

Contents

<i>Acknowledgments</i>	<i>page viii</i>
1 Introduction	1
1.1 Small Summaries for Big Data	1
1.2 Preliminaries	4
1.3 Summaries in Applications	13
1.4 Computational and Mathematical Tools	19
1.5 Organization of the Book	25
PART I FUNDAMENTAL SUMMARY TECHNIQUES	27
2 Summaries for Sets	29
2.1 Morris Approximate Counter	29
2.2 Random Sampling	34
2.3 Weighted Random Sampling	37
2.4 Priority Sampling	46
2.5 k Minimum Values (KMV) for Set Cardinality	48
2.6 HyperLogLog (HLL) for Set Cardinality	55
2.7 Bloom Filters for Set Membership	61
3 Summaries for Multisets	68
3.1 Fingerprints for Testing Multiset Equality	69
3.2 Misra–Gries (MG)	73
3.3 SpaceSaving	80
3.4 Count-Min Sketch for Frequency Estimation	84
3.5 Count Sketch for Frequency Estimation	92
3.6 (Fast) AMS Sketch for Euclidean Norm	98
3.7 L_p Sketch for Vector Norm Estimation	102
3.8 Sparse Vector Recovery	105

3.9	Distinct Sampling/ ℓ_0 Sampling	112
3.10	L_p Sampling	116
4	Summaries for Ordered Data	120
4.1	Q-Digest	121
4.2	Greenwald–Khanna (GK)	130
4.3	Karnin–Lang–Liberty (KLL)	135
4.4	Dyadic Count Sketch (DCS)	143
	PART II ADVANCED SUMMARIES AND EXTENSIONS	151
5	Geometric Summaries	153
5.1	ε -Nets and ε -Approximations	153
5.2	Coresets for Minimum Enclosing Balls	158
5.3	ε -Kernels	163
5.4	k -Center Clustering	168
5.5	The (Sparse) Johnson–Lindenstrauss Transform	171
6	Vector, Matrix, and Linear Algebraic Summaries	176
6.1	Vector Computations: Euclidean Norm and Inner Product Estimation	176
6.2	ℓ_p Norms and Frequency Moments	179
6.3	Full Matrix Multiplication	181
6.4	Compressed Matrix Multiplication	183
6.5	Frequent Directions	186
6.6	Regression and Subspace Embeddings	190
7	Graph Summaries	192
7.1	Graph Sketches	192
7.2	Spanners	197
7.3	Properties of Degree Distributions via Frequency Moments	199
7.4	Triangle Counting via Frequency Moments	200
7.5	All-Distances Graph Sketch	202
8	Summaries over Distributed Data	207
8.1	Random Sampling over a Distributed Set	208
8.2	Point Queries over Distributed Multisets	209
8.3	Distributed Ordered Data	216
9	Other Uses of Summaries	218
9.1	Nearest-Neighbor Search	218
9.2	Time Decay	225

	Contents	vii
9.3	Data Transformations	231
9.4	Manipulating Summaries	238
10	Lower Bounds for Summaries	242
10.1	Equality and Fingerprinting	244
10.2	Index and Set Storage	245
10.3	Disjointness and Heavy Hitters	247
10.4	Gap Hamming and Count Distinct, Again	250
10.5	Augmented Index and Matrix Multiplication	251
	<i>References</i>	253
	<i>Index</i>	267