### Dichiarazione Sostitutiva di Certificazione (art. 46 DPR n. 445/2000) Dichiarazione Sostitutiva dell'Atto di Notorietà (art. 47 DPR n. 445/2000)

Il sottoscritto Giulio Ermanno Pibiri

- o nato a Bagno a Ripoli (FI) il 13/07/1990
- o attualmente residente a Prato (PO) in via Guizzelmi, n. 16, CAP 59100
- o reperibile al numero di telefono +39 340 9100050

visto il D.P.R. 28 dicembre 2000, n. 445 concernente "T.U. delle disposizioni legislative e regolamentari in materia di documentazione amministrativa" e successive modifiche ed integrazioni;

vista la Legge 12 novembre 2011, n. 183 ed in particolare l'art. 15 concernente le nuove disposizioni in materia di certificati e dichiarazioni sostituitive;

consapevole che, ai sensi dell'art. 76 del DPR 445/2000, le dichiarazioni mendaci, la falsità negli atti e l'uso di atti falsi sono punite ai sensi del Codice penale e delle leggi speciali vigenti in materia, dichiara sotto la propria responsabilità: che quanto dichiarato nel seguente *Curriculum Vitae et Studiorum*, comprensivo delle informazioni sulla produzione scientifica, corrisponde a verità.

# Giulio Ermanno Pibiri

### Curriculum Vitae et Studiorum

### Contact Information

National Research Council of Italy (CNR)

Institute of Science and Information Technologies "A. Faedo" (ISTI)

Via G. Moruzzi 1, 56124 Pisa, Italy

Email giulio.pibiri@di.unipi.it

Email giulio.ermanno.pibiri@isti.cnr.it

Personal page http://pages.di.unipi.it/pibiri

GitHub profile https://github.com/jermp

### Personal Information

Place of birth Bagno a Ripoli (Florence), Italy

Date of birth 13 July 1990

### Education

### 01/11/2015 - PhD in Computer Science (INF/01).

31/10/2018

- o University of Pisa, Pisa, Italy
- o Thesis: Space- and Time-Efficient Data Structures for Massive Datasets (Defended on 08/03/2019)
- o Grade: Excellent
- Supervisor: Rossano Venturini (https://rossanoventurini.github.io)

#### 2012 - 2014

### Master Degree in Computer Science & Networking (class LM18).

- o University of Pisa and Scuola Superiore Sant'Anna, Pisa, Italy
- o Thesis: Dynamic Elias-Fano Encoding (Defended on 06/03/2015)
- o Grade: 110/110 summa cum laude
- Supervisor: Rossano Venturini (https://rossanoventurini.github.io)

### 2009 – 2012 Bachelor Degree in Computer Engineering (class L08).

- University of Florence, Florence, Italy
- o Thesis: Quantum Computation & Grover's Algorithm (Defended on 09/10/2012)
- o Grade: 110/110 summa cum laude
- Supervisor: Gabriele Vezzosi (http://www.dma.unifi.it/~vezzosi)

### 2004 - 2009

### High School Diploma.

- o Liceo Scientifico Statale Guido Castelnuovo, Florence, Italy
- o Grade: 100/100.

### Research Interests

Keywords Data Structures, Data Compression, Indexing, Efficiency

Short Description

The research activity focuses on devising compressed data structures and algorithms to index and search large quantities of data. The proposed solutions are available as research papers and optimized software libraries (written in C++).

Research Problems

Compressed Dictionaries for K-Mers (BIOINF 2022); Time Series Compression (SPIRE 2021); Minimal Perfect Hashing (SIGIR 2021); Prefix-Sums (SPE 2020); Rank/Select Queries (IN-FOSYS 2021); Query Auto-Completion (SIGIR 2020); Bitmap Compression (DCC 2021); Indexing of Semantic Relations (TKDE 2020, ICDE 2021); Indexing and Estimation of Language Models (SIGIR 2017, TOIS 2019); Inverted Index Compression (TOIS 2017, EBDT 2018, WSDM 2019, TKDE 2019, CSUR 2020); Succinct and Dynamic Ordered Sets of Integers (CPM 2017).

### Research Positions

# 15/03/2021 -

### Postdoctoral Research Fellow in Computer Science.

- o Institute of Science and Information Technologies "A. Faedo" (ISTI), National Research Council of Italy (CNR), Pisa, Italy
- o Research grant issued on the European project ACCORDION with theme "Tecniche algoritmiche per compressione, indicizzazione e ricerca di grandi quantità di dati e progettazione di relative librerie software open source" (Protocollo n. 0000901/2021, 09/03/2021, ISTI 004/2021 - PI)
- o Reference publications: SIGIR 2021, SPIRE 2021, BIOINF 2022

### 01/11/2018 -28/02/2021

### Postdoctoral Research Fellow in Computer Science.

- o Institute of Science and Information Technologies "A. Faedo" (ISTI), National Research Council of Italy (CNR), Pisa, Italy
- o Research grant issued on the European project BIGDATAGRAPES with theme "Compressione, indicizzazione e ricerca su grandi collezioni di dati semantici" (Protocollo n. 0003847, 24/10/2018, ISTI 014/2018 - PI)
- o Reference publications: TKDE 2019, TKDE 2020, SIGIR 2020, SPE 2020, INFOSYS 2021, DCC 2021, ICDE 2021

### 01/11/2015 - PhD Student in Computer Science.

- 31/10/2018 o University of Pisa, Pisa, Italy
  - o Thesis: Space- and Time-Efficient Data Structures for Massive Datasets
  - o Supervisor: Rossano Venturini (https://rossanoventurini.github.io)
  - o Reference publications: TKDE 2019, TOIS 2019, WSDM 2019, SIGIR 2017, CPM 2017, TOIS

Part of the research was conducted abroad (6 months):

- o 01/05/2018 01/10/2018
  - The University of Melbourne, School of Computing and Information Systems, Melbourne, Australia
  - Supervisor: Alistair Moffat (https://people.eng.unimelb.edu.au/ammoffat)
  - Worked on fast dictionary-based decoding of compressed inverted index data.
  - Reference publication: WSDM 2019
- o 01/04/2018 30/04/2018
  - RIKEN Advanced Intelligence Project (AIP), Tokyo, Japan
  - Supervisor: Yasuo Tabei (https://sites.google.com/site/yasuotabei)
  - Worked on various problems, such as, string similarity search, trie indexing, rank/select indexes, and sparse matrix multiplication.
  - Reference publication: INFOSYS 2021

## **Projects**

#### 01/09/2021 -Italian PON project OK-INSAID.

present

- o Institute of Science and Information Technologies "A. Faedo" (ISTI), National Research Council of Italy (CNR), Pisa, Italy
- o Role: Investigator

#### 15/03/2021 -European project ACCORDION.

- present o Institute of Science and Information Technologies "A. Faedo" (ISTI), National Research Council of Italy (CNR), Pisa, Italy
  - Role: Investigator

#### 01/11/2018 -European project BIGDATAGRAPES.

28/02/2021

- o Institute of Science and Information Technologies "A. Faedo" (ISTI), National Research Council of Italy (CNR), Pisa, Italy
- o Role: Task Leader for the task 3.3 "Big Data Indexing"

### 01/06/2017 – European project LIGA.

31/10/2018

- o Institute of Science and Information Technologies "A. Faedo" (ISTI), National Research Council of Italy (CNR), Pisa, Italy
- o Role: Principal Developer

# Teaching Activity

Teacher for Theory and Practice of Data Compression (5 CFU), PhD Program in Ingegneria A.Y. 2021/2022 dell'Informazione, University of Pisa, Italy

https://github.com/jermp/data\_compression\_course

- A.Y. 2019/2020 Teacher for Algorithmics and Laboratory - Corso B (code 008AA, 3 CFU), Bachelor Degree in Computer Science, University of Pisa, Italy
- A.Y. 2018/2019 Assistant for Algorithmics and Laboratory - Corso A (code 008AA, 3 CFU), Bachelor Degree in Computer Science, University of Pisa, Italy
- A.Y. 2017/2018 Assistant for Competitive Programming and Contests (code 645AA, 6 hours), Master Degree in Computer Science, University of Pisa, Italy
- A.Y. 2016/2017 Assistant for Competitive Programming and Contests (code 645AA, 6 hours), Master Degree in Computer Science, University of Pisa, Italy
- Teacher for Algorithmics and Laboratory Corso di recupero (code 008AA, 3 CFU), Bachelor A.Y. 2015/2016 Degree in Computer Science, University of Pisa, Italy
- A.Y. 2015/2016 Assistant for Algorithmics and Laboratory - Corso A (code 008AA, 3 CFU), Bachelor Degree in Computer Science, University of Pisa, Italy

### Awards and Grants

- 2021 Young Researcher Award for the year 2020, issued by ISTI-CNR.
- 2020 Young Researcher Award for the year 2019, issued by ISTI-CNR.
- 2017 SIGIR Student Travel Grant, issued by ACM SIGIR.
- 2015 PhD Scholarship, issued by the University of Pisa, Department of Computer Science.
- 2015 Master Degree Award for the A.Y. 2013/2014, issued by Scuola Superiore Sant'Anna.
- 2015 Best Master Thesis Award in Theoretical Computer Science, issued by the Italian chapter of the European Association for Theoretical Computer Science (EATCS).

### **Organizing Committees**

- 2020 The 28-th edition of the Annual European Symposium on Algorithms (ESA 2020).
- 2019 The 30-th edition of the International Symposium on Combinatorial Pattern Matching (CPM 2019).
- 2017 The 24-th International Symposium on String Processing and Information Retrieval (SPIRE 2017).
- 2016 The 39-th ACM International SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016). (As student volunteer.)

### **Program Committees**

- 2022 The 45-th ACM International SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022).
- 2022 The 44-th European Conference on Information Retrieval (ECIR 2022).
- 2022 The 15-th International ACM Conference on Web Search and Data Mining (WSDM 2022).
- 2021 The 15-th IEEE International Conference on Application of Information and Communication Technologies (AICT 2021).
- 2021 The 30-th ACM International Conference on Information and Knowledge Management (CIKM 2021).
- 2021 The 44-th ACM International SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021).
- 2021 The 43-rd European Conference on Information Retrieval (ECIR 2021).
- 2021 The 14-th International ACM Conference on Web Search and Data Mining (WSDM 2021).
- 2020 The 29-th ACM International Conference on Information and Knowledge Management (CIKM 2020).
- 2020 The 43-rd ACM International SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020).
- 2019 The 42-nd ACM International SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019).

## Reviewing Activity

2016 – present I am or have been an anonymous reviewer for the following conferences/journals.

### **Conferences**

- o SIGIR ACM Conference on Research and Development in Information Retrieval
- o WSDM ACM Conference on Web Search and Data Mining
- o WWW The Web Conference
- o CIKM ACM Conference on Information and Knowledge Management
- o CPM Annual Symposium on Combinatorial Pattern Matching
- o DCC IEEE Data Compression Conference
- o ECIR European Conference on Information Retrieval
- o ESA European Symposium on Algorithms
- o SPIRE String Processing and Information Retrieval
- o ISAAC International Symposium on Algorithms and Computation

#### **Journals**

- TALG ACM Transactions on Algorithms
- o INFOSYS Information Systems
- o SPE Software: Practice and Experience
- o Algorithmica
- o MDPI Algorithms

### **Talks**

### **Invited Keynote**

10/05/2022 On Weighted K-Mer Dictionaries. TUDASTIC 2022 (TUtorials on DAta Structures for Text Indexation and Compression). Lille, France.

### Conferences

- 20/05/2022 Sparse and Skew Hashing of K-Mers. RECOMB-seq 2022. La Jolla, California, USA.
- 04/10/2021 TSXor: A Simple Time-Series Compression Algorithm. SPIRE 2021. Lille, France (Virtual event).
- 14/09/2021 PTHash: Revisiting FCH Minimal Perfect Hashing. IIR 2021. Polytechnic University of Bari, Bari (Virtual event).
  - 07/2021 PTHash: Revisiting FCH Minimal Perfect Hashing. SIGIR 2021. Montreal, Canada (Virtual event).
  - 04/2021 Compressed Indexes for Fast Search of Semantic Data. ICDE 2021. Chania, Greece (Virtual event).
  - 03/2021 Fast and Compact Set Intersection through Recursive Universe Partitioning. DCC 2021. Snow Bird, USA (Virtual event).
- 27/07/2020 Efficient and Effective Query Auto-Completion. SIGIR 2020. Xi'An, China (Virtual event).
- 17/09/2019 Compressed Indexes for Fast Search of Semantic Data. IIR 2019. Department of Information Engineering, Padova, Italy.
- 12/02/2019 Fast Dictionary-based Compression for Inverted Indexes. WSDM 2019. Melbourne Exhibition Center, Melbourne, Australia.
- 10/08/2017 Efficient Data Structures for Massive N-Gram Datasets. SIGIR 2017. Keio Plaza Hotel, Tokyo, Japan.
- 06/06/2017 Efficient Data Structures for Massive N-Gram Datasets. IIR 2017. Università della Szizzera Italiana, Lugano, Switzerland.

### **Seminars**

- 19/05/2022 Sparse and Skew Hashing of K-Mers. ISTI-CNR, Pisa, Italy (Virtual event).
- 22/12/2021 *Minimal Perfect Hashing and K-Mer String Dictionaries*. "Ca' Foscari" University of Venice, Venice, Italy (Virtual event).
- 16/11/2021 PTHash: Revisiting FCH Minimal Perfect Hashing. ISTI-CNR, Pisa, Italy (Virtual event).
- 04/03/2021 Efficiency for Real-World Applications. ISTI-CNR, Pisa, Italy (Virtual event).
- 07/06/2019 Ordered Set Problems. ISTI-CNR, Pisa, Italy.
- 01/02/2019 Indexing Compressed Data for Fast Retrieval. University of Pisa, Pisa, Italy.
- 29/10/2018 Effective Web Graph Representations. University of Pisa, Pisa, Italy.
- 17/05/2018 On Optimally Partitioning Variable-Byte Index Data. RMIT University, Melbourne, Australia.
- 10/04/2018 Elias-Fano Encoding: a powerful tool for data structure design. RIKEN AIP, Tokyo, Japan.
- 21/06/2016 Elias-Fano Encoding: succinct representation of monotone integer sequences with search operations. University of Pisa, Pisa, Italy.

### Publications

### **Journal Papers**

BIOINF 2022 Giulio Ermanno Pibiri, Sparse and Skew Hashing of K-Mers. 2022. Bioinformatics (BIOINF),

pages 9.

Scimago Rating: Q1

DOI: 10.1093/bioinformatics/btac245

ISSN: 1460-2059

INFOSYS 2021 Giulio Ermanno Pibiri and Shunsuke Kanda, Rank/Select Queries over Mutable Bitmaps. 2021.

Information Systems (INFOSYS), pages 15.

Scimago Rating: Q2

DOI: 10.1016/j.is.2021.101756

ISSN: 0306-4379

CSUR 2020 Giulio Ermanno Pibiri and Rossano Venturini, Techniques for Inverted Index Compression.

2020. ACM Computing Surveys (CSUR), pages 36.

Scimago Rating: **Q1**DOI: 10.1145/3415148
ISSN: 0360-0300

SPE 2020 Giulio Ermanno Pibiri and Rossano Venturini, Practical Trade-Offs for the Prefix-Sum Problem.

2020. Software: Practice and Experience (SPE), pages 29.

Scimago Rating: **Q2** DOI: 10.1002/spe.2918 ISSN: 0038-0644

TKDE 2020 Raffaele Perego, Giulio Ermanno Pibiri and Rossano Venturini, Compressed Indexes for Fast

Search of Semantic Data. 2020. IEEE Transactions on Knowledge and Data Engineering

(TKDE), pages 12. Scimago Rating: **Q1** 

DOI: 10.1109/TKDE.2020.2966609

ISSN: 1041-4347

TKDE 2019 Giulio Ermanno Pibiri and Rossano Venturini, On Optimally Partitioning Variable-Byte Codes.

2019. IEEE Transactions on Knowledge and Data Engineering (TKDE), pages 12.

Scimago Rating: Q1

DOI: 10.1109/TKDE.2019.2911288

ISSN: 1041-4347

TOIS 2019 Giulio Ermanno Pibiri and Rossano Venturini, Handling Massive N-Gram Datasets Efficiently.

2019. ACM Transactions on Information Systems (TOIS), pages 41.

Scimago Rating: **Q1** DOI: 10.1145/3302913 ISSN: 1046-8188

TOIS 2017 Giulio Ermanno Pibiri and Rossano Venturini, Clustered Elias-Fano Indexes. 2017. ACM

Transactions on Information Systems (TOIS), volume 2, pages 33.

Scimago Rating: **Q1** DOI: 10.1145/3052773 ISSN: 1046-8188

### **Conference Papers**

SPIRE 2021 Andrea Bruno, Franco Maria Nardini, Giulio Ermanno Pibiri, Roberto Trani, and Rossano Venturini, *TSXor: A Simple Time Series Compression Algorithm.* 2021. International Symposium on String Processing and Information Retrieval (SPIRE), 8 pages.

GGS Rating: B

DOI: 10.1007/978-3-030-86692-1\_18

ISBN: 9783030866921

SIGIR 2021 Giulio Ermanno Pibiri and Roberto Trani, *PTHash: Revisiting FCH Minimal Perfect Hashing*. 2021. ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 10.

GGS Rating: A++

DOI: 10.1145/3404835.3462849

ISBN: 9781450380379

DCC 2021 Giulio Ermanno Pibiri, Fast and Compact Set Intersection through Recursive Universe Partitioning. 2021. IEEE Data Compression Conference (DCC), pages 10.

GGS Rating: A-

DOI: 10.1109/DCC50243.2021.00037

ISBN: 9781665403337

SIGIR 2020 Simon Gog, Giulio Ermanno Pibiri and Rossano Venturini, *Efficient and Effective Query Auto-Completion*. 2020. ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 10.

GGS Rating: A++

DOI: 10.1145/3397271.3401432

ISBN: 9781450380164

WSDM 2019 Giulio Ermanno Pibiri, Matthias Petri, Alistair Moffat, Fast Dictionary-based Compression for Inverted Indexes. 2019. ACM Conference on Web Search and Data Mining (WSDM), pages 9.

GGS Rating: A+

DOI: 10.1145/3289600.3290962

ISBN: 9781450359405

SIGIR 2017 Giulio Ermanno Pibiri and Rossano Venturini, *Efficient Data Structures for Massive N-Gram Datasets*. 2017. ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 10.

GGS Rating: A++

DOI: 10.1145/3077136.3080798

ISBN: 9781450350228

CPM 2017 Giulio Ermanno Pibiri and Rossano Venturini, *Dynamic Elias-Fano Representation*. 2017. Annual Symposium on Combinatorial Pattern Matching (CPM), pages 14.

GGS Rating: B

DOI: 10.4230/LIPIcs.CPM.2017.30

ISBN: 9783959770392

### **Posters**

ICDE 2021 Raffaele Perego, Giulio Ermanno Pibiri and Rossano Venturini, *Compressed Indexes for Fast Search of Semantic Data.* 2021. IEEE International Conference on Data Engineering (ICDE), pages 2.

GGS Rating: A++

DOI: 10.1109/ICDE51399.2021.00248

ISBN: 9781728191850

### **Chapters and Theses**

2019 Giulio Ermanno Pibiri. *Space- and Time-Efficient Data Structures for Massive Datasets.* 2019. Ph.D. Thesis, University of Pisa, 210 pages.

EBDT 2018 Giulio Ermanno Pibiri and Rossano Venturini, *Inverted Index Compression*. 2018. Encyclopedia of Big Data Technologies (EBDT), pages 8.

DOI: 10.1007/978-3-319-63962-8\_52-1

ISBN: 9783319639628

### **Pre-Prints**

- 2021 Giulio Ermanno Pibiri and Roberto Trani, *Parallel and External-Memory Construction of Minimal Perfect Hash Functions with PTHash*. 2021. ArXiv, https://arxiv.org/abs/2106.02350, pages 12.
- 2020 Giulio Ermanno Pibiri and Rossano Venturini, *Succinct Dynamic Ordered Sets with Random Access.* 2020. ArXiv, https://arxiv.org/abs/2003.11835, pages 15.
- 2019 Giulio Ermanno Pibiri. *On Implementing the Binary Interpolative Coding Algorithm*. 2019. Tech Report, 8 pages.

### Software

GitHub profile All software is open-source and available from https://github.com/jermp.

### **Data Structures**

Efficient C++ implementations of the following data structures (see also related publications):

- o Inverted Indexes (TOIS 2017, TKDE 2019, WSDM 2019, SIGIR 2020, CSUR 2020)
- o Tries (SIGIR 2017, TOIS 2019, TKDE 2020)
- o Compressed Bitmaps (DCC 2021)
- Mutable Bitmaps with Rank/Select (INFOSYS 2021)
- o Segment-Trees and Fenwick-Trees (SPE 2020)
- o Minimal Perfect Hash Functions (SIGIR 2021)
- o K-Mer Dictionaries (BIOINF 2022)

A more detailed list follows below.

sshash SSHash: A compressed, weighted, associative, exact dictionary for k-mers.

Reference publication: BIOINF 2022

**pthash** PTHash: Fast and compact minimal perfect hash functions.

Reference publications: SIGIR 2021, ArXiv 2106.02350.

rank\_select Mutable bitmaps with support for Rank and Select queries.

Reference publication: INFOSYS 2021.

psds A range of tree-shaped data structures for maintaining prefix-sums, including:

- o binary Segment-Tree (top-down and bottom-up),
- o b-ary Segment-Tree,
- o Fenwick-Tree,
- o b-ary Fenwick-Tree.
- o blocked Fenwick-Tree,
- o truncated Fenwick-Tree.

Reference publication: SPE 2020.

autocomplete Efficienct and effective autocompletion framework, based on forward/inverted indexes, succinct

RMQ, and string dictionaries (Front-Coding and tries).

Reference publication: SIGIR 2020.

2i\_bench A benchmarking suite for inverted index data structures, featuring the following compressors:

- o Elias-Fano and partitioned Elias-Fano,
- o Opt-PFor-Delta,
- o Binary Interpolative,
- o QMX,
- Simple family,
- o Variable-Byte family, including Opt-VByte,
- o Gamma, Delta, Rice, Zeta,
- o DINT.

Reference publication: CSUR 2020.

**interp** An efficient implementation of the Binary Interpolative Coding algorithm.

**s\_indexes** Compressed bitmap indexes that support fast intersection and union.

Reference publication: DCC 2021.

rdf\_indexes Trie-based indexes for semantic data like RDF triples.

Reference publication: TKDE 2020.

dint DINT: fast and compact dictionary-based decoder for inverted lists.

Reference publication: WSDM 2019.

opt\_vbyte Optimal partitioning of inverted lists compressed using binary vectors and point-wise encoders,

like Variable-Byte.

Reference publication: TKDE 2019.

tongrams Fast language model queries and estimation in compressed space.

Reference publications: SIGIR 2017, TOIS 2019.

### clustered\_indexes

Clustered Elias-Fano inverted indexes. Reference publication: TOIS 2017.

### Miscellanea

### essentials

A C++ library providing essential core utilities for data structure design and benchmarking. More precisely:

- o benchmarking facilities, including: messages displaying local time, configurable timer class, function to prevent code elision by compiler, simple creation and printing of json documents;
- o functions to serialize-to and load-from disk data structures,
- o functions to compute the number of bytes consumed by data structures,
- o support for creating, removing, and iterate inside directories,
- o transparent support for contiguous memory allocation.

cmd\_line

Command line parser for C++17. It offers all handy features in just 150 lines of code.

mm\_file

A self-contained, header-only, implementation of memory-mapped files in C++ for both reading and writing.

### Languages

Native Italian

CEFR level: C2

English **Fluent**  CEFR level: C1

2018 TOEFL iBT in English.

100 (HIGH level)

2008

First Certificate in English (Level B2).

University of Cambridge, Cambridge, United Kingdom