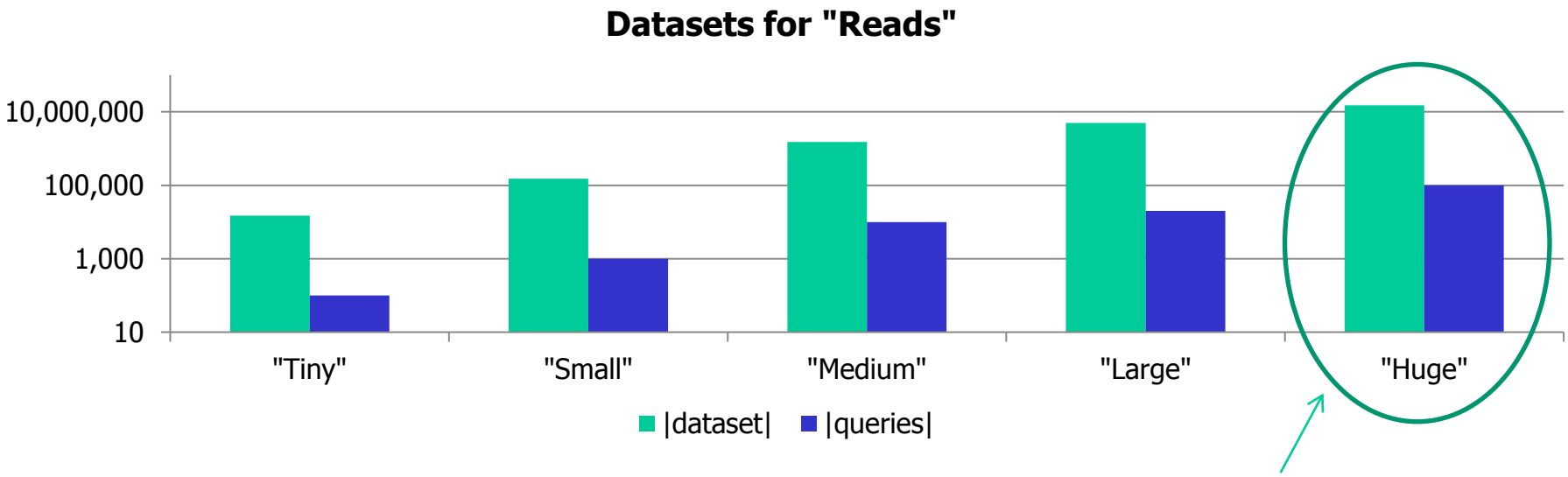
EDBT/ICDT 2013 Joint Conference
March 18-22, 2013 - Genoa, Italy

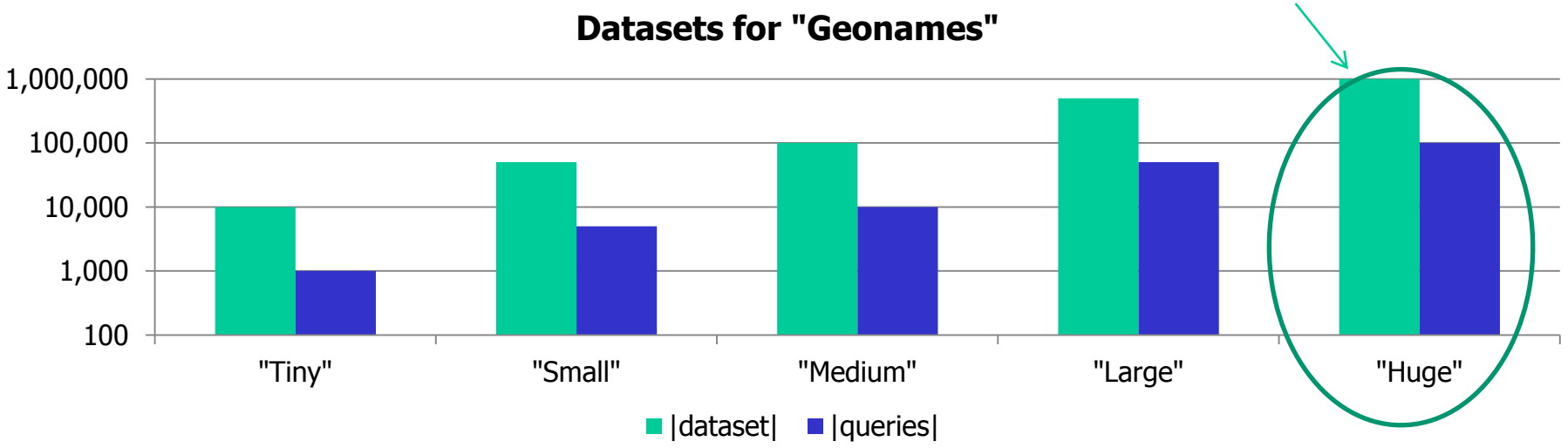
S⁴: International Competition on Scalable String Similarity Search & Join

Sebastian Wandelt, Humboldt-Universität zu Berlin

Datasets



Original evaluation dataset



Competing Teams in the beginning

Team	Authors	Affiliation	Paper name
1	Yu Jiang, Dong Deng, Jiannan Wang, Guoliang Li, Jianhua Feng	Tsinghua University, China	Efficient Parallel Partition-based Algorithms for Similarity Search and Join with Edit Distance Constraints
2	Jan Hentschel, Thomas Meyer, Thomas Rommel	Universität Magdeburg, Germany	Trying to outperform well-known indices with a sequential scan
3	Alexander Tiskin	University of Warwick, UK	Efficient High-Similarity String Comparison: The Waterfall Algorithm
4	Stefan Gerdjikov, Stoyan Mihov, Petar Mitankin, Klaus U. Schulz	Sofia University, Bulgaria	WallBreaker - overcoming the wall effect in similarity search
5	Enrico Siragusa, David Weese, Knut Reinert	FU Berlin, Germany	Scalable string similarity search / join with approximate seeds and multiple backtracking
6	Akhil Arora, Shashwat Mishra, Tejas Gandhi, Arnab Bhattacharya	IIT Kanpur, India	Efficient Edit Distance based String Similarity Search using Deletion Neighborhoods
7	Manish Patil, Xuanting Cai, Sharma V. Thankachan, Rahul Shah, David Foltz, Seung-Jong Park	Louisiana State University, USA	Approximate String Matching by Position Restricted Alignment
8	Jianbin Qin, Xiaoling Zhou, Wei Wang, Chuan Xiao	University of New South Wales, Australia	Efficient algorithms for edit similarity queries
9	Mitsuki Kimura, Atsuhiro Takasu, Jun Adachi	National Institute of informatics, Japan	FPI:Frequent Patterns Indexing for Approximate String Searches
10	Jiaying Wang, Xiaochun Yang, Bin Wang	Northeastern University, China	Cache-Aware Parallel Approximate String Search and Join Using BWT

Teams: general approach

- Team 1: partitioning and pruning
- Team 2: sequential search (no “real” index)
- Team 3: Bit-parallel LCS computation
- Team 4: Compact acyclic directed word graphs
- Team 5: radix/suffix trees and filtration
- Team 6: deletion neighborhoods / hashing
- Team 7: q-gram indexing with filtering
- Team 8: Trie-index with filtering
- Team 9: variable-length n-grams
- Team 10: BWT / cache-aware implementation

Competing programs

Team	Competitor	Program / Parameter
1	1_A	search-1
1	1_B	search-2
2	2_A	geo/dna
3	3_A	tsearch
4	4_A	wallbreaker … 16 y 5 3
4	4_B	wallbreaker ... 16 n 5 3
5	5_A	search --online -t 8
5	5_B	search --seed-length 4/10 -t 8
5	5_C	search --seed-length 5/13 -t 8
5	5_D	search --seed-length 6/15 -t 8
6	6_A	a.out
7	7_A	simsearch
8	8_A	unsw_dbw_edbt_simsearch
9	9_A	search
10	10_A	indexsearch

Evaluation

- Track 1: searching reads -

Track 1 (reads): Results „Huge“

Competitor	Indexing time (s)	Search time (s)	Ranking	Peak mem (GB)
1_A	107.964	312.105	3	24.8
1_B	100.91	924.664	7	24.9
3_A	2.042	30898.026	9	15.6
4_A	2251.815	232.504	1	37.8
4_B	1754.473	248.962	2	37.0
5_B	192.212	580.778	4	13.6
5_C	193.899	761.174	5	13.7
5_D	193.712	900.251	6	13.7
7_A	2710.906	1587.776	8	20.1
10_A	465.608	42866.558	10	40.6

Evaluation

- Track 1: searching geonames -

Track 1 (geonames): Results

Competitor	Indexing time (s)	Search time (s)	Ranking	Peak Mem (GB)
1_A	1.933	59.947	2	1.5
1_B	1.673	46.803	1	1.5
3_A	0.243	109.571	5	6.5
4_A	39.717	69.175	4	13.3
4_B	39.874	67.336	3	12.7
5_B	3.127	4903.032	11	1.6
5_C	3.17	4387.356	10	1.3
5_D	3.148	3097.004	9	1.2
6_A	1206.345	248.294	7	24.7
8_A	2.028	445.513	8	1.1
10_A	1.56	137.517	6	0.6

Track 1: Combined Results

	Searching reads		Searching geonames		Sum	
Competitor	**Search time (s)**	**Ranking**	**Search time (s)**	**Ranking**	**Search time (s)**	**Ranking**
1_A	312.105	3	59.947	2	372.052	3
1_B	924.664	7	46.803	1	971.467	4
3_A	30898.026	9	109.571	5	31007.597	8
4_A	232.504	1	69.175	4	301.679	1
4_B	248.962	2	67.336	3	316.298	2
5_B	580.778	4	4903.032	11	5483.81	7
5_C	761.174	5	4387.356	10	5148.53	6
5_D	900.251	6	3097.004	9	3997.255	5
10_A	42866.558	10	137.517	6	43004.075	9

Track 1: Combined Results

- The **winner** of Track 1 is:

Team 4 with **Program 4_A**

Evaluation

- Track 2: joining reads -

Track 2 (reads): Results

	Join time (s)					
Competitor	**k=0**	**k=4**	**k=8**	**k=12**	**k=16**	**Ranking**
1_A	9	57	223	5064	82637	1
1_B	9	57	536	13857		
3_A	15531	345600				
4_A	2257	2478	4872	13248	149344	2
4_B	1764	1884	4544	12844		
5_B	30	741	7602	155421		
5_C	31	724	6713			
5_D	31	738	4494			
10_A	328	71853				

Evaluation

- Track 2: joining geonames -

Track 2 (geonames): Results

	Join time (s)					
Competitor	**k=0**	**k=1**	**k=2**	**k=3**	**k=4**	**Ranking**
1_A	1	2	6	50	346	1
1_B	1	2	7	54	353	2
3_A	588	564	656	848	1700	5
4_A	41	45	81	441	945	4
4_B	40	42	79	418	942	3
5_B	11	78	1719			
5_C	11	37	726	11463		
5_D	11	33	786			
8_A	3	21	218	3339	21230	
10_A	11	29	199	1913		

Track 2: Combined Results

- Both datasets are dominated by Team 1
- The **winner** of Track 2 is:

Team 1 with **Program 1_A**

Results Overview

- Track 1 (search):
  - Reads: Team 4 with 4_A
  - Geonames: Team 1 with 1_B
  - OVERALL: **Team 4 with 4_A**
- Track 2 (join):
  - Reads: Team 1 with 1_A
  - Geonames: Team 1 with 1_A
  - OVERALL: **Team 1 with 1_A**