Fun with FUN

Fabien Mathieu, Sébastien Tixeuil

Fun with FUN

Fabien Mathieu
Swapcard, Paris, France

Sébastien Tixeuil[1]
Sorbonne Université, CNRS, LIP6, Paris, France

Abstract

The notions of *scientific community* and *research field* are central elements for researchers and the articles they publish. We propose to explore the evolution of the FUN conference community since its creation from the articles listed in DBLP, authors, program committees, and advertised themes, by means of a novel symmetric embedding, and carefully crafted software tools. Our results make it possible on the one hand to better understand the evolution of the community, and on the other hand to easily integrate new themes or researchers during future editions.

2012 ACM Subject Classification Information systems → Similarity measures; Information systems → Information extraction

Keywords and phrases Natural Language Processing, Relevance Propagation, Bibliometry, Community, Scientific Fields

Digital Object Identifier 10.4230/LIPIcs.FUN.2022.21

Supplementary Material *Software (Source Code)*: `https://github.com/balouf/conference_analysis/tree/main/FUN`; archived at `swh:1:dir:8b96e766c426c12f0b9a7901b52230bc09d04a8b`

Acknowledgements This work was done at LINCS (`https://www.lincs.fr/`).

1 Introduction

Understanding the progress of science by studying how new scientific knowledge is created is an important societal challenge. In particular, there are many studies regarding how scientific knowledge spreads through scientific literature. For example, the field of bibliometrics is concerned with measuring the properties of the research corpus, and leads to measures of importance, based on notions such as the number of citations of an article, the impact factor of a journal, and an author's h-index. Furthermore, the social aspects associated with scientific research, such as the *sociology of scientific knowledge*, including the social structures and processes of scientific activity, as well as its political aspects, have also been the subject of studies [6]. The reader interested in such topics can benefit from the recent textbook by Wang and Barabási [12] and references therein.

In this article, we are interested in the analysis of scientific communities, from both a spatial and temporal point of view. The spatial (that is, community structure and/or habits) dimension is often assessed via the researchers' co-publications graph [7] to obtain scientific communities through clustering, but also thematically, for example by examining the vocabulary used in the main text of published articles [3]. The temporal dimension accounts for the evolution of the featured aspects measured over time. Most often, the temporal dimension concentrates on research impact or on evolution of collaborating relations [12].

Our focus in this paper is orthogonal. We focus on a known scientific community (here, the community involved in the many editions of the FUN conference), and aim to assess the evolution of its researchers (featured as conference authors or as PC members) and its associated themes (as described in the calls for papers, but also those associated to the PC

[1] Corresponding author.

members and authors of various editions) over the lifespan of the conference. This simple motivation raises several intriguing questions: how to map a researcher to her research themes? how to map research themes to researchers? how to represent the evolution of mappings over time?

Our contribution, for which the source code of the implementation is public, is threefold:

1. we present a new symmetric embedding augmented with a random walk mechanism to obtain a powerful cross mapping mechanism;
2. we design a methodology to collect and iterate modifications on data from various sources about a given conference or journal in IT to obtain spatio-temporal information about the venue, using the previously defined embedding;
3. we collect data for the FUN conference, apply our methodology, and describe our findings about the evolution of the conference since its creation.

We argue that our methodology is not only valuable to observe the past of a scientific community, but also to envision its future (*e.g.* by strategically choosing future PC member with respect to promoted themes).

The rest of the paper is organized as follows. Section 2 describes our new embedding. Section 3 presents our methodology to apply the embedding to a scientific venue, using the FUN conferences as running example. Section 4 describes the actual outputs of our methodology applied to FUN. Finally, Section 5 provides some concluding remarks.

2 Embedding documents and words, and vice versa

Searching for documents and extracting relevant information out of them is a task everyone faces on a regular basis. Common examples include looking for a relevant Internet page, digging a crucial e-mail from thousands of unread messages, or finding out relevant papers for a given topic. An effective search typically relies on a precise and well-organized search engine. The majority of current search engine techniques combine a keyword search with structural information (ontologies, relationships between elements) to order the documents in a corpus by relevance.

To increase the search performance, a standard approach consists in *embedding* documents or keywords, i.e. representing them by vectors in some space. A popular such example is *Term Frequency, Inverse Document Frequency* (TF-IDF) (further described in Section 2.1). In this section, we propose a novel approach that extends TF-IDF to enhance the quality and the flexibility of the embedding. This new approach will be instrumental in the analysis of the FUN community. This section is organized as follows: Section 2.1 presents the TF-IDF embedding; Section 2.2 introduces our symmetrized version called *Term Frequency - Inverse Document & Term Frequency* (TF-IDTF); Section 2.3 adds a random-walk approach on top of TF-IDTF to refine the quality of the embedding.

2.1 Texts are made of words

TF-IDF is a metric commonly used in language processing to estimate the importance of a word based on its frequency in a document and its rarity in a corpus. The metric postulates that a document can be represented by the set of the words it contains and that a rare word in the corpus is more important that a common one.

In details, consider a corpus X made of n documents. These documents are made of words. Let Y be the set of the m (unique) words that are present in the corpus. We can build a simple bipartite graph $G = (X \cup Y, E)$ of components X and Y by creating an (undirected)

edge between $x \in X$ and $y \in Y$ if, and only if, x contains the word y. If $x.y$ denotes the number of occurrences of y in x (the frequency of y in x), TF-IDF attributes the following weight to the edge (x, y):

$$w(x, y) = x.y \log\left(\frac{n}{d(y)}\right), \text{ where } d(.) \text{ is the degree in } G. \tag{1}$$

Using Equation (1), one can associate to each document x the m-dimensional vector $\vec{x}$ on Y defined by $\vec{x}_y = w(x, y)$. This is called an *embedding* of X into Y. This embedding is usually *sparse*: if the corpus is large and various, a typical document contains a fraction of the available words, so most components of its embedding are null.

Embeddings have many uses, including the possibility to measure the *cosine similarity* between two documents. If x_1 and x_2 are two documents of X, their cosine similarity is:

$$\text{sim}(x_1, x_2) = \frac{\vec{x_1}.\vec{x_2}}{||\vec{x_1}||_2.||\vec{x_2}||_2}. \tag{2}$$

Despite its simplicity and the loss of meaning induced by reducing a text to a *bag of words*, the embedding induced by Equation (1) is surprisingly effective in practice. Several explanations have been proposed to explain the success of TF-IDF. One of the most elegant ones comes from information theory [1]: assume that I need to find one specific document among a large set. I know that the document contains a specific word, so I can restrict the search to the documents that contain it. If only the desired document contains the word, the search has been made trivial by the knowledge of the word. If all documents contain the word, the knowledge is useless. In general, if one expresses the quantity of information, in the sense of information theory, brought by the knowledge of the presence of the word, we get the *Inverse Document Frequency* (IDF) of the word (the logarithmic term in Equation (1)).

In other words, $w(x, y)$ can be seen as the product of the strength of the relationship between x and y (estimated by $x.y$) by the quantity of information given by y (expressed by the inverse document frequency).

2.2 Words are defined by texts

The graph G we consider in Section 2.1 represents an inclusion relationship, which is asymmetric. However, the graph itself is a simple graph, and contains no information to distinguish the *documents* part X from the *words* part Y. For example, if we represent each document x of X by a unique word y' (not necessarily from Y), and each word y of Y by a document x' made by assembling the representatives of all documents that contain y, we end up with the original graph G except that X and Y are reversed. Based on this symmetry between X and Y, it is natural to investigate what happens if one switches them in Equation (1).

This would introduce $\log\left(\frac{m}{d(x)}\right)$ (the logarithm of the ratio between the total number of unique words in X and the total number of unique words in x).

This *Inverse Term Frequency* (ITF) can be interpreted as a dual version of IDF: while IDF relies on the assumption that a document is defined by the words it contains, and favors scarcity (of words), ITF relies on the assumption that a word is defined by its context (the documents that contain it), and favors conciseness (of documents). So, if a document is concise and contains only a few words, knowing that a word belongs to that document gives a lot of information for separating that specific word from the others. Conversely, knowing that a word belongs to a lengthy document with many distinct words gives little information about the meaning of that specific word.

Based on this observation, we introduce a refinement of TF-IDF that we call TF-IDTF. TF-IDTF attributes the following weight to the edge (x, y):

$$w(x,y) = (1 + \log(x.y)) \log\left(\frac{1+n}{1+d(y)}\right) \log\left(\frac{1+m}{1+d(x)}\right) \text{ if } y \in x, \text{ 0 otherwise.} \quad (3)$$

Compared to Equation (1), the ITF is added to the formula, to convey the information that document x brings about word y. Equation (3) also introduces the following classical modifications (cf. for example the TF-IDF implementation in the scikit-learn library [10]):

- $x.y$ is logarithmically smoothed to limit possible over-representation when a word is widely used in a document.
- A shift of one unit is introduced in the expressions of the inverse frequencies. It corresponds to the addition of a fictitious document containing all the words, and of a fictitious word appearing in all the documents. These additions smooth the weights.

The introduction of the ITF does not drastically change the resulting embedding of documents. In fact, the embedding of one document is just scaled by its ITF, and in particular the cosine similarity is unaffected. The real change is that the new weights provides a new dual embedding, not on documents but on *words*: one can associate to each word y the n-dimensional vector $\vec{y}$ on X defined by $\vec{y}_x = w(x, y)$. With this embedding, two words are considered close if they are often co-occurrent in documents, with more importance given to co-occurrence in short documents. This allows for example to identify words that belong to the same lexical field. Just like the IDF weighting prevents frequent words from polluting the embedding of documents, the ITF weighting prevents lengthy documents from polluting the embedding of words.

The two embeddings (document and word) can be unified by considering the $n \times m$ matrix W defined by $W_{x,y} = w(x, y)$: each line (resp. each column) of W represents the embedding of a document (resp. a word) in Y (resp. in X).

Now that we have the tools to compare two documents or two words, we will see how to compare words and documents.

2.3 The friend of my friend is my friend

One major caveat of embeddings like TF-IDF (or TF-IDTF) is that they are blind to synonyms: if one document on graphs uses the term *node* and another the term *vertex*, these two words will make the embeddings of the documents less similar although it should be the opposite. Another issue is the impossibility to directly compare a word and a document beyond the term frequency metric.

To address these issues, we propose to refine the embedding of documents and words by considering a random walk on G. The idea is that through our symmetric document embedding, any vector on X, seen as a weighted set of documents, can be turned into a vector on Y. Reciprocally, any vector on Y, seen as a weighted set of words, can be turned into a vector on X by the word embedding. In addition to allowing translations between embeddings, this process may enable *idea associations*: if one document on graphs uses the terms *node*, *edge*, and *graph* and another the terms *vertex*, *edge* and *graph*, going back and forth between words and documents will uncover the similarity between *node* and *vertex* as words that belong to documents that contain the words *edge* and *graph*. With this approach, it is possible to find out the similarity between two documents that use the same lexical field, even if they have no word in common (so, their similarity is 0 according to Equation (2)). This mitigates (without fully nullifying) the impact of synonyms.

This idea of extracting information from graph exploration is highly reminiscent of the *random surfer* model used in *PageRank*, the algorithm behind the Google search engine [2]. In fact, PageRank has already been used in language processing to produce extractive summaries of documents [9]. Note that in the case of extractive summaries, the walk transitions from documents to words are often weighted according to IDF to avoid pollution of frequent words. However, transitions from words to documents are usually chosen uniformly, which can induce pollution from lengthy documents. As we do not want the pure length of a document to be a competitive advantage in the random walk, we propose to use TF-IDTF to handle both types of transitions (from documents to words, and from words to documents).

For the actual random walk computation, we propose to use D-Iteration [5], a PageRank variation adapted to the exploration of the neighborhood of a part of a graph. Intuitively, D-iteration consists in diffusing a finite quantity of evanescent fluid on the graph vertices from an initial distribution (like a word or a document) and measuring the quantity of fluid flowing through the vertices.

The D-Iteration algorithm takes as parameters a stochastic matrix A (which represents a random walk on a graph), a vector Z that represents the start of the walk, and an attenuation coefficient $\alpha \in (0, 1)$. In our case, the matrix A derives from W, Z derives from the words or documents one wants to analyze, and α is a parameter to be chosen carefully.

In detail, A is a matrix of size $(n + m) \times (n + m)$ defined by $A = \begin{pmatrix} 0_{n,n} & S(W) \\ S(W^t) & 0_{m,m} \end{pmatrix}$, where $S(M)$ denotes the stochastic renormalization of M, which consists of dividing each non-zero line of a positive matrix M by the sum of its elements. A defines a random walk on G where each edge is chosen proportionally to its TF-IDTF weight.

If z represents a set of words (or documents) one wants to analyze, possibly weighted, it can be represented by a vector Z on $X + Y$ whose components are equal to 1 (or their weight) if they correspond to an element of z, and 0 otherwise.

The D-Iteration algorithm associates to z a vector $P(z)$ defined by :

$$P(z) = \sum_{k \geq 1} \alpha^k Z A^k. \tag{4}$$

Each term of the sum corresponds to a random walk of length k from Z, weighted by α^k. This implies that the average length of the random steps that $P(z)$ aggregates is $\frac{1}{1-\alpha}$. In other words, the attenuation coefficient α controls the span of the graph exploration, which is subject to a trade-off: when α is close to 0, the walk lengths are close to 1, and $P(z)$ converges to a straight TF-IDTF embedding of z. Conversely, when α is close to 1, the walk lengths are long. This enables idea associations, but also "blurs" the result: the underlying Markov chain, while not being ergodic (the graph is bipartite, hence periodic with period 2), is irreducible on its connected components, which means that all information but the starting bipartite component are forgotten on long walks.

The vector $P(z)$ has multiple uses. One possibility is to look at the components with the greatest values to get the documents and words that are the closest to z. Another possibility is to treat $P(z)$ as an embedding of z in the space $X + Y$. Using a random walk of finite length to represent a weighted subset of vertices in a graph is not a new approach [4, 11]. However, as stated above, the bipartite property of the graph creates a natural asymmetry depending on whether the starting point of the walk is a document or a word. For example, if we start from a document, walks of even length always return documents, and walks of odd length always return words. This induces a natural distortion that makes it difficult to compare a document with a word: by design, the embedding of a word (resp. a document)

obtained with Equation (4) systematically gives more weights to documents (resp. words). To mitigate this effect, we renormalize the components of $P(z)$ on documents and words so that the total weight on words is always equal to the total weight on documents.

3 Embedding FUN

Section 2 exposed a generic way to embed arbitrary documents and their content. In this section, we expose how to use and adapt those general mathematical tools to analyze the FUN community. The originality of our approach lies in the fact that instead of trying to compute the supposed importance of a community or a theme compared to another, we want to characterize the links of similarity that connect them. For this purpose, we employ the methodology summarized in Figure 1:

- From the *Digital Bibliography & Library Project* (DBLP) database, we extract a community subset centered around FUN;
- From the community subset, we build a graph between researchers and words that allows to represent them in the same space, and thus to compare them;
- By calculating the cosine similarity between the vectors, we investigate the links between authors, program committees, and themes.

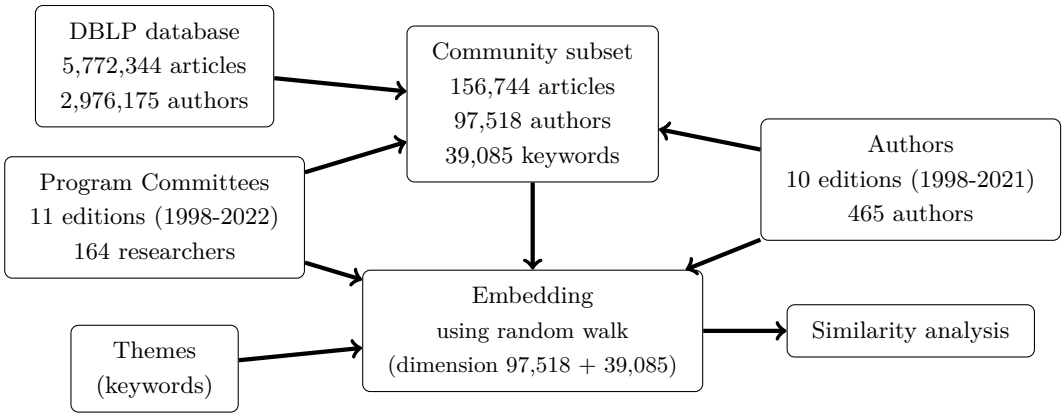


Figure 1 Our methodology in a nutshell.

The rest of the section is organized as follows: Section 3.1 presents DBLP and the actual dataset that is used; Section 3.2 describes the use of authors instead of words to describe a scientific paper and the extraction of a community graph; Section 3.3 explains how to merge author and word descriptions altogether; lastly, Section 3.4 presents the implementation of our methodology. The actual analysis is presented in Section 4.

3.1 Judging a book by its cover

The DBLP project indexes English-language publications in the IT field. The database, publicly available at the address `https://dblp.uni-trier.de/xml/dblp.xml.gz`, contains the majority of bibliographical references in IT. In particular, for each bibliographical reference, the database contains its title and authors. At the time this article was written, the database referenced 5,772,344 articles written by 2,976,175 unique authors.

Sadly, DBLP does not provide in its downloadable version any information on the content of the articles beyond their title. In particular, the paper abstract, introduction, and keywords are missing. At first glance, it seems highly insufficient to carry out a relevant analysis,

which is true if one wants to analyze one single article. Luckily, our approach uses articles by batches, like all articles written by a researcher, so we can hope that quantity will compensate the noisy semantic quality conveyed by titles to identify highlighting trends.

However, for the particular case of the FUN conference, the hypothesis that a paper title conveys information is hindered. Indeed, FUN articles are typically written in a fun and amusing way, and it reflects on their titles. What themes can we infer from titles like *Kings, Name Days, Lazy Servants and Magic* or *Urban Hitchhiking*? For this reason, if one wants to study the FUN community, it is probably better to consider the people of that community instead of the titles of the articles they publish in FUN (but still consider the titles of the papers they write outside FUN).

3.2 Articles are made of authors, and vice versa

Traditionally, the elements of an embedding are called *features*, as they yield a description of a document. Obviously, the authors of an article give information on that article. For example, they can hint at the topic of the article. If one uses authors instead of words as features of the articles, can we re-interpret TF-IDTF?

- We assign a term frequency of 1 for all authors, which means that we assume that all authors have the same importance for a considered article. This is of course debatable, but given the limited amount of information available, this is the only sensible choice, and we can hope that this is true *in average*.
- The IDF term means that for a given article, more weight is given to authors that have few publications. This reduces the natural bias towards prolific authors. We emphasize that this is no indication of the intrinsic value of people, but just a measure of how much an article can be characterized by the presence of a given author. A starting researcher has a big weight as it narrows down the corpus of articles to a small subset. Paul Erdös (about 1,500 publications) has low weight as the corpus reduction is smaller. Didier Raoult has 4 articles referenced in DBLP so he has a big weight on the corpus of IT publications; on the other hand with more than 3,100 publications referenced in PubMed[2], he has a very low weight on the corpus of medical publications.
- The ITF term means that for a given author, more weight is given to the articles that have few authors. Observe that this reduces the natural bias towards articles with many authors. The interpretation, in terms of information theory, is that with less co-authors, or no co-author at all, an article is more representative of the production of a researcher. Conversely, it is likely that an article with dozens of co-authors provides little information on the profile of one single author.

All things considered, we can see that all the reasoning behind the introduction of TF-IDTF also makes sense if one considers authors instead of words.

Based on this observation, we carried out a filtering of the complete database centered on the FUN community. For this, we collected all the program committees and authors of the different editions, and manually disambiguated each of them with respect to their DBLP entry (homonyms or near homonyms have specific entries in the database). For each edition, we computed $P(z)$ using a default value $\alpha = 1/2$ (cf Equation (4))[3], and we aggregated the

[2] `https://pubmed.ncbi.nlm.nih.gov/`

[3] The value $\alpha = 1/2$ is chosen empirically. It is based on the quality of the results roughly assessed by the authors of this paper. To give a sense of comparison, extractive summaries typically use smaller values for α, usually less than 0.1 (see e.g. the parameter $1 - d$ used by Otterbacher et al. [9]), to remain very close to the starting point, while the original PageRank algorithm uses the empirical value $\alpha = 0.85$ to

most valued articles according to each embedding. This allowed us to select 156,744 articles representing the FUN community and its (large) neighborhood. The goal of this reduction was twofold: first, improving computation time by working on a smaller dataset; second, improving the corpus quality with respect to the FUN community by removing irrelevant publications from unrelated fields.

3.3 The words of my papers are my words

From the articles of the FUN community, we can extract two bipartite graphs with stochastic transitions: one between articles and words, one between articles and researchers. As said before, we have little interest in the articles themselves, which convey little information, so it is natural to combine the two graphs into a new bipartite graph (with stochastic transitions) between researchers and words. The new graph links each researcher to all words she used in her articles, with more weights on the rare words that appear on articles with few co-authors. Conversely, each word is linked to all researchers that have used it, with more weight on the non-prolific researchers that prefer small titles. This graph will be used to perform the embedding for our analysis.

We did not take into account the years of publication when constructing the graphs. In particular, this means that each author is analyzed throughout their entire career, even if their own research interests have evolved.

3.4 Our implementation

To perform the actual analysis of the FUN community, we used a Python package called *Gismo* (*Generic Information Search with a Mind of its Own*) [8]. Gismo performs most of the "heavy lifting", from the DBLP interface to the building of embeddings, and allows to focus on the FUN part. The code for the FUN part is available at the following address: `https://github.com/balouf/conference_analysis/tree/main/FUN`.

The repository mainly contains two *Jupyter Notebook* files, one dedicated to the creation of the community subset and associated embedding, the other to the actual similarity analysis. Some technical details (word pre-processing, incorporation of multi-words like *stochastic geometry*, ...) are left out of this paper but are available in the notebooks.

4 Results: a brief history of FUN

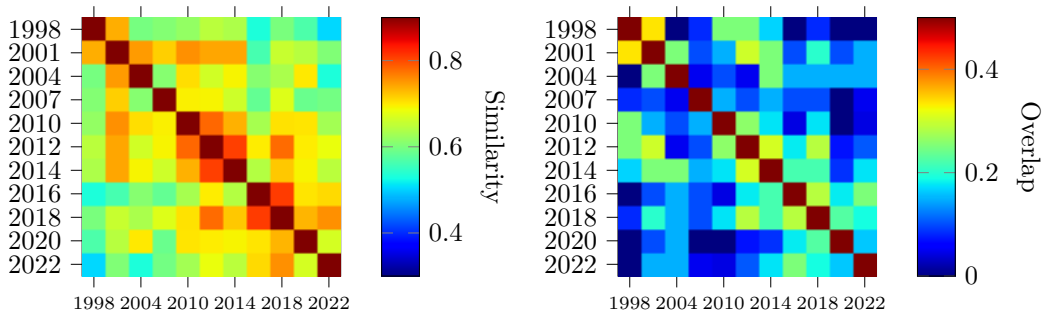
We propose to conduct our analysis in two parts. Section 4.1 studies the evolution of the PC members and authors over the different editions, while Section 4.2 focuses on the dynamics of FUN themes as put forward in the call for papers.

4.1 It's a small world, after all

Figure 2a presents the matrix of similarities of scientific communities induced by the PC members along the different editions of FUN since its creation. A warm color indicates strong similarity, while a cool color indicates weak similarity. The following observations can be made. First, the heatmap is generally warm, meaning that the scientific community induced by the PC members remains similar along time. Second, we can distinguish pairs of FUN

give more room to exploration.

editions that are very strongly similar: 1998-2001, 2001-2004, 2010-2012, 2012-2014, 2016-2018, and 2018-2020. By contrast, the most dissimilar consecutive editions are 2004-2007. Two editions are remarkable, 2001 is highly similar to the next five editions (until 2014), while 2016 is quite dissimilar to early editions. Some of those observations can be explained looking at Figure 2b that describes the overlap matrix of PC members along editions (that is, the proportion of PC members common between two editions). We observe again that pairs of consecutive years have (relatively) high overlap: 1998-2001, 2001-2004, 2010-2012, 2012-2014, 2016-2018, and 2018-2020. Also, the pair 2004-2007 has the less overlap in PC members. However, the particular edition of 2001 show very little overlap with the 2010 edition, despite being similar from a scientific community point of view.



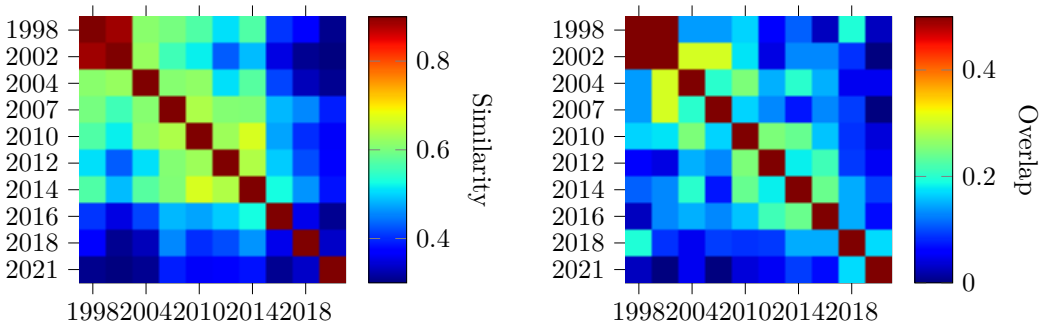
(a) Similarity matrix.

(b) Overlap matrix.

Figure 2 Evolution of the FUN community through its PC members along FUN conference editions.

Arguably, a scientific venue such as FUN can also be analyzed through its authors (that is, the scholars that publish papers appearing in the conference). Figure 3a presents the matrix of similarities of scientific communities induced by the authors along the different past editions of FUN. There, the years correspond to the years of publication of the papers, so the latest available data at the time of writing is from FUN 2020, published in 2021. We observe that the heatmap is much cooler overall than the one generated from the PC members. A possible explanation of the lesser intensity of similarity between PC members induced communities and author induced communities could come from the average pool size. Program Committees are smaller that the authors that publish in a given edition, and smaller groups tend to have more consistency. We also observe some clusters of consecutive editions with similar authors: 1998-2002 mainly, and to some extent 2004-2007-2010 and 2007-2010-2012-2014. However, editions become quickly dissimilar with other editions further away in time, implying that authors induced communities do not last. Again, those results are partly explained by Figure 3b that presents the overlap of FUN authors. It confirms that the turnover of authors is slightly higher than that of PC members, although some continuity is sometimes preserved: editions 1998-2002 have a huge overlap that fades afterwards, and occasionally we observe periods with significant overlap of authors, like 2002-2004-2007, 2010-2012-2014-2016, or 2018-2020.

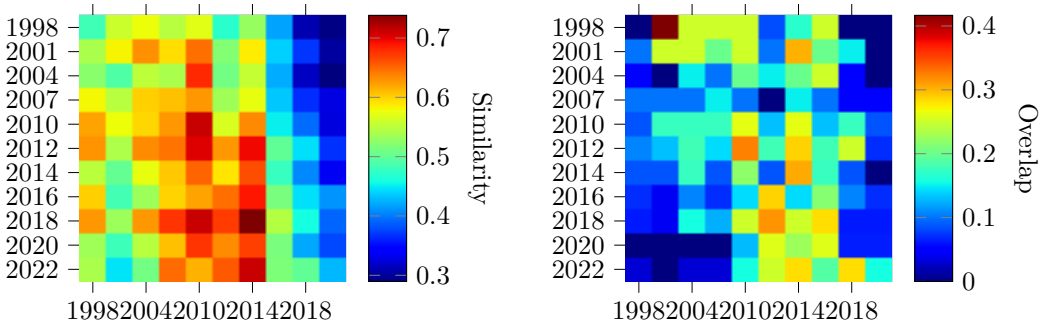
At this point, one might wonder about explaining the evolution of a scientific community. Figure 4a presents the similarity of the community induced by the PC members (on the y axis) versus the community induced by the authors (on the x axis). One striking observation is that some author years (column-wise) are highly similar to many previous PC (2010, 2012, and 2014), while some others (2002 and 2018-2021) are less so (note that the coolest colors here are still quite similar in absolute value). Another interesting observation is that as lines



(a) Similarity matrix. **(b)** Overlap matrix.

Figure 3 Evolution of the FUN community through its authors along FUN conference past editions.

go down (*i.e.*, when the PC members evolve), the similarity gets higher, which means that the PC members are selected according to the authors of previous editions interests. This is confirmed by Figure 4b that presents the overlap between authors and PC members along the years. For example, the PC members of FUN 2022 are heavily selected from the authors of FUN 2012 to 2020, but less so from authors of the previous editions. As we go back in time for PC members selection, we also go back to the earlier editions of the conference to select them from the authors' pool. A few outliers are worth mentioning: the PC members of 1998, 2001 (the first two editions), and 2012 are longstanding contributors to the conference as authors (roughly a span of 20 years).



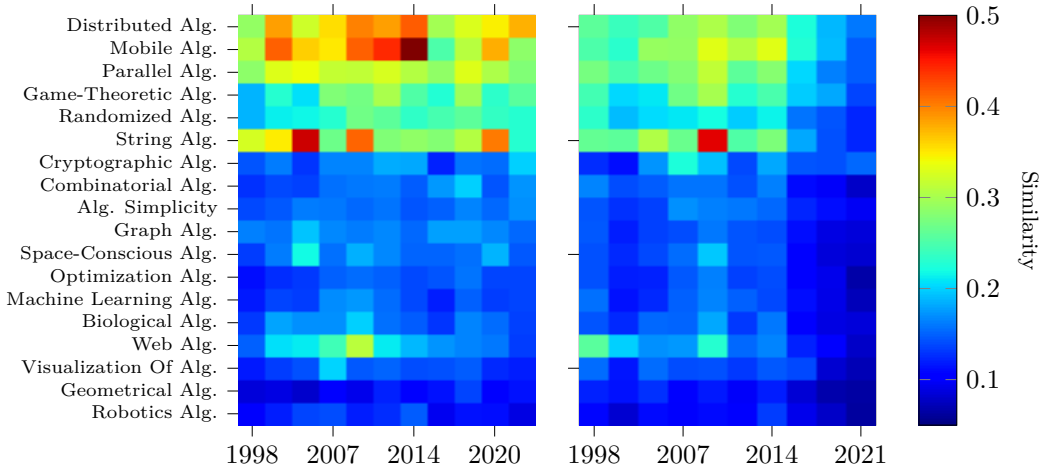
(a) Similarity matrix. **(b)** Overlap matrix.

Figure 4 Evolution of the equilibrium between PC members and authors along FUN conference editions.

4.2 Hot topics

To analyze the evolution of the FUN themes, we considered three editions: 1998, 2010, and 2022. For each edition, we compared the similarity between the advertised themes, the PC member of all editions, and the authors of all past editions. Results are displayed in Figures 5–7.

Figure 5 focuses on the 2022 edition. The main observation is that most of the similarity with the FUN community (PC members and authors alike) is concentrated in about one third of the themes, which perhaps show a lack of equilibrium between the community and the selected themes.



(a) PC members. **(b)** Authors.

Figure 5 Evolution of the themes presented in the call for papers of FUN 2022. Themes are ordered by similarity with the 2022 PC members.

Some outlier themes are worth mentioning. *Mobile algorithms* and *Distributed Algorithms* are rotating as themes highly similar with PC members since the second FUN edition, and although they seem to have peaked in 2014, they are still going strong. Also, *Game-theoretic algorithms* is gaining popularity along the years (still when comparing with PC members). *String algorithms* is intricate, as it goes in burst in some editions (2004, 2010, 2020 for PC members, only 2010 for authors) but not others. Then, *web algorithms* used to be well represented by both PC and author induced communities (with a peak in 2010) but seems to have periclited.

Another outlier is the *Space-conscious algorithms* theme that exhibit a burst in some early editions of FUN (2004 for PC members, 2010 for authors), and mostly disappears afterwards.

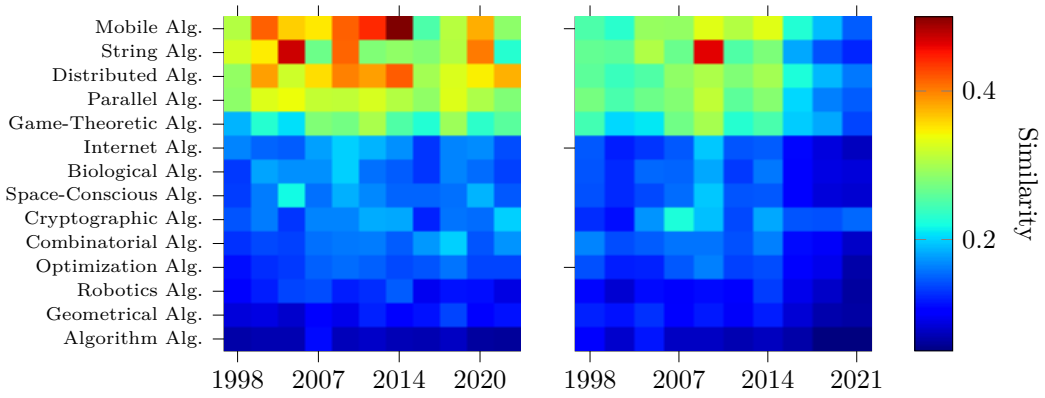
Another trend worth mentioning is that the similarity between themes and authors seems to decrease with time, possibly indicating that as time passes, compliance with the advertised themes is less required by the selection process.

Figure 6 focuses on the 2010 edition. The trend is globally the same as in Figure 5, which is not surprising as many themes are common to both editions. We still observe a concentration on five main themes (*Mobile algorithms*, *String algorithms*, *Distributed Algorithms*, *Parallel algorithms*, and *Game-theoretic algorithms*).

One interesting outlier is the missing one: "web algorithms", which peaked in 2010, is not present in the advertised. This seems to indicate that when published papers themes are put forward too late (that is, once the momentum is gone), it may be difficult to attract good submissions on those topics later.

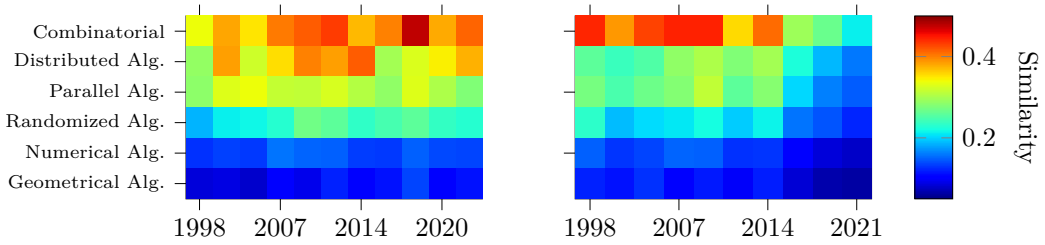
By contrast, in Figure 7, we observe that the first edition has two thirds of themes that are similar to the FUN communities induced by its PC members and authors. Arguably, the low number of proposed themes (6) could be an explanation.

However, in the FUN 1998 edition, the theme *Combinatorial* seems to be an outlier, as it is a single word (versus all other themes being pairs of words), which translated to increased similarity (for example, the pair *Combinatorial algorithms* is dissimilar throughout the years, as shown by Figures 5 and 6), hence the real proportion for FUN 1998 is half-half.



(a) PC members. **(b)** Authors.

Figure 6 Evolution of the themes presented in the call for papers of FUN 2010. Themes are ordered by similarity with the 2010 PC members.



(a) PC members. **(b)** Authors.

Figure 7 Evolution of the themes presented in the call for papers of FUN 1998. Themes are ordered by similarity with the 1998 PC members.

5 The future of FUN

We have shown how, from the only knowledge of the members of the program committees, the conference authors, the calls for contributions, and the DBLP database, it is possible to analyze the themes and communities of a conference and to observe their evolution.

Beyond the analytical aspect, our approach can also be used to assist in the development of a program committee, in particular to avoid a weak similarity between the scientific community induced by the program committee and the themes promoted, as seen in Section 4. The source code that accompanies this article (see Section 3.4) includes in particular tools to obtain suggestions from PC members relevant to (possibly new) themes, by setting a renewal rate (that is, picking a suitable proportion of PC members from previous ones).

The same tool can be used for more individual purposes: suppose a researcher wants to write a paper about a FUN theme for the next edition of the conference, but is a bit ignorant on some topic or technique that would be instrumental in carrying out the research. Our implementation can be used to suggest suitable collaborators for the task, possibly issued from the FUN community.

References

1 A. Aizawa. An information-theoretic perspective of TF-IDF measures. *Info. Proc. & Manag.*, 39(1):45–65, 2003. doi:10.1016/S0306-4573(02)00021-3.

2 S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117, 1998. doi:10.1016/S0169-7552(98)00110-X.

3 Graham Cormode, S. Muthukrishnan, and Jinyun Yan. Scienceography: The study of how science is written. In Evangelos Kranakis, Danny Krizanc, and Flaminia L. Luccio, editors, *Fun with Algorithms - 6th International Conference, FUN 2012, Venice, Italy, June 4-6, 2012. Proceedings*, volume 7288 of *Lecture Notes in Computer Science*, pages 379–391. Springer, 2012. doi:10.1007/978-3-642-30347-0_37.

4 Bruno Gaume and Fabien Mathieu. PageRank Induced Topology for Real-World Networks. working paper or preprint, May 2016. URL: https://hal.archives-ouvertes.fr/hal-01322040.

5 D. Hong, T. D. Huynh, and F. Mathieu. D-Iteration: diffusion approach for solving PageRank. *CoRR*, abs/1501.06350, 2015. arXiv:1501.06350.

6 John H. Marburger III, Julia I. Lane, Stephanie S. Shipp, and Kaye Husbands Fealing, editors. *The Science of Science Policy: A Handbook*. Stanford University Press, 2011. URL: http://www.sup.org/books/title/?id=18746.

7 Michael Kuhn and Roger Wattenhofer. The theoretic center of computer science. *SIGACT News*, 38(4):54–63, 2007. doi:10.1145/1345189.1345202.

8 Fabien Mathieu. Generic Information Search with a Mind of its Own (Gismo). URL: https://gismo.readthedocs.io/en/latest/.

9 J. Otterbacher, G. Erkan, and D. Radev. Using random walks for question-focused sentence retrieval. In *HLT/EMNLP*, pages 915–922, 2005. URL: https://www.aclweb.org/anthology/H05-1115.

10 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

11 Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008. doi:10.1073/pnas.0706851105.

12 Dashun Wang and Albert-László Barabási. *The Science of Science*. Cambridge University Press, 2021. doi:10.1017/9781108610834.