vispubdata.org: A Metadata Collection about IEEE Visualization (VIS) Publications

Petra Isenberg, Florian Heimerl, Steffen Koch, Tobias Isenberg, Panpan Xu, Charles D Stolper, Michael Sedlmair, Jian Chen, Torsten Möller, John Stasko

vispubdata.org: A Metadata Collection about IEEE Visualization (VIS) Publications

Petra Isenberg, Florian Heimerl, Steffen Koch, Tobias Isenberg, Panpan Xu, Charles D. Stolper, Michael Sedlmair, Jian Chen, Torsten Möller, John Stasko

Abstract—We have created and made available to all a dataset with information about every paper that has appeared at the IEEE Visualization (VIS) set of conferences: InfoVis, SciVis, VAST, and Vis. The information about each paper includes its title, abstract, authors, and citations to other papers in the conference series, among many other attributes. This article describes the motivation for creating the dataset, as well as our process of coalescing and cleaning the data, and a set of three visualizations we created to facilitate exploration of the data. This data is meant to be useful to the broad data visualization community to help understand the evolution of the field and as an example document collection for text data visualization research.

Index Terms—Visualization, publication data, citation data.

1 Introduction and Motivation

Over the past decades, the IEEE Visualization (VIS) conference series and its constituent conferences (Vis, InfoVis, VAST, and SciVis) have become the prime venues to present scientific work on visualization techniques, systems, and experiences. The 2014 meeting in Paris marked the 25th anniversary of the event and provided a reason to reflect on its history (e. g., [16], [20]). For the development of exhibits and websites commemorating the anniversary and to support other ongoing work, several researchers collected datasets of the publication history of the member conferences. This article reports on the acquisition, cleaning, curation, and visualization of a dataset [7] that provides metadata on all the papers that have appeared in the conferences.

Datasets such as the one described in this article are highly relevant to the work of researchers who want to carry out a variety of tasks for which an understanding of a research field's history and current trends is essential. An example task frequently undertaken is to understand the state of the art for one's own publications. Similarly, researchers serving in an editorial role often need to find appropriate reviewers for grant proposals and paper submissions. One could also envision using such data to plan a conference by trying to predict visitors' interest in specific topics and supporting them with personalized schedules according to their interests and previous publications. We have no doubt that there exist many more applications for which this dataset can be useful.

- *Petra Isenberg is with Inria, France. E-mail: petra.isenberg@inria.fr.*
- *Florian Heimerl is with Univ. Stuttgart, Germany. E-mail: florian.heimerl@vis.uni-stuttgart.de.*
- *Steffen Koch is with Univ. Stuttgart, Germany. E-mail: steffen.koch@vis.uni-stuttgart.de.*
- *Tobias Isenberg is with Inria, France. E-mail: tobias.isenberg@inria.fr.*
- *Panpan Xu is with Hong Kong University of Science and Technology, China. E-mail: pxu@cse.ust.hk.*
- *Charles Stolper is with Georgia Tech, USA. E-mail: chadstolper@gatech.edu.*
- *Michael Sedlmair is with Univ. Vienna, Austria. E-mail: michael.sedlmair@univie.ac.at.*
- *Jian Chen is with University of Maryland, Baltimore County, USA. E-mail: jichen@umbc.edu.*
- *Torsten Möller is with Univ. Vienna, Austria. E-mail: torsten.moeller@univie.ac.at.*
- *John Stasko is with Georgia Tech, USA. E-mail: stasko@cc.gatech.edu.*

Manuscript received October 5, 2016.

Besides these introspective tasks, the dataset can provide value for evaluating some of the visualization community's own approaches. For instance, the use of visualization to support retrieval, exploration, and analysis of textual data is an increasingly important topic in visualization research [11]. Typically, the field of visualization provides support for other domains to better understand their data, rather than generate data of its own. Yet, visualization publications are one of the rare exceptions where 'data' is produced by the visualization community itself. Due to the intricate knowledge visualization researchers have of their own community, the dataset is valuable for use in assessing text visualization approaches without relying on the help of external domain experts. The creation of datasets for testing known as well as developing new methods and approaches has some tradition in the field of Visualization.

In the context of numerous challenges, artificial and real-world datasets have been created and many of them were and are still used beyond the tasks and scopes defined by the corresponding contests [2], [15]. In particular the InfoVis Challenge 2004 [4] already made the metadata of a subset of information visualization papers available. This dataset, while now outdated, has in the past sparked and supported many research efforts.[1] Another earlier dataset of IEEE Visualization papers was created by Voegele [18].

The dataset we introduce is certainly not large in terms of "Big Data," (Fig. 1 shows a graph of the three conferences' publication counts) but its complexity with respect to the concepts described within articles as well as its heterogeneity regarding the provided bibliographic information and citation relations, makes it an interesting test base for different visualization approaches.

While portions of our metadata collection are available from other websites [10], [20] and have been used in prior publications [5], [8], [9], [17], we came together to provide cleaned and complete data for all IEEE VIS subconferences and to make the data publicly available. Our goal is to foster research on our field, its specific topics, authors, history, and development. As a point of reference, this article describes the dataset itself, our process of data acquisition and preparation (e. g., data cleaning), and briefly mentions some existing visualization systems that use our data.

1. https://tinyurl.com/infovis2004contest

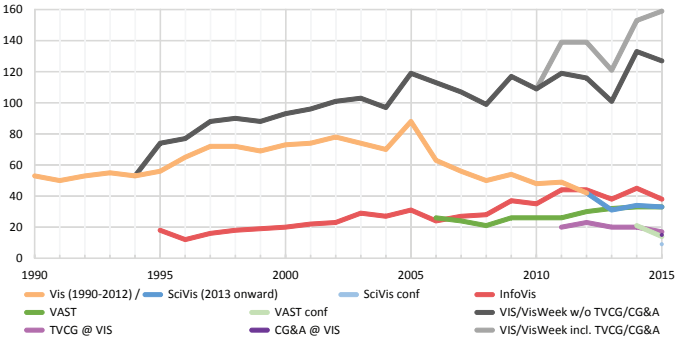

Fig. 1. Paper counts per year and publication category/conference. *Image is in the public domain.*

While the dataset provides a large cross-section of data visualization research in general, a significant amount of visualization-focused research has also been published in conference and journal venues outside of IEEE VIS. The dataset thus does not represent a "complete" view of the field. This narrower approach of focusing on one particular venue, however, allows us to be more precise about the metadata included in the collection.

2 Dataset Description

Our dataset includes metadata on papers that appeared at the IEEE VIS conference series from 1990–2015. In this section, we introduce the data format and the information stored about each paper. We then discuss peculiarities of the dataset to better illustrate the choices we made during data collection.

2.1 Collected Data

We chose a simple format for the dataset in order to fit it into one table or spreadsheet. Each row in the table corresponds to one IEEE VIS publication, i. e. InfoVis, SciVis, VAST, or Vis. A publication is characterized by the following attributes/columns:

- **A**: The **conference** in which the paper appeared: InfoVis, SciVis, VAST, or Vis. See remarks on the name changes of IEEE VIS conference series in the following section.
- **B**: The **year** in which a paper was presented at the conference. This should not be confused with the year of publication, as, for example, the papers presented at VIS'15 will be published in January 2016 but are listed here under 2015.
- **C**: The **title** of the paper as it appeared on the paper.
- **D**: The paper's **DOI** pointing to a digital library entry. Two papers were not featured in a digital library (see Sect. 3.5) and received a fake but syntactically valid DOI string starting with 10.0000/ and a unique 8 character suffix, making entries in column D a unique identifier for each paper.
- **E**: A **link to the paper** in the IEEE digital library—based on the DOI. If no DOI was present in IEEE Xplore, a link may be present pointing to the paper in other document repositories, such as the ACM or IEEE Computer Society digital library.
- **F**: The **number of the paper's first page** in the printed proceedings or the journal special issue.
- **G**: The **number of the paper's last page** in the printed proceedings or the journal special issue.
- **H**: This field contains an **X** if an entry in the dataset is either a **capstone, keynote, panel, or poster** and thus is not considered a peer-reviewed full paper.
- **I**: The **paper type**: one of C (conference paper), J (journal paper; for details see Sect. 2.2), M (miscellaneous: capstone, keynote, panel, or poster).
- **J**: The **abstract** of the paper.
- **K**: The **author names** (typically *lastname, initial*; however, forms such as "*firstinitial. lastname*" also exist), separated by a semicolon. Authors are ordered as they appear on the paper.
- **L**: The organizational **affiliation of the first author**.
- **M**: The **author IDs** according to the IEEE Xplore library, if we could find them.
- **N**: A copy of field K but with **author names** cleaned and **duplicates removed**, unifying different listings/writings of the same author's name to one unique string (see Sect. 3.4).
- **O**: A **list of references** this article makes to other IEEE VIS papers, using the unique identifiers of Column D. We do not include citations to papers outside of the conference series.
- **P**: A list of **author keywords** supplied on the paper PDF.

In total, we collected data on 2 752 publication items from 4 890 unique (sanitized) authors. We specifically did not capture data such as external citation counts and Google Scholar IDs because no automated way to reliably acquire that information exists. In addition, such data is constantly changing and we would thus only be able to provide a temporary snapshot.

2.2 Data Peculiarities

To explain some of our choices during data collection as well as differences of our dataset to the printed proceedings, it is important to understand the historic evolution of the IEEE VIS conference series as represented in Fig. 2. The IEEE Conference on Visualization started in 1990 under the name IEEE Visualization (Vis). It quickly grew and in 1995 the IEEE Symposium on Information Visualization (InfoVis) was held for the first time. The symposium was renamed to the IEEE Information Visualization Conference in 2007 but kept the acronym InfoVis.[2] In 2006, the Symposium on Visual Analytics Science and Technology (VAST) joined and was renamed to VAST Conference in 2010. The original IEEE Visualization Conference changed its name in 2013 to the IEEE Scientific Visualization Conference using the acronym SciVis.

For a number of years (2008–2012) the three conferences ran together as parallel tracks of a large joint meeting using the umbrella term VisWeek. Then, in 2013, the name VisWeek was dropped in favor of VIS as the common acronym for all three conferences. We chose to only use the acronyms InfoVis, SciVis, VAST, and Vis in the dataset—within this article we use the abbreviation VIS to refer to all conference tracks from 1990 to 2015.

Each conference track began by publishing its full papers in regular conference proceedings and later (Vis and InfoVis in 2006, VAST in 2011) switched to publishing all (Vis, InfoVis) or a subset of papers (VAST) as journal articles in *IEEE Transactions on Visualization and Computer Graphics* (TVCG). In addition, starting in 2014, the VAST conference featured two types of papers: regular papers published in the conference proceedings and top rated submissions that appear as articles in the TVCG journal. In 2015, SciVis also included papers that were published as either conference or TVCG papers. We chose to mark the conference-only papers in the dataset in field I, knowing the distinction between conference and journal papers can be critical in determining, e. g., paper acceptance rates.

2. InfoVis is not to be confused with the International Conference on Information Visualisation. The latter uses the acronym IV.

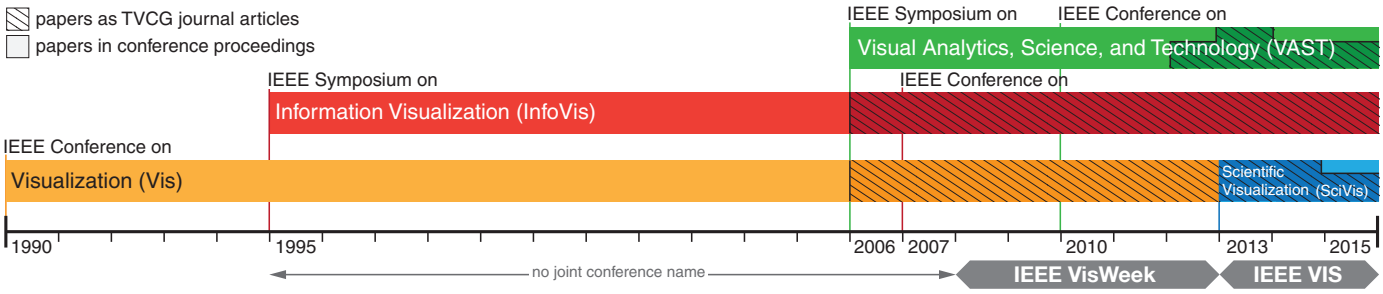

Fig. 2. History of conference/symposia names as well as publication types for IEEE Vis, SciVis, InfoVis, and VAST. *Image is in the public domain.*

3 Data Acquisition and Preparation

We used a combination of several data preparation and matching techniques to automatically extract as much data about the publications as possible. In the process, we introduced multiple checks and methods to assure high data quality. Finally, we carried out an extensive manual review of the automatically extracted data.

3.1 Data Collection Methodology

We started by collecting all electronic proceedings since 1990. For most of the material we had direct access to conference CDs, DVDs, and memory sticks from the respective years through our research institutions. Missing documents, partially caused by mismatches between the official proceedings and the available PDFs, were acquired by download.[3] Once we had the complete collection, we processed each PDF by applying *pdftotext*, which is part of the *Poppler PDF library*.[4] The result is a plain text file containing the textual content of the PDF file, including metadata such as title, authors, and abstract as well as the reference list. Unfortunately, as text extraction from PDFs is a challenging problem, we found that *pdftotext* was not able to process all PDFs adequately. The two column format, in particular, seemed to pose a problem for text extraction, often resulting in missing material, including entire reference lists in the resulting text files. To maintain a high quality of the final dataset, we decided to additionally apply optical character recognition (OCR) software to the source PDFs, similar to the methodology used by Heimerl et al. [6]. We used Nuance's Omnipage[5] version 18 for this purpose, which can extract text plus information about the layout directly from PDFs. We stored the output for each source PDF as plain text and as XML. The latter contains layout information in addition to the text. We thus had four versions of each publication: the original PDF, two plain text versions, and one XML version.

As a second step, we applied ParsCit [3] to the two text and one XML versions of the extracted electronic proceedings. ParsCit accepts plain text files as well as the Omnipage XML format as input. It is a system based on machine learning that parses the input documents, automatically recognizes and splits different sections of a document, and outputs its results as an XML file. This includes all metadata of a publication contained in the source file such as authors, titles, and abstracts. In addition, it is able to detect and extract the reference list of each document, including the bibliographic data of the references: i. e., titles, authors, publication venues, and others.

For creating the dataset we were especially interested in the papers' titles, authors, abstracts, and citation links to other IEEE VIS publications. We found that ParsCit is most reliable when extracting titles of citations compared to other citation metadata. For this reason, we took the two text and the XML version of each paper and generated a list of titles for each reference in the paper. Then, we selected the list with the highest number of entries to achieve a high recall and discarded the results of the text extraction methods that were not able to extract all references. To obtain all citations within IEEE VIS, we matched all extracted titles to the titles of publications in our dataset and stored the corresponding paper IDs to be included as citation links.

3.2 Data Cleaning

While the creation of this dataset would have been very difficult without automatic extraction, the techniques we used came at the cost of reduced accuracy. Text extraction from PDFs, directly as well as by OCR, is prone to errors. Possible errors include wrong characters, tables or figures erroneously extracted as text, spurious white space, and omissions of entire sections. On top of this, ParsCit tends to introduce additional inaccuracies. In addition to problems during pre-processing and extraction, inconsistencies introduced by abbreviations, orthographic variations, typos, etc. had to be taken care of. To create a coherent dataset, unique data attributes such as author names and paper titles had to be generated. Consequently, additional processing and manual effort was necessary at various steps to cope with these problems. We reviewed the three different results for each publication, and adjusted them manually by selecting the most correct version and revising it. In cases where there was no output at all or the existing output was too faulty, we manually extracted the metadata.

To ensure the completeness of the publication dataset, we created a second list of titles from the proceedings' tables of content and checked it against those of the papers from the electronic proceedings as described in Sect. 3.3 below. This not only helped us to identify incorrectly extracted metadata, but also showed that some publications were accidentally missing in the proceedings. We also revised extracted titles from the reference lists by applying the same process as described above for document metadata. In the future we consider to use GROBID[6] for metadata extraction and to compare its accuracy to our current approach.

3.3 Extraction of Internal Citation Links

We chose to include only citations within the dataset, i. e., only references to papers published at one of the VIS conferences. One

3. Note that some discrepancies exist in paper count to other summaries of IEEE VIS papers; e. g., those published at https://github.com/steveharoz/Vis-Acceptance-Rates—we used the printed proceedings to verify our dataset and consistently included a specific paper type called case studies, which in the early years of the Vis conference was a separate submission category.

4. https://poppler.freedesktop.org/

5. http://www.nuance.com/for-individuals/by-product/omnipage/

6. https://github.com/kermitt2/grobid

reason for limiting the citation relation to this subset is that citation data is inherently noisy due to errors and omissions being made. Since we did not have access to a comprehensive dataset of cleaned publication references outside of the IEEE VIS conferences, we did not have a dataset to match these references against and produce reliable results. The second reason is that our goal was to create a coherent dataset that captures the scientific development of the IEEE VIS conferences. For this goal, citation links within the same venue are an important indicator. Yet, while internal citations can be a valuable measure of scientific development, they are clearly limited. Unfortunately without the investment of a massive amount of manual labor to clean other citation links by hand, external links can probably not be reliably included in the dataset at this point.

To match the citations within the dataset, we compared all titles from the longest, automatically extracted and manually checked and revised reference list of each publication to all titles within the dataset. For comparing title strings, we first transformed both strings x and y to lower case. Then, we applied a similarity measure based on Levenshtein's edit distance [12]:

$$sim_{Levenshtein}(x,y) = 1 - \frac{dist_{Levenshtein}(x,y)}{max(|x|,|y|)}$$

Levenshtein quantifies the distance between two strings by expressing it as the number of character insert, delete, and exchange operations necessary to convert one into the other. In order to transform this distance into a similarity in the range between 0 and 1, we divided the distance by the maximum length of both of the input strings, which constitutes its possible maximum, and subtracted the result from 1. For each reference title, we considered the most highly ranked title in the dataset above a threshold of 0.6 as the candidate for a match. We experimentally determined the threshold 0.6 to reduce the effort for manual revision, while still maintaining a high recall. Matching titles this way helped us to deal with typos, orthographic variations, variations in punctuation, and general noise in the data.

After we automatically identified all reference candidates in the data, we manually reviewed them to remove the numerous false positives that this procedure generated. The quite large number of false positives was mainly due to very similar titles, sometimes only differing by one word. An example for such cases are subsequent publications of the same authors that constitute, e. g., the evolution or the application of a previously developed technique. With the manual revision of the automatically detected candidates, we were able to create a mapping with high recall as well as high precision. However, a risk remained that ParsCit was not able to capture all listed references for a publication. To address this concern, we inspected some samples and did not encounter any omissions, but we cannot guarantee the completeness of the citation links.

3.4 Name Resolution

One challenge in making the dataset useful to all is the preponderance of authors' names being listed in multiple ways. For instance, a hypothetical author Mary Jane Smith may have her name recorded on different papers as "Smith, Mary", "Smith, M.J.", "Smith, M.", or some other variant. To accurately identify an individual consistently, we wanted to resolve these differences. This process is known as *de-duplication*—a known challenge in scientometrics [1]. Ideally, a bibliographic system would employ unique author IDs to avoid this problem, but such IDs did not exist in our case.[7]

To help with de-duplication on the visualization publication dataset, we used the Jigsaw visual analytics system [5].[8] We first transformed the CSV-stored dataset into Jigsaw's datafile format (XML-based) via a simple text translation program. In doing so, we created an AUTHOR entity type for each author on one of the papers. We then read the data into Jigsaw, opened its List View, and viewed the AUTHOR entity list. It shows each unique author name, sorted alphabetically by default.

Next, we exported the list to a text file for further text processing. We then looked for author names beginning with the same first string token (last name usually) and same first letter of the second string token (first name usually), e.g., "Smith, Mary" and "Smith, M.". We filtered out all names without such duplication, and thus kept candidate names to be unified. We then manually examined this list of unification candidates to decide which needed fixing.

Unfortunately, four other common problematic cases could occur and would not be found by that process. First, an author's last name might appear differently. Typically, this occurred due to special characters such as accents or umlauts being used in specific cases. Second, authors may sometimes use a nickname different than their given first name on papers, thus the first letter of the second string token would not match. One example is the frequent IEEE VIS author Eduard Gröller whose students name him the "Meister" and include his name as M. Gröller or M. E. Gröller but whose last name is also often spelled Groller or even Groeller. Third, our dataset included most Asian names not always in "lastname, firstname" format, but sometimes also in "firstname lastname" format. Finally, some authors have changed their names at some point in time, for example due to a marriage.

All four of these cases as well as simple name logging and transcription errors required manual resolution. We thus viewed the names in Jigsaw's List View which simplified finding such problems. Nonetheless, some errors were undetected and we have relied and continue to rely on the community to report any problems.

3.5 Comparison to Other Datasets

To check the data for errors we also compared our papers to the table of contents of the respective proceedings stored in the IEEE Xplore digital library. If a paper was found we looked up its IEEE Xplore article number as well as DOI and saved it with our data. If a paper was not found, we searched for a similar title and marked the two papers as potential matches. In the same way we also searched in the other direction for paper titles from the digital library in our dataset. We manually inspected all papers that were not found and the suggested potential matches to find errors in either our extraction or the digital library. Where data was missing or unclear, we looked up papers again in the printed proceedings if available or made a triage of several sources such as an author version, digital library version, or the TOC from the proceedings.

During this process, we found many errors in the IEEE Xplore library and the printed proceedings, making our dataset a more reliable source as an IEEE VIS publication history. At the time of writing, 47 entries from the VIS proceedings (i. e., 1.7%) are entirely missing from the IEEE Xplore digital library—most of these (45), however, are available in the IEEE Computer Society digital library. Two additional entries exist in the IEEE digital library but the DOI provided for them leads to an error message by IEEE Xplore. For many of the earlier years, the entire front matter or much of

7. The IEEE Xplore digital library includes some author IDs but these are not consistent across the years and do not exist for all papers.

8. http://www.cc.gatech.edu/gvu/ii/jigsaw/

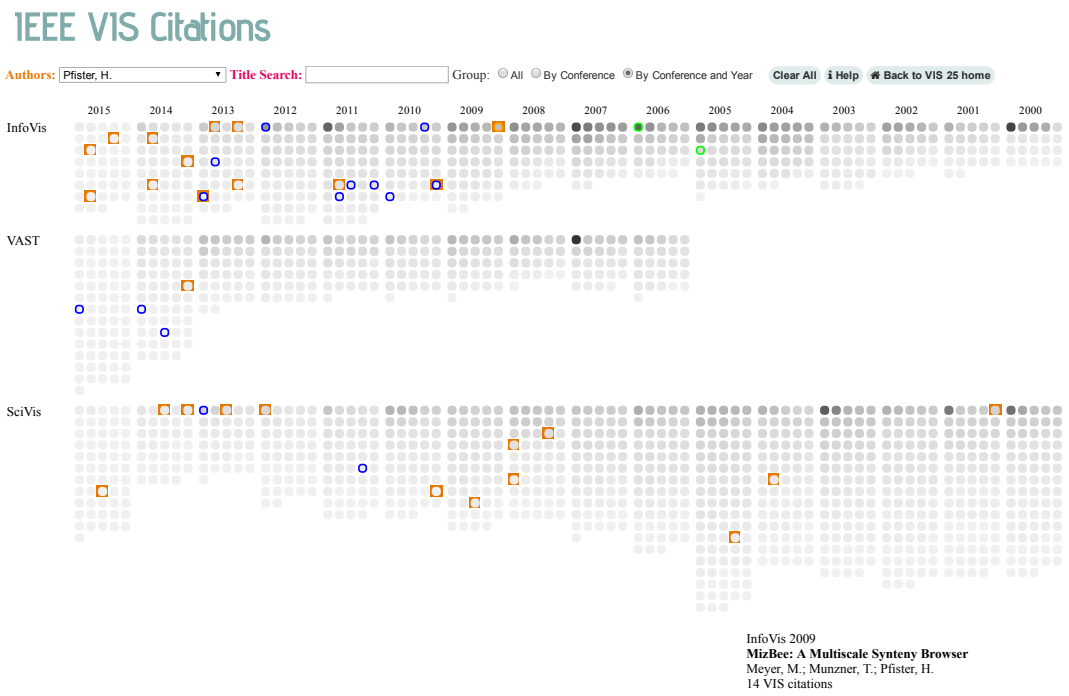

Fig. 3. Screenshot of CiteVis2, showing the author Hanspeter Pfister selected with his papers depicted in orange. The papers citing his 2009 InfoVis paper are shown in blue and the papers cited by the 2009 paper are in green.

it is also missing from IEEE Xplore, along with "official" submission and acceptance numbers as provided by other digital libraries such as ACM's. We also found entries in IEEE Xplore that link to the wrong paper PDFs (DOIs: 10.1109/INFVIS.1997.636794, 10.1109/TVCG.2007.70587) and papers that were missing from the official CDs (DOI: 10.1109/INFVIS.1999.801869). In the printed proceedings of Vis 1998 we even found a whole session (pages 463–481) to be missing from the table of contents. Finally, we found several instances in which data in the IEEE Xplore entries were incorrect or incomplete (e. g., titles, page numbers, etc.).

Our dataset is also similar to the one provided for the 2004 InfoVis contest [4] in terms of the data fields it contains. Yet, the earlier dataset is much smaller and had a different capture process. In it, the researchers worked with the ACM Digital Library and they included all of the references on each paper, even those outside the IEEE InfoVis Conference. The two key components included in both datasets are the list of references made in each paper (our set is more narrow) and the inclusion of unique (de-duped) authors. However, many references were not actually resolved to specific paper IDs in the dataset. As pointed out above, we also engaged in a manual cleaning pass making sure that no papers were missed, that titles are correct, and the links to the original pdfs work and point to the right place. Similarly, the older dataset did not appear to have the level of unique author identification that ours has.

4 Visualization and Exploration Tools

The visualization publication dataset contains a rich trove of information about papers at the three conferences over the years. We now show examples of tools we built and to explore the dataset. Other tools that have already been built using the dataset are listed on our website vispubdata.org. In addition, several visualization courses (e. g., at Jacobs University Bremen, Wayne State University, Université de Fribourg) have used the dataset in teaching.

As an examples exploration tools for the dataset, we created three visualizations (CiteVis2, CiteMatrix, and VISLists) to present different aspects of the public dataset [20]. They focus, respectively, on citations between papers including citation counts, aggregate citations across conferences and years, and individual author's contributions across conferences and years. The tools presented here are not the main contribution of this article and therefore their descriptions are purposefully short—the interested reader can try the tools at http://www.cc.gatech.edu/gvu/ii/citevis/VIS25/ .

4.1 CiteVis2

The publication dataset including citations provides the potential to communicate a great deal of information about the VIS conferences. A natural set of questions arises concerning conference papers that are highly cited and presumably impactful in the visualization community. Furthermore, the data make it possible to explore individual author's papers and citations. Which papers from different conferences and years are the most highly cited? For a particular paper, which papers does it cite and which cite it? In what years and conferences has a particular person authored papers, and how well cited are that person's papers? To answer these questions and more, we developed the CiteVis2 visualization (Fig. 3).

CiteVis2 draws inspiration from the original CiteVis system [17] that shows InfoVis Conference papers year-by-year and their citation counts and links. CiteVis2, like the original, represents each paper as a small gray circle. The circle's darkness indicates the total number of citations it has received. When the viewer moves the mouse over one of the circles, the visualization lists the paper's details (conference, year, title, authors, and citation count) at the bottom. It also highlights the papers citing that paper in blue and the papers it cites in green. Individual authors can be selected at the top and their papers will be highlighted with an orange box.

Because the dataset includes the other three IEEE VIS conferences' papers and is much larger that that of the original CiteVis, it required a slighly different layout. We chose to show the papers in a grid, ordered by the number of citations each paper has received.

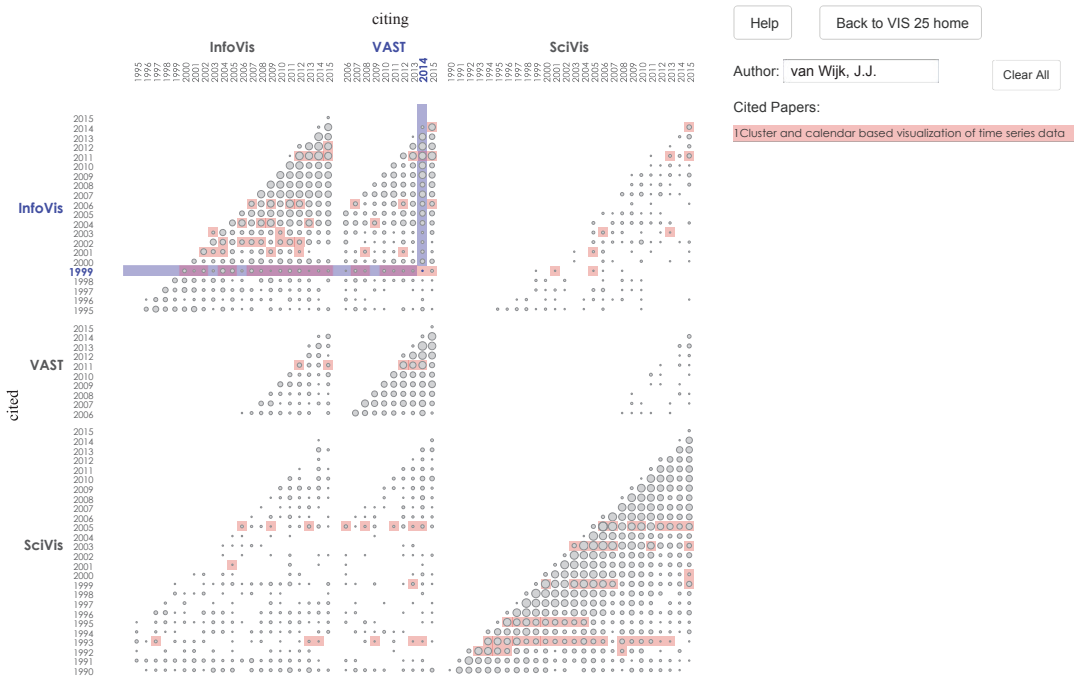

Fig. 4. Screenshot of CiteMatrix, with a blue highlight showing how many VAST 2014 papers (column) cite InfoVis 1999 papers (row). Also shown in pink are citations to papers from Jarke van Wijk.

The grid is organized depending on the mode selected at the top: by conference and year, by conference, or as one collection.

In Fig. 3, the author Hanspeter Pfister is selected and his papers are highlighted in orange. He is one of a few researchers to have published papers in all three conferences. The mouse is over his 2009 InfoVis paper about the MizBee system that has similarly received citations from subsequent papers in all three conferences (indicated by papers colored in blue).

4.2 CiteMatrix

The analysis of the citation links can reveal insights beyond individual papers and authors [14], [19]. It can extend to the level of connections between different subareas of visualization research, hinting at the influence of one area on another. Do papers in different conferences frequently cite each other? Will the number of citations to earlier work decline over time and if so, how? To answer these questions and others, we developed CiteMatrix (Fig. 4) that shows an overview of the aggregate citations among VIS conference papers (InfoVis, SciVis, VAST, and Vis) for all years of the conference.

CiteMatrix aggregates citation links by year and paper venues: columns show venue/year of papers that make a citation and rows show venue/year of papers that are being cited. The circle sizes in the matrix are logarithmically proportional to the total number of citations from papers in the venue/year column to papers in the venue/year row. For instance, Fig. 4 shows CiteMatrix depicting (blue highlight) citations from VAST 2014 papers to InfoVis 1999 papers. Only one citation occurred, to the paper "Cluster and calendar based visualisation of time series data," indicated to the right.

Glancing at the visualization, one can notice denser citation links within each conference itself, confirming a likely intuition that the papers of each conference mostly cite earlier papers of that same conference. In addition, one notices that VAST papers frequently cite InfoVis papers and the reciprocal pattern is true as well, but to a lesser degree. SciVis papers seldom cite VAST papers, however, as can be seen in the middle right. On the whole, papers tend to cite relatively recent papers. This pattern is most evident in SciVis where the circle sizes fade off along the diagonal. In InfoVis, papers from about 2004 onward are strongly cited.

Through interaction, CiteMatrix also facilitates the exploration of detailed information besides the aggregated citation links. When a user hovers over or clicks on one of the circles, the papers being cited and the number of received citations are displayed to the right. Viewers can use this feature to discover papers being cited long after they have been published, or the specific papers from one of the conferences that are frequently cited by those in another. Viewers can also search for individual authors—CiteMatrix then highlights the circles of venue/year combinations when that person authored a paper. In Fig. 4, the circles highlighted in pink indicate papers by Jarke van Wijk. Observe how a paper (or papers) he published in SciVis 1993 have been cited in almost every subsequent year of that conference—evident by the long pink stripe in the lower right.

4.3 VISLists

Because the publication dataset includes author information on each paper, it can be used to examine the contributions of different researchers across the various years of all three conferences. A variety of questions then can be posed and answered: Which researchers have published the most papers within and across the conferences? Who has been the most prolific recently? Who are the most frequent co-authors of a particular person? To answer these questions and more, we created the VISLists visualization (Fig. 5).

VISLists, adapted from the List View of the Jigsaw system [5], shows three lists of items: authors, conferences, and years. Each item has a small "Frequency count" bar in front of it that represents a count of how many papers match that item. One can quickly see which conferences and years have contained more papers. The default ordering of items is alphabetical/numerical, but a button above the list allows sorting by frequency.

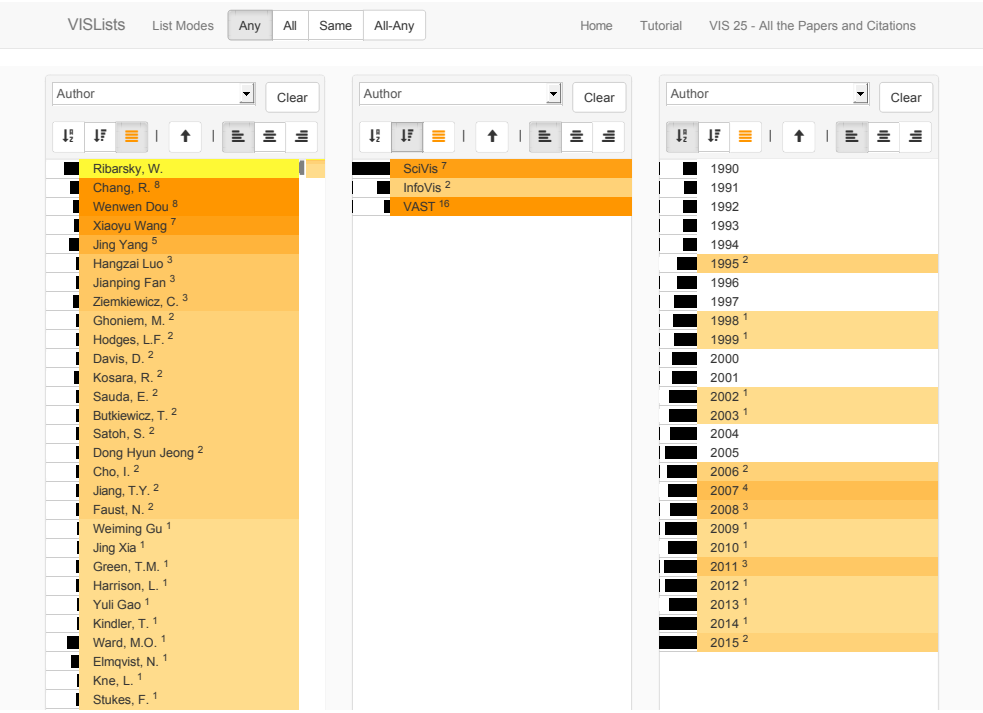

Fig. 5. Screenshot of VISLists, showing lists of authors, conferences, and years, with the author William Ribarsky selected.

Clicking on an item selects it so its label is drawn with a yellow background. Any other items in the three lists "connected" to that item through some paper in the collection then are drawn using an orange background. For instance, selecting an author highlights (colors orange) the person's co-authors in the Author list, and conferences and years of the person's papers in those two lists as well. A darker shade of orange indicates a more frequent connection—exact counts (i. e., how many papers connect the items) are shown behind each connected item and are relative to the current selection. A third sorting order for each list is by connection strength, so the darkest orange items are shown highest in the list.

Multiple items within or across lists can be selected concurrently. When this happens, different interpretations of being "connected" are possible. We identified four distinct modes of multi-selection—each can be selected via buttons at the top of the tool:

- **Any**—Items that have a relationship to any one or more of the selected items are considered to be connected (like a logical "OR" mode).
- **All**—Items that have a relationship to all of the selected items, though not necessarily through the same paper, are considered to be connected (like a logical "AND" mode).
- **Same**—Items that have a relationship to all of the selected items via some exact same paper are considered to be connected.
- **All-Any**—Items that have a relationship to at least one selected entity from each different list are considered to be connected (like performing an "AND" between lists but an "OR" within a list).

These modes facilitate different types of analytical questions. To find the authors who published papers in all of a set of years, one can select those years in **All** mode. To find the years, conferences and co-authors of papers written by two researchers together, one needs to select those researchers in **Same** mode. To find the authors of papers from a particular conference during a range of years, one has to select the conference and those years using **All-Any** mode.

Fig. 5 shows a selection for William Ribarsky. The conference list is sorted by frequency, the year list numerically, and the author list by connection strength. The conference and year lists thus show Ribarsky's publications and the author list identifies his most frequent co-authors and how many papers they wrote together.

5 Limitations

The dataset only contains entries from the IEEE VIS conferences, and consequently also only citations within this body of work (as explained in Sect. 3.3). As mentioned in the introduction, although the dataset captures a sizable part of publications on visualization, papers from venues such as the EuroVis and PacificVis conferences as well as non-conference journal papers in TVCG and articles in other journals such as IEEE's *Computer Graphics and Applications*, Sage's *Information Visualization*, or Springer's *Journal of Visualization* are not included. This means, in particular, that it is impossible to judge the overall contribution of any given researcher to the field solely based on our dataset—for that a more comprehensive data collection would be necessary.

Other limitations relate to the captured data itself. As discussed in Sect. 3.5, we did our best to verify the publication status, the page numbers, the DOI links, etc., but some errors may still have escaped our attention. Our process of name sanitizing (Sect. 3.4) may also not have detected all cases of where a single person is referred to with two different names or spellings of the same name. Exploring more advanced approaches for author disambiguation (e. g., semi-supervised learning [13]) are another avenue of potential future work. For all of these issues we hope that the community will assist us in detecting and, consequently, removing these errors.

6 Lessons Learned

Our work is further evidence to our previous experience that the data in digital libraries in general and the one in the IEEE Xplore

library in particular is not without errors. The omission of almost 2% of the proceedings from the IEEE Xplore digital library greatly surprised us, a fact that is not beneficial for the visibility of work in our community. Consequently, each author should be careful to verify any citation data obtained from online sources such as IEEE Xplore. Digital libraries also should provide a better channel to report such data errors—in getting our edits and corrections to the IEEE Xplore library and having them promptly acted upon.

If the digital libraries as well as metadata collectors such as Google Scholar provided API access to their publication databases, data collection and consolidation efforts such as ours could be greatly improved. However, even with better access to existing libraries our work complements the data collected by institutions such as IEEE or ACM. By creating a dataset outside of their closed formats we can (a) give free edit access to whoever wants to help in maintaining the data, (b) choose data formats of our own, (c) easily extend the data to conferences not covered under the heading of a particular sponsor (e. g., EuroVis is sponsored by Eurographics) and, thus, maintain a central hub for visualization-related papers. Furthermore, improving existing digital libraries can be a big effort requiring major political, bureaucratic, and technical effort. Even small changes can take months or even years to be realized.

In addition, because the data surrounding visualization papers is maintained by organizations such as the IEEE, the level of community involvement in logging and documenting information about papers is limited. In domains such as high energy physics where research publications are typically in the public domain, more extensive tracking of papers, paper meta information, and citations has occurred. The INSPIREHEP repository,[9] for example, provides extensive information about papers and has even inspired new visualizations communicating publication impact [14].

As for the actual data collection, we found that the most difficult aspects were to identify unique author names as well as to extract the citations correctly. If the unique identifiers that exist for both these elements (DOI for papers and, for example, ResearcherID or ORCID for researchers/authors)[10] were used consistently in the metadata, a lot of problems would disappear.

One important recommendation for conference organizers, authors, viewers, as well as any future collaborators of ours is to continuously check the cleanliness of the data, both in our dataset and in the digital libraries. Perhaps better data collection tools could be provided at the conference level to avoid data entry errors.

7 Summary and Conclusion

We introduced a carefully curated dataset of the IEEE VIS conference publications. This dataset can help scientists to better understand history and trends in the area of data visualization. In addition, we hope that the dataset will be used for the development of new methods of text visualization and the evaluation of research approaches. We would like to encourage others to support our effort in keeping the dataset up to date and in correcting errors. The online data spreadsheet can be downloaded, copied online, and individual fields can be commented on. Our website (http://www.vispubdata.org/) will be kept up to date when changes to the data are made. Of course, the dataset does not have to be limited to IEEE VIS publications. It could be complemented with the publications of other conferences and journals from the field of visualization as well as with additional data fields, such as a total citation count. The methods for data cleaning described in this article might be useful for such tasks. Last, but not least, we would appreciate comments, suggestions, and feedback, including those on how the dataset is used.

9. https://inspirehep.net/

10. http://www.doi.org/, http://wokinfo.com/researcherid/, and http://orcid.org/

Acknowledgments

We thank Anand Sainath and Sakshi Pratap who implemented VISLists. This work was supported, in part, by the DFG priority program 1335 “Scalable Visual Analytics” as well as National Science Foundation grants NSF IIS-1302755 and NSF MRI-1531491.

References

[1] M. Bilgic, L. Licamele, L. Getoor, and B. Shneiderman. D-Dupe: An interactive tool for entity resolution in social networks. In *Proc. VAST*, pp. 43–50. IEEE Computer Society, Los Alamitos, 2006. doi: 10.1109/VAST.2006.261429

[2] K. Cook, G. Grinstein, and M. Whiting. The VAST challenge: History, scope, and outcomes: An introduction to the special issue. *Information Visualization*, 13(4):301–312, Oct. 2014. doi: 10.1177/1473871613490678

[3] I. G. Councill, C. L. Giles, and M.-Y. Kan. ParsCit: An open-source CRF reference string parsing package. In *Proc. LREC*, pp. 2764–2767. European Language Resources Association, 2008.

[4] J.-D. Fekete, G. Grinstein, and C. Plaisant. IEEE InfoVis 2004 contest: The history of InfoVis. Dataset and Web site: http://www.cs.umd.edu/hcil/iv04contest/, 2004. Visited 12/2015.

[5] C. Görg, Z. Liu, J. Kihm, J. Choo, H. Park, and J. Stasko. Combining computational analyses and interactive visualization for document exploration and sensemaking in Jigsaw. *IEEE Transactions on Visualization and Computer Graphics*, 19(10):1646–1663, Oct. 2013. doi: 10.1109/TVCG.2012.324

[6] F. Heimerl, Q. Han, S. Koch, and T. Ertl. CiteRivers: Visual analytics of citation patterns. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):190–199, Jan. 2016. doi: 10.1109/TVCG.2015.2467621

[7] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko. Visualization publication dataset. Dataset and website: http://vispubdata.org/. Published Jun. 2015.

[8] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller. Toward a deeper understanding of visualization through keyword analysis. Technical Report RR-8580, Inria, France, Aug. 2014. Also on arXiv (# 1408.3297).

[9] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller. Visualization as seen through its research paper keywords. *IEEE Transactions on Visualization and Computer Graphics*, 23(1), Jan. 2017. To appear. doi: 10.1109/TVCG.2016.2598827

[10] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, T. Möller, N. Kircanski, and J. Schlereth. KeyVis. URL: http://keyvis.org/, 2014. Visited 12/2015.

[11] K. Kucher and A. Kerren. Text visualization techniques: Taxonomy, visual survey, and community insights. In *Proc. PacificVis*, pp. 117–121. IEEE, April 2015. doi: 10.1109/PACIFICVIS.2015.7156366

[12] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[13] G. Louppe, H. Al-Natsheh, M. Susik, and E. Maguire. Ethnicity sensitive author disambiguation using semi-supervised learning. arXiv.org preprint 1508.07744, 2016.

[14] E. Maguire, J. M. Montull, and G. Louppe. Visualization of publication impact. In *EuroVis 2016 – Short Papers*, pp. 103–107. Eurographics Association, Goslar, Germany, 2016. doi: 10.2312/eurovisshort.20161169

[15] C. Plaisant, J.-D. Fekete, and G. Grinstein. Promoting insight-based evaluation of visualizations: From contest to benchmark repository. *IEEE Transactions on Visualization and Computer Graphics*, 14(1):120–134, 2008. doi: 10.1109/TVCG.2007.70412

[16] T.-M. Rhyne, J. Stasko, H. Hagen, B. Bach, and S. Huron. VIS25 timeline. URL: http://www.aviz.fr/~bbach/vis25timeline/, 2014. Visited 12/2015.

[17] J. Stasko, J. Choo, Y. Han, M. Hu, H. Pileggi, R. Sadana, and C. D. Stolper. CiteVis: Exploring conference paper citation data visually. In *Posters of IEEE InfoVis*, 2013.

[18] K. Voegele. Annotated bibliography of the visualization conference proceedings. In *Proc. IEEE Visualization*, p. xxii. IEEE Computer Society, Los Alamitos, 1995. doi: 10.1109/VIS.1995.10003

[19] L. Xie. Visualizing citation patterns of computer science conferences. Blog post: http://cm.cecs.anu.edu.au/post/citation_vis/, Aug. 2016.

[20] P. Xu, C. D. Stolper, A. Sainath, and J. Stasko. VIS 25—All the papers and citations. Website: http://www.cc.gatech.edu/gvu/ii/citevis/VIS25/, 2014. Includes CiteVis2, CiteMatrix, and VISLists. Visited 12/2015.