This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2865022, IEEE Transactions on Visualization and Computer Graphics

VIS Author Profiles: Interactive Descriptions of Publication Records Combining Text and Visualization

Shahid Latif and Fabian Beck

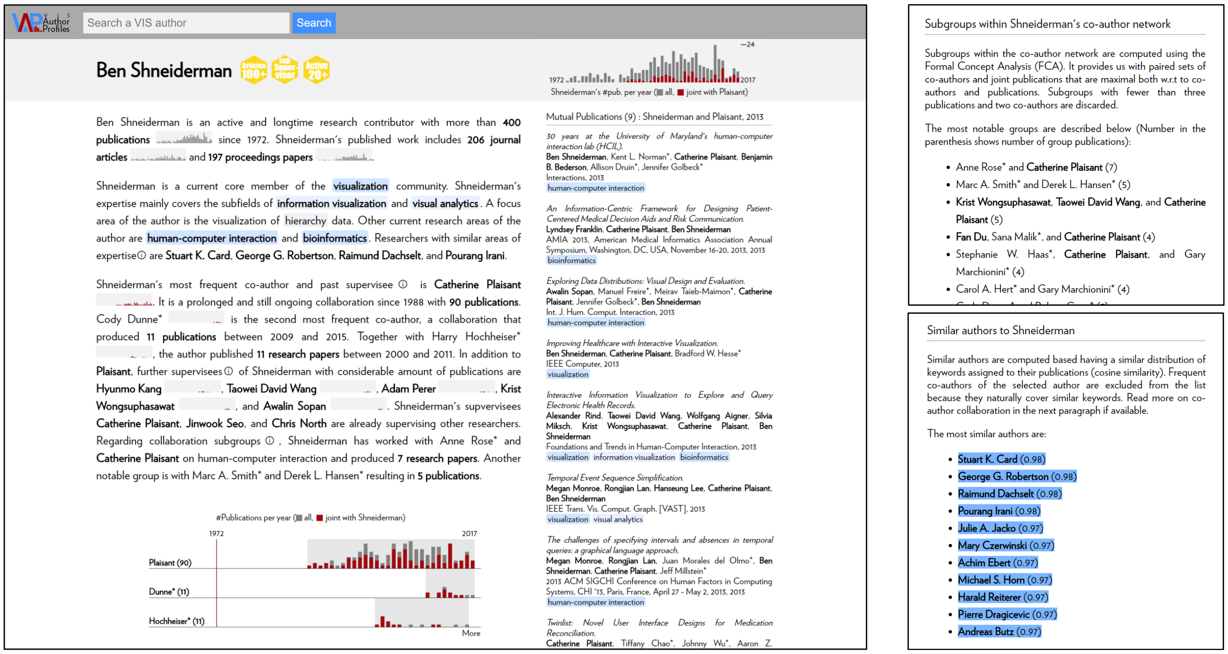

Fig. 1. Profile of author *Ben Shneiderman*. The text consists of three sections describing general information, research areas, and collaboration relationships. The visualization below provides information on joint work with co-authors on a timeline. The sidebar shows details on demand, whereas the top-right bar chart displays the temporal distribution of publications. Badges at the top summarize achievements. The cut-outs on the right are two different versions of the sidebar (list of collaboration groups and similar authors).

Abstract—Publication records and collaboration networks are important for assessing the expertise and experience of researchers. Existing digital libraries show the raw publication lists in author profiles, whereas visualization techniques focus on specific subproblems. Instead, we look at publication records from various perspectives mixing low-level publication data with high-level abstractions and background information. This work presents VIS Author Profiles, a novel approach to generate integrated textual and visual descriptions to highlight patterns in publication records. We leverage template-based natural language generation to summarize notable publication statistics, evolution of research topics, and collaboration relationships. Seamlessly integrated visualizations augment the textual description and are interactively connected with each other and the text. The underlying publication data and detailed explanations of the analysis are available on demand. We compare our approach to existing systems by taking into account information needs of users and demonstrate its usefulness in two realistic application examples.

Index Terms—Natural language generation, document visualization, interactive documents, sparklines, digital libraries.

1 Introduction

Publication records contain rich information and play an important role in assessing the expertise and experience of researchers, for instance, when recruiting faculty members, forming a program committee, or finding potential collaborators. Existing digital library systems show relevant author-centric information, but only add little abstraction to raw publication records. For instance, they abstract publication meta data to co-author relationships but only provide them as a list rather than an explorable co-author visualization. Users have to go through different views and apply various filters to gather the required information. As an alternative, visualizations that show publication and author data have been suggested. Although existing visualizations provide high-level abstractions about author profiles, their focus is often narrower, or they grow difficult to read and complex when adding more information.

In contrast, we suggest combining visualizations with natural language text and leverage the advantages of both representations. Natural language is easy to understand and self-explaining, it provides great flexibility for integrating various context, and can express implicit data explicitly. When showing larger sets of items and numbers, however, visualization can provide an overview and rich insights. For an interplay, both representations need to be closely integrated. We achieve this by interactively linking textual and visual data descriptions as well as augmenting the text with *sparklines* [31], small visualizations that are integrated in line with the text.

• *S. Latif and F. Beck are with paluno, University of Duisburg-Essen, Germany. E-mail: {shahid.latif, fabian.beck}@paluno.uni-due.de.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2865022, IEEE Transactions on Visualization and Computer Graphics

We describe a novel way of presenting publication records and related analysis results for scientific authors. Our system *VIS Author Profiles (VAP)*, a Web-based visual analytics tool, generates author profiles in the form of interactive reports (Figure 1). The text describes general statistics, research topics, and collaboration networks. In addition, interactive visualizations allow for understanding trends and extended collaboration relationships. Our main contributions are:

- the identification of author-centric information needs for relevant academic usage scenarios that are not sufficiently supported by existing systems (Section 2),
- the extensions of a template-based text generation approach for the integrated production of interactive, visualization-enriched documents (Section 4),
- a system for the generation of profiles for authors of the visualization community, including analysis and summary of publication records, research topics, and collaborations (Section 5), and
- a realistic demonstration of the approach in two scenarios (Section 6) as well as a discussion of applicability and extendibility of the approach (Section 7).

The novelty of the approach lies in the joint and integrated generation of natural language text and visualization. Our solution is a visual analytics system in the form of interactive visual reports that builds on algorithmic data analysis and preprocessing. The focus, however, is the comprehensible dissemination of results and storytelling. Although our primary goal is to provide explanations, the proposed system also supports exploratory analysis to some extent as a secondary objective.

The interactive system is available at `https://vis-tools.paluno.uni-due.de/vap/` and the supplemental material of the paper contains an interactive appendix with additional information on results.

2 Analyzing Publication Records

We first discuss the application of analyzing publications records: for which author types and scenarios such analysis is targeted, how we frame the scope, and what information an analyst requires.

2.1 Author Personas

Our work is focused on authors of scientific papers, a group ranging from PhD students who have just published their first paper to senior researchers who have authored hundreds of publications. We aim at creating meaningful reports for all these researchers. In detail, we discern the following categories of authors—we call these categories *personas* in the sense that they represent a certain role or stereotype:

- **Student:** A student (Bachelor, Master, or PhD level) who contributes to research projects within a study program or dependent employment under the supervision of experienced researchers.
- **Researcher:** A postdoctoral researcher, assistant professor, research scientist, or lecturer who is conducting first independent research and might start supervising students.
- **Senior Researcher:** An associate or full professor, senior research scientist, or senior lecturer who can build on years of experience in research and supervising.
- **Occasional Contributor:** An outsider to the studied scientific community who occasionally contributes to academic work within the community.

We do not claim that every author can be unambiguously assigned to a certain persona. However, this list helps us structure the set of authors and make sure that our approach finally produces meaningful descriptions for authors of various levels of experience. Although we do not automatically identify the persona of a researcher, the association of researchers to personas is indirectly linked to their experience such as the number of published articles, active years in research, and whether they have supervised other researchers.

2.2 Scenarios

Publication lists of researchers reflect their accumulated formal contributions to science. They not only provide information on research activity, but also on the topics the authors have worked on and their collaboration network. We focus on two scenarios that build on publication records and require an author-centric view. Use cases like literature search or the analysis of a research field are out of scope for this paper.

S1 – Recruiting: The overarching question of every recruiting process is how well the candidate is qualified for the job. In academia, this is closely linked to the level of academic experience and achievements. In addition to a CV, the publication record of the candidate is a main source of information. Similar scenarios that can be considered a variant are the admission to a funding program as well as assessing researchers as part of a grant or tenure evaluation.

S2 – Assessing Expertise: In contrast to recruiting, where specific people apply, assessing the expertise of researchers for a task or role is open for additional suggestions—in addition to understanding the data of a single researcher, users also want to explore related researchers. Typical examples are to find a reviewer for a specific paper, to select candidates for a program committee, or to look for potential collaborators or supervisors. In all cases, expertise with respect to certain research topics as well as experience regarding academic work are key criteria to invite or contact a researcher.

2.3 Scope and Data

All sciences and research fields share that they publish articles, papers, and books. The specific publication culture, however, differs largely between fields, from monograph-oriented areas to purely peer-reviewed-based systems. We focus in our work on publications from computer science, which is our own research area and where we have a good understanding of the publication culture. DBLP makes available the required publication meta data.[1] For investigating research topics, we want to make use of the keywords that the authors assigned to the papers. This information is not available in DBLP or any general data collection for all computer science. Hence, we decided to focus on the visualization area, where the *Visualization Publication Data Collection* provides such information.[2] We restrict the generated author profiles to authors listed in this data set, but also integrate DBLP data to add publications of these authors that appeared in non-visualization venues. This narrower focus also allows us to consider specifics of the visualization community, for instance, its subdivision into the fields of *scientific visualization*, *information visualization*, and *visual analytics*.

The combined data includes 5,086 authors and 128,961 publications till August 29, 2017. For the retrieval of research topics, we further enrich the data by categorizing the most frequent publication venues (i.e., journals, conferences, workshops, etc.) into research communities of computer science. A classification of 688 venues with 56 keywords provided a community assignment for 61,469 publications. We enrich these high-level keywords with further keywords extracted from paper titles based on a manually created mapping of 26 typical terms (e.g., *visualizing* $\rightarrow$ *visualization*). To identify subtopics within the visualization community, we leverage the author-assigned keywords. Since they are inconsistent, we use the mapping that the *KeyVis* project provides to map them to a standardized set of keywords.[3] We select a subset of such aggregated keywords that—in our opinion—best reflects certain subareas of the visualization community and map these to simpler terms preserving their original meaning as far as possible.

2.4 Information Needs

For the above scenarios, we derive information needs that users might have with respect to the described publication data. A list of publications may give rich insights about the research topics a researcher is working on as well as an overview on the researcher's collaborators. Further, publication statistics and temporal evolution of publication activity hint at experience levels and academic achievements. While

[1] https://dblp.dagstuhl.de/
[2] http://www.vispubdata.org
[3] http://keyvis.org/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2865022, IEEE Transactions on Visualization and Computer Graphics

citation and biographic information would provide more details, we restrict ourselves to publication data because of its wide, reliable, and standardized availability. Therefore, we do not list any information needs that would require such additional data sources.

First, general statistics provide an idea how actively a researcher is publishing, which is a central criterion for recruiting (S1). In computer science, both certain journals and conference proceedings can be considered premier publication venues. On average, however, the journal articles have an estimated higher contribution because of their greater length and due to the fact that proceedings might also comprise short papers and workshop contributions. Hence, discriminating journal articles from proceedings papers hints at potential contributions of the publications. The temporal distribution of publications indicates academic age and level of experience, which is not only relevant for recruiting (S1) but also for finding an expert with sufficient experience (S2). A special publication is the PhD thesis, where the author independently worked on a topic and set a first academic milestone. First-author publications can be considered particularly relevant when hiring early-career researchers (S1), because these might best express the research interests and abilities of the author (assuming that, like commonly applied in practical computer science, the author sequence reflects contributions but does not follow alphabetic order).

IN 1 – General Information

IN 1.1 *What is the **number of publications**, overall and discerned by publication type?*

IN 1.2 *What is the publication span and **temporal publication distribution**?*

IN 1.3 *When was the **PhD** thesis published and what is the ratio of **first-author** publications (for early-career researchers)?*

For selecting experts (S2), it is important to understand their areas of experience on different levels of abstraction—we discern between research communities (e.g., *visualization*), subfields (e.g., *scientific visualization*), and focus areas (e.g., *flow visualization*). Also, the temporal evolution is relevant for distinguishing a researcher's current and previous research direction. Research topics can also lead to other researchers with similar expertise—imagine you invited the researcher to do a review, but the researcher has declined, and you need to find somebody with a similar focus. Frequent co-authors, who naturally share similar research interests, are excluded here but will be discussed as part of the next group of information needs.

IN 2 – Research Topics

IN 2.1 *What is the main **research community** and connections to other research communities?*

IN 2.2 *What are **subfields** and focus areas within the main community?*

IN 2.3 *What is the **evolution of topics**?*

IN 2.4 *Who are other authors contributing to **similar topics** (excluding frequent co-authors)?*

Finally, co-author relationships help to see how a researcher is connected in a community. This information is important when searching for references of a candidate (S1), judging on influence, academic inheritance and experience in supervising (S1), or searching for similar experts (S2). A special relationship that can be estimated from author sequence in publications is the supervisor (often, one of the last authors) and supervisee (often, first author). Former supervisees already supervising other researchers indicate certain influence of the supervisors. Moreover, the co-authors may form noteworthy subgroups; for instance, an author might frequently publish on a certain topic with a specific subset of co-authors.

IN 3 – Collaboration Network

IN 3.1 *Who are the **main collaborators** and what is the temporal distribution of joint work?*

IN 3.2 *Who are or were **supervisors** and are the collaborations still ongoing?*

IN 3.3 *Who are or were **supervisees**, are the collaborations still ongoing, and are the former supervisees already supervising?*

IN 3.4 *What are **subgroups** of co-authors who have frequently worked together on certain topics?*

3 Existing Systems and Related Work

Various digital library systems already allow for interactive exploration of publication data of scientific authors; we evaluate them against the information needs discussed above and show that they do not yet satisfy the needs sufficiently. Complementary to this, visualizations provide representations of the same data. This section also discusses this previous work from the application domain and introduces techniques to integrate text and visualization like intended in our approach.

3.1 Digital Library Systems

We categorize views of existing digital library systems for exploring scientific literature into *publication-centric* and *author-centric* views. We focus our discussion on the latter because these are more closely related to the scope of our approach and exclude systems that do not have a dedicated author profile page (e.g., *IEEE Xplore*). *Google Scholar* is probably the most widely used publication-centric system for online literature search. Besides the main search interface, it has a profile page for each (registered) author containing information on affiliation, research topics, list of publications, co-authors, and citations. *Microsoft Academic* is a competing system and has similar layout for author profiles. The author profiles of *ACM Digital Library*, *DBLP*, *ResearchGate*, and *Semantic Scholar* use faceted browsing [35] to subselect the publications along certain facets such as the publication type, topic, co-author, or publication venue. *Google Scholar*, *Scopus*, and *Semantic Scholar* show citation or publication frequency on a timeline. *AMiner* provides other visualizations, including a stream graph of research topics, a Kiviat diagram of publication metrics, and an ego-centric (simplified) co-author network. *Semantic Scholar* features a visualization of academic impact. But none of them uses generated natural language text as part of the author profile.

We evaluate if the information needs (Section 2.4) are already fulfilled by existing systems. The evaluation considers the availability of features, but not their usability or data quality. This provides an objective and reproducible comparison, whereas rating quality attributes would necessarily be subjective. Table 1 summarizes the outcomes of the evaluation; an extended interactive version of the table is part of the interactive appendix and provides explanations for every rating.

Whereas all systems show a list of publications in temporal order, not all discern publication types (**IN 1.1**) or show publications distributions aggregated on a timeline (**IN 1.2**). No system highlights first author publications, but PhD theses (though rarely contained in the data sets) can at least be retrieved in those systems that allow for filtering by publication type (**IN 1.3**). Research communities (**IN 2.1**) can be identified in most cases, indirectly derived from aggregated venue information (e.g., *DBLP*) and author-selected keywords (e.g., *Google Scholar*), or directly as automatically mined subject areas (e.g., *Scopus*). Similarly, more detailed information on subfields is indirectly or directly available in most systems (**IN 2.2**). In contrast, the analysis of the evolution of research topics is only supported in *AMiner* and *Semantic Scholar* by providing a timeline (**IN 2.3**). *AMiner* is the only system that indicates similar researchers (**IN 2.4**). Finally, the systems mostly list frequent collaborators, but—except for *Semantic Scholar*—without temporal information (**IN 3.1**). Only *AMiner* explicitly highlights supervisors and supervisees of an author (**IN 3.2**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2865022, IEEE Transactions on Visualization and Computer Graphics

Table 1. Assessment of fulfillment of information needs in author profile pages of existing digital library systems; degree of fulfillment: *no* ○○○, *partly* ●○○, *largely* ●●○, and *yes* ●●●. See interactive appendix for a more detailed version of the table.

Information Need (IN)	General Information			Research Topics				Collaboration Network			
	1.1 number of publications	1.2 temporal pub. distribution	1.3 PhD and first-author	2.1 research community	2.2 subfields	2.3 evolution of topics	2.4 similar topics	3.1 main collaborators	3.2 supervisors	3.3 supervisees	3.4 subgroups
ACM Digital Library	●●○	●○○	○○○	●○○	●●○	○○○	○○○	●○○	○○○	●●○	○○○
AMiner	●○○	●○○	○○○	●●●	●●●	●●●	●●●	●●○	●●●	●●●	○○○
DBLP	●●●	●○○	●●○	●●○	○○○	○○○	○○○	●●○	○○○	○○○	●●○
Google Scholar	●○○	●○○	○○○	●●○	●●○	○○○	●○○	●●○	○○○	○○○	○○○
ResearchGate	●●●	●○○	●●○	●●●	●●●	○○○	●○○	●●○	○○○	○○○	○○○
Scopus	●○○	●●●	○○○	●●●	●○○	○○○	○○○	●●○	○○○	○○○	○○○
Semantic Scholar	●●●	●●●	●●○	●●○	●●○	●●○	●○○	●●●	●○○	●○○	●○○

and **IN 3.3**). None of the systems considerably supports the identification of collaboration groups (**IN 3.4**) except for DBLP.

In conclusion, there are systems like *DBLP* and *Google Scholar* that are easy to use, but only fulfill a smaller fraction of the information needs. In contrast, other systems like *AMiner* and *Semantic Scholar* already meet a good share of the information needs and might be extensible towards the remaining ones. However, these systems feel already overloaded with views, tables, and visualizations—adding more information would further clutter the display. Our suggestion is the usage of natural text to describe a large variety of information in an easy-to-understand and compact way. We aim at fully supporting all information needs listed above.

3.2 Author Visualizations

In our approach, we focus on individual authors and want to summarize their publication activity and co-author collaboration networks. From a perspective of social network analysis, we can consider this to be an ego-centric perspective. Some ego-centric (dynamic) graph visualizations have studied evolving co-author relationships. Huang and Huang [17] show years as rings subdivided by color-coded co-authors. Reitz [29] uses links in an ego-centric node-link diagram to encode temporal distributions of the joint work. In contrast, Shi et al. [30] extend the ego node to a timeline and connect co-authors to points in time where joint work appeared. Similarly, *EgoNetCloud* [23] uses a single timeline to which the co-authors can be connected or arranged in independent node-link components along the timeline. *egoSlider* [34] also shows author-specific timelines, where co-authors cluster vertically at points in time and are linked through time steps. *MENA* [15] represents the ego-centric network as a dynamic graph in small multiples. Fung et al. [10] suggest a botanically inspired tree visualization to summarize in each branch the collaborations of a time span with co-authors encoded as leaves. In our approach, we include a visualization related to the discussed ego-centric visualizations to show the collaboration network. But it is only a small part of the interactive document—we also integrate other versatile information to the profile description, for instance, covering research topics and a summary of general author information.

In addition to ego-centric author visualizations, various other visualizations also show author information, for instance, in the form of a co-author graph [1, 16, 21]. These systems often integrate information on research topics, in the coloring of the nodes [21] or as topic–collaboration nodes [1]. *CiteWiz* [8] combines co-author networks with keyword co-occurrence networks and citation impact visualizations. *PivotSlice* [36] use faceted exploration to investigate an author's publications based on keywords, citation, and venue information. *PivotPaths* [7] links author nodes to paper nodes to keyword nodes and supports an interactive exploration. *SurVis* [3] associates word clouds of keywords and authors in a faceted browsing approach for literature collections. For further approaches on visualization of scientific literature datasets, we refer to the recent survey of Federico et al. [9]. In general, we are not aware of any approach that augments such visualizations with generated textual descriptions.

3.3 Integration of Generated Text and Visualization

Natural language generation (NLG) is the field that investigates how to produce natural language text from data and abstracted information [12, 28]. We use a simple template-based generation approach, which—unlike more advanced approaches—does not build a full grammar model of the sentences. We chose this approach because it already provides sufficient flexibility and power for the intended purpose. Moreover, it is easier to design and control.

While there exist various generation approaches [12], only few have investigated the joint generation of text and visualizations as part of multimodal documents (i.e., documents that integrate text and graphics). Previous work scatters among many domains, for instance, user and maintenance instructions for physical devices [32], weather forecasts [26], or learning analytics in an education scenario [27]. Furthermore, the generation of textual content with respect to existing diagrams is related [6, 18, 19, 25], but here the visualization is already given and not generated in coordination with the text. There are only few approaches that take into account the close integration and interactive linking between the textual and visual content. *Method Execution Reports* [4] describes the execution behavior of the method of a program; it integrates simple graphics into the text and makes them explorable to some extent. *PersaLog* [2] advocates the personalization of news consisting of text and interactive visualization; the system focuses on how personalized content can be authored but does not particularly discuss the close integration of text and visualizations. In contrast to these systems, our approach studies a different area of application and goes at least one step ahead regarding the close and interactive integration of textual and visual content as part of an interactive document.

Embedding *sparklines*, "data-intense, design-simple, word-sized graphic[s]" [31] also called *word-sized* [5] or *word-scale graphics* [14], into the text is a means for achieving a close integration of text and visualizations. We use interactive sparklines [5, 14, 22] that can be explored on demand and are linked to text and regular visualizations. Others have already demonstrated that sparklines can be leveraged for literature data analysis [3, 5, 24]. Note that the term *sparkline* is not only used for small line charts, but—in accordance to the original definition of Tufte [31]—refers to any word-sized graphic; the sparklines used in this paper are all small bar charts.

4 Generating Interactive Visual Reports

Before we introduce the actual content of the interactive author profiles, we provide some technical background on the generation process. We use template-based natural language generation to automatically produce text from data that we then connect with visualizations and augment with interactions. The selection of relevant items is a frequent problem that we are facing in the text generation process, for which we suggest a generic solution.

4.1 Generation Pipeline

The first step in our generation pipeline is the pre-processing of *DBLP* publication records. We integrate this data with the *Visualization Publication Data Collection* and enrich it with keywords as discussed in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2865022, IEEE Transactions on Visualization and Computer Graphics

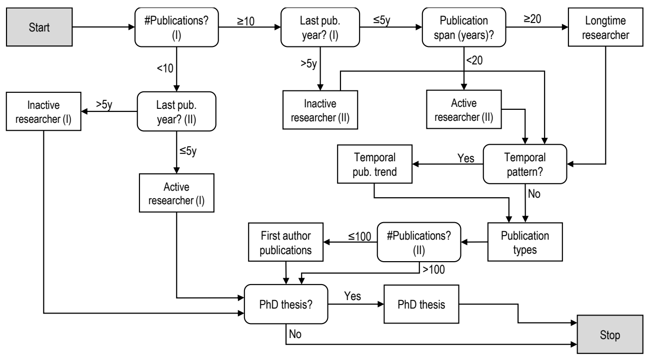

Fig. 2. Decision graph explaining the flow of text generation for the first paragraph of a profile. Rectangular nodes represent *text* vertices whereas nodes with rounded corners are *decision* vertices. The traversal of any path from *start* to *end* node produces a meaningful paragraph.

Section 2.3. All pre-processing is done in Java. The front end is written in HTML and JavaScript. For producing the visualizations, we use Scalable Vector Graphics (SVG) and D3js. The text generation templates are implemented as part of the front-end code in JavaScript.

4.2 Template-based Text Generation

Template-based systems work with well-formed, pre-written phrases with gaps in them and produce the output text when these gaps are filled with data. For instance, an announcement generation system at a train station can be considered as a very simple example of template-based system; the template "[*train*] will leave for [*city*] at [*time*]," where the gaps [*train*], [*city*], and [*time*] are filled from a data table might produce: "ICE577 will leave for Frankfurt at 13:30."

To build the templates, we followed an informal iterative approach. In every iteration, we drafted a text fragment based on an author's publication record. We implemented a base version of it as a template and then kept on refining and fine-tuning the template by testing over many random authors belonging to various personas and special cases. We continued the iterations until the text covered all information needs discussed in Section 2.4. With this approach, we received quick results and continuously tested the generated text. Step by step, we also integrated interactions and visualizations in a similar fashion.

We follow a similar approach for generating natural language text as applied in *Method Execution Reports* [4]. Directed acyclic decision graphs (Figure 2 gives an example) generate text from the parameterized templates. An author's profile consists of three fixed paragraphs (one for each group of information needs) and we define a decision graph per paragraph. The sequence of text fragments (usually, a text fragment represents a sentence or phrase) within a paragraph is fixed. In the decision graph, *start* and *stop* vertices mark the beginning and end of the text generation process, *text* vertices (rectangular nodes) add a new text fragment to the paragraph when traversed, and finally *decision* vertices (rounded rectangular nodes) determine the path based on conditional statements. The path is deterministic and any traversal from *start* to *stop* vertex results in a paragraph. Hence, the text fragments need to be designed to form well-formed sentences regarding all possible paths.

Our approach is flexible and produces grammatically correct sentences if all conditions are carefully checked. In the templates, we take into account the already generated text and connect to previous sentences with appropriate conjunctions. The use of numerals (e.g., one, two) in place of numbers (if a paragraph only contains small numbers less than 10) and rounding down larger numbers to nearest fifties make the text more natural to read. We use adjectives to characterize the objects we are describing (e.g., *long-lasting collaboration*) and consider different tenses for a correct referral to time spans.

4.3 Integration of Text and Graphics

A novel aspect of our approach is the coherent integration and linking of jointly generated text and graphics. A first step of integration is the embedding of sparklines in the lines of the text, which already visually connects the visualized data to the related phrase. Moreover, the use of regular-sized visualizations, which are interactively linked to the text and sparklines, allows detailed exploration of data. Our visualization–text interactions are adaptations from the previous work [5, 22], which describes the linking of text, sparklines, and regular visualizations. For instance, clicking on linked text fragments highlights relevant parts of visualization and shows related data in a side panel. All sparklines are also linked to a larger visualization—on click, the information presented in the sparkline is shown as an overlay in the visualization. With respect to layout transitions for sparklines [14], this can be considered an *offsetting* solution (in contrast to *in-place* and *growing* transitions). We further use info icons ⓘ to mark the availability of additional explanations and present this information on click.

We include interactive sparklines as part of the text templates; they can be considered as parameters (gaps) in the template. For instance, when the vertex named *First author publications* (Figure 2) is visited, a sentence about researcher's first author publications including the corresponding sparkline, showing temporal distribution of these publications, is produced (e.g., third sentence in Figure 3).

4.4 Selection of Data Items

Throughout the analysis of publication records, we come across large lists of data items that need to be sliced to a reasonable length for presenting in the text. We assume that every item has an importance value attached that can be represented as a numeric value. For instance, in a list of co-authors (data items), the number of joint publications is considered as the importance value. Similarly, in a list of keywords, the frequency of each keyword might be the importance value. The selection problem may sound trivial—one could just select the top x items. However, if we as human authors of a text would select a number of important items, we do not restrict ourselves to a fixed number, but choose a good cut-off point dynamically. We try to avoid that the list grows too long, but we also do not cut off at a position where the distance to the next item is small. Hence, we need to select a to b items from a sorted list $L = (l_i)_{i=1}^{n}$ of n numeric items ($n > b$, $l_i \geq l_{i+1}$). We select cut-off index $c \in [a, b]$ where the difference of list elements $l_{c+1} - l_c$ is maximal. However, there can be several maximal differences—in that case, we pick the smallest index, formally:

$$c = \min\left(\underset{k \in [a, b+1]}{\arg\max}\, l_{k+1} - l_k\right) \quad (1)$$

Finally, the list is cut after element l_c and hence only contains the top c elements. In the following, we refer to this procedure as Equation 1.

5 VIS Author Profiles

VIS Author Profiles (VAP) is a Web-based visual analytics tool that generates profiles for authors of the visualization community describing their publication record. It is designed to fulfill the information needs discussed in Section 2.4 (**IN 1** – General Information, **IN 2** – Research Topics, and **IN 3** – Collaboration Network). In order to create a coherent and easy-to-use interface, we applied the following set of design principles throughout the development process.

- *Make it self-explaining:* We want to leverage natural language text to explain the data and analysis results. When the explanations would make the main text become lengthy, we at least make additional information available on demand.

- *Make algorithms and data transparent:* We provide background on our algorithms and make underlying data relevant to a context available on demand to allow users validate the textual descriptions and build trust in our descriptions.

- *Better say nothing than say something wrong:* The explicit description in text holds us responsible for what we say; therefore, we leave out descriptions if uncertainty is too high. Uncertainty decreases with a higher quantity of information (e.g., better data

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2865022, IEEE Transactions on Visualization and Computer Graphics

availability) or a higher quality of information (e.g., more reliable heuristics). We calibrated the parameters when to omit a certain finding as part of our iterative fine-tuning process.

Figure 1 shows the user interface of *VAP*. The central panel is used to display the textual description of an author profile and is divided into three paragraphs describing (i) general information, (ii) research topics, and (iii) the co-author network. The *co-author publication timeline* visualization at the bottom provides insights into the joint work by presenting all co-authors and their yearly publications. The right sidebar is reserved for displaying details on demand such as additional explanations and publication records. The *ego publication timeline* at the top right provides temporal distribution of individual, joint, and topic-filtered publication records with respect to the selected profile author. Enlarged versions of the sparklines are also displayed in this bar chart. The text produced in boldface characters is interactive and allows for exploring the underlying publication records by presenting them in the sidebar. Author names in the publication list and anywhere on the page are links to their profiles. Please note that the authors, not available in the *Visualization Publication Data Collection*, are not explorable through our tool and are marked with asterisk (*). An info icon ⓘ in the text indicates that users can explore additional information by clicking and loading this information in the sidebar.

To provide a quick overview of a researcher's experience, we use digital badges and display them next to the author's name in the header as shown in Figure 1. It is a concept applied in computer games and often used for *gamification* (i.e., to make a non-game interface or tasks more enjoyable or increase the motivation of users through integrating game aspects). These badges are indicators of accomplishments and skills. We award gold, bronze, and silver badges based on various levels of experience in terms of number of published papers, length of active publication time, and supervision of other researchers.

For instance, the golden sup-supervisor badge (third from left) indicate that the supervisees of the author have also started supervising (i.e., is following an academic career) and silver article badge (fourth from left) highlights an accomplishment of publishing thirty research papers or more. Badges are intended to roughly match the author personas (cf. Section 2.1): *students* can realistically earn bronze badges, but as soon as authors have a first silver badge, they can be considered *researchers*, and *senior researchers* for a first golden badge. Only *occasional contributors* are harder to link to the badges, as they might be active for a long time, but do not publish many papers.

5.1 General Information

The first paragraph of the main text gives an overview of the publication statistics of the author and aims at fulfilling the general information needs (**IN 1**). Figure 3 shows the text for author *Benjamin Bach* as an example of a *researcher*. It is generated by traversing the decision graph shown in Figure 2 according to the following path: *Start* → *#Publications? (I)* → *Last pub. year? (I)* → *Publication span (years)?* → *Active researcher (II)* → *Temporal pattern?* → *Temporal pub. trend* → *Publication types* → *#Publications? (II)* → *First author publications* → *PhD thesis?* → *PhD thesis* → *Stop*.

In general, the first sentence starts by reporting the total number of publications (**IN 1.1**) and highlights the current status as active if the last publication appeared no earlier than five years ago (nodes *Active researcher (I)* and *(II)* in Figure 2) or span of research if researcher is not active anymore (nodes *Inactive researcher (I)* and *(II)*). Active longtime contribution ($\geq$ 20 years), like for author *Ben Shneiderman* (cf. Figure 1), is also indicated (**IN 1.2**; node *Longtime researcher*). The adjacent sparkline shows the temporal distribution of publications in more detail (**IN 1.2** – publication span).

We manually analyzed the publication behavior of a set of authors across all personas to extract the most commonly occurring temporal patterns, which were then implemented and detected automatically. For

Benjamin Bach is an active researcher since 2010 and has published **25 research papers**, where most contributions appeared since 2015 (14 publications). The publications include **12 journal articles** and **13 proceedings papers**. Out of the total 25 publications, the author published 14 articles as first author. The author received a PhD degree from University of Paris-Sud, Orsay, France with the dissertation published in 2014 and titled "Connections, changes, and cubes : unfolding dynamic networks for visual exploration".

Fig. 3. Excerpt (general information) from Benjamin Bach's profile.

Weiskopf is a current core member of the **visualization** community. Weiskopf has contributed to all **information visualization**, **visual analytics**, and **scientific visualization**. Focus areas of the author are volume visualization, graph visualization, flow visualization, multimedia visualization, and the evaluation of visualization. Other current research areas of the author are **computer graphics**, **software engineering**, and **bioinformatics**. Researchers with similar areas of expertiseⓘ are **Quang Vinh Nguyen**, **Jarke J. van Wijk**, **Jonathan C. Roberts**, **Heidrun Schumann**, and **Mikael Jern**.

Fig. 4. Excerpt (research topics) from Daniel Weiskopf's profile.

each detected pattern (**IN 1.2** – temporal publication distribution), we add a clause (node *Temporal pub. trend*). As two of the most frequent and notable patterns, we currently highlight if authors published more than half the papers in any third of their publication history and exceptional peak years greater than twice of second maximum value in the time series of yearly publications. Author *Bach* has a clearly growing publication rate and hence most publications appeared in the last third.

Next, we discern the publications as journal articles and proceedings papers (node *Publication types*) with respective sparklines attached (**IN 1.1** – publication types).

Considering the importance of publications as first author for early-career researchers, the third sentence states the number and temporal distribution of such publications as text and in a sparkline respectively (**IN 1.3** – first author publications; node *First author publications*). For senior researchers, this number is not as important anymore and hence skipped (cf. Figure 1).

Finally, information about the author's PhD (dissertation title, institution, and year of publication) is described in the last sentence of this paragraph (**IN 1.3** – PhD thesis; node *PhD thesis*)—unfortunately, this data is only available for a fraction of the authors in DBLP. Profiles authors with one or only few publications have a largely reduced version of this paragraph in their profile.

5.2 Research Topics

The following second paragraph aims at satisfying the information needs corresponding to the research communities and evolution of topics (**IN 2**). We analyze the publications enriched with venue specific keywords and author specified keywords as described in Section 2.3. Figure 4 shows an excerpt of the profile of *senior researcher Daniel Weiskopf*. Since we restrict ourselves to the *visualization* community, this paragraphs starts with the description of authors status within the visualization community (**IN 2.1** – research community). We discern between core member, member, and contributor depending on the number of research papers that are classified under the *visualization* keyword. We describe the author as *active* if the most recent publication appeared with in the last two years. Next, we discuss relevant subfields for the visualization community: *information visualization*, *scientific visualization*, and *visual analytics*, which appear in the order of frequency in that they are assigned to the author's publications (**IN 2.2** – subfields). The focus areas within visualization research are reported in the following sentence (**IN 2.2** – focus areas).

Then, other research areas are listed the author has contributed to, again in the order of keyword frequency (**IN 2.1** – connection to other communities). We discern current and past research topics (since the authors of the examples in Figure 1 and 4 are still actively contributing to all listed communities, this part is skipped here). In

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2865022, IEEE Transactions on Visualization and Computer Graphics

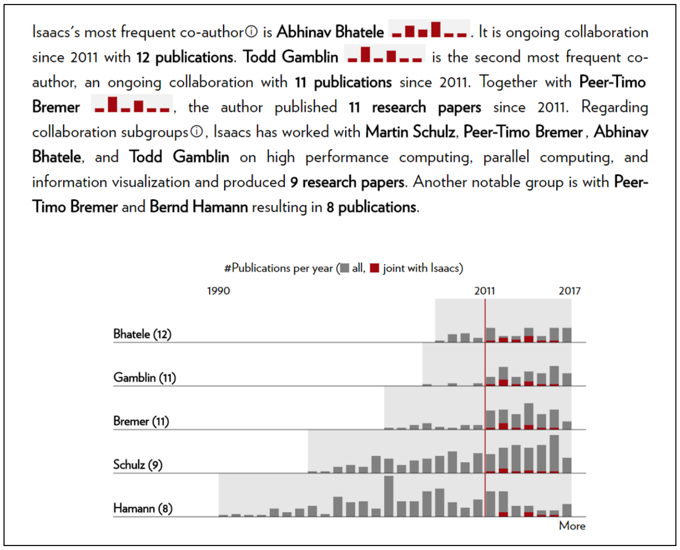

Fig. 5. Excerpt (co-author network) from Katherine E. Isaacs's profile including the *co-author publication timeline*. Gray bars indicate the number of publications per year of the co-author, while red bars show joint publications with Isaacs.

both cases, we limit the list to three to six keywords according to Equation 1. The communities are highlighted in blue, subfields in light blue, and research topics in light gray. Those keywords marked in bold font are linked with publications and are listed in the sidebar on click. Their temporal evolution is shown as an overlay on the *ego publication timeline* (**IN 2.3** – evolution of topics).

For the identification of authors having similar research interests (**IN 2.4**), we compare the research interests of the profile author with all other authors within the visualization community. We compute their cosine similarity based on keyword frequency vectors. Since keyword coverage is not optimal in our data, we restrict the similarity computation to authors with at least 30 publications to avoid marking less similar authors as similar. Since frequent collaborators are likely to share most research interests, we also exclude co-authors that share more than two publications with the profile author from the comparison. We use again Equation 1 to cut short the list to the three to five most similar authors in the last sentence of this paragraph. A longer list of similar authors is available on demand (cf. Figure1, bottom right).

5.3 Co-Author Network

To analyze the collaboration network formed within the co-authors (**IN 3**), we explore pairwise collaborations as well as groups of researchers working together. Figure 5 shows the textual description of the co-author network of *researcher Katherine E. Isaacs* along with the *co-author publication timeline*.

This paragraph first describes the pairwise collaborations among the most frequent co-authors (**IN 3.1** – main collaborators). The frequent three to eight co-authors are selected using Equation 1 from the list of all co-authors based on the amount of joint publications. The most frequent collaborations are described along with their current status (e.g., still ongoing, ended). The use of adjectives such as *prolonged* and *long-lasting* highlight if the span of collaboration is more than 15 years. If the collaboration has ended, we mark it with the total span of years along with the number of joint publications. Adjacent sparklines show the temporal distribution of joint work. Users can load the comparative view of joint publications as a proportion of the profile author's overall publications in the *ego publication timeline* as shown in Figure 1. Further exploration of the joint publications is possible by interacting with this chart—the publications of a selected year appear in the right sidebar sorted in alphabetical order of their titles.

In the scientific community, two roles of an author are of significant importance; i.e., supervisor and supervisee (**IN 3.2** and **IN 3.3**). To identify these roles, we use a heuristic method with two assumptions: (i) the name of the supervisor appears at the last positions in an author sequence of a publication and (ii) the name of supervisee appears at the first position. All co-authors of the profile author who started publishing at least five years prior to the profile author are added to a list of potential supervisors. Then, the potential supervisors who appear more frequently before the profile author in their joint publications than after are discarded. Finally, a potential supervisor is identified as a supervisor if, in fifty percent of their joint publications, the potential supervisor appears at the last position. Analogously, supervisees are computed. We highlight the author roles while describing their collaborations with adjacent sparklines as introduced above. For *senior researchers* like author *Ben Shneiderman* (cf. Figure 1), the list of supervisees is quite long. We, therefore, list more supervisees in the following sentence after cutting down the list to a maximum of five further supervisees by applying Equation 1 and using their number of joint publications with the profile author as rating. Sparklines again provide details on the temporal distribution of joint work. To find if the supervisees of the profile author are already assuming the role of supervisors (**IN 3.3**), we look for the supervisees of the supervisees and describe this in the fifth sentence. Since our approach could not detect a clear supervisor or supervisees for *Katherine E. Isaacs*, her profile text does not comment on this (cf. Figure 5).

To identify meaningful subgroups among the co-authors who have frequently worked together with the profile author on the same publications (**IN 3.4**), we employ *formal concept analysis* (FCA) [11, 33]. FCA provides paired sets of co-authors and joint publications (*formal concepts*) that are maximal both regarding co-authors and publications; that means, there is no other co-author that has also contributed to all identified publications and there is no other publication that has been co-authored by all identified co-authors. To focus on higher-order groups of co-authors and substantial collaboration, we discard formal concepts with less than two co-authors ($n < 2$) and less than three publications ($m < 3$). For all other formal concepts, we compute a score $\sqrt{n} \cdot m$ that rates both larger groups and groups with more publications higher. We apply a mitigating transform on the number of co-authors $\sqrt{n}$ because otherwise large groups of co-authors with only few publications might dominate the scores. This ranking results in an ordered list of groups of co-authors. However, groups in the list might significantly overlap with respect to co-authors, for instance, one being a subgroup of the other. Since we want to avoid listing similar groups, we sweep through the list and discard all groups that share a high similarity with any of the previously listed groups (Jaccard coefficient of $\geq 1/3$ comparing the two sets of co-authors). Among the list of groups sorted based on score, we select the top two groups according to Equation 1 for presenting in the profile text along with the number of joint publications and research topics. A longer list of collaboration groups and the group publications is available on demand and is presented in the sidebar.

Whereas only the most frequent collaborations are described in the text, the diagram showing the *co-author publication timeline* below the text (cf. Figure 5) allows for exploration of all collaborations in detail. It displays distributions of the co-author's publications per year (gray) as well as those joint with the profile author (red) on a timeline. The timeline begins with the starting year of the most senior author in the visible co-authors with a vertical red line pointing at the starting year of profile author. The active spans of co-authors are marked with light gray background. The *More* button allows for expanding the list. The number in parenthesis after the co-author's name shows the number of joint publications. The gray and red bars are selectable, and the respective publications are displayed in the sidebar. This visualization provides an easy way to explore temporal variations in the collaborations, possible shift of collaborators from one co-author to another, and the span of joint work.

6 Application Examples

Mirroring the two initial motivating scenarios (cf. Section 2.2), we explore two use cases for our approach. They demonstrate and test the approach in a realistic setting.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2865022, IEEE Transactions on Visualization and Computer Graphics

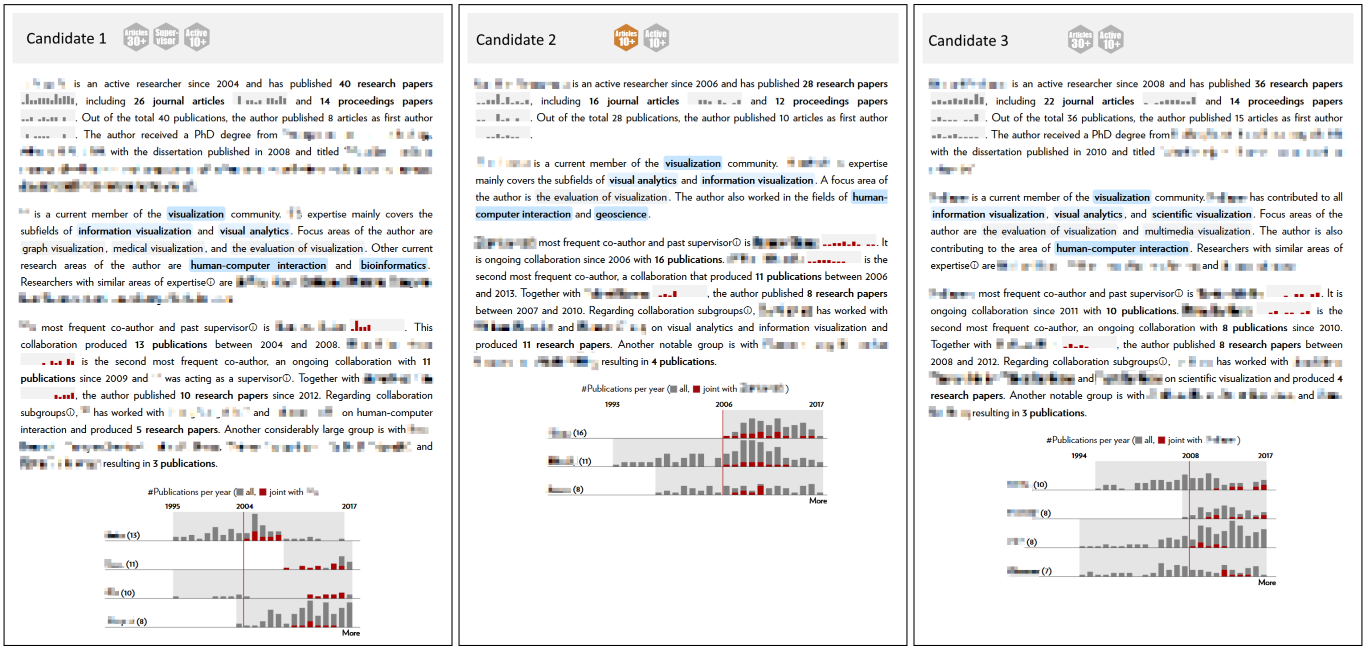

Fig. 6. Anonymized main text and co-author visualization of the job candidates' profiles (Application Example S1).

S1 – Recruiting

In this use case, we simulate a recruiting process for an academic position. To have a specific scenario that fits our scope, we assume to recruit an associate-level professor for a position in visualization research, embedded into a research institute on *human-computer interaction* at a university with a certain focus on *geoscience*. The candidate should hold a PhD degree and afterwards have already developed an independent research agenda. Experience in supervising and interdisciplinary research would be counted as an advantage. Considering the specialization of our hypothetic institute and university, cross-links to *human-computer interaction* and *geoscience* would be a good fit. Further evaluation criteria are an active role in publishing with a special focus on visualization as well as having a broad collaboration network. Since we are evaluating personal achievements of individual researchers, we anonymize the discussed reports.

Data Sampling: To draft realistic candidates, we randomly sample the authors in our data set (in a real scenario, these would be the submitters of an application). For each candidate, we identify the category to which the author belongs based on the author personas. We only include authors who fall between categories *researcher* and *senior researchers* as candidates because they are at the targeted stage of career. This simulates a preselection where less fitting candidates are excluded. We restrict our sample to three candidates for sake of brevity. Figure 6 lists the respective anonymized profiles.

Results: The badges give a first impression on the candidates, having a total number of publications (**IN 1.1**) between 10 and 30 (bronze badge, Candidate 2) or between 30 and 100 (silver badge, Candidate 1 and 3), and being active (**IN 1.2**) at least ten years (silver badge). Distinguished from the other candidates, Candidate 1 already has documented experience in supervising (**IN 3.3**, silver badge). The first paragraph provides more detailed information like exact publication counts (28–40, **IN 1.1**), year of the first publication (2004–2008, **IN 1.2**) and temporal publication distribution (stable to growing, **IN 1.2**), as well as the balance between journal and proceedings publications (slight preference on journal articles for all candidates, **IN 1.1**)—the candidates perform relatively similar with respect to these measures. All candidates also published a good share of publications as first authors, indicating a successful doctoral (and postdoctoral) work phase (**IN 1.3**). With respect to topics described in the second paragraph, all candidates are frequent and recently active contributors of the visualization community (**IN 2.1**), with similar focus (**IN 2.2**) on *information visualization* and *visual analytics* (Candidate 3 also has some experience in *scientific visualization*). They all share an interest in *the evaluation of visualization* as well as *human-computer interaction* (**IN 2.1**), while Candidate 1 seems to have published most in the latter as can be retrieved when clicking on the keyword. Candidate 2 has also considerably published in the desired secondary field of *geoscience*. The co-author visualizations of all three candidates show that they have been publishing frequently with various experienced researchers (**IN 3.2**), but did not rely on a single main collaborator (**IN 3.1**) or subgroup of co-authors (**IN 3.4**). Candidate 1 already has a supervisee among the top collaborators (**IN 3.3**).

Conclusion: Judging from their generated profiles, Candidate 1 might be the most promising applicant due to experience and fit of topics. However, the differences are nearly balanced. Candidate 2 would even better fit topic-wise, while Candidate 3 is youngest with respect to academic age and has a more quickly growing record of publications. Hence, a recommendation would be to invite all three candidates for a job interview. As is common in some countries, a follow up step could be to find independent reviewers for the candidates among the authors with similar research interests (**IN 2.4**).

S2 – Assessing Expertise

The second application example plays through organizing a new workshop on a visualization topic, *biological visualization* for the sake of demonstration. The workshop should cover various topics of biological visualization, including scientific visualization like protein representations as well as visual analytics approaches and information visualizations for biological networks. One important task is finding candidates for the program committee, which is an instance of the described scenario of assessing expertise (S2). We aim at inviting a small committee of 10 candidates. They should have good expertise in the area and academic experience at least on *researcher* level.

Data Sampling: We start with randomly sampling authors to find currently active researchers who have both contributed to *visualization* and *bioinformatics*, until we have collected a starting set of six seed authors. In a real setting, this starting set could as well come from random sampling, but also from personal recommendations or from keyword searches in digital libraries. We use the author similarity (**IN 2.4**) and co-author relationships (**IN 3.1**) to expand the list of candidates (not following transitive connections because we would sample many too similar candidates). This procedure resulted in a set

Table 2. Suggested program committee candidates for an imaginary event on biological visualization (Application Example S2). Personas: R – *researcher* and SR – *senior researcher*; subfields: IV – *information visualization*, SV – *scientific visualization*, and VA – *visual analytics*.

Candidate	Persona	Visualization	Subfield	Bioinformatics
Ghassan Hamarneh	SR	+		+++
Jing Hua	R–SR	++	IV, SV	+++
Igor Jurisica	R–SR	++	IV	+++
Silvia Miksch	SR	+++	IV, VA	+++
Thorsten Möller	SR	+++	IV, SV	+
Klaus Mueller	SR	+++	IV, SV, VA	++
Adam Perer	R	+++	IV, VA	++
Falk Schreiber	SR	+++	IV, VA	+++
Roberto Thern	R–SR	++	IV, VA	++
William M. Wells III	SR	+		+++

of 22 authors, from which we select the 10 most suitable candidates. We restrict ourselves to only using *VAP* as a source of information.

Results: For every sampled author, we look at the badges and read the first text paragraph to assign a persona mainly based on publication count (**IN 1.1**), academic age (**IN 1.2**), year of PhD graduation (if available, **IN 1.3**), and supervisor experience (**IN 3.2**). Then, we check the visualization expertise (**IN 2.1**): Clicking on the keyword, we retrieve the set of relevant publications. Considering frequency, quality of venue, and recency (**IN 2.3**), we rate the experience in three levels (+, ++, and +++). We also list the visualization subfield covered by the authors as described in the text (**IN 2.2**). Further, we also judge the experience with respect to *bioinformatics* by studying the respective publications (**IN 2.1**). Table 2 presents the resulting invitation list, while an extended list of all sampled authors and all related reports are available as part of the interactive appendix document.

Conclusion: The approach helped quickly finding committee candidates that are sufficiently experienced and well balanced between the two targeted communities. While some candidates are equally active in both fields, we have selected a few candidates who are specialists in one field with only occasional contributions in the other. Also, the candidates cover all subfields of visualization. The procedure furthermore yielded four secondary candidates (see interactive appendix) who could be invited if some of the primary candidates would decline.

7 Discussion

We have demonstrated that our approach fulfills the required information needs, but there are also other aspects to discuss that are either strengths or limitations of the approach. Thinking beyond the specific example of the presented system, we also make suggestions for the application and generalization of the approach to other domains.

Selection of Information: A strength of our approach is that it only shows information available in a sufficient quality and quantity. For instance, we do not speculate on research topics of an author if keyword information is only available for a single publication. Of course, such checks can be included also into other approaches. However, in a tabular or visual representation such missing information might cause confusion. In text, it seems more natural to only focus on those things that are most relevant and reliable. One might argue that this selection takes away the user's control, but we counterbalance by providing explanations and making the underlying data available on demand.

Data Scope: The scope of our work is restricted to publication data. We did not yet include related information such as affiliations, citation information, or awards. Whereas reliable data available is a challenge, adding paragraphs for this information to the author profiles would be a natural extension. Unlike table-based or visualization approaches, where adding more data aspects could result in a more cluttered representation, the growth of a textual description usually does not reduce how self-explaining or understandable it is. In some cases, author profiles are inaccurate because publications of several people with the same name are mixed—DBLP often (not always) provides a correct name disambiguation reflected in additional indices attached to the author name.

Evaluation: The assessment of existing approaches and the application examples demonstrate that our approach fills a relevant gap and produces useful results in realistic scenarios. The first example (S1 – Recruiting) demonstrates the usefulness of our approach in an explanatory scenario. The second example (S2 – Assessing Expertise) shows the exploratory analysis aspect of our tool that can be used to identify researchers working on similar topics. However, we cannot yet claim that our approach performs better than a purely table-based or visualization-based representation. For such comparison, we will need to first develop the other representations showing equivalent information and optimize each as we have optimized the presented approach. Then, we can perform a user study comparing the three representations. While a quantitative study could answer which of the representations is best in terms of accuracy and answer times, a qualitative approach could also reflect on how self-explaining the representations are and how the users work with them. Furthermore, textual explanations might influence the way users interpret the visualizations. Whereas a recent study [20] investigates the effect of diagram titles on the interpretation of visualizations, the impact of longer textual explanations on the accompanying visualizations remains yet to be explored.

Applicability: In our approach, we use the generated textual representation as a primary entity to demonstrate the capabilities of this rarely leveraged representation. However, we do not argue that our representation needs to replace conventional list-based author records or author visualizations. On the contrary, our technique would be simple to apply in various contexts, for instance, as an introductory text of an author profile in a digital library or as a detail description of an author node in a co-author network visualization. Following this idea, our technique could enrich existing approaches without the threat that the additional information would clutter the interface. Moreover, the generated texts are rather likely to make a formal representation more self-explaining and understandable.

Generalizability: We have developed a specific solution for a specific data set and domain. Despite some tailoring for the visualization community, most of the generated text would still work for other DBLP publication data. Increasing with difference in publication culture, more adaptation would become necessary when switching to other fields. For instance, some research communities might use alphabetic ordering of authors that renders our current algorithm unable to identify supervisor relationships. Still, the seniority check in our approach (supervisor is at least five years senior than the supervisee) holds across publication cultures and might convey this information to some extent. Whereas developing a system uniting various sciences might be challenging, tailored solutions for specific fields are straightforward to derive. Only the phrasing of the text and calibration of algorithms might need to be adapted. In a similar way, we can also build profile pages for actors, musicians, software developers, or other people taking part in public life. The only requirement is that these people work together and contribute to specific work items.

8 Conclusion

We presented a novel approach to use natural language text combined with visualizations for presenting an author-centric analysis of publication records. This visualization–text integration led to the development of *VIS Author Profiles*, a Web-based visual analytics tool to explore the profiles of researchers. It provides an easy-to-use and self-explanatory interface for understanding publication statistics, research interests, and collaboration networks. A distinguishing feature is the interactive linking of text and visualizations where text explicitly describes the most important patterns and visualization allows for contextual exploration of publication records. We demonstrated the approach in two realistic scenarios. Although designed for a specific domain, it can be extended to other scientific communities and excerpts of the generated text might be integrated into other systems and visualizations.

Acknowledgments

Fabian Beck is indebted to the Baden-Württemberg Stiftung for the financial support of this research project within the Postdoctoral Fellowship for Leading Early Career Researchers.

REFERENCES

[1] M. Abdelaal, F. Heimerl, and S. Koch. ColTop: Visual topic-based analysis of scientific community structure. In *Proceedings of the 2017 International Symposium on Big Data Visual Analytics*, BDVA, pp. 1–8. IEEE, 2017. doi: 10.1109/BDVA.2017.8114622

[2] E. Adar, C. Gearig, A. Balasubramanian, and J. Hullman. PersaLog: Personalization of news article content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 3188–3200. ACM, 2017. doi: 10.1145/3025453.3025631

[3] F. Beck, S. Koch, and D. Weiskopf. Visual analysis and dissemination of scientific literature collections with SurVis. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):180–189, 2016. doi: 10.1109/TVCG.2015.2467757

[4] F. Beck, H. A. Siddiqui, A. Bergel, and D. Weiskopf. Method Execution Reports: Generating text and visualization to describe program behavior. In *Proceedings of the 5th IEEE Working Conference on Software Visualization*, pp. 1–10. IEEE, 2017. doi: 10.1109/VISSOFT.2017.11

[5] F. Beck and D. Weiskopf. Word-sized graphics for scientific texts. *IEEE Transactions on Visualization and Computer Graphics*, 23(6):1576–1587, 2017. doi: 10.1109/TVCG.2017.2674958

[6] S. Demir, S. Carberry, and K. F. McCoy. Summarizing information graphics textually. *Computational Linguistics*, 38(3):527–574, 2012. doi: 10.1162/COLI_a_00091

[7] M. Dörk, N. Henry Riche, G. Ramos, and S. Dumais. PivotPaths: Strolling through faceted information spaces. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2709–2718, 2012. doi: 10.1109/TVCG.2012.252

[8] N. Elmqvist and P. Tsigas. CiteWiz: a tool for the visualization of scientific citation networks. *Information Visualization*, 6(3):215–232, 2007. doi: 10.1057/palgrave.ivs.9500156

[9] P. Federico, F. Heimerl, S. Koch, and S. Miksch. A survey on visual approaches for analyzing scientific literature and patents. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2179–2198, 2017. doi: 10.1109/TVCG.2016.2610422

[10] T.-L. Fung, J.-K. Chou, and K.-L. Ma. A design study of personal bibliographic data visualization. In *Proceedings of the 2016 IEEE Pacific Visualization Symposium*, PacificVis, pp. 244–248. IEEE, 2016. doi: 10.1109/PACIFICVIS.2016.7465279

[11] B. Ganter and R. Wille. *Formal concept analysis: mathematical foundations*. Springer Science & Business Media, 1st ed., 2012. doi: 10.1007/978-3-642-59830-2

[12] A. Gatt and E. Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018. doi: 10.1613/jair.5477

[13] P. Goffin, W. Willett, J. D. Fekete, and P. Isenberg. Exploring the placement and design of word-scale visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2291–2300, Dec 2014. doi: 10.1109/TVCG.2014.2346435

[14] P. Goffin, W. Willett, J.-D. Fekete, and P. Isenberg. Design considerations for enhancing word-scale visualizations with interaction. In *Posters of the Conference on Information Visualization*. IEEE, 2015.

[15] Q. He, M. Zhu, B. Lu, H. Liu, and Q. Shen. MENA: Visual analysis of multivariate egocentric network evolution. In *Proceedings of the 2016 International Conference on Virtual Reality and Visualization*, ICVRV, pp. 488–496. IEEE, 2016. doi: 10.1109/ICVRV.2016.88

[16] N. Henry, J.-D. Fekete, and M. J. McGuffin. NodeTrix: a hybrid visualization of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1302–1309, 2007. doi: 10.1109/tvcg.2007.70582

[17] T.-H. Huang and M. L. Huang. Analysis and visualization of co-authorship networks for understanding academic collaboration and knowledge domain of individual researchers. In *Proceedings of the 2006 International Conference on Computer Graphics, Imaging and Visualisation*, pp. 18–23. IEEE, 2006. doi: 10.1109/CGIV.2006.20

[18] J. Hullman, N. Diakopoulos, and E. Adar. Contextifier: automatic generation of annotated stock visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, pp. 2707–2716. ACM, 2013. doi: 10.1145/2470654.2481374

[19] P. Jordan, N. L. Green, C. Thomas, and S. Holm. TBI-Doc: Generating patient & clinician reports from brain imaging data. In *Proceedings of the 8th International Natural Language Generation Conference*, pp. 143–146. Association for Computational Linguistics, 2014.

[20] H.-K. Kong, Z. Liu, and K. Karahalios. Frames and slants in titles of visualizations on controversial topics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 438. ACM, 2018. doi: 10.1145/3173574.3174012

[21] T. Kurosawa and Y. Takama. Co-authorship networks visualization system for supporting survey of researchers future activities. *Journal of Emerging Technologies in Web Intelligence*, 4(1):3–14, 2012. doi: 10.4304/jetwi.4.1.3-14

[22] S. Latif, D. Liu, and F. Beck. Exploring interactive linking between text and visualization. In *EuroVis 2018 - Short Papers*, pp. 91–94. The Eurographics Association, 2018. doi: 10.2312/eurovisshort.20181084

[23] Q. Liu, Y. Hu, L. Shi, X. Mu, Y. Zhang, and J. Tang. EgoNetCloud: Event-based egocentric dynamic network visualization. In *Proceedings of the 2015 IEEE Conference on Visual Analytics Science and Technology*, VAST, pp. 65–72. IEEE, 2015. doi: 10.1109/VAST.2015.7347632

[24] I. Milligan. Illusionary order: Online databases, optical character recognition, and Canadian history, 1997-2010. *Canadian Historical Review*, 94(4):540–569, 2013. doi: 10.3138/chr.694

[25] V. O. Mittal, S. F. Roth, J. D. Moore, J. Mattis, and G. Carenini. Generating explanatory captions for information graphics. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, IJCAI, pp. 1276–1283. Morgan Kaufmann Publishers Inc., 1995.

[26] A. Ramos-Soto, A. J. Bugarin, S. Barro, and J. Taboada. Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data. *IEEE Transactions on Fuzzy Systems*, 23(1):44–57, 2015. doi: 10.1109/TFUZZ.2014.2328011

[27] A. Ramos-Soto, B. Vazquez-Barreiros, A. Bugarín, A. Gewerc, and S. Barro. Evaluation of a data-to-text system for verbalizing a learning analytics dashboard. *International Journal of Intelligent Systems*, 32(2):177–193, 2017. doi: 10.1002/int.21835

[28] E. Reiter, R. Dale, and Z. Feng. *Building natural language generation systems*. MIT Press, 2000.

[29] F. Reitz. A framework for an ego-centered and time-aware visualization of relations in arbitrary data repositories. *arXiv preprint arXiv:1009.5183*, 2010.

[30] L. Shi, C. Wang, Z. Wen, H. Qu, C. Lin, and Q. Liao. 1.5D egocentric dynamic network visualization. *IEEE Transactions on Visualization and Computer Graphics*, 21:624–637, 2015. doi: 10.1109/TVCG.2014.2383380

[31] E. R. Tufte. *Beautiful Evidence*. Graphics Press, 1st ed., 2006.

[32] W. Wahlster, E. André, W. Finkler, H.-J. Profitlich, and T. Rist. Plan-based integration of natural language and graphics generation. *Artificial Intelligence*, 63(1-2):387–427, 1993. doi: 10.1016/0004-3702(93)90022-4

[33] R. Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. In *Ordered Sets*, pp. 445–470. Springer, 1982. doi: 10.1007/978-94-009-7798-3_15

[34] Y. Wu, N. Pitipornvivat, J. Zhao, S. Yang, G. Huang, and H. Qu. egoSlider: Visual analysis of egocentric network evolution. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):260–269, 2016. doi: 10.1109/TVCG.2015.2468151

[35] K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI, pp. 401–408. ACM, 2003. doi: 10.1145/642611.642681

[36] J. Zhao, C. Collins, F. Chevalier, and R. Balakrishnan. Interactive exploration of implicit and explicit relations in faceted datasets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2080–2089, 2013. doi: 10.1109/TVCG.2013.167