



# Non-IID Outlier Detection with Coupled Outlier Factors

Guansong Pang

Supervisors:

Longbing Cao, Principal

Ling Chen

Huan Liu, Arizona State University

*This thesis is presented as part of the requirements for the conferral of the degree:*

Doctor of Philosophy

University of Technology Sydney

Faculty of Engineering and Information Technology

April 2019

*To my beloved parents, Yuxiang Pang and Shangxing Li.  
To my wife Lisha and my son Louis.*

## Certificate of Original Authorship

*I, Guansong Pang, declare that this thesis, submitted in fulfillment of the requirements for the award of the degree: Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.*

*This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.*

*This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.*

---

***Guansong Pang***

*April 1, 2019*



# Acknowledgments

I am greatly indebted to my principal supervisor, Professor Longbing Cao, one of the best mentors I have ever met. Over the past three years, Longbing spent tremendous time on teaching me how to think critically, do solid research and write ‘beautiful’ technical papers. He has been very nice, thoughtful and supportive to me in that he gave me large freedom to exclusively explore interesting research problems and publish high-quality papers at the early stage of my candidature while prepared me for an independent academic career in my senior candidature by involving me in different professional activities (e.g., conference/workshop organization, program committee of leading conferences), conference tutorials, student mentoring, research proposals, and industry project meetings. Longbing has been very fruitful, critical and constructive in making comments on research. I will not forget his massive advice on my designs and manuscripts, which are extremely important and helpful for sharpening my research designs and paper writings. In addition to research, Longbing has been a great life mentor of me for sharing many wonderful and undesirable experiences to encourage me when I was down and to remind me pitfalls when I was in triumph. I cannot thank him enough for all his supports.

I would like to thank my co-supervisor Dr. Ling Chen and external supervisor Professor Huan Liu for their helpful and insightful comments on both of my research designs and papers. I would like to particularly thank Ling for her nice and warm greetings on many festivals and public holidays. Ling has been always very supportive and understanding for many aspects in my research and daily life I seek suggestions and help from her. I am very grateful to Huan for not only having nice and thorough discussions over my research whenever I met him, but also suggesting interesting directions for my work, offering me great advice on job applications and career planning.

Besides my supervisors, I would also like to thank the panel members of my candidature assessment, Dr. Wei Liu, Professor Jinyan Li, Dr. Haiyan Lu, and Professor Bogdan Gabrys, for their constructive comments and great encouragement.

I thank all my friends and colleagues in the Data Science Lab at AAI for being critical and insightful to my research in our weekly research meetings and being very nice to me whenever I met them. I particularly thank: Chengzhang Zhu for many wonderful discussions about our research and careers, Songlei Jian for the nice research discussions and collaborations, Shoujin Wang and Thac Do for their companion and story sharing in our quiet lab. I am very grateful to visiting professors Defu Lian, Wenpeng Lu, and Lizhen Wang for their constructive comments to my research and great help in my career planning. I really appreciate the collaboration with Hongzuo Xu. It was a pleasure to

work with him and I learned many helpful mentoring skills from this experience. I would also like to thank the other friends in the lab who enrich my research life and/or give me consistent encouragement: Wei Wang, Liang Hu, Qi Zhang, Ke Liu, Longxiang Shi, Wenfeng Hou, Qing Liu, Usman Naseem, Yan Xing, Jingyu Shao, Frank Xu, Dr. Allen Lin, Professor Quanguai Zhang, and Dr. Lei Gu. It was always a great time to celebrate every of our success in the lab and to gather together to have fun and enjoy the holidays in festivals.

I would like to express my special thanks to Professor Shengyi Jiang who introduced me to the academic community and was always there to give me advice whenever I seek help from him. I am also very thankful to Dr. Huidong Jin, Professor Kai Ming Ting, and Dr. David Albrecht for their enormous time and effort on my previous research which lays an important foundation to my PhD research. I also thank everyone who has inspired me and helped me in my research.

I would also like to thank the UTS Advanced Analytics Institute, School of Software, and Graduate Research School for their quality administration service and financial support for my domestic and international conference travels. Additionally, I thank Michele Mooney for her excellent proofreading of this thesis.

Last but not least, I wholeheartedly thank my beloved mother and father for their everlasting love, amazing encouragement and strong support throughout my life. I am also specially grateful to my brothers and sisters who are very supportive in taking care of all the family matters during my graduate study. Special thanks to my dear wife Lisha Zhang for being patient and extremely supportive during this tough period and for planning my career and making decisions on our future with me. I am so fortunate to have her in my life. I would also like to thank my son Louis (Zhizhong) Pang who colors my life with his cheerfulness, lovely and gorgeous smiles, and chaos of trouble. This thesis is dedicated to them.

# Publications During Candidature

Papers that have been published or accepted for publication:

1. **Guansong Pang**, Longbing Cao, Ling Chen, and Huan Liu. “Learning Representations of Ultrahigh-dimensional Data for Random Distance-based Outlier Detection”, In: *24th SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-18)*. London, UK. Accepted (long presentation).
2. Songlei Jian, **Guansong Pang**, Longbing Cao, Kai Lu, and Hang Gao. “CURE: Flexible Categorical Data Representation by Hierarchical Coupling Learning”. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*. Accepted.
3. **Guansong Pang**, Longbing Cao, Ling Chen, Defu Lian and Huan Liu. “Sparse Modeling-based Sequential Ensemble Learning for Effective Outlier Detection in High-dimensional Numeric Data”, In: *32nd AAAI Conference on Artificial Intelligence (AAAI-18)*. AAAI Press, pp. 3892-3899. New Orleans, US.
4. **Guansong Pang**, Longbing Cao, Ling Chen and Huan Liu. “Learning Homophily Couplings from Non-IID Data for Joint Feature Selection and Noise-Resilient Outlier Detection”. In: *26th International Joint Conference on Artificial Intelligence (IJCAI-17)*. AAAI Press, pp. 2585-2591.
5. Songlei Jian, Longbing Cao, **Guansong Pang**, Kai Lu and Hang Gao. “Embedding-based Representation of Categorical Data by Hierarchical Value Coupling Learning”. In: *26th International Joint Conference on Artificial Intelligence (IJCAI-17)*. AAAI Press, pp. 1937-1943.
6. **Guansong Pang**, Hongzuo Xu, Longbing Cao and Wentao Zhao. “Selective Value Coupling Learning for Detecting Outliers in High-Dimensional Categorical Data”. In: *26th ACM International Conference on Information and Knowledge Management (CIKM-17)*, Long paper track. ACM, pp. 807-816. Singapore.
7. **Guansong Pang**, Kai Ming Ting, David Albrecht and Huidong Jin. “ZERO++: Harnessing the Power of Zero Appearances to Detect Anomalies in Large-Scale Data Sets”. *Journal of Artificial Intelligence Research (JAIR)* 57, pp. 593–620, 2016.
8. **Guansong Pang**, Longbing Cao, and Ling Chen. “Outlier Detection in Complex Categorical Data by Modeling Feature Value Couplings”. In: *25th International*

*Joint Conference on Artificial Intelligence (IJCAI-16)*. AAAI Press, pp. 1902–1908, 2016. New York City, US.

9. **Guansong Pang**, Longbing Cao, Ling Chen and Huan Liu. “Unsupervised Feature Selection for Outlier Detection by Modeling Hierarchical Value-Feature Couplings.” In: *2016 IEEE International Conference on Data Mining (ICDM-16)*, Long paper track. IEEE, pp. 410-419. Barcelona, Spain.
10. **Guansong Pang**, Kai Ming Ting, and David Albrecht. “LeSiNN: Detecting anomalies by identifying Least Similar Nearest Neighbours”. In: *2015 IEEE 15th International Conference on Data Mining Workshops (ICDMW-15)*. IEEE, pp. 623–630, 2015.

Papers that are under review:

1. **Guansong Pang**, Longbing Cao, and Ling Chen. “Outlier Detection with Non-IID Outlier Factors”. *Data Mining and Knowledge Discovery (DMKD)*. Under review.
2. **Guansong Pang** and Longbing Cao. “Building Optimal Heterogeneous Univariate Outlier Ensembles”. *IEEE Transactions on Neural Networks and Learning System (TNNLS)*. Under review.



# Abstract

Outliers are data objects which are rare or inconsistent compared to the majority of objects. Outlier detection is one of the most important tasks in data mining due to its wide applications in various domains, such as finance, information security, healthcare and earth science. Most existing outlier detection methods assume that the outlier factor (i.e., outlierness scoring measure) of the entities (e.g., feature values, features, data objects) in a data set is Independent and Identically Distributed (IID), but this assumption is violated by many real-world applications where the outlierness of an entity is coupled with that of some other entities, leading to the failure of detecting sophisticated outliers. This issue is intensified in more challenging environments, e.g., noisy and/or high-dimensional data sets. To address this challenge, this thesis considers three key questions: what are the coupling relations between different outlier factors? how can we effectively and efficiently model these couplings? and how can we leverage these couplings to address challenging outlier detection problems?

Our explorations result in the following four key contributions. (i) This thesis introduces a *new outlier detection task*, non-IID outlier detection in multidimensional data, which opens a new research direction for tackling real-world complex outlier detection problems. (ii) We introduce the *first architecture* for the non-IID outlier detection task, which provides principled approaches to learn the outlierness interdependence at different levels from feature values, features, to data objects. The architecture breaks down the general coupling learning into a series of important finer-grained components: basic coupling relation, coupling capacity, coupling utility, and coupling passage manners, providing feasible ways to learn sophisticated couplings between outlier factors with efficient models. (iii) We propose principled frameworks and their instantiations under the non-IID outlier detection architecture to learn different types of couplings. Supported by extensive theoretical analysis and empirical experiments on diverse real-world data sets, these designs are shown to be scalable and effective in *addressing some notoriously challenging problems*, including outlier detection in non-IID data, data with many noisy features, or high-dimensional data. (iv) This thesis also introduces a set of *seminal work* on unsupervised feature selection for outlier detection in both categorical data and numeric data, including innovative feature selection methods that capture pairwise or full feature interactions and joint feature selection and outlier detection methods. Our proposed approaches are able to effectively compute the outlierness of features, which enables outlying feature selection and substantially improves the efficacy of subsequent outlier detection on data with high dimensionality or many noisy features.

Our extensive empirical results show that the average accuracy improvement of our non-IID outlier detectors over state-of-the-art IID outlier detectors ranges from 4% up to 18% on a large collection of real-world data sets; the maximum accuracy improvement on single data sets can be more than 50%, in which state-of-the-art IID detectors only obtain an accuracy of being nearly equivalent to a random guess. This significant accuracy improvement can have great business value, e.g., the prevention of millions of dollars loss in credit card fraud detection, enabling safer digital environments by mitigating malicious programs or network intrusions, or saving life by having early detection of fatal diseases. This thesis also offers much more interpretable outlier detection solutions by enabling outlier detection in highly relevant and substantially smaller feature subsets.

# Contents

<b>Dedication</b>	<b>ii</b>
<b>Declaration</b>	<b>iii</b>
<b>Acknowledgments</b>	<b>iv</b>
<b>List of Publications</b>	<b>vi</b>
<b>Abstract</b>	<b>viii</b>
<b>List of Tables</b>	<b>xv</b>
<b>List of Figures</b>	<b>xviii</b>
<b>I Research Background and Foundation</b>	<b>1</b>
<b>1 Introduction</b>	<b>2</b>
1.1 Research Motivation . . . . .	2
1.2 Contributions . . . . .	3
1.2.1 Value-level Coupled Outlier Factors . . . . .	4
1.2.2 Feature/object-level Coupled Outlier Factors . . . . .	6
1.2.3 Summary . . . . .	6
1.3 Organization . . . . .	7
<b>2 Preliminaries and Foundation</b>	<b>9</b>
2.1 Common Symbols . . . . .	9
2.2 Non-IID Outlier Detection . . . . .	9
2.3 Experiment Approach . . . . .	11
2.3.1 Data Preparation . . . . .	11
2.3.2 Detection Performance Evaluation . . . . .	12
2.3.3 Data Indicator for Outlier Detection . . . . .	13
<b>3 Literature Review</b>	<b>17</b>
3.1 Traditional Outlier Detection Methods . . . . .	17
3.1.1 Methods for Numeric Data . . . . .	17
3.1.2 Methods for Categorical Data . . . . .	19

3.2	Non-IID Outlier Detection Methods . . . . .	20
3.2.1	Non-IID Learning . . . . .	21
3.2.2	Outlier Detection with Non-IID Outlier Factors . . . . .	21
3.3	High-dimensional Outlier Detection . . . . .	22
3.3.1	Full-space-based Methods . . . . .	22
3.3.2	Subspace-based Methods . . . . .	22
3.3.3	Feature Selection-based Methods . . . . .	23
3.4	Summary . . . . .	24
<b>II Value-level Coupled Outlier Factors</b>		<b>25</b>
<b>4</b>	<b>Conditional Cascade of Outlier Factors</b>	<b>27</b>
4.1	Introduction . . . . .	27
4.2	The Proposed CUOT Framework . . . . .	29
4.2.1	Value Outlierness Initialization . . . . .	30
4.2.2	Outlierness Influence Between Values . . . . .	31
4.2.3	Value Graph Construction . . . . .	31
4.2.4	Value Outlierness Estimation . . . . .	32
4.3	A CUOT Instance: CBRW . . . . .	32
4.3.1	Mode-based Initial Outlierness . . . . .	32
4.3.2	Conditional Probability-based Outlierness Influence . . . . .	33
4.3.3	Directed and Attributed Value Graph . . . . .	34
4.3.4	Biased Random Walks for Learning Value Outlierness . . . . .	35
4.3.5	The Algorithm and Its Time Complexity . . . . .	36
4.3.6	Applications of CBRW . . . . .	37
4.4	Theoretical Analysis . . . . .	39
4.4.1	Convergence Analysis . . . . .	39
4.4.2	Modeling Homophily Outlying Behaviors . . . . .	40
4.4.3	Stability w.r.t. Parameter $\alpha$ . . . . .	41
4.5	Experiments and Evaluation . . . . .	41
4.5.1	Outlier Detectors and Their Parameter Settings . . . . .	41
4.5.2	Data Sets . . . . .	42
4.5.3	Outlier Detection Performance . . . . .	42
4.5.4	Performance of Outlying Feature Selection . . . . .	46
4.5.5	Scalability Test . . . . .	49
4.5.6	Sensitivity Test w.r.t. the Damping Factor $\alpha$ . . . . .	50
4.5.7	Convergence Test . . . . .	51
4.6	Summary . . . . .	51
<b>5</b>	<b>Selective Conditional Cascade of Outlier Factors</b>	<b>53</b>
5.1	Introduction . . . . .	53
5.2	The Proposed SelectVC Framework . . . . .	55
5.2.1	Value Subset Evaluation Function $\psi$ . . . . .	56

5.2.2	Selective Value Coupling-based Scoring Function $\phi$ . . . . .	56
5.2.3	Stationary Criterion . . . . .	57
5.3	The SelectVC Instance: POP . . . . .	57
5.3.1	Specifying $\psi$ Using Top- $k$ Outlying Value Selection . . . . .	57
5.3.2	Specifying $\phi$ by Partial Outlierness Propagation . . . . .	57
5.3.3	$\ell_1$ -Norm Stationary Criterion . . . . .	59
5.3.4	The Algorithm and Its Time Complexity . . . . .	59
5.4	Theoretical Analysis . . . . .	60
5.4.1	Quality of the Stationary Vector $\mathbf{q}^*$ . . . . .	60
5.4.2	Handling Distance Concentration Effect . . . . .	60
5.4.3	Guidelines for Setting $k$ . . . . .	61
5.5	Experiments and Evaluation . . . . .	62
5.5.1	Experiment Environment . . . . .	62
5.5.2	Data Sets . . . . .	62
5.5.3	Effectiveness in Real-world Data . . . . .	63
5.5.4	Significance of Partial Outlierness Propagation . . . . .	65
5.5.5	Significance of Joint Value Selection and Outlier Scoring . . . . .	66
5.5.6	Scalability Test . . . . .	67
5.5.7	Sensitivity Test . . . . .	68
5.5.8	Convergence Test . . . . .	68
5.6	Summary . . . . .	69
<b>6</b>	<b>Binary Cascade of Outlier Factors</b> . . . . .	<b>71</b>
6.1	Introduction . . . . .	71
6.2	The Proposed WrapperOD Framework . . . . .	73
6.2.1	Fast Outlier Scoring Function . . . . .	73
6.2.2	Outlier Ranking Evaluation . . . . .	74
6.2.3	Feature Subset Generation . . . . .	74
6.3	A WrapperOD Instance: HOUR . . . . .	74
6.3.1	Specifying $\phi_S$ with Homophily Couplings . . . . .	74
6.3.2	Specifying $J$ with Average Score Margin . . . . .	75
6.3.3	Recursive Search of Feature Subset $\mathcal{S}$ . . . . .	75
6.3.4	The Algorithm and Its Time Complexity . . . . .	76
6.4	Theoretical Analysis . . . . .	76
6.4.1	Robustness w.r.t. Noisy Features . . . . .	76
6.4.2	Theoretical Bound . . . . .	78
6.5	Experiments and Evaluation . . . . .	78
6.5.1	Data Sets . . . . .	78
6.5.2	Experiment Environment . . . . .	78
6.5.3	Effectiveness in Real-world Data Sets . . . . .	79
6.5.4	Sensitivity Test . . . . .	81
6.5.5	Scalability Test . . . . .	82
6.6	Summary . . . . .	82

<b>7</b>	<b>High-order Cascade of Outlier Factors</b>	<b>83</b>
7.1	Introduction . . . . .	83
7.2	The Proposed HOCOF Framework . . . . .	85
7.3	A HOCOF Instance: SDRW . . . . .	86
7.3.1	Pointwise Mutual Information-based Outlierness Influence . . . . .	86
7.3.2	Refining the Value Graph with Subgraph Densities . . . . .	87
7.3.3	Noise-tolerant Random Walks for Learning Value Outlierness . . . . .	89
7.3.4	The Algorithm and Its Time Complexity . . . . .	89
7.4	Theoretical Analysis . . . . .	90
7.4.1	Closed-form Solution . . . . .	90
7.4.2	Handling Noisy Features . . . . .	91
7.5	Experiments and Evaluation . . . . .	92
7.5.1	Effectiveness of Outlier Detection . . . . .	92
7.5.2	Justification of Algorithmic Components . . . . .	94
7.5.3	Outlying Feature Selection Performance . . . . .	95
7.5.4	Scalability Test . . . . .	97
7.6	Summary . . . . .	98
<b>III</b>	<b>Feature/object-level Coupled Outlier Factors</b>	<b>100</b>
<b>8</b>	<b>Two-way Couplings of Feature-level Outlier Factors</b>	<b>102</b>
8.1	Introduction . . . . .	102
8.2	The Proposed CUFS Framework . . . . .	104
8.2.1	Value Graph Construction . . . . .	105
8.2.2	Feature Graph Construction . . . . .	105
8.2.3	Feature Subset Selection . . . . .	106
8.3	A CUFS Instance: DSFS . . . . .	107
8.3.1	Specifying Functions $\delta$ , $\eta$ and $g$ for the Value Graph . . . . .	107
8.3.2	Specifying Functions $\delta^*$ , $\eta^*$ and $h$ for the Feature Graph . . . . .	108
8.3.3	The Search Strategy . . . . .	109
8.3.4	The Algorithm and Its Time Complexity . . . . .	110
8.4	Theoretical Analysis . . . . .	110
8.5	Experiments and Evaluation . . . . .	112
8.5.1	Data Sets . . . . .	112
8.5.2	Baselines and Settings . . . . .	113
8.5.3	Feature Reduction Rate . . . . .	113
8.5.4	Performance of Different Subsequent Outlier Detectors . . . . .	114
8.5.5	Comparison to Feature Weighting-based Contenders . . . . .	116
8.5.6	Scalability Test . . . . .	117
8.6	Summary . . . . .	118

<b>9</b>	<b>Sequential Couplings of Object-level Outlier Factors</b>	<b>119</b>
9.1	Introduction . . . . .	119
9.2	The Proposed SEMSE Framework . . . . .	121
9.3	A SEMSE Instance: CINFO . . . . .	122
9.3.1	Building a Sequential Ensemble . . . . .	122
9.3.2	Aggregating a Set of Sequential Ensembles . . . . .	125
9.3.3	The Algorithm and Its Time Complexity . . . . .	125
9.4	Theoretical Analysis . . . . .	126
9.4.1	Upper Bound for Outlier Thresholding . . . . .	126
9.4.2	Optimal Feature Subsets w.r.t. Outlier Scoring $\phi$ . . . . .	126
9.4.3	Obtaining Good Outlier Scores by Subsampling . . . . .	127
9.5	Experiments and Evaluation . . . . .	127
9.5.1	Data Sets . . . . .	127
9.5.2	Experiment Environment . . . . .	127
9.5.3	Effectiveness in Real-world Data . . . . .	128
9.5.4	Comparison to State-of-the-art Competitors . . . . .	129
9.5.5	Resilience to Noisy Features . . . . .	130
9.5.6	Scalability Test . . . . .	131
9.6	Summary . . . . .	131
<b>IV</b>	<b>Conclusions and Future Directions</b>	<b>133</b>
<b>10</b>	<b>Conclusion</b>	<b>134</b>
10.1	Learning Couplings of Outlier Factors . . . . .	134
10.2	Significance of Non-IID Outlier Detection . . . . .	136
<b>11</b>	<b>Vision and Future Work</b>	<b>137</b>
11.1	Broadening Non-IID Outlier Detection . . . . .	137
11.1.1	Further Exploration of Coupled Outlier Factors . . . . .	137
11.1.2	Heterogeneous Outlier Factors . . . . .	137
11.1.3	Exploration of Coupled Heterogeneous Outlier Factors . . . . .	138
11.2	Selection of IID/non-IID Outlier Detection Methods . . . . .	138
11.2.1	Data Indicators for Measuring the IID/Non-IID Information . . . . .	138
11.2.2	Automatic Selection or Combination of IID/Non-IID Methods . . . . .	138
<b>A</b>	<b>Codes and Data Sets</b>	<b>139</b>
	<b>Bibliography</b>	<b>140</b>

# List of Tables

1.1	Thesis Overview (Main Body) . . . . .	4
2.1	Common Symbols and Their Description. Caligraphic letters for sets. Bold capital letters for matrices. Lowercase bold letters for vectors. . . . .	9
4.1	A Summary of 15 Data Sets and Their Complexity Evaluation Results. The following acronyms are used for brevity: Bank Marketing = BM, aPascal = APAS, Internet Advertisements = AD, Contraceptive Method Choice = CMC, Solar Flare = SF, Reuters10 = R10, CoverType = CT, and Linkage = LINK. The data sets are ordered by the average rank in the last column. . . . .	43
4.2	AUC Performance of CBRW <sub>od</sub> , its Two Variants and Five Contenders on the 15 Data Sets. ‘o’ indicates out-of-memory exceptions, while ‘•’ indicates that we cannot obtain the results within two months. The middle horizontal line roughly separates complex data from simple data based on average rank in Table 4.1. The best performance for each data set is boldfaced. The p-value of the null hypothesis rejected at the 1% or 5% confidence level is underlined. . . . .	44
4.3	Data Complexity Evaluation Results of Data Sets with Feature Subsets Selected by CBRW <sub>fs</sub> (denoted as CBRW) and ENFW, Using the Results on Original Data Sets as a Baseline. The last row ‘Simp.’ indicates the average simplification percentage compared to the baseline on the original data. . . . .	48
4.4	AUC Performance of MarP and iForest Using Feature Selection Methods CBRW <sub>fs</sub> , ENFW, RADM and Their Baseline FULL Using the Full Feature Set. . . . .	48
4.5	Two Key Properties of a Value Graph. Data is sorted by clustering coefficient. ‘o’ indicates out-of-memory exceptions. . . . .	51
5.1	A Summary of Data Sets Used and Indicator Quantization Results. $\kappa_{live} = \frac{\kappa_{rel}(\mathcal{U}) - \kappa_{rel}(\mathcal{V})}{\kappa_{rel}(\mathcal{V})}$ describes the level of irrelevant value couplings per data. The middle horizontal line roughly separates data sets with high $\kappa_{live}$ from that with low $\kappa_{live}$ . . . . .	63



5.2	AUC Performance of POP, POP <sup>+</sup> and Their Competitors: Five Full Space- or Subspace-based Outlier Detectors. CBRW runs out of memory on high-dimensional data <i>R8</i> and <i>WebKB</i> . ABOD runs out-of-memory on large data <i>w7a</i> and <i>CelebA</i> . . . . .	64
5.3	AUC Results of POP and the Combinations of CBRW, ZERO, iForest and LOF with Three Feature Selection Methods POFS, CBFS and DSFS on the 12 Data Sets. The performance of ABOD using POFS, CBFS or DSFS is similar to that of CBRW, ZERO, and iForest. We therefore omit the results of ABOD to fit the table well. . . . .	67
6.1	AUC and $P@n$ Performance on 15 Data Sets. Data is sorted by $\kappa_{fnl}$ . ‘ $\nabla$ ’ indicates the feature reduction rate of HOUR. FPOF runs out of memory in four high-dimensional data sets. . . . .	79
6.2	AUC and $P@n$ Performance Comparison between HOUR and the Combination of CBRW and CompreX with CBFS (Denoted by $\dagger$ ) and DSFS (Denoted by $\ddagger$ ). . . . .	80
6.3	Data Complexity Evaluation Results on $\mathcal{F}$ , $\mathcal{S}$ , $\mathcal{S}'$ and $\mathcal{S}''$ . $\mathcal{F}$ is the original feature set. $\mathcal{S}$ , $\mathcal{S}'$ and $\mathcal{S}''$ are feature subsets retained by HOUR, CBFS and DSFS, respectively. . . . .	81
7.1	Conceptual Comparison of CBRW and SDRW . . . . .	86
7.2	AUC Performance of SDRW <sub>od</sub> and Its Six Contenders on the 15 Data Sets. ‘o’ indicates out-of-memory exceptions, while ‘•’ indicates that we cannot obtain the results within two months. Simiar to the empirical analysis for CBRW in Chapter 4, we separate <i>complex</i> data from <i>simple</i> data sets based on average rank in Table 4.1. The best performance for each data set is boldfaced. . . . .	93
7.3	AUC Performance of SDRW <sub>od</sub> , CBRW <sub>od</sub> and Their Variants Created by Removing One or Two Components. The best performance within CBRW/SDRW is boldfaced. . . . .	95
7.4	Complexity Quantification of Data Sets with Feature Subsets Selected by SDRW <sub>fs</sub> , CBRW <sub>fs</sub> , ENFW and FULL. The last row shows the percentage of the average complexity reduction compared to the baseline FULL. We use SD = SDRW <sub>fs</sub> , CB = CBRW <sub>fs</sub> , EN = ENFW, and FU = FULL to concisely present the results. . . . .	96
7.5	AUC Performance of MarP and iForest Using SDRW <sub>fs</sub> , CBRW <sub>fs</sub> , ENFW, RADM, and FULL. . . . .	97
8.1	Feature Selection Results on Data Sets with Different Characteristics. The data sets are sorted by $\kappa_{fnl}$ . The middle horizontal line roughly separates data sets with many noisy features (i.e., $\kappa_{fnl} > 35\%$ ) from the other data sets. $RED = \frac{D-D'}{D}$ (%) denotes the reduction rate by DSFS. . . . .	114

8.2	AUC Performance of the Three Detectors with or without DSFS. The three baseline detectors are MarP, CompreX and FPOF. Their editions using DSFS are MarP*, CompreX* and FPOF*, respectively. IMP indicates the AUC improvement of the detectors combined with DSFS. ‘o’ indicates out-of-memory exceptions. ‘●’ indicates that we cannot obtain the results within four weeks, i.e., 2,419,200 seconds. . . . .	115
8.3	Runtime of the Three Detectors with or without DSFS. Three baseline detectors are MarP, CompreX and FPOF. Their editions using DSFS are MarP*, CompreX* and FPOF*, respectively. SU indicates the runtime speedup of the detectors combined with DSFS. . . . .	115
8.4	AUC Performance Comparison of the Three Detectors Using ENFW and DSFS respectively. IMP denotes the improvement of DSFS over ENFW. . .	117
9.1	Feature Reduction and AUC Performance of CINFO-enabled LeSiNN and iForest (denoted by LeSiNN* and iForest*). $D$ is the original feature number. $D'$ and $D''$ are the average numbers of features retained by LeSiNN* and iForest*, respectively. The average iteration for sequential ensembles per data is 2 to 5. . . . .	128
9.2	AUC Performance of CINFO, RegFS, FB, and CARE Empowered LeSiNN and iForest. ‘NA’ indicates the execution cannot be completed in two weeks.	130
A.1	Data Sources. . . . .	139

# List of Figures

1.1	Conceptual Architecture of Outlier Detection with Coupled Outlier Factors.	4
4.1	Conditional Cascade Couplings of Value-level Outlier Factors. $u_i$ represents a value. . . . .	28
4.2	The Proposed CUOT Framework. $F_i$ denotes an individual feature. $\mathcal{S}_j$ is a feature subset that contains a pair of features. $\mathcal{V}$ denotes the entire value set. $\delta$ computes an initial outlierness of feature values. $\eta$ considers the inter-feature value couplings that highlight the homophily relations between outlying values. $\mathbf{M}_\eta$ is a $ \mathcal{V}  \times  \mathcal{V} $ matrix whose entries are determined by $\eta$ . $\mathbf{G}$ denotes the value-value graph, $\omega$ is an edge weighting function based on $\delta$ and $\eta$ , and $\phi$ is the value outlierness learning function on the value graph. . . . .	30
4.3	Scale-up Test Results of the Five Detectors w.r.t. Data Size and Dimensionality. Logarithmic scale is used in the vertical axis. Note that FPOF runs out-of-memory when the number of features reaches 80. . . . .	49
4.4	Sensitivity Test Results w.r.t. the Parameter $\alpha$ . . . . .	50
4.5	Convergence Test Results. . . . .	51
5.1	Comparison of Selective Conditional Cascade Couplings and the Conditional Cascade Couplings Presented in Chapter 4. . . . .	54
5.2	The SelectVC Framework for Estimating Value Outlierness Based on Selective Value Couplings. The outlierness of data objects can then be obtained using value outlierness. SVC is short for Selective Value Couplings. . . . .	55
5.3	Scalability Test Results. ABOD and CBRW run out of memory when the number of objects reaches 25,000 and the number of features reaches 8,000, respectively. . . . .	68
5.4	Sensitivity Test Results of POP w.r.t. $k$ . . . . .	69
5.5	Convergence Test Results . . . . .	69
6.1	Comparison of Binary Cascade Couplings and the Conditional Cascade Couplings Presented in Chapter 4. . . . .	72
6.2	The Proposed WrapperOD Framework . . . . .	73

6.3	Representative AUC Performance of HOUR w.r.t. $k$ . HOUR performs stably in most of the other data sets. The dashed line shows HOUR's performance with $k = outlier\%$ . . . . .	81
6.4	Scale-up Test w.r.t. Data Size and Dimensionality. FPOF runs out of memory when dimensionality reaches 80. . . . .	82
7.1	High-order Cascade Couplings. The top subgraphs denote a collection of high-order couplings between multiple values, capturing certain local interactions. We aim to use this type of local high-order couplings w.r.t. each node (e.g., $u_2$ ) or edge to augment the outlieriness estimation in the original value graph that only captures pairwise low-order couplings . . . . .	84
7.2	The Proposed HOCOF Framework. $\{SG_1, SG_2, \dots\}$ is a set of subgraphs derived from the value graph $G$ . . . . .	85
7.3	Scale-up Test Results of the Seven Detectors w.r.t. Data Size and Dimensionality. Logarithmic scales are used in both axes. Note that FPOF runs out-of-memory when the number of features reaches 80. . . . .	98
8.1	Two-way Couplings of Feature-level Outlier Factors. The outlieriness of features is inferred from the value-level outlieriness interdependence. . . . .	103
8.2	The Proposed CUFS Framework. VCA and FCA are short for Value Coupling Analysis and Feature Coupling Analysis, respectively. . . . .	104
8.3	Scale-up Test Results of DSFS against ENFW w.r.t. Data Size and the Number of Features. . . . .	117
9.1	Sequential Couplings of Object-level Outlier Factors. $\phi_S$ denotes the outlieriness scoring performed on the feature subset $S$ . Therefore, the sequential couplings indicate that the feature subset $S_t$ is successively determined by the outlieriness scoring function $\phi$ on the feature subset $S_{t-1}$ , which helps iteratively enhance the feature subset selection for the $\phi$ function. . . . .	120
9.2	Our SEMSE Framework. $\mathbf{y}$ contains outlier scores of all data objects. $\eta$ , $\psi$ and $\phi$ are functions for outlier thresholding, fragmentary sparse modeling, and outlier scoring, respectively. . . . .	122
9.3	AUC Performance on Data with Different Levels of Noisy Features. 'ORG' denotes the bare LeSiNN/iForest. All methods obtain AUC of one with more than 32% relevant features. . . . .	130
9.4	Runtime of CINFO and Its Competitors Using LeSiNN. 'ORG' denotes the bare LeSiNN. Logarithmic scales are used. Similar trends can be expected when using iForest as the outlier detector, since LeSiNN and iForest have similar time complexities. . . . .	131