>
</tr>
<tr>
<td>Predicted Q</td>
</tr>
<tr>
<td><strong>Class Recall</strong></td>
</tr>
</tbody>
</table>

Table 6.1: Cross Validation results for Stars (S) and QSOs (Q) for RF Models I, II and III. There were 1,468 QSOs and 3,643 stars in the training set.
6.1 Modelling

The cross-validation tests the model accuracy whilst being constrained to the training set. In our case the training set is not fully representative of the general population of Pan-STARRS objects, as discussed in Section 6.1.1. However, without a broad spectroscopic catalogue with which to test our results, the cross-validation does provide a useful test, as does the ‘blind test’ which we describe in Section 6.3.1.

Pichara et al. (2012) extended the work of Kim et al. (2011) by using a boosted version of a random forest classifier (Breiman, 2001) to classify QSOs in both EROS-2 and MACHO data sets. They utilised some new features, including parameters of a continuous autoregressive model. The accuracy of their model in both EROS-2 and MACHO training sets is \( \sim 90\% \) precision and \( \sim 86\% \) recall. An \( F \)-score indicator was used to present the results of their cross-validation; this calculates the harmonic mean of precision and recall values as follows:

\[
F\text{-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

For the EROS-2 sample, they report an \( F \)-score of 0.868 for their boosted random forest model with their new features, and when applying an SVM model with their new features to the same data set they report an \( F \)-score of 0.855. For the MACHO sample, they report an \( F \)-score of 0.877 for their boosted random forest model with their new features, and when applying an SVM model with their new features to the same data set they report an \( F \)-score of 0.824. Our RF Model I cross-validation gives an improved precision rate of 98.6\% and recall rate of 96.9\%, with \( F \)-score=0.978. It is unfair to directly compare our cross-validation results with those from a MACHO sample since they are very different data sets; the ratio of quasars to stars is significantly lower in the MACHO data set. However, it would be an interesting test to apply our features to a MACHO data set in future.

6.1.3 SVM Model

We conduct our SVM modelling on RapidMiner using a LibSVM module with C-SVC SVM type, a linear kernel and \( \epsilon=0.001 \) for the tolerance of termination criterion. Three models are trained using the listed attributes for the training set:

- SVM Model I: all features and colours; including features for all filters \( g, r, i, z, y \)
- SVM Model II: \( g \)-band only features
- SVM Model III: colours only
Table 6.2 shows the cross-validation results for SVM Models I-III. Notice that the cross-validation results for SVM Model I shows the highest accuracy recall and precision rate for QSOs (with 97.7% and 98.3% respectively). This is the same as we observed for RF Model I in Table 6.1, where we use a combination of features and colours.

### SVM Cross Validation Table

<table>
<thead>
<tr>
<th>SVM Model</th>
<th>Predicted S</th>
<th>True S</th>
<th>Predicted Q</th>
<th>True Q</th>
<th>Class Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM Model I</td>
<td>3618</td>
<td>34</td>
<td>25</td>
<td>1434</td>
<td>99.1%</td>
</tr>
<tr>
<td>Class Recall</td>
<td>99.3%</td>
<td>97.7%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM Model II</td>
<td>3586</td>
<td>95</td>
<td>57</td>
<td>1373</td>
<td>97.4%</td>
</tr>
<tr>
<td>Class Recall</td>
<td>98.4%</td>
<td>93.5%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM Model III</td>
<td>3395</td>
<td>279</td>
<td>248</td>
<td>1189</td>
<td>92.4%</td>
</tr>
<tr>
<td>Class Recall</td>
<td>93.2%</td>
<td>81.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 6.2:** Cross Validation results for Stars (S) and QSOs (Q) for SVM Models I, II and III. There were 1,468 QSOs and 3,643 stars in the training set.

The QSO recall rate for SVM Model III is particularly low, at only 81.0%, whereas for the other models (including all three RF models) the QSO recall rate is at least 92.0%. For a direct model comparison using the same attributes, the QSO recall rate is 95.9% for RF Model III. Also, the QSO precision rate for SVM Model III is only 82.7%, whereas for RF Model III it is significantly higher at 96.8%. The cross-validation implies that when using only colours as attributes, RF is a more successful classification method than SVM.

Kim et al. (2011) reported a precision rate of 75.0% and recall rate of 82.8% for their SVM QSO classification model using a MACHO training set. These rates are lower than for our SVM Models I and II, however, their recall rate is higher than for our SVM Model III. Many of our features overlap since our features are an extension of their feature list. Their MACHO data set consisted of 40 million light curves for objects in the Large Magellanic Cloud; with 58 known QSOs, 1,629 variable stars and 4,288 non-variables used in their training set. Our data sets are very different;
we expect a larger ratio of QSOs to stars in our Pan-STARRS data set and we have significantly more QSOs to train our model with.

In Sections 6.3 and 6.4, these models are applied to our light curve data set for the nine Pan-STARRS MDFs in order to extract a large sample of likely QSOs, but first we will analyse the strongest classification features.

### 6.2 Identifying the Best Features

The weights outputted by the linear SVM kernel model are the coordinates of vectors orthogonal to the optimal hyperplane. The highest positive coefficients represent features that have been weighted up during the optimisation, hence these were the most useful features during the modelling for class separation. Therefore the absolute value of the kernel model weights tells us how important the features were considered to be during the SVM modelling. Figure 6.2 shows the kernel model weights for an SVM model created using only our $g$-band features. Higher weighted features such as Stetson $K(SAC)$, Med$_{mag}$ and Eta(SAC) were considered better for optimal linear separation than lower weighted features such as N and Skewness.

The feature with the highest SVM kernel weight is Stetson $K(SAC)$, followed by Med$_{mag}$ and Eta(SAC) and so these features were deemed to be the most successful at separating QSOs from stars during the SVM modelling. In Figure 6.3, the normalised counts of various $g$-band feature distributions are shown for objects in the training set. We know the types of these objects from their spectra; stars are shown in blue and QSOs are shown in red. The top six features from the SVM kernel weight histogram are shown, as well as the bottom two features. For the top six features (e.g. Stetson $K(SAC)$ and Med$_{mag}$ etc.) there is a clear distinction between the feature distributions for stars and QSOs, hence a combination of these features are useful for our classification purposes.

For a distribution which is roughly symmetric around 0, Stetson $K$ quantifies how much the distribution deviates from 0; it will be larger for distributions that spend a lot of time away from 0 and where many data points are located near the maximum and minimum values. The slotted autocorrelation was very efficient for separating QSOs from stars, since QSOs tended to show strong deviation from 0 whereas stars generally did not (see Figures 5.4-5.12). By applying Stetson $K$ to the SAC output we now have a very efficient feature to measure this deviation from 0 which is characteristic of QSOs.

As discussed in Section 6.1.1, QSOs with faint magnitudes were under-represented by our training set. This will have been recognised and exploited during
6.2 Identifying the Best Features

Figure 6.2: Bar plot of the absolute value of SVM kernel model weights for $g$-band features. The highest SVM kernel weight is indicative of the success of the feature during the classification modelling.

the modelling, and as a result the median magnitude of each light curve in each filter, Med$_{mag}$, is highly weighted as a feature that can distinguish QSOs from stars (see Figure 6.3, top right).

The feature $\eta$ quantifies the extent to which successive magnitudes, $m_0, \ldots, m_N$, are independent within each light curve; allowing us to check for trends in the data. For example, $\eta$ is small in the event of a positive serial correlation, and large in the event of a negative serial correlation (e.g. von Neumann, 1941; Press, 1969; Shin, Sekora & Byun, 2009; Kim et al., 2011; Woźniak et al., 2004). Kim et al. (2011) show that $\eta$ is relatively small for QSOs, Be stars and long period variables that have positive serial correlation. We modified $\eta$ to make it invariant between sampling distributions, then applied it to the SAC output to look for trends between successive SAC data points; we can see in Figure 6.3 that the $\eta$(SAC) distribution for QSOs extends to lower values than for stars, and that it is a useful feature to separate the two classes.

Kim et al. (2011) show that Cusum is typically larger for QSOs, Be stars, long period variables and microlensing events, and smaller for non-variables and other
periodic variables such as RR Lyraes, Cepheids, and eclipsing binaries. We modified Cusum for invariance between sampling distributions, and in Figure 6.3 it is shown that QSOs have a distinctly larger and narrower Cusum distribution than stars in our training set. The standard deviation, \( \sigma \), is also shown to be typically larger for QSOs than stars, this is due to the light curve variability of QSOs.

For the bottom two ranked features, N and Skewness, there is very little distinction between the distributions for stars and QSOs. This is not surprising for N; since this is the number of \( g \)-band exposures for each object, there should be little distinction in N between stars and QSOs because Pan-STARRS images over a large region of the sky without discrimination. We took care to remove spurious exposures without removing data points showing intrinsic variability, but it’s likely that QSOs may have had more variable data points removed than stars, hence the slightly higher normalised counts for stars with \(~200-240\) exposures. Otherwise these distributions are very similar. This feature was included in the modelling process in case the number of data points was correlated with any of the resulting features, so that the SVM training could utilise N alongside other features.

Skewness also appears to be relatively ineffective at distinguishing QSOs from stars in Figure 6.3. A Normal distribution would have skewness value zero. In Figure 6.4 we show the Skewness distribution in higher resolution (i.e. binned according to smaller intervals), and it is apparent that the distribution for QSOs is slightly more centered around zero; i.e. the QSO \( g \)-band magnitudes for each light curve tend to be closer to a normal distribution than those of stars. This difference is only slight and Skewness has not been ranked highly as an effective classification feature.

In Figure 6.5 (top left), Stetson\_K(SAC) (the feature with the highest SVM kernel weight) \( r \)-band values are plotted against Med\_mag (the feature with the second highest SVM kernel weight) \( r \)-band values for objects in the training set. Indeed, we can see the success of the top two features for separating QSOs from stars, as they mostly occupy different regions in this two dimensional plot.

In Figure 6.5 (top right), Stetson\_K(SAC) \( r \)-band values are plotted against Eta(SAC) (the feature with the third highest SVM kernel weight) \( r \)-band values. Again we see a good separation between the regions occupied by QSOs and stars. Skewness, N, Kurtosis and Anderson Darling, i.e. some of the worst features identified in Figure 6.2, are plotted in Figure 6.5 (bottom left and right); we can see that these features are less successful at separating stars from QSOs in these two dimensional plots. Various other features are plotted for comparison in Figures 6.6 and 6.7.

Some of the most successful features for the classification modelling are
6.2 Identifying the Best Features

Figure 6.3: Normalised counts of various $g$-band feature distributions for Stars (blue) and QSOs (red) in the training set. The top six features and bottom two features from the SVM kernel weight bar plot are shown.
new ones that we have created. In order of success, they are Stetson\_K(SAC), Eta(SAC), $\sum \text{SAC} < e^{-1}(\text{SAC})$, Median($\text{Cusum}\_20(\text{SAC})$), Sigma(SAC), $\text{N}(\text{SAC} > e^{-1.2})$, SNR, $\text{N}(\text{SAC} < e^{-1.2})$ and $\sum \text{SAC} > e^{-1}(\text{SAC})$. Cusum is also one of the strongest features; whilst we did not create the original feature (it was originally created by Ellaway (1978) for neuroscience studies), we did modify it to make it invariant between sampling patterns, and it is our modified version that we use here.

We also tried to analyse the most successful features independently from the RF model using out-of-bagging methods. During this process, one feature column is randomly permuted to provide essentially white-noise for that feature for the RF training, with all other features columns remaining the same. The class recall percentage is then compared with the original recall percentage; for features which are most important to the RF training, the class recall should be significantly lower, and this difference gives a measure of their importance. However, because we only have two classification types, we found that randomly permuting each feature column was still frequently assigning each object to the same class type and that the class type was largely unaffected even for our best separating features. Dubath et al. (2011) and Kim et al. (2014) found that after a certain number of features (approximately 22 features), adding more features does not dramatically improve results in RF modelling. We were using 26 $g$-band features in this test so this may be why results were unaffected by altering individual feature columns. We decided that the

Figure 6.4: Normalised count of the $g$-band Skewness distribution for Stars (blue) and QSOs (red) in the training set with higher resolution binning.
6.2 Identifying the Best Features

Figure 6.5: Various r-band features are plotted for Stars (blue) and QSOs (red) in the training set. Top left: Stetson\textsubscript{K}(SAC) is plotted against Med\textsubscript{mag}. Top right: Stetson\textsubscript{K}(SAC) is plotted against Eta(SAC). Bottom: Skewness is plotted against N (left) and Anderson Darling is plotted against Kurtosis (right).
Figure 6.6: More two dimensional feature plots. Various $r$-band features are plotted for Stars (blue) and QSOs (red) in the training set.
Figure 6.7: More two dimensional feature plots. Various $r$-band features are plotted for Stars (blue) and QSOs (red) in the training set.
SVM output weights are sufficient for a measure of the success of each feature.

In order to test to what extent our best new features improve the classification model, we re-ran the RF cross-validation using various subsets of $g$-band features. Table 6.3 shows the results. Our top 3 new features shown in Figure 6.2 (Stetson$_{K(SAC)}$, Eta(SAC) and Cusum) perform reasonably well alone in the cross-validation with 88.5% recall and 90.8% precision. The top 3 features from Figure 6.2 that have been used in previous studies are Med$_{mag}$, Med$_{mag}$$_{err}$ and Sigma; these have 78.3% recall and 81.9% precision and so do not perform well on their own in the cross-validation. Med$_{mag}$ and Med$_{mag}$$_{err}$ are arguably not true variability features as they do not encode much time series variability information. The top 3 variability features used in previous work are Sigma, Sigma$_{mean}$ and Con; these features alone do not perform well in the cross-validation. Cross-validation of the top 6 features (Stetson$_{K(SAC)}$, Med$_{mag}$, Eta(SAC), Med$_{mag}$$_{err}$, Cusum and Sigma) showed that a combination of our top new features with the previous top features gives the highest precision and recall rates. A combination of the top 6 variability features (Stetson$_{K(SAC)}$, Eta(SAC), Cusum, Sigma, Sigma$_{mean}$ and Con) perform well in the test, however, not as well as the top 6 when Med$_{mag}$ and Med$_{mag}$$_{err}$ are included. These results show that our best new or modified features, Stetson$_{K(SAC)}$, Eta(SAC) and Cusum, are crucial to our optimised classification models.

### 6.3 Random Forest Results

Our three RF models were run on all 253,106 light curves. We compare results with the full spectroscopic catalogue, where we have spectral types for 2,048 QSOs and 5,116 stars (approximately five sevenths of which were used during the training), in order to evaluate our models.

The RF models output a predicted type for each light curve; QSO (Q) or star (S). The model also outputs two confidence values (between 0 and 1) that the object is a QSO (Q) and star (S). These two confidences sum to 1 and the type with the greater confidence is the predicted type; e.g. if the QSO confidence is $\geq 0.6$ and the star confidence is $\leq 0.4$ then the predicted type will be a QSO. There are 2,048 spectroscopically confirmed QSOs (which we call ‘true Qs’) and 5,116 spectroscopically confirmed stars (which we call ‘true Ss’), and for these objects we compare predicted objects with their known spectroscopic type to evaluate the success of each RF Model I-III.

Table 6.4 shows the number of each type predicted by each model and, for
6.3 Random Forest Results

<table>
<thead>
<tr>
<th>Analysis of Top Features</th>
<th>True S</th>
<th>True Q</th>
<th>Class Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Top 6 Features</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stetson, K(SAC), Med_mag, Eta(SAC), Med_mag_err, Cusum, Sigma</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted S</td>
<td>3545</td>
<td>126</td>
<td>96.6%</td>
</tr>
<tr>
<td>Predicted Q</td>
<td>98</td>
<td>1342</td>
<td>93.2%</td>
</tr>
<tr>
<td><strong>Class Recall</strong></td>
<td>97.3%</td>
<td>91.4%</td>
<td></td>
</tr>
</tbody>
</table>

| **Top 6 *Variability* Features** |        |        |                 |
| Stetson, K(SAC), Eta(SAC), Cusum, Sigma, Sigma_mean, Con |        |        |                 |
| Predicted S              | 3529   | 138    | 96.2%           |
| Predicted Q              | 114    | 1330   | 92.1%           |
| **Class Recall**         | 96.9%  | 90.6%  |                 |

| **Our Top 3 Features**   |        |        |                 |
| Stetson, K(SAC), Eta(SAC), Cusum |        |        |                 |
| Predicted S              | 3512   | 169    | 95.4%           |
| Predicted Q              | 131    | 1299   | 90.8%           |
| **Class Recall**         | 96.4%  | 88.5%  |                 |

| **Previous Top 3 Features** |        |        |                 |
| Med_mag, Med_mag_err, Sigma |        |        |                 |
| Predicted S              | 3388   | 318    | 91.4%           |
| Predicted Q              | 255    | 1150   | 81.9%           |
| **Class Recall**         | 93.0%  | 78.3%  |                 |

| **Previous Top 3 *Variability* Features** |        |        |                 |
| Sigma, Sigma_mean, Con |        |        |                 |
| Predicted S              | 3348   | 353    | 90.5%           |
| Predicted Q              | 295    | 1115   | 79.1%           |
| **Class Recall**         | 91.9%  | 76.0%  |                 |

Table 6.3: Analysis of Top Features: RF Cross Validation results for Stars (S) and QSOs (Q) for combinations of the top features in $g$-band only. There were 1,468 QSOs and 3,643 stars in the training set.

objects with known type, compares these predictions with the known true types. These results are used to calculate recall and precision rates for each model to evaluate the success of the predictions. For all RF Models I-III, precision and recall rates for QSOs and stars are very high. RF Model I, where all features and colours were used as training attributes, shows the highest recall and precision rates for QSOs and stars; this is the same as we saw in the cross-validation results in Section 6.1.2.

For RF Model II there are only marginally more false QSO predictions than for RF Model III (37 instead of 34); however, notice that RF Model III predicts a much higher number of QSOs (24,060 as opposed to 13,554). We seek a model with a high
6.3 Random Forest Results

<table>
<thead>
<tr>
<th>RF Evaluation Table</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
</tr>
<tr>
<td><strong>RF Model I</strong></td>
</tr>
<tr>
<td>Predicted S</td>
</tr>
<tr>
<td>Predicted Q</td>
</tr>
<tr>
<td><strong>Class Recall</strong></td>
</tr>
<tr>
<td><strong>RF Model II</strong></td>
</tr>
<tr>
<td>Predicted S</td>
</tr>
<tr>
<td>Predicted Q</td>
</tr>
<tr>
<td><strong>Class Recall</strong></td>
</tr>
<tr>
<td><strong>RF Model III</strong></td>
</tr>
<tr>
<td>Predicted S</td>
</tr>
<tr>
<td>Predicted Q</td>
</tr>
<tr>
<td><strong>Class Recall</strong></td>
</tr>
</tbody>
</table>

**Table 6.4:** Predictions from RF Model I (all features and colours were used as attributes), RF Model II (g-band feature attributes only), and RF Model III (colours attributes only). For each model, this table shows the number of predicted QSOs, the number of predicted stars, the number of predicted QSOs with spectroscopic classification as QSO (Predicted Q, True Q), the number of predicted QSOs with spectroscopic classification as star (Predicted Q, True S), the number of predicted stars with spectroscopic classification as star (Predicted S, True S), and the number of predicted stars with spectroscopic classification as QSO (Predicted S, True Q). Precision and recall rates are shown for Stars (S) and QSOs (Q) predicted from RF Models I, II and III using the 2,048 true QSOs and 5,116 true stars from the spectroscopic catalogue.

accuracy for predictions, but also a high completeness rate; the higher the number of predicted QSOs, the more likely it is that the model is predicting a large number of false positives.

We analyse the confidence levels outputted with our model predictions. We concentrate on RF Model I since this has the best cross-validation results. Figure 6.8 (left) shows a normalised histogram of QSO prediction confidences for both the spectroscopic sample and the full unknown sample; these histograms are normalised since the full unknown sample is much larger than the spectroscopic sample. Reassuringly, objects with known QSO type are usually predicted with confidence 1. This is the same for the analysis for known stars in Figure 6.8 (right).

The number of true QSOs with prediction confidences less than 1 decreases sharply. For the unknown objects that are predicted to be QSOs, we extrapolate that by choosing a set of QSO predictions with high confidence we will find a high
6.3 Random Forest Results

Figure 6.8: Normalised histogram of RF Model I outputted confidences for QSOs (left) and stars (right) for predicted objects from the known (spectroscopic) sample (bold) and predicted objects from the unknown sample (dotted).

Figure 6.9: Histogram and table showing RF Model I outputted confidences for QSO predictions from the full unknown sample.
number of true QSOs. However, by taking a high confidence cutoff we are likely to miss some true QSOs with lower confidence predictions and so accuracy is gained at the expense of completeness. Figure 6.9 shows the number of objects outputted as QSOs at each confidence level. Of the known QSOs, 1,789 of them were predicted as QSOs with confidence 1 by RF Model I, 149 were predicted as QSOs with confidence 0.9, 47 were predicted as QSOs with confidence 0.8, and only 39 were predicted as QSOs with confidence $\leq 0.7$; 24 were falsely predicted as stars. There were only 2 objects predicted to be QSOs with confidence 1 that were actually stars (these were both visually classified as variable K-type stars).

6.3.1 RF Blind Test

We trained the models using only approximately five sevenths of the spectroscopic catalogue and Table 6.4 showed us results for a mixture of known objects that were used in the training and objects that were not used in the training. It is to be expected that the models will do well at predicting objects that were used in the training process. We call the subset with known type that were not used in the training process a ‘blind sample’ as these objects provide a true blind test of the success of each model.

There were 5,111 objects used in the training set (1,468 QSOs and 3,643 stars) and 7,164 objects in the full spectroscopic catalogue (2,048 QSOs and 5,116 stars). We now analyse the 2053 objects (580 QSOs and 1,473 stars) in the blind sample. See Table 6.5 for the blind test results for the three RF models. Most of the objects that were incorrectly predicted in the full spectroscopic sample were in the blind sample subset. Again, the highest precision and recall rates occur in RF Model I, where we have 97.9% precision and 96.0% recall rates for QSOs.

Figure 6.10 (left) shows the confidence distribution of true Qs that were correctly predicted in the blind test as Qs in bold and true Ss that were incorrectly predicted as Qs in dotted lines. RF Model I generally predicts known QSOs in the blind sample correctly with high confidence (usually confidence 1), and at lower confidences there are not as many true QSOs. At lower confidences the predictions are more likely to be incorrect. This is further evidence that if we take a high QSO confidence cutoff we are more likely to get a larger number of true QSOs, without sacrificing too many true QSOs which may have a lower prediction confidence rate.

Our blind test results show that all three RF models perform extremely well when predicting objects that were chosen from our spectroscopic sample but not used in the training. We extrapolate that these models would also be successful
6.3 Random Forest Results

<table>
<thead>
<tr>
<th>RF Blind Test Evaluation Table</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>RF Model I</td>
</tr>
<tr>
<td>Predicted S</td>
</tr>
<tr>
<td>Predicted Q</td>
</tr>
<tr>
<td>Class Recall</td>
</tr>
<tr>
<td>RF Model II</td>
</tr>
<tr>
<td>Predicted S</td>
</tr>
<tr>
<td>Predicted Q</td>
</tr>
<tr>
<td>Class Recall</td>
</tr>
<tr>
<td>RF Model III</td>
</tr>
<tr>
<td>Predicted S</td>
</tr>
<tr>
<td>Predicted Q</td>
</tr>
<tr>
<td>Class Recall</td>
</tr>
</tbody>
</table>

Table 6.5: RF blind test evaluation table: precision and recall rates are shown for Stars (S) and QSOs (Q) predicted from RF Models I, II and III using the 580 true QSOs and 1,473 true stars from the spectroscopic blind sample.

...at predicting a larger sample of unknowns with similar properties to those in the training set. However, our training set is not completely representative of our full Pan-STARRS sample, as was illustrated in Figure 6.1. In the next section we seek to understand how these models perform when classifying QSOs for the full sample.

6.3.2 RF Predicted QSOs

The Random Forest models predict the following number of QSOs:

- RF Model I: 14,352 QSOs
- RF Model II: 13,554 QSOs
- RF Model III: 24,060 QSOs

We want to know how much the inclusion of our features as attributes improves a colours-only analysis and to understand if each model may be predicting different populations of QSOs. The predictions for each model are presented in a Venn diagram in Figure 6.11. The intersection of these three models impressively recalls 1,979 of the 2,048 known QSOs, with very little contamination; only 2 stars and 3 galaxies are falsely predicted as QSOs. The intersection of all three models also
Figure 6.10: Blind test results. Normalised histogram of RF Model I outputted confidence for the QSOs (left) and stars (right) for spectroscopically confirmed objects (bold) and falsely predicted objects (dotted).

predicts 7,816 QSOs, which is a realistic number for our sample. Whilst the colours-only model does predict an extra 13 known QSOs, an unrealistic number of 12,535 extra QSOs are predicted.

In Figure 6.12 we analyse the 12,535 QSOs predicted by RF Model III but not RF Models I or II in order to understand why the colours-only model predicts too many QSOs. The green points show all predicted QSOs and the red points show those predicted as QSOs with confidence 1. Upon comparison with Figure 6.1 (left), many of the predicted objects lie in the fainter $g$-band magnitude region where our training set does not have many objects. We think that because the training set is not adequately representing this region, RF Model III is unable to adequately model objects in this colour-magnitude region and the model fails to predict reliably for this population of unknowns. Whilst a colours-only analysis fails to adequately classify QSOs using RF modelling, we show that the inclusion of our variability features leads to a more reliable classification model.

Figures 6.13 and 6.14 compare colour-colour and colour-magnitude plots for RF Models I, II and III. We expect QSOs to inhabit the regions in these plots where the QSOs from the spectroscopic sample were shown in Figure 6.1. Predicted QSOs from RF Model I and the intersection of predicted QSOs from RF Models II and III mostly inhabit the same regions as the training set QSOs. However, those predicted using colours and not variability tend to inhabit the fainter $g$-band region where the training set is incomplete (Figure 6.14, bottom left). As discussed above, we expect
that the colours model is unable to predict QSOs reliably in this region because of the lack of faint \( g \)-band objects in the training set. Those predicted using variability features and not colours tend to have the faintest \( g \)-band magnitudes (Figure 6.14, bottom right), suggesting that the variability model is also affected by the lack of faint objects in the training set. Also, these objects mostly inhabit a region at the centre-right of the \( u - g \) and \( g - r \) colour-colour plot where we would typically expect to see M-type stars (Figure 6.13, bottom right). M-type stars are sometimes variable, and it is possible that variable stars are being confused with QSOs by the variability model. The most convincing population of predicted QSOs occurs when both colour and variability features are utilised.

Figure 6.15 shows the distribution of the 3,180 QSOs predicted by RF Model I but not RF Model III in colour-magnitude and colour-colour plots and also their distribution in the two dimensional Stetson_\( K(SAC) \) and Med_mag \( r \)-band feature space.
Figure 6.12: The top two plots show the 12,535 QSOs predicted by RF Model III but not RF Models I or II in green, with confidence 1 predictions over-plotted in red, and the whole sample in black. Top Left: $g$-band magnitude and $u - g$ colour-magnitude plot. Top Right: $u - g$ and $g - r$ colour-colour plot. Bottom: Normalised histogram of confidences for these 12,535 QSO predictions in dashed red line, with prediction confidences for the 13 known QSOs in this sample in bold.
Figure 6.13: $u - g$ and $g - r$ colour-colour plots. Top left: QSOs predicted by RF Model I, top right: QSOs predicted by both RF Model II and RF Model III, bottom left: QSOs predicted by RF Model III but not RF Model II, bottom right: QSOs predicted by RF Model II but not RF Model III.
Figure 6.14: $g$-band magnitude and $u - g$ colour-magnitude plots. Top left: QSOs predicted by RF Model I, top right: QSOs predicted by both RF Model II and RF Model III, bottom left: QSOs predicted by RF Model III but not RF Model II, bottom right: QSOs predicted by RF Model II but not RF Model III.
Figure 6.15: The top two plots show the 3,180 QSOs predicted by RF Model I but not RF Model III in green, with confidence 1 predictions over-plotted in red, and the whole sample in black. Top Left: $g$-band magnitude and $u - g$ colour-magnitude plot. Top Right: $u - g$ and $g - r$ colour-colour plot. Centre Left: $r$-band magnitude Stetson_K(SAC) and Med_mag plot for QSOs predicted with all confidence levels. Centre Right: $r$-band magnitude Stetson_K(SAC) and Med_mag plot for QSOs predicted with confidence 1. Bottom: Normalised histogram of confidences for these 3,180 QSO predictions in dashed red line, with prediction confidences for the 22 known QSOs in this sample in bold.
6.4 SVM Results

Our SVM models were run on all 253,106 light curves. In this section, we compare results with the full spectroscopic catalogue (with spectral types for 2,048 QSOs and 5,116 stars; approximately five sevenths of which were used during the training) in order to evaluate our results. As with the RF models, the SVM models output a predicted type for each light curve along with a confidence level. The SVM confidences are outputted to higher precision than the RF confidences (which were only given to one decimal place). The two outputted confidences sum to 1.

Table 6.6 shows the number of each type predicted by each model and, for objects with known type, compares these predictions with the known true types. We calculate recall and precision rates for each model to evaluate the success of the predictions. SVM Model I, where all features and colours were used as training attributes, shows the highest recall and precision rates for QSOs and stars; this is the same as we saw in the cross-validation results in Section 6.1.3.

Notice that SVM Model III predicts a much higher number of QSOs than SVM Model II (50,540 as opposed to 13,292). As with RF Model III, we suspect that the colours model is predicting too many QSOs and it is likely that many of these predicted objects are actually stars; the SVM model is predicting more than twice as many QSOs as RF Model III (which predicted 24,060).

We now analyse the confidence levels outputted with our model predictions,
### SVM Evaluation Table

<table>
<thead>
<tr>
<th>SVM Model</th>
<th>Total Predicted</th>
<th>True S</th>
<th>True Q</th>
<th>Class Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM Model I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted S</td>
<td>241,479</td>
<td>5085</td>
<td>47</td>
<td>99.1%</td>
</tr>
<tr>
<td>Predicted Q</td>
<td>11,627</td>
<td>31</td>
<td>2001</td>
<td>98.5%</td>
</tr>
<tr>
<td>Class Recall</td>
<td></td>
<td>99.4%</td>
<td>97.7%</td>
<td></td>
</tr>
<tr>
<td>SVM Model II</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted S</td>
<td>239,814</td>
<td>5029</td>
<td>141</td>
<td>97.3%</td>
</tr>
<tr>
<td>Predicted Q</td>
<td>13,292</td>
<td>87</td>
<td>1907</td>
<td>95.6%</td>
</tr>
<tr>
<td>Class Recall</td>
<td></td>
<td>98.3%</td>
<td>93.1%</td>
<td></td>
</tr>
<tr>
<td>SVM Model III</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted S</td>
<td>202,566</td>
<td>4729</td>
<td>387</td>
<td>92.4%</td>
</tr>
<tr>
<td>Predicted Q</td>
<td>50,540</td>
<td>387</td>
<td>1661</td>
<td>81.1%</td>
</tr>
<tr>
<td>Class Recall</td>
<td></td>
<td>92.4%</td>
<td>81.1%</td>
<td></td>
</tr>
</tbody>
</table>

**Table 6.6:** Predictions from SVM Model I (all features and colours were used as attributes), SVM Model II (g-band feature attributes only), and SVM Model III (colours attributes only). For each model, this table shows the number of predicted QSOs, the number of predicted stars, the number of predicted QSOs with spectroscopic classification as QSO (Predicted Q, True Q), the number of predicted QSOs with spectroscopic classification as star (Predicted Q, True S), the number of predicted stars with spectroscopic classification as star (Predicted S, True S), and the number of predicted stars with spectroscopic classification as QSO (Predicted S, True Q). Precision and recall rates are shown for Stars (S) and QSOs (Q) predicted from SVM Models I, II and III using the 2,048 true QSOs and 5,116 true stars from the spectroscopic catalogue.

Concentrating on SVM Model I since this has the best classification results of SVM Models I-III. Figure 6.16 (left) shows a normalised histogram of confidences for QSO predictions for both the spectroscopic sample and the full unknown sample. Objects with known QSO type are usually predicted with high confidence. This is the same for the analysis for known stars in Figure 6.16 (right).

The number of true types with confidence less than 1 decreases sharply. For the unknown objects which are predicted to be QSOs, we extrapolate that by choosing a set of objects predicted with QSO confidence 1, we will find a high number of true QSOs; however, by taking a high cutoff we are likely to miss a set of true QSOs. Figure 6.17 shows the number of objects outputted as QSOs at each confidence level.
6.4 SVM Results

Figure 6.16: Normalised histogram of SVM Model I outputted confidences for QSOs (left) and stars (right) for predicted objects from the known (spectroscopic) sample (bold) and predicted objects from the unknown sample (dotted).

Figure 6.17: Histogram and table showing SVM Model I outputted confidences for QSO predictions from the full unknown sample.
6.4 SVM Results

6.4.1 SVM Blind Test

The blind test results for the three SVM models are shown in Table 6.7 and Figure 6.18. Again, the highest precision and recall rates occur in SVM Model I, where we have 97.7% precision and 96.7% recall rates for QSOs. These are very close to the RF Model I blind test results which had 97.9% precision and 96.0% recall rates. Figure 6.18 (left) shows the confidence distribution of true Qs that were correctly predicted in the blind test as Qs in bold and true Ss that were incorrectly predicted as Qs in dotted lines. SVM Model I generally predicts known QSOs in the blind sample correctly with high confidence, and at lower confidences there are not as many true QSOs.

<table>
<thead>
<tr>
<th></th>
<th>True S</th>
<th>True Q</th>
<th>Class Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM Model I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted S</td>
<td>1460</td>
<td>19</td>
<td>98.7%</td>
</tr>
<tr>
<td>Predicted Q</td>
<td>13</td>
<td>561</td>
<td>97.7%</td>
</tr>
<tr>
<td>Class Recall</td>
<td>99.1%</td>
<td>96.7%</td>
<td></td>
</tr>
<tr>
<td>SVM Model II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted S</td>
<td>1439</td>
<td>47</td>
<td>96.8%</td>
</tr>
<tr>
<td>Predicted Q</td>
<td>34</td>
<td>533</td>
<td>94.0%</td>
</tr>
<tr>
<td>Class Recall</td>
<td>97.7%</td>
<td>91.9%</td>
<td></td>
</tr>
<tr>
<td>SVM Model III</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted S</td>
<td>1335</td>
<td>109</td>
<td>92.5%</td>
</tr>
<tr>
<td>Predicted Q</td>
<td>138</td>
<td>471</td>
<td>77.3%</td>
</tr>
<tr>
<td>Class Recall</td>
<td>90.6%</td>
<td>81.2%</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.7: SVM blind test evaluation table: precision and recall rates are shown for Stars (S) and QSOs (Q) predicted from SVM Models I, II and III using the 580 true QSOs and 1,473 true stars from the spectroscopic blind sample.

6.4.2 SVM Predicted QSOs

The SVM models predict the following number of QSOs:

- SVM Model I: 11,627 QSOs
- SVM Model II: 13,292 QSOs
- SVM Model III: 50,540 QSOs
6.4 SVM Results

Figure 6.18: Blind test results. Normalised histogram of SVM Model I outputted confidence for the QSOs (left) and stars (right) for spectroscopically confirmed objects (bold) and falsely predicted objects (dotted).

The predictions for each model are presented in a Venn diagram in Figure 6.19. The intersection of these three models recalls 1,562 of the 2,048 known QSOs, which is significantly less than the 1,979 recalled by the intersection of the three RF models. Whilst the colours-only model does predict an extra 26 known QSOs, an unrealistic number of 40,495 extra QSOs are predicted.

In Figure 6.20 we analyse the 40,495 QSO objects predicted by RF Model III but not RF Models I or II in order to understand why the colours-only model predicts too many QSOs. The green points show all predicted QSOs and the red points show those predicted as QSOs with confidence 1. We see a similar trend to that shown in Figure 6.12, where many of the predicted objects lie in the fainter $g$-band magnitude region; our training set does not adequately represent this region. There is no preference towards higher or lower confidences for predicted QSOs in this sample. The prediction of too many QSOs in this region appears to be independent of the machine learning model used, and is more a consequence of using only colours as attributes when this colour region is not represented in the training set.

There are 7,902 objects that are predicted as QSOs by both SVM Model II and SVM Model III, 5,390 QSO candidates predicted by SVM Model II but not SVM Model III and 42,628 QSO candidates predicted by SVM Model III but not SVM Model II. In Figures 6.21 and 6.22 we show colour magnitude and colour-colour plots which compare the distributions of objects predicted as QSOs by SVM Model I (top left), objects predicted as QSOs by both SVM Model II and SVM Model III
Figure 6.19: QSO predictions summary for each SVM model. Overlapping regions show QSO predictions common to both models. The text in red shows known types from the spectroscopic catalogue.

(top right), objects predicted as QSOs by SVM Model III but not SVM Model II (bottom left), and objects predicted as QSOs by SVM Model II and SVM Model III (bottom right). QSOs predicted by SVM Model I, and both SVM Model II and SVM Model III, tend to lie in the region where we expect to find QSOs on colour plots. In the case of objects predicted by SVM Model II but not SVM Model III, i.e. where variability features alone were used to train the model, many of the predicted QSOs lie in regions where we would expect to find M-type stars; as with the RF results in Section 6.3.2, we think that variable stars may be confusing the models in the case where variability features but not colours are used to train the models. A combination of features and colours as training attributes gives the most robust results.

Figure 6.23 shows the distribution of the 2,127 QSOs predicted by SVM Model I but not SVM Model III in colour plots and also their distribution in the two
Figure 6.20: The top two plots show the 40,495 QSOs predicted by SVM Model III but not SVM Models I or II in green, with confidence 1 predictions over-plotted in red, and the whole sample in black. Top Left: $g$-band magnitude and $u - g$ colour-magnitude plot. Top Right: $u - g$ and $g - r$ colour-colour plot. Bottom: Normalised histogram of confidences for these 40,495 QSO predictions in dashed red line, with prediction confidences for the 26 known QSOs in this sample in bold.
**Figure 6.21**: $u - g$ and $g - r$ colour-colour plots. Top left: QSOs predicted by SVM Model I, top right: QSOs predicted by both SVM Model II and SVM Model III, bottom left: QSOs predicted by SVM Model III but not SVM Model II, bottom right: QSOs predicted by SVM Model II but not SVM Model III.
Figure 6.22: $g$-band magnitude and $u - g$ colour-magnitude plots. Top left: QSOs predicted by SVM Model I, top right: QSOs predicted by both SVM Model II and SVM Model III, bottom left: QSOs predicted by SVM Model III but not SVM Model II, bottom right: QSOs predicted by SVM Model II but not SVM Model III.
6.5 QSOs Predicted by both RF and SVM

In this section we describe a catalogue comprising objects that were predicted as QSOs by both RF Model I and SVM Model I with high confidence, since these are most likely to be true QSOs. This catalogue is intended to be used as a candidate list for TDSS. Figure 6.24 shows histograms of the number of QSOs predicted by both models with various RF and SVM confidence Q cutoffs. There are 5,749 QSOs predicted by both RF Model I with confidence $\geq 0.9$ and SVM Model I with confidence $\geq 0.9$, and 7,090 predicted with confidence $\geq 0.8$. There are 1,988 true QSOs predicted by both RF Model I and SVM Model I (out of a sample of 2,048 true QSOs altogether); 1,396 true QSOs predicted with both Q confidences set to $\geq 0.9$, 1,752 true QSOs predicted with both Q confidences set to $\geq 0.8$, 1,912 true QSOs predicted with both Q confidences set to $\geq 0.7$, 1,962 true QSOs predicted with both Q confidences set to $\geq 0.6$.

Figure 6.25 shows colour-colour and colour-magnitude distributions for objects predicted as QSOs by both RF Model I and SVM Model I with confidence higher than 0.7 and 0.9. Upon comparison with the colour-magnitude and colour-colour plots in Figure 6.1, we can see that by taking a combination of both SVM and RF predicted QSOs, all with high confidence, that these distributions follow those of true QSOs very well. Only one known star was falsely predicted as a QSO with confidence greater than 0.8; this object was a K-type variable star.
Figure 6.23: These plots show the 2,127 QSOs predicted by SVM Model I but not SVM Model III in green, with confidence > 0.9 predictions over-plotted in red, and the whole sample in black. Top Left: $g$-band magnitude and $u-g$ colour-magnitude plot. Top Right: $u-g$ and $g-r$ colour-colour plot. Centre Left: $r$-band magnitude Stetson_K(SAC) and Med_mag plot for QSOs predicted with all confidence levels. Centre Right: $r$-band magnitude Stetson_K(SAC) and Med_mag plot for QSOs predicted with confidence > 0.95. Bottom: Normalised histogram of confidences for these 2,127 QSO predictions in dashed red line, with prediction confidences for the 367 known QSOs in this sample in bold.
Figure 6.24: RF Model I confidences (top) and SVM Model I (bottom) for various RF and SVM Confidence Q cutoffs.
Figure 6.25: Colour-colour (top) and colour-magnitude (bottom) plots for predicted QSOs with RF Model I and SVM Model I Q confidences $\geq 0.7$ in green and Q confidences $\geq 0.9$ in red.
6.6 Summary and Discussion

We used SVM and RF supervised learning algorithms to train QSO classification models for a Pan-STARRS data set. Known QSOs and stars from a spectroscopic catalogue were used as a training set. Our models classified objects as either QSOs or stars using features derived from a time series analysis of Pan-STARRS light curves; these features characterise variability properties.

We presented many new features that characterise and quantify variability behaviour associated with QSOs. Some of the most successful features for classification in our models were new or modified ones that we have presented in this work. In order of success, they are Stetson$_K$(SAC), Eta(SAC), Cusum, $\sum SAC < e^{-1}(SAC)$, Median(Cusum$_{20}$(SAC)), Sigma(SAC), N(SAC$< e^{-1.2}$), SNR, N(SAC$< e^{-1.2}$) and $\sum SAC > e^{-1}(SAC)$. Our modified version of Cusum (Ellaway, 1978; Kim et al., 2011) worked well as a classification feature (it was the fifth best feature). Our modified version of Eta (von Neumann, 1941; Press, 1969; Kim et al., 2011) was one of the least successful features; however, from it we created a new feature, Eta(SAC), which applied our modified sampling-invariant version of Eta to the SAC output. Eta(SAC) was our third most successful classification feature. When we re-designed the Stetson$_K$ feature (Stetson, 1996) to apply it to our SAC output, we found that this was the strongest of all of the QSO classification features. Our top three new and modified features, Stetson$_K$(SAC), Eta(SAC) and Cusum, worked so well under cross-validation that they gave a QSO recall rate of 88.5% and precision rate of 90.8% when used alone to train a RF model in a single filter.

Our RF Model I cross-validation showed a precision rate of 98.6% and recall rate of 96.9%, with $F$-score=0.978. Our cross-validation results for SVM Model I showed a precision rate of 98.3% and recall rate of 97.7%, with $F$-score=0.980. A direct comparison with other samples, e.g. MACHO and EROS-2 (Kim et al., 2011; Pichara et al., 2012), is not possible since our data sets are very different, although in future it would be useful to apply our features to the MACHO data set.

The successfully high precision and recall rates from cross-validation on our training sets were tainted by both models predicting too many QSOs in a colour-only analysis. The training set was mostly colour-selected, had an unrealistically high ratio of QSOs to stars, and the QSOs were shown to have brighter $g$-band magnitudes than the full sample. Also our training set did not contain galaxies and we expect a small portion of our sample ($< 1\%$) to be contaminated by galaxies. Thus the training set did not fully represent our full Pan-STARRS sample. As a result, colours-only classification models were unable to adequately model objects in...
under-represented colour-magnitude regions and we showed that they predicted too many QSOs in these regions. Models using only variability features also failed to predict well in under-represented colour-magnitude regions, and we thought these models may confuse some variable stars, such as M-type stars, with QSOs. The most convincing population of predicted QSOs occurred when both colours and variability features were utilised; with the inclusion of our new features we are confident that our models are effective classifiers. We showed that the number of false positives can be significantly reduced by taking a cutoff in the QSO prediction confidence outputted by the models.

The expected colours of high redshift QSOs are redder than those of low redshift QSOs, and so classification by colour is restricted to specific redshift ranges. Both RF and SVM models trained using variability features as well as colours predicted a population of QSOs that colours-only models did not; some of these predictions are fainter with redder colours and could be higher redshift QSOs. SDSS colours for QSOs in the redshift range $2 < z < 3.5$ overlap with those of stars and so a colours-only analysis would not allow us to identify such a population of QSOs, but the inclusion of our variability features may permit a reliable classification for intermediate redshift QSOs. We would require a spectroscopic sample of such objects in our training set to investigate this further.

By taking a combination of both SVM and RF predicted QSOs, all with high confidence, we made a catalogue of likely QSOs to be used as a TDSS candidate list and showed that their colour-magnitude and colour-colour distributions follow those of true QSOs very well. There were 7,090 QSOs predicted by both models with confidences $\geq 0.8$, of these 1,752 true QSOs (of 2048 known QSOs) were correctly predicted. Only one known star was falsely predicted as a QSO with these constraints; this object was a K-type variable star.

The goal of our classification models was to identify a large and complete sample of active galaxies with high precision. Spectroscopic follow-up of the large sample that we have proposed with high confidence will allow future studies to further investigate the internal properties of active galaxies in more detail. We can tackle questions such as what triggers AGN activity, which internal processes cause the variability we observe, and what is the relationship between this variability and black hole mass. A complete sample would also have broader implications for studies into galaxy evolution and cosmological models, since the QSO luminosity density function appears to scale with the global SFR luminosity density and the merger rate as a function of redshift (discussed in Section 1.3). A lack of completeness in active galaxies identified by our models would lead to underestimates in the QSO number.
density and could impact resulting evolutionary models. The incompleteness identified from our current models appears to be a direct result of the under-representation of faint objects in our training set, and our next step is to use a broader spectroscopic training set to reduce the resulting incompleteness.

The following is a summary of the key points from our analysis:

- All of our classification models performed impressively well in the blind test. This implies that with a training set which better represents the general population of Pan-STARRS light curves a very complete and reliable set of QSO predictions should be produced.
- RF consistently performs better than SVM to classify QSOs.
- Our new variability features perform better than features used in previous classification studies.
- A combination of features and colours works best; both in the blind test and in predicting likely QSOs outside of the range of the training set.
Chapter 7

Conclusions

7.1 Summary and Discussion

This thesis has presented research into the role of active and merging galaxies in galaxy evolution. We have utilised a range of data sets from cutting-edge astronomical instruments and employed modern computer science methods to analyse our data.

In Chapter 1, we introduced the standard paradigm of structure formation in the framework of the ΛCDM cosmological model. We reviewed relevant scientific literature in cosmology and evaluated the proposed model of hierarchical formation through galaxy mergers (within a ΛCDM cosmological framework) from the viewpoint of theoretical studies, computer simulations and observational data from telescopes. We described active galaxies and discussed their star formation and variability trends and their role in some models of galaxy evolution. Galaxy mergers were then introduced, with a literature review on star formation in mergers as a function of environment and mass (including previous research into major and minor close pairs). This chapter presented and familiarised the reader with the essential background topics that were prerequisites for the subsequent thesis chapters.

In Chapter 2, the various telescopes and surveys from which we utilised data were described. Technical specifications and limitations, as well as science goals, were provided for each instrument. Various techniques for measuring star formation were described, and we explained our choice to use the NUV waveband for our research into galaxy mergers. The main advantage of using NUV photometry is its sensitivity to recently triggered star formation. Various methods were introduced to classify galaxy samples; morphologically (e.g. CAS and visual classification), spectroscopically and photometrically. Colour-magnitude diagrams were presented, and their usefulness at distributing galaxy samples into starforming ver-
sus non-starforming regions was discussed. Methods to estimate the environment density in which mergers are located were also introduced.

In Chapter 3, we introduced the close pair and wide pair data sets and described how these were extracted from the SDSS DR7 and cross-matched with the GALEX GR4/GR5 data base for NUV measurements. Galaxy masses and environment densities were calculated for each close pair and they were classified according to the predominant mechanism of excitation by a BPT analysis. SSFRs were derived from NUV luminosities. Close pairs and wide pairs with mass ratio $<1/3$ were categorised as major mergers, and the remaining pairs were classified as minor mergers.

In Chapter 4, we investigated how the NUV-derived SSFR in close pair galaxies evolves as mergers progress. We considered various mass and environment parameters, and also tested for observational evidence of AGN activity being triggered in close pairs. SSFR enhancements were summarised in Table 4.3. We found enhancements in SSFR for close pairs compared to non-close pair galaxies. Due to the phenomenon of cosmic downsizing we expect to see more SF in low mass galaxies in the local Universe (Cowie et al., 1996; Terlevich, López & Terlevich, 2007; Faber et al., 2007). This enhancement was particularly pronounced for low stellar mass close pair galaxies (by an average factor of $5.1 \pm 0.7$ increase), with high stellar mass close pairs showing a lower rate of SSFR enhancement (by an average factor of $2.0 \pm 0.9$ increase). We found that SSFR enhancements were particularly high for close pairs in field environments. In the low mass sample we found an average rise in SSFR by a factor of $2.4 \pm 0.7$ for pairs in the field and an average rise by a factor of $3.3 \pm 0.9$ for pairs in groups. For high mass pairs we saw an average rise by a factor of $2.5 \pm 0.7$ increase in SSFR in field environments. This is likely to be due to the higher gas fraction available in low density environments to fuel SF, since tidal fields and ram pressure stripping can reduce gas fractions in higher density environments.

There was a larger rate of increase in SSFR in the primary progenitor sample as pairs progress from the widest to smallest separation than in the secondary progenitor sample; although overall the secondaries on average showed more star formation at all separations. There is no other literature showing a general and significant impact on rSF in the primary progenitor of minor mergers; it is likely that we are seeing this effect for the first time thanks to the sensitivity of the NUV-waveband to rSF.

We also showed evidence that both the primary and secondary progenitors in minor mergers on average show enhancements in rSF, particularly for very low mass ratios (when the primary is at least three times more massive than the secondary)
and at the lowest separation (i.e. as the close pairs are reaching an advanced state of merging). Perhaps in a minor merger where the masses are very different, the smaller progenitor stirs up gas in the primary more efficiently than if they are of similar masses. This trend is not shown in the existing literature; it is statistically significant in our research and it is likely that it has only now been detected due to the sensitivity of the NUV waveband to rSF.

We saw strong evidence that merging can cause a change in emission line processes, leading to an evolution in a galaxy’s location in the BPT diagram. However, based on our BPT analysis, where Seyferts are the only category to definitely harbour AGN activity, we saw little evidence that mergers are triggering AGN activity during the close pairs stage of merging. Very small increases were seen in the Seyfert fraction in pairs at very low separation, paralleled with strong increases in the Transition fraction; this suggests that AGN activity may increase but it may be overwhelmed by star formation in low separation close pairs. We found no significant evidence of increased AGN activity in major mergers over minor mergers in our sample. We propose that, if AGN activity is ignited in some interacting massive galaxies as theoretically predicted, this process may lead to another class of AGN activity, or take place at the post-merger stage once the merging black holes have coalesced.

In Chapter 5, we introduced machine learning classification methods such as Support Vector Machines and Random Forest classifiers and reviewed recent literature where modern computer science methods have been used in astrophysics; particularly for QSO classification. Our Pan-STARRS light curve data set was introduced; this boasts optical \((g,r,i,z,y)\) light curves for 253,106 objects spanning nine medium deep fields, each covering a 7°-squared region of the sky. Exposures which were likely to be unreliable due to imaging problems were removed, and only light curves with at least 25 remaining exposures in each filter were used.

We described various statistical features and calculated their value for each light curve. These features characterise variability behaviour such as light curve periodicity, amplitude and autocorrelation features and were utilised in order to separate QSOs from non-variable objects and variable stars. Some features were modified to use slotted autocorrelation instead of standard autocorrelation, allowing us to account for differences in sampling patterns so that these features were suitably invariant between medium deep fields.

In Chapter 6, we introduced our Support Vector Machine (SVM) and Random Forest (RF) QSO classification algorithms which exploit our time series variability features to optimally classify QSOs in our photometric sample. We trained our mod-
els using a set of spectroscopically confirmed QSOs and stars and compared models that were trained using both variability features and colours with models trained using features only and colours only. We showed that models trained using our features predict a population of likely QSOs that colours alone do not, this highlights the importance of including variability features in our classification models.

Some of the most successful features for classification in our models were new or modified ones that we have presented in this work. In order of success, they are Stetson$_K$(SAC), Eta(SAC), Cusum, $\sum SAC < e^{-1}(SAC)$, Median($Cusum_{20}(SAC)$), Sigma(SAC), $N(SAC < e^{-1.2})$, SNR, $N(SAC < e^{-1.2})$ and $\sum SAC > e^{-1}(SAC)$. Our modified version of Cusum (Ellaway, 1978; Kim et al., 2011) worked well as a classification feature (it was the fifth best feature). Our modified version of Eta (von Neumann, 1941; Press, 1969; Kim et al., 2011) was one of the least successful features; however, from it we created a new feature, Eta(SAC), which applied our modified sampling-invariant version of Eta to the SAC output. Eta(SAC) was our third most successful classification feature. When we re-designed the Stetson$_K$ feature (Stetson, 1996) to apply it to our SAC output, we found that this was the strongest of all of the QSO classification features. Even when used alone in model training, our top three new features, Stetson$_K$(SAC), Eta(SAC) and Cusum, showed impressive classification results.

Cross-validation and a blind test on the training set showed impressive results, however, these results were tainted by models often predicting too many QSOs in the full sample. We attribute this to the training set not being fully representative of our Pan-STARRS sample, since the models worked very well for objects within the training set during the blind test. The training set was mostly colour-selected, had an unrealistically high ratio of QSOs to stars, and the QSOs were shown to have brighter $g$-band magnitudes than the full sample. Colours-only classification models were unable to adequately model objects in under-represented colour-magnitude regions and we showed that they predicted too many QSOs in these regions. The most convincing population of predicted QSOs occurred when both colours and variability features were utilised; with the inclusion of our new features we are confident that our models are effective classifiers and would be further improved when trained with a broader training set. We showed that the number of false positives can be significantly reduced by taking a cutoff in the QSO prediction confidence outputted by the models.

Both RF and SVM models trained using variability features as well as colours predicted a population of QSOs that colours-only models did not; some of these predictions are fainter with redder colours and could be higher redshift QSOs. SDSS
colours for QSOs in the redshift range $2.2 < z < 3.5$ overlap with those of stars and so a colours-only analysis would not allow us to identify such a population of QSOs (Fan, 1999; Ross et al., 2012), but the inclusion of our variability features may have permitted a reliable classification for intermediate redshift QSOs. We would require a spectroscopic sample of such objects in our training set to investigate if our models have correctly predicted intermediate redshift QSOs.

By taking a combination of both SVM and RF predicted QSOs, all with high confidence, we made a catalogue of likely QSOs to be used as a TDSS candidate list and showed that their colour-magnitude and colour-colour distributions follow those of true QSOs very well. There were 7,090 QSOs predicted by both models with confidences $\geq 0.8$, of these 1,752 true QSOs (of 2,048 known QSOs) were correctly predicted. Only one known star was falsely predicted as a QSO with these constraints; this object was a K-type variable star.

7.2 Outlook

This thesis presented the first research where NUV luminosity-derived specific star formation rates have been used to study galaxy close pairs, allowing us to study merging galaxies with a new level of sensitivity to star formation. Various close pair scenarios were analysed and enhancements in recent star formation were measured and compared. For the first time, evidence was shown for a higher rate of star formation in the more massive galaxy in a minor merger. We showed a significant evolution in the emission-line behaviour of close pair galaxies which continues to evolve as close pairs draw closer together; indicating that merging may be linked with black hole accretion and AGN activity.

Future space missions with more advanced technology will be able to probe further into these changes in star formation rate and AGN activity, and soon-to-be-launched telescopes such as the Advanced Technology Large-Aperture Space Telescope (ATLAST) and the James Webb Space Telescope (JWST) will provide further insights. ATLAST has been proposed to run from 2025 to 2035 to replace the Hubble Space Telescope. It will have an 8 to 16.8 metre UV-optical-NIR telescope with a sensitivity limit up to 2000 times better than that of the Hubble Space Telescope. This will allow exciting extensions to be made to the close pairs work in this thesis which used UV-derived star formation rates, and optical-NIR measurements can also be employed to analyse large samples of close pairs in depth.

JWST will hopefully be launched by 2018 and is planned to be an infrared telescope with a 6.5 metre primary mirror; this will be ideal for studying light from
very early galaxies (that has been redshifted into the infrared) to a high degree of
sensitivity. These large diameter telescopes will provide insight into star formation
in merging galaxies to higher precision and at higher redshift, allowing us to analyse
how merger activity has influenced galaxy growth as a function of time. The EU-
CLID spacecraft is planned to launch in 2020 and looks promising for future studies
of galaxy evolution. EUCLID aims to measure the shapes and redshifts of galaxies
and galaxy clusters up to redshift $\sim 2$, to analyse how galaxies and cosmic structures
have evolved.

The Pan-STARRS mission is ongoing and high resolution light curves will be
made available for many more objects in the years to come. Future wide-field time-
resolved surveys such as the Large Synoptic Survey Telescope (LSST) will allow for
even better light curve data sets. The LSST is due to conduct a ten year survey that
is planned to commence in January 2022. More exposures in light curves will allow
for more reliable variability features to be calculated and tighter constraints to be
made when identifying active galaxies. Such a data set would permit higher precision
photometric classifications using time-series methods and would improve our ability
to distinguish active galaxies from a greater subset of objects; such as various galaxy
types and stellar types. Spectroscopic surveys are constantly providing new spectra
for objects, and will continue to provide more spectral types to allow our training
set to be improved and our predictions to be further tested. This will allow further
refinements to be made to our classification algorithms to optimally classify larger
samples of active galaxies from photometric surveys.
Acknowledgements

This PhD would not have been possible without the help, contribution and support of many others. I am extremely grateful for the support and encouragement from Moira Scott, Kevin Glasgow, Michael Scott, John Urquhart and other friends and family who have encouraged me throughout.

I am very grateful to Sugata Kaviraj for supervising me throughout the PhD and for patiently guiding me from having no experience in astrophysics research all the way to the end of this thesis. I am also very thankful to Stephen Warren for his help and encouragement throughout the PhD and particularly at the later stages when writing this thesis. The PhD has led to many travels and adventures for which I am immensely grateful, and the work has been very rewarding.

I would like to thank Pavlos Protopapas for generously and patiently investing much of his time in improving my previously non-existent computer science and programming skills; his motivation as a supervisor was an essential force behind this thesis. I would also like to thank Paul Green for always being on-hand to offer useful ideas and comments, and for being a real pleasure to work with. Thanks to Marc Sarzi, Dave Clements and Yvonne Unruh for their very useful comments on the thesis. Alfredo Carpineti, Pablo Heise, Dae-Won Kim and Karim Pichara have also been a joy to work with, as have many other graduate students and staff at Imperial College and Harvard University; there are too many to mention but I have enjoyed the company of them all.

The research presented in this thesis was supported by an STFC studentship at Imperial College London, a Smithsonian Astrophysical Observatory PreDoctoral Fellowship at the Harvard-Smithsonian Center of Astrophysics, and a Research Fellowship from Harvard University’s Institute for Applied Computational Science.

During the PhD, I also spent three very enjoyable and educational months at Kyoto University thanks to a Fellowship from their Bilateral International Exchange Program. Although the work conducted there was a side project and is not presented in this thesis, I am very grateful to my supervisor Professor Kouji Ohta, his graduate student, Kiyo Yabe, and the rest of the Astrophysics group at Kyoto University.
(especially Murray Brightman and Kohei Ichikawa) for being such fantastic and welcoming hosts throughout my stay. I also had the pleasure of observing at both the UKIRT and JCMT telescopes, and am grateful to all who were involved in these trips.

Sugata Kaviraj (SK) and I are grateful to the anonymous referee of the paper Scott & Kaviraj (2014) for many useful comments that improved the original manuscript. We thank Daniel Mortlock for his help with statistical calculations, Marc Sarzi for his publicly available GANDALF code and Mike Blanton for his publicly available KCORRECT code. SK acknowledges research fellowships from the 1851 Royal Commission, Imperial College London and Worcester College Oxford and support from the BIPAC institute at Oxford.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, University of Cambridge, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofísica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

GALEX is a NASA Small Explorer, launched in April 2003. We gratefully acknowledge NASA’s support for construction, operation, and science analysis for the GALEX mission, developed in cooperation with the Centre National d’Etudes Spatiales of France and the Korean Ministry of Science and Technology.

Much of this work was enabled using the PS1 System operated by the PS1 Science Consortium (PS1SC) and its member institutions. The PS1 Surveys have been made possible through the combinations of the Institute for Astronomy at the University of Hawaii, The Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg, and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, the University of Durham, the University of Edinburgh, the
Queen’s University of Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Network, and the National Central University of Taiwan.

References

Bennett K. P., Campbell C., 2000, SIGKDD Explorations, 2, 1
Bertone G., Hooper D., Silk J., 2005, Physics Reports, 405, 279


Breiman L., 2001, Machine Learning, 45, 5


Cappellari M., Bacon R., Bureau M., Damen M. C., Davies R. L., de Zeeuw P. T., Emsellem E.,
Cappellari M., Emsellem E., Krajnović D., McDermid R. M., Serra P., Alatalo K., Blitz L., Bois
Cappellari M., Scott N., Alatalo K., Blitz L., Bois M., Bournaud F., Bureau M., Crocker A. F.
Casey C. M., Berta S., Béthermin M., Bock J., Bridge C., Bundykiewicz J., Burgarella D., Chapin
Cicone C., Maiolino R., Sturm E., Graciá-Carpio J., Feruglio C., Neri R., Aalto S., Davies R. et al.,
Cole S., Percival W. J., Peacock J. A., Norberg P., Baugh C. M., Frenk C. S., Baldry I., Bland-
Curtis H. D., 1918, Publications of Lick Observatory, 13, 9
de Vaucouleurs G., 1959, Handbuch der Physik, 53, 275
215


Dietterich T., 2000, in Machine Learning, 40th edn., pp. 139–158


Famaey B., McGaugh S. S., 2012, Living Reviews in Relativity, 15, 10


Fan X., 2012, Research in Astronomy and Astrophysics, 12, 865


Hsu C.-W., Chang C.-C., Lin C.-J., 2003, A practical guide to support vector classification. Tech. rep., Department of Computer Science, National Taiwan University


Hubble E., 1929, Proceedings of the National Academy of Science, 15, 168


Iocco F., Pato M., Bertone G., Jetzer P., 2011, Journal of Cosmology and Astroparticle Physics, 11, 29


Lynden-Bell D., 1969, Nat, 223, 690


Milgrom M., 2013, Physical Review Letters, 111, 041105
Oort J. H., 1932, Bulletin of the Astronomical Institutes of the Netherlands, 6, 249


Planck Collaboration, 2013, ArXiv e-prints


Rees M. J., 1990, Science, 247, 817


Sodrê, Jr. L., Cuevas H., 1994, Vistas in Astronomy, 38, 287
Steinmetz M., Navarro J. F., 2002, New Astronomy, 7, 155
Wozniak P. R., 2000, Acta Astronomica, 50, 421


Yuan Y., Shaw M. J., 1995, Fuzzy Sets and Systems, 69, 125


Zwicky F., 1933, Helvetica Physica Acta, 6, 110