The SIGACT Theoretical Computer Science Genealogy: Preliminary Report

Ian Parberry*
Department of Computer Sciences
University of North Texas

David S. Johnson†
AT&T Bell Laboratories

May 4, 2004

Abstract

The SIGACT Theoretical Computer Science Genealogy, which lists information on earned doctoral degrees of theoretical computer scientists, is currently in the process of being published on the World-Wide Web. We describe the document, its applications, and some simple statistics.

1 Introduction

The SIGACT[1] Theoretical Computer Science Genealogy lists information on earned doctoral degrees (thesis adviser, university, and year) of theoretical computer scientists worldwide. The genealogy was initially published in print form over a decade ago, and included a listing of the entire genealogy (Johnson [1]). However, the genealogy has since doubled from 554 entries listing 665 names to 1196 entries listing 1369 names, making it impractical to print the entire genealogy in the archival literature. Instead, the genealogy will be published electronically over the World-Wide Web as a collection of `html` files. A preliminary version is currently available [6]. An added bonus is that it is now possible to explore the intellectual ancestry of individuals in the genealogy by following a series of hypertext pointers.

The Theoretical Computer Science Genealogy is intended as an informational tool. Its main application is undoubtedly entertainment, but it does have more formal uses. At various times in the past, Program Directors at the National Science Foundation have used the genealogy to avoid possible conflicts of interest caused by having a funding proposal refereed by the doctoral adviser or student of the investigator. We envisage editors of refereed journals using it for the same purpose.

The remainder of this document is divided into four sections. The first describes the organization of the World-Wide Web version of the TCS Genealogy. The second describes the text database from which the `html` files are generated. The third describes some simple statistics about the TCS genealogy that can easily be obtained from the `html` files. The fourth describes the work remaining to be done before the genealogy is ready for a full release.

2 Organization

The World-Wide Web version of the TCS Genealogy is divided into a large number of files so that users who browse only a fraction of the genealogy will not have to wait while large amounts of unnecessary data are transferred across the Internet. The overall structure of the genealogy is shown in Figure 1 (many of the links are condensed or omitted to enhance readability). The major parts of the genealogy are the main index, the submission details page, the online form, the text file page, the statistics page, the name index, the letter indices, the university index, the country indices, the university indices, the year index, the decade indices, and the main database. Each of these is described briefly in a separate subsection below.

2.1 The Main Index

The main index is the first thing that the user sees, and is therefore very brief. It contains links to the SIGACT page [3], the submission details page, the text file page, the statistics page, the name index, the university index, and the year index (see Figure 2).

2.2 The Submission Details Page

The submission details page contains information on how to submit an update, what information is needed in an update, and what qualifications are necessary for entry into the genealogy. Basically, a person must have made a contribution of some kind to theoretical computer science, loosely defined as at least one of the following:

*Author's address: Department of Computer Sciences, University of North Texas, P.O. Box 13886, Denton, TX 76203–3886, U.S.A. Electronic mail: `ian@ponder.csci.unt.edu`. URL: `http://hercule.csci.unt.edu/ian`.

†Author's address: AT&T Bell Laboratories, 600 Mountain Avenue Rm. 2D–150, Murray Hill, NJ 07974, U.S.A. Electronic mail: `dsj@research.att.com`.

[1] SIGACT is the acronym for the ACM Special Interest Group on Algorithms and Computation Theory. More information about SIGACT is available on the World-Wide Web [3].

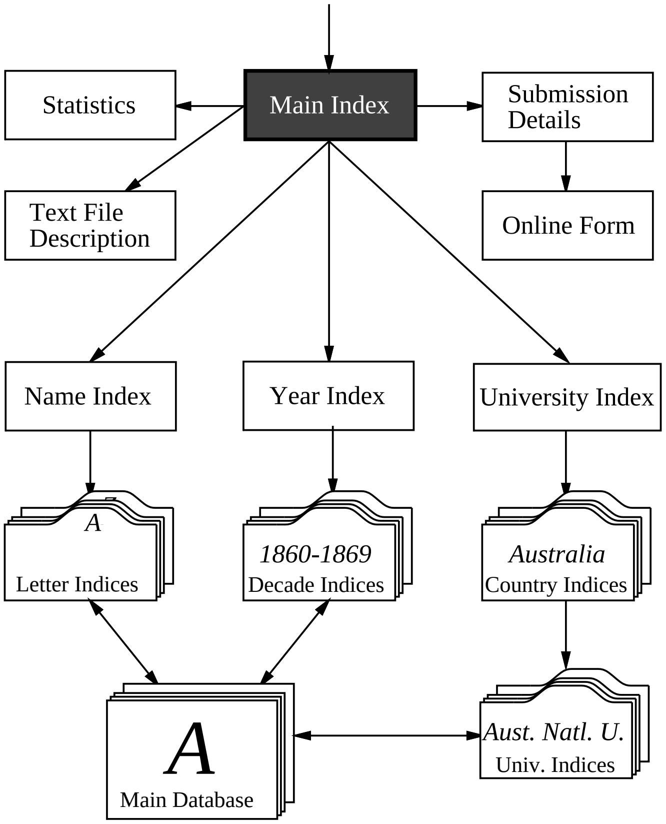
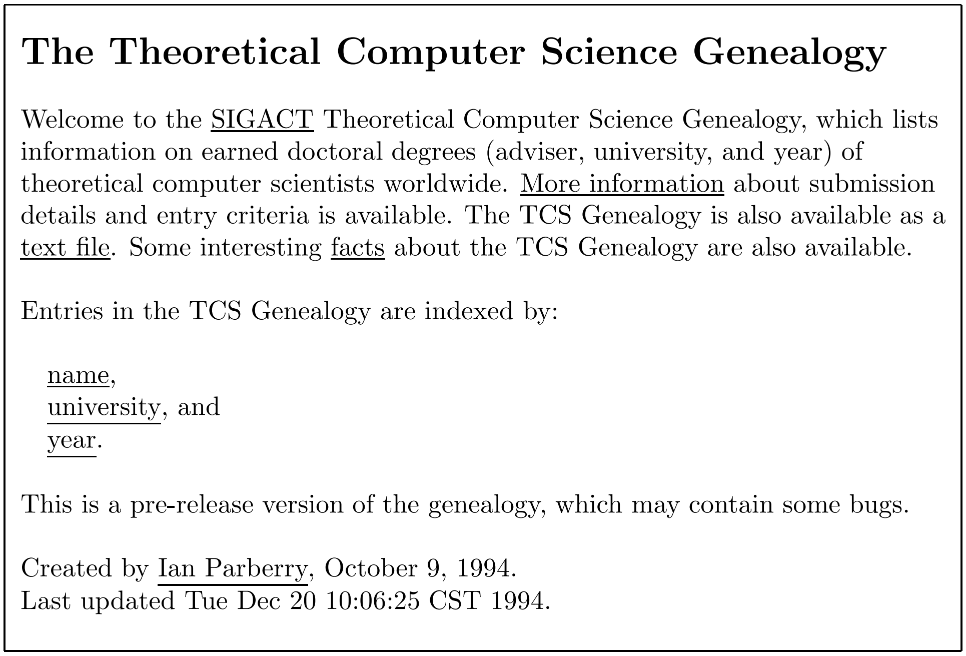


Figure 1: Flowchart showing main `html` files and the primary links.



The Theoretical Computer Science Genealogy

Welcome to the SIGACT Theoretical Computer Science Genealogy, which lists information on earned doctoral degrees (adviser, university, and year) of theoretical computer scientists worldwide. More information about submission details and entry criteria is available. The TCS Genealogy is also available as a text file. Some interesting facts about the TCS Genealogy are also available.

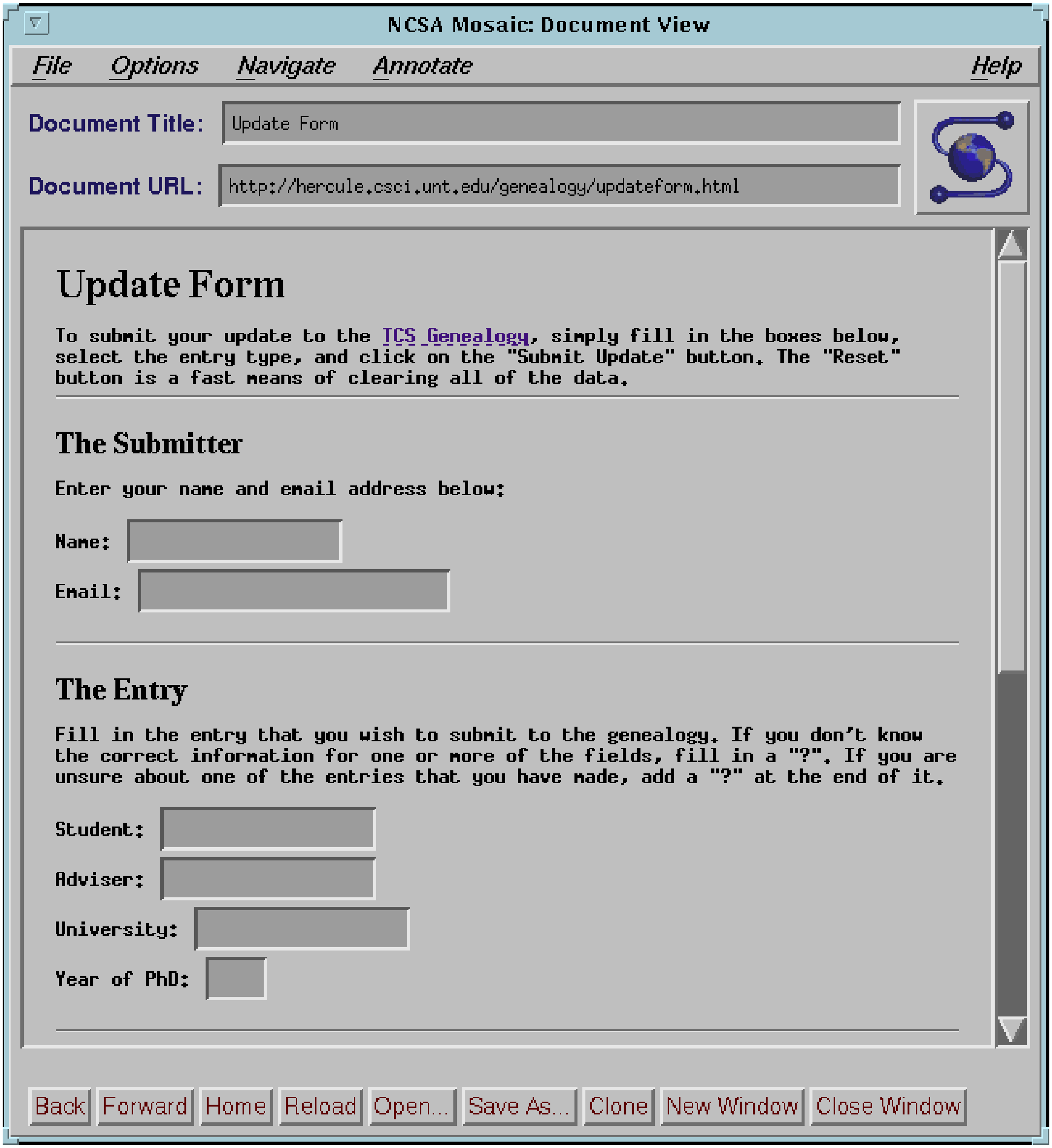
Entries in the TCS Genealogy are indexed by:

name,
university, and
year.

This is a pre-release version of the genealogy, which may contain some bugs.

Created by Ian Parberry, October 9, 1994.
Last updated Tue Dec 20 10:06:25 CST 1994.

Figure 2: The main index. Underlining indicates hypertext links.


NCSA Mosaic: Document View
File
Options
Navigate
Annotate
Help
Document Title:
Update Form
Document URL:
http://hercule.csci.unt.edu/genealogy/updateform.html

Update Form

To submit your update to the TCS Genealogy, simply fill in the boxes below, select the entry type, and click on the "Submit Update" button. The "Reset" button is a fast means of clearing all of the data.

The Submitter

Enter your name and email address below:

Name:

Email:

The Entry

Fill in the entry that you wish to submit to the genealogy. If you don't know the correct information for one or more of the fields, fill in a "?". If you are unsure about one of the entries that you have made, add a "?" at the end of it.

Student:

Adviser:

University:

Year of PhD:

Figure 3: Screen shot of the fill-out form using NCSA Mosaic for X windows.

1. an article published in refereed theoretical computer science journal,
2. a conference paper in a leading theoretical computer science conference,
3. regular attendance at a leading theoretical computer science conference,
4. being sufficiently famous that most readers will recognize one, or
5. an ancestor of an existing entry.

Except for people qualifying under (5), one must have officially received one's PhD before one can be entered into the database.

The submission details page also provides access to the online form.

2.3 The Online Form

The online form lets users submit entries to the genealogy using browsers that support fill-out forms. Figure 3 shows a screen shot of the top of the form using NCSA Mosaic. Before the World-Wide Web version of the genealogy was conceived, entries were submitted by sending email to `pedigree@hercule.csci.unt.edu`. For consistency, the online form automatically emails completed forms to the same address. Updates are not fully automatic, however. Each entry must be processed by hand to ensure consistency (for example, Richard Karp has been referred to in various updates as R. Karp, R. M. Karp, Richard M. Karp, and Dick Karp) and perform error-checking (for example, spelling, and checking that the fields were entered in the correct order).

2.4 The Text File Page

The text file page explains the format of the text version of the genealogy, and allows `ftp` access to the text files.

2.5 The Statistics Page

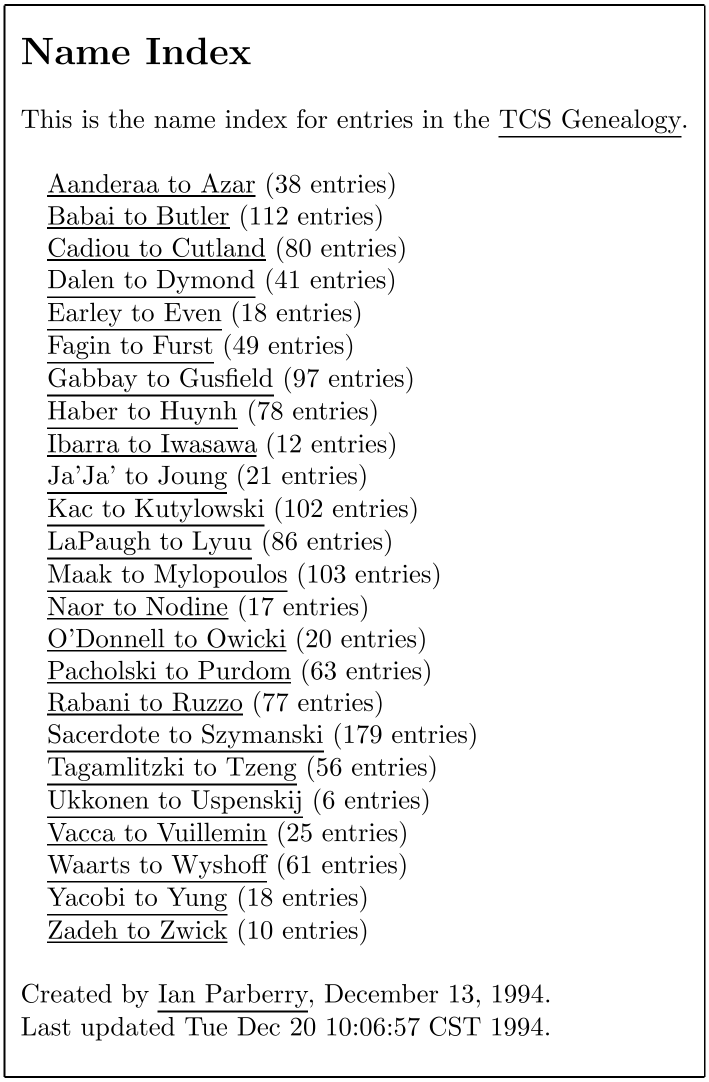
The statistics page lists a few very simple statistics about the genealogy that were gathered automatically.

2.6 The Name Index

The name index contains hypertext links to the letter index files (see Figure 4).

2.7 The Letter Indices

There is a letter index file for each letter of the alphabet. The letter index file for the letter "A", for example, contains a hypertext link to the main database entry for each person whose last name begins with the letter "A".



Name Index

This is the name index for entries in the TCS Genealogy.

Aanderaa to Azar (38 entries)
Babai to Butler (112 entries)
Cadiou to Cutland (80 entries)
Dalen to Dymond (41 entries)
Earley to Even (18 entries)
Fagin to Furst (49 entries)
Gabbay to Gusfield (97 entries)
Haber to Huynh (78 entries)
Ibarra to Iwasawa (12 entries)
Ja'Ja' to Joung (21 entries)
Kac to Kutylowski (102 entries)
LaPaugh to Lyuu (86 entries)
Maak to Mylopoulos (103 entries)
Naor to Nodine (17 entries)
O'Donnell to Owicki (20 entries)
Pacholski to Purdom (63 entries)
Rabani to Ruzzo (77 entries)
Sacerdote to Szymanski (179 entries)
Tagamlitzki to Tzeng (56 entries)
Ukkonen to Uspenskij (6 entries)
Vacca to Vuillemin (25 entries)
Waarts to Wyshoff (61 entries)
Yacobi to Yung (18 entries)
Zadeh to Zwick (10 entries)

Created by Ian Parberry, December 13, 1994.
Last updated Tue Dec 20 10:06:57 CST 1994.

Figure 4: The name index. Underlining indicates hypertext links.

2.8 The University Index

The university index allows access to the main database according to the the university that granted the doctoral degree. It contains hypertext links to the country indices.

2.9 The Country Indices

There is a country index for each country mentioned in the genealogy. Each country index contains hypertext links to the university indices for the universities in that country.

2.10 The University Indices

There is a university index for each university mentioned in the genealogy. Each university index gives the full name and geographic location of a university, and hypertext links to the main database entries of its doctoral graduates.

2.11 The Year Index

The year index allows access to the main database by year of graduation. It contains hypertext links to the decade indices.

2.12 The Decade Indices

There is a decade index for each decade mentioned in the genealogy. Each decade index has a section for each year in the corresponding decade. Each year section contains hypertext links to the main database entries of doctoral candidates who graduated in that year.

2.13 The Main Database

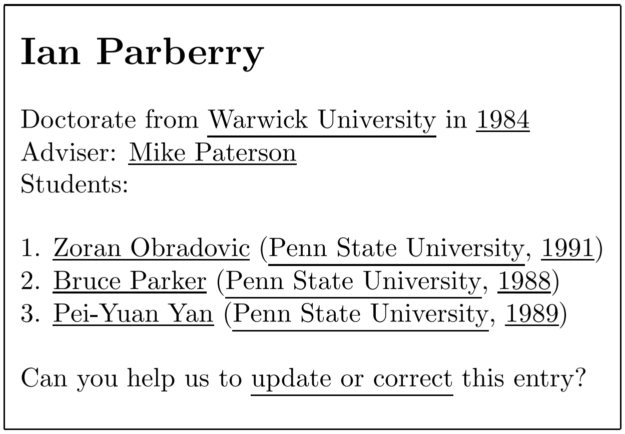
The main database consists of 26 `html` files, one for each letter of the alphabet. The database file for the letter “A”, for example, contains the entry for each person whose last name begins with the letter “A”. Each entry lists the person’s name, the university from which they received their doctorate, and the year in which the degree was granted, followed by a list of their doctoral students, and the universities and years in which their doctoral degrees were granted. Each of these pieces of information is a cross-reference to information in another part of the database.

For example, Figure 5 shows the entry for the first author of this paper. The first line lists his name. The second line states that he obtained his degree from Warwick University in 1984. The text “Warwick University” is a hypertext link to the index for Warwick University, and the text “1984” is a hypertext link to the index for the year 1984. The third line states that his adviser is Mike Paterson. The text “Mike Paterson” is a hypertext link for the main database entry for Mike Paterson (where the browser will see Ian Parberry listed as one of his students). The succeeding lines list Ian Parberry’s doctoral students, with hypertext links to their main database entries, and to the indices for the university and year of their respective doctoral degrees. The last line contains a link to the submission details page.

3 The Text Database

The text version of the database consists of two files, the database file, and the university file. Each is described below in a separate subsection. The text files are the canonical version of the genealogy. The hypertext version of the TCS Genealogy is created automatically from the text files by a Unix shell script (using `sed` and `grep`) written by the first author.

3.1 The Database File



Ian Parberry

Doctorate from Warwick University in 1984
Adviser: Mike Paterson
Students:

1. Zoran Obradovic (Penn State University, 1991)
2. Bruce Parker (Penn State University, 1988)
3. Pei-Yuan Yan (Penn State University, 1989)

Can you help us to update or correct this entry?

Figure 5: The main database entry for Ian Parberry. Underlining indicates hypertext links.

The database file contains the main database. It consists of a header, followed by the entries. Each line of the header begins with the character “#”. Each entry consists of four fields separated by a single tab character. The fields are, from left to right:

1. the student’s name,
2. the name of the student’s thesis adviser,
3. an acronym for the university granting the doctoral degree (see below), and
4. the year the degree was granted.

A student with multiple doctoral degrees has one entry for each. A student with multiple advisers for a single doctoral degree also has multiple entries (one for each adviser), but the university and year are the same.

A field consisting solely of the character “?” indicates that the information in that field is unknown. The “?” character is also used to indicate that the information provided in a field may be incorrect. An entry for a person without a doctoral degree (which is included when he or she has served as a thesis adviser on doctoral degrees) has the string “`---`” (three hyphens) in the adviser, university, and year fields.

3.2 The University File

The university file maps acronyms to universities. Each entry consists of an acronym, followed by the character “=”, followed by the name, city or town, state or province (if applicable), and country of a university (separated by commas).

4 Statistics

Since the database is maintained electronically, it is relatively easy to gather some simple statistics. The remainder of this section is divided into two subsections. The first contains statistics about the database files, and the second contains statistics about the TCS Genealogy itself.

The information reflects the state of the genealogy as of December 20, 1994.

Note that statistics from the TCS Genealogy do not necessarily reflect the whole of the theoretical computer science community. Much of the information in the original database was obtained by personal solicitation from the second author (in person or via email), and despite his intent to be as universal as possible, the information he obtained probably reflects at least a slight bias toward those areas (both geographic and technical) with which he was most familiar, as well as the school (MIT) that he attended. Subsequent entries are biased in different ways. So far they have for the most part been obtained as a result of general solicitations, rather than individual arm-twisting, and so people who do not normally read or respond to such solicitations have a higher probability of being absent. We hope to rectify this in the near future.

4.1 The Database Files

The genealogy consists of 240 html files, which are cross-referenced using a total of 11126 hypertext links (HREFs), and take up a total of 1.185 MB of file space.

4.2 The Data

The TCS genealogy contains entries for 1369 scientists with last names starting with 24 of the 26 letters of the alphabet (the exceptions are “Q” and “X” — we may be able eventually to get up to 26 letters, since there are three authors whose names begin with “X” and one whose name begins with “Q” in the STOC/FOCS bibliography [2]). The most frequent letter is “S”, with 179 entries. A frequency graph is shown in Figure 6.

The genealogy contains entries from 141 universities in 24 countries. Most entries are from the US (see Table 1). A total of 30 universities have at least 10 entries (see Table 2). As expected, MIT has more entries than any other university.

The number of entries in each decade grows rapidly from the 1940s through the 1970s (see Figure 7). The entries before the 1950s are mainly ancestors of theoretical computer scientists. A closer examination of the data since 1960 (see Figure 8) reveals that the number of entries per year has roughly leveled out since the early 1970s.

5 Remaining Work

A small amount of work remains to be completed before the WWW genealogy is ready for full public release. Some things are currently done incorrectly, including the following.

- There is no distinction between official advisers, unofficial advisers, and co-advisers.

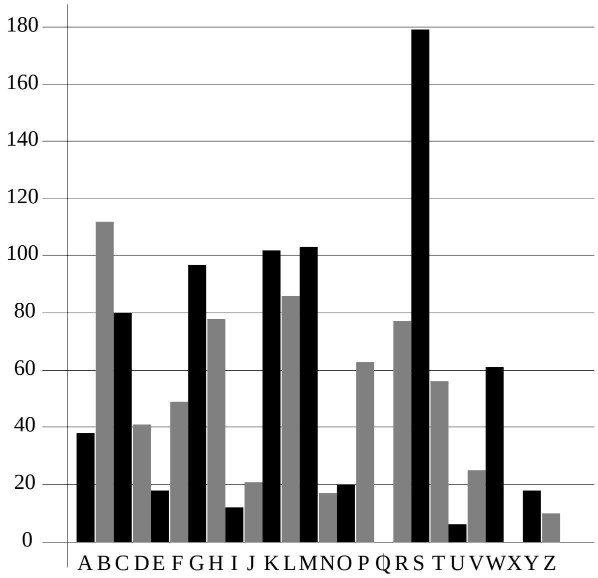


Figure 6: Number of names in the TCS Genealogy starting with each letter of the alphabet.

Country	Count
Australia	1
Austria	3
Belgium	1
Bulgaria	1
Canada	6
Denmark	1
England	6
Finland	3
France	3
Germany	18
Hungary	2
Israel	5
Italy	2
Japan	1
Norway	1
Poland	3
Prussia	1
Russia	5
Scotland	1
Spain	2
Sweden	2
Switzerland	2
The Netherlands	4
USA	67

Table 1: Number of universities mentioned in the TCS Genealogy by country.

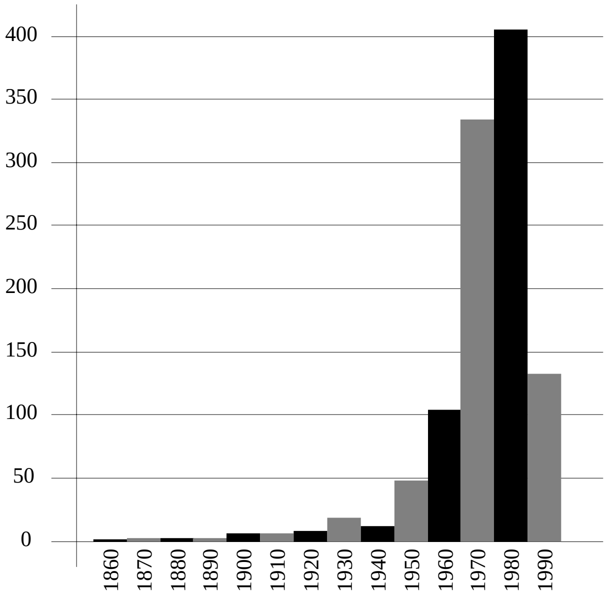


Figure 7: Number of entries in the TCS Genealogy graduating in each decade.

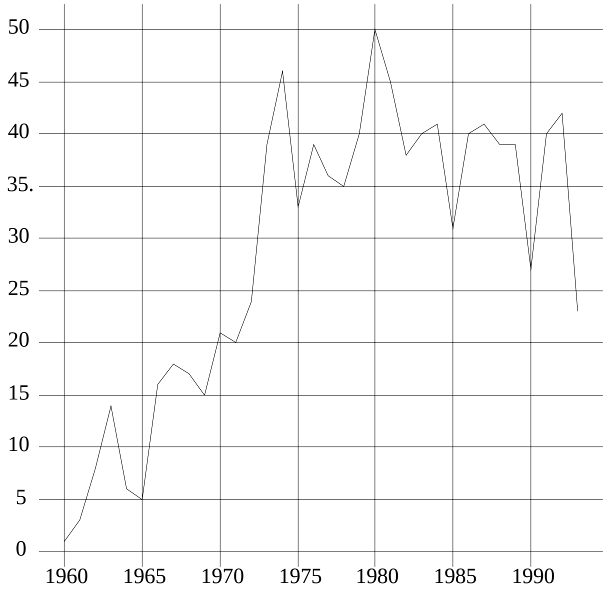


Figure 8: Number of entries in the TCS Genealogy graduating in each year from 1960.

University	Count
Columbia University	10
Edinburgh University	10
University of Maryland	10
UCLA	10
Brown University	11
Georg-August-Universitat Gottingen	11
University of Michigan	11
Warsaw University	11
Purdue University	12
University of Turku	12
University of Southern California	12
Yale University	12
Utrecht University	13
University of Chicago	15
University of Minnesota	16
Hebrew University	19
Weizmann Institute	20
University of Wisconsin	20
University of Waterloo	21
Penn State University	25
University of Toronto	25
University of Washington	25
University of Illinois at Urbana-Champaign	35
Carnegie Mellon University	36
Harvard University	55
Stanford University	68
Cornell University	69
Princeton University	70
University of California at Berkeley	77
MIT	94

Table 2: Number of entries from universities that have at least ten entries.

- Dual doctorates are not handled properly (the genealogy currently contains two dual doctorates, Andrew Yao and Leonid Levin).
- Accents in foreign names are omitted.
- Compound last names (such as Meyer auf der Heide, and van Emde Boas) are not alphabetized correctly.

Until then, a pre-release version is available [6]. Please feel free to browse it and report any errors, bugs, or updates to the first author.

Some additional features to be added at a later date include:

- Create one `html` file for each person, rather than one for each letter of the alphabet. This will decrease downloading time substantially.
- Add links to the home pages of people who have them. A list of such links is already available in the TCS Virtual Rolodex [5]. All that remains is to integrate them with the genealogy.

- Allow the inclusion of small pictures of each individual in the genealogy.
- The student-supervisor relationships form a DAG. Provide the ability to do online queries on the DAG, including properties such as connected components, paths, cycles, least common ancestors, and graph drawing.

The final version of this report, to be published in *SIGACT News* (see [4]), will include more information on the database, including issues that were covered in the original report [1] such as directed and undirected cycles, and connected components. This information will be computed automatically from the main database. We also plan to develop methods for drawing "family trees" in postscript format. Finally, as mentioned in Section 4, the authors plan to start soliciting genealogical information from individual members in the theoretical computer science community, starting with names mentioned in the STOC/FOCS bibliography [2], and attendee lists from recent theory conferences.

References

[1] D. S. Johnson. The genealogy of theoretical computer science. *SIGACT News*, 16(2):36–44, 1984. Reprinted in *Bulletin of the EATCS*, (25):198–211, 1985.

[2] D. S. Johnson (Editor). STOC/FOCS Bibliography (Preliminary Version). ACM Press, 1991.

[3] I. Parberry. ACM SIGACT. A WWW document with URL `http://sigact.acm.org/sigact`, 1994.

[4] I. Parberry. SIGACT News. A WWW document with URL `http://sigact.acm.org/sigactnews`, 1994.

[5] I. Parberry. The Theoretical Computer Science Virtual Rolodex. A WWW document with URL `http://sigact.acm.org/tcs-rolodex`, 1995.

[6] I. Parberry. The Theoretical Computer Science Genealogy. A WWW document with URL `http://sigact.acm.org/genealogy`, 1994.