

TABLE OF CONTENTS

	Certificate	i
	Declaration	ii
	Acknowledgements	iii
	Abstract	v
	Table of Contents	vii
	List of Tables	x
	List of Figures	xii
Chapter I	Introduction	1
	1.1 Data Mining	1
	1.2 Privacy in Data Mining	1
	1.2.1 Phases in Privacy Preserving	2
	1.2.2 Data Anonymization	6
	1.3 Privacy Issues in Relational Datasets	8
	1.3.1 Linkage of Records	9
	1.3.2 Linkage of Attributes	10
	1.3.3 Linkage of Tables	12
	1.3.4 Attacks of Probabilistic Nature	13
	1.4 Privacy Issues in Network Datasets	13
	1.4.1 Definitions and Notations	16
	1.4.2 Disclosure of Identity	16
	1.4.3 Active Attacks and Passive Attacks	17
	1.4.4 Revealing Link Structure	19
	1.4.5 Link Structure Derivation of the Total Network	20
	1.4.6 Content Disclosure	20
	1.5 Privacy versus Cryptography	21
	1.6 Data Utility	22
	1.7 Motivation	23
	1.8 Thesis Statement	24
	1.9 Thesis Organization	26
Chapter II	Literature Review	27
	2.1 Introduction	27
	2.1.1 Data Partitioning Methods	29
	2.1.2 Cryptography-Based Methods	29
	2.1.3 Generative-Based Methods	30
	2.1.4 Data Modification Methods	30
	2.1.5 Noise Addition Methods	31
	2.1.6 Space Transformation Methods	31
	2.1.7 Data Restriction Methods	31

2.1.8	Blocking-Based Methods	32
2.1.9	Sanitization-Based Methods	32
2.1.10	Data Ownership Techniques	33
2.1.11	k-anonymity	33
2.1.12	l-Diversity	36
2.1.13	t-Closeness	38
2.1.14	Anonymization Operations	39
2.1.14.1	Generalization and Suppression	39
2.1.14.2	Anatomization and Permutation	42
2.1.14.3	Data swapping	43
2.1.14.4	Data Randomization	44
2.1.14.5	Additive Noise	44
2.1.15	Synthetic data generation	45
2.1.16	Matrix Decomposition Methods	47
2.1.16.1	Uniformly Distributed Noise	47
2.1.16.2	Normally Distributed Noise	48
2.1.16.3	Singular Value Decomposition	48
2.1.16.4	Non-negative Matrix Factorization	49
2.1.17	Wavelet Based Data Distortion	50
2.2	Data Utility Measures for Relational Datasets	50
2.2.1	General Purpose Metrics	51
2.2.2	Special Purpose Metrics	53
2.2.3	Trade-off Metrics	55
2.2.4	Classification Metric	57
2.2.5	Discernibility Metric	57
2.2.6	Minimal Distortion Metric	58
2.3	Privacy Preserving in Network Datasets	59
2.3.1	k-Anonymity and Graph Randomization	60
2.3.2	k-Anonymity and Minimal Edge Changes	61
2.3.3	k-Anonymity and Isomorphic Graphs	62
2.3.4	Preventing Link Re-identification Attacks	64
2.3.5	Privacy-Preserving Link Analysis	64
2.3.6	Random Perturbation for Relationship Protection	65
2.3.7	Cryptographic Protocols for Private Relationships	66
2.3.8	Synthetic Graph Generation	66
2.4	Data Utility Measures in Network Datasets	67
2.5	Summary	67
Chapter III	Finding and Fine Tuning Data Utility in Relational Dataset	68
3.1	Introduction	68
3.2	System Model	68
3.2.1	Input Collection	69
3.2.2	Pre-Analysis	70

3.2.3	Data Perturbation	70
3.2.4	Classification/Clustering Accuracy Evaluation	70
3.2.5	Iteration/Threshold Check	70
3.2.6	Mathematical Background for Data Utility	71
3.3	Data Utility Computation	72
3.4	Data Utility based Privacy: Iterative Version	73
3.5	Data Utility based Privacy: Threshold Version	74
3.6	Implementation and Dataset	76
3.6.1	Lung Cancer Dataset	76
3.6.2	Credit History Dataset	76
3.7	Experimental Results and Analysis	77
3.8	Summary	85
Chapter IV	Finding and Fine Tuning Data Utility in Graph Datasets	87
4.1	Introduction	87
4.2	System Model	87
4.2.1	Input Collection	88
4.2.2	Pre-Analysis	88
4.2.3	Data Perturbation	89
4.2.4	Graph Statistics Evaluation	89
4.2.5	Iteration Threshold/Check	89
4.2.6	Mathematical Background for Data Utility	90
4.3	Data Utility Computation	91
4.4	Data Utility based Privacy: Iterative Version	92
4.5	Data Utility based Privacy: Threshold Version	93
4.6	Implementation and Dataset	95
4.6.1	Enron Dataset	95
4.6.2	Orkut Profiles Dataset	95
4.6.3	Implementation	96
4.7	Experimental Results and Analysis	97
4.8	Summary	105
Chapter V	Conclusion & Future Scope of Work	107
5.1	Conclusion	107
5.2	Future Scope of Work	110
Appendices	Appendix A : Graph Definitions	112
	Appendix B : Reference	114
	Appendix C : e-Websites/Downloads	121
List of Publications		122