

Contents

Acknowledgements	iii
Contents	vi
1 Introduction	1
1.1 Motivation	4
1.2 Related Work	8
2 Preliminaries	13
2.1 General Notations	13
2.2 Data Streams	14
2.3 Clustering	17
2.4 Chernoff Bounds with Limited Independence	19
3 Sampling Data Streams	21
3.1 The Unique Element (UE) Data Structure	22
3.2 The Distinct Elements (DE) Data Structure	23
3.3 A Sample Data Structure using Totally Random Hash Functions	23
3.4 A Sample Data Structure using Random Number Generators	25
3.5 A Sample Data Structure using Pairwise Independent Hash Functions	29
4 Sampling Geometric Data Streams and Applications	33
4.1 Sampling Geometric Data Streams	34
4.2 ϵ -Nets and ϵ -Approximations in Data Streams	35
4.3 Random Sampling with Neighborhood Information	36
4.4 Estimating the Weight of a Euclidean Minimum Spanning Tree	39
5 The Coreset Method	47
5.1 Definitions	47
5.2 Coresets for k-Median	54
5.3 Coresets for k-Means	60
5.4 Coresets for Oblivious Optimization Problems	65
5.5 Constructing Solutions on the Coreset	69
5.6 Coresets via Sampling	76

6	Coresets in Data Streams	93
6.1	Insertions	95
6.2	Deletions	97
6.3	Maximum Spanning Tree	99
7	A Kinetic Data Structure for MaxCut	103
7.1	Kinetic Tournament Trees	104
7.2	Approximating the Bounding Cube	104
7.3	The Kinetic Data Structure for MaxCut	105
8	An Efficient k-Means Implementation using Coresets	111
8.1	Definitions and Notations	112
8.2	The Algorithm	114
8.3	Experiments	120
9	Counting Motifs in Data Streams	133
9.1	Counting Triangles in Adjacency Streams	134
9.2	Counting Triangles in Incidence Streams	142
9.3	Counting Cliques of Arbitrary Size	149
9.4	Counting $K_{3,3}$ in Incidence Streams	152
10	Conclusions	157
	Bibliography	159