

© 2023 IEEE. This is the author's version of the article that has been published in IEEE Transactions on Visualization and Computer Graphics. The final version of this record is available at: [xx.xxxx/TVCG.201x.xxxxxxx](https://doi.org/10.1109/TVCG.2023.3327387)

# InnovationInsights: A Visual Analytics Approach for Understanding the Dual Frontiers of Science and Technology

Yifang Wang , Yifan Qian , Xiaoyu Qi , Nan Cao\* , Dashun Wang\* 

**Abstract**— Science has long been viewed as a key driver of economic growth and rising standards of living. Knowledge about how scientific advances support marketplace inventions is therefore essential for understanding the role of science in propelling real-world applications and technological progress. The increasing availability of large-scale datasets tracing scientific publications and patented inventions and the complex interactions among them offers us new opportunities to explore the evolving dual frontiers of science and technology at an unprecedented level of scale and detail. However, we lack suitable visual analytics approaches to analyze such complex interactions effectively. Here we introduce *InnovationInsights*, an interactive visual analysis system for researchers, research institutions, and policymakers to explore the complex linkages between science and technology, and to identify critical innovations, inventors, and potential partners. The system first identifies important associations between scientific papers and patented inventions through a set of statistical measures introduced by our experts from the field of the Science of Science. A series of visualization views are then used to present these associations in the data context. In particular, we introduce the *Interplay Graph* to visualize patterns and insights derived from the data, helping users effectively navigate citation relationships between papers and patents. This visualization thereby helps them identify the origins of technical inventions and the impact of scientific research. We evaluate the system through two case studies with experts followed by expert interviews. We further engage a premier research institution to test-run the system, helping its institution leaders to extract new insights for innovation. Through both the case studies and the engagement project, we find that our system not only meets our original goals of design, allowing users to better identify the sources of technical inventions and to understand the broad impact of scientific research; it also goes beyond these purposes to enable an array of new applications for researchers and research institutions, ranging from identifying untapped innovation potential within an institution to forging new collaboration opportunities between science and industry.

**Index Terms**—Science of Science, Innovation, Academic Profiles, Patent Data, Publication Data, Visual Analytics

## 1 INTRODUCTION

Science is central to improving the human condition [15, 37]. Not only has science long been recognized as the engine for long-run economic growth and prosperity, but also it has been essential to creating critical solutions to confront emergent threats to humanity, from climate change to the COVID-19 pandemic. While scientific research propels both fundamental understanding and practical applications [6, 31, 65, 77], there has been a lack of visual analytics approaches to explore the complex linkages (i.e., the dual frontiers) between scientific advances and technical inventions. Here we introduce *InnovationInsights*, which represents an initial step toward filling this crucial gap.

A better understanding of the dual frontiers of science and technology informs a diverse range of stakeholders, from researchers and research institutions to policymakers and private companies, helping them identify gaps and opportunities for innovation while facilitating more rapid and effective knowledge transfer. For example, research institutions such as universities aim to discover untapped innovation potentials and forge new collaboration and partnership opportunities between science and industry. Companies seek to stay abreast of the latest scientific breakthroughs to drive the creation of new applications.

The availability of large-scale datasets [3, 47, 49] tracing scientific publications and patented inventions and the complex interactions among them has created new opportunities to tackle this research question. Here we build on the Science of Science literature (SciSci [31, 65]), which provides descriptive insights into the connections between science and technology [6, 47, 77]. While SciSci has furthered our understanding of the uses of science both within and outside of science, it

also highlights the numerous challenges to interactively explore the complex interactions among multiple entities, from inventors and inventions to scientists and their publications. Meanwhile, existing studies in the visualization community have primarily focused on papers [46, 53] or patents [8, 41, 43] separately, rarely examining their interconnections.

Here we hypothesize that visual analytics may provide an effective means to meeting these new analytical demands. Designing such a system requires us to overcome several major challenges: (1) The complexity of the data, which encompasses networks, multi-dimensional attributes, hierarchical structures, and temporal features, poses visualization challenges for effectively navigating these complex interactions; (2) The multi-dimensional nature of the entities and complex linkages among them, coupled with multiple levels of granularity, pose analytical challenges for quantitative measures of the science-technology interface. (3) The massive amounts of underlying data, including papers, paper citations, patents, patent citations, and paper-to-patent citations, pose scalability challenges for the system. (4) Given the long pathways in knowledge transfer, we need new predictive models to accompany visual approaches to identify innovation gaps and opportunities.

To tackle these challenges, we first characterize the problem domain and define a set of statistical measures in collaboration with our domain experts. We then develop a prediction model to estimate the extent to which scientific advances may propel future technological applications, allowing us to systematically identify a list of innovators whose work holds considerable potential for commercialization. Based on these measures and models, we develop a visualization system with multi-dimensional views into the complex relationships between science and technology. We introduce the *Interplay Graph*, which enables interactive exploration of the dual frontiers of science and technology in a scalable manner. Our contributions are summarized as follows:

- We formulate the domain of visual analysis of dual frontiers of science and technology and propose a novel design to visualize the complex interactions between science and technology.
- We design and develop *InnovationInsights*, which to the best of our knowledge, is the first visual analysis system to explore rich interactions between upstream scientific research and downstream technological development.
- We conducted comprehensive evaluations, including case studies and expert interviews, to demonstrate the effectiveness of our system.

\* Yifang Wang, Yifan Qian, and Dashun Wang are with The Center for Science of Science and Innovation, Northwestern University. E-mail: {yifang.wang, yifan.qian1, dashun.wang}@kellogg.northwestern.edu.  
• Xiaoyu Qi and Nan Cao are with Intelligent Big Data Visualization Lab, Tongji University. E-mail: {qixiaoyu, nan.cao}@tongji.edu.cn.  
• \* Dashun Wang and Nan Cao are the co-corresponding authors.

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

Moreover, we engage with a premier research university and its institutional leaders as a trial run for the developed visual analytics systems, helping key stakeholders uncover new insights for gaps and untapped potentials for innovation in real-world settings.

Overall, the system we developed serves as an initial but crucial step toward using visual analytics to bridge the ivory tower and the real world, helping amplify the real-world impact of scientific research while significantly advancing the R&D success of research institutions.

## 2 RELATED WORK

The related literature spans three domains, including the science of science, visualization of scientific data, and graph and tree visualization.

### 2.1 The Science of Science

The increasing availability of large-scale data tracing nearly all phases of scientific production and use has fueled the emergence of an interdisciplinary field, SciSci, to explore opportunities and premises to accelerate scientific discoveries. Despite the rapid progress in this field, the bulk of the literature has focused on the impact of science within science, ranging from the unfoldings of scientific careers [48] to the scientific impact of papers [66], to scientific collaborations [74].

More recently, studies have made initial attempts to quantify the broad impact of science [6, 49, 77, 78], aiming to better understand the interface between science and various facets of human society, from policy-making [78] to public perception [77]. In particular, the interaction between science and technology is a critical area of focus, with studies testing theories that emphasize the connections between patents and prior scientific advances. Ahmadpoor and Jones [6] conducted the first systematic analyses into *the dual frontier of science and technology*, finding that advances that sit directly at the science-technology interface are significantly more impactful within their respective domains. Marx and Fuegi [49, 50] created a large-scale dataset, “Reliance on Science”, to trace citations from patents to papers. Yin et al. [77] introduced an index to quantify the extent to which papers from a scientific field are consumed by patents. Cao et al. [16] specifically targeted the HCI community and studied the impact of HCI papers on the industry. Overall, these efforts contribute to a data-driven understanding of the dual frontiers. Yet, the statistical nature of these studies highlights the lack of visualization approaches to studying this important problem. This is especially true given the substantial challenges in identifying patterns and extracting insights hidden beneath complex data structures.

In this paper, we aim to provide an interactive visual analytics approach to help experts analyze the complex interactions between science and technology more systematically and effectively.

### 2.2 Scientific Data Visualization

Large-scale scientific data represent a familiar domain in the visualization community, with several comprehensive reviews written on the subject [10, 18, 28]. Here we examine pertinent studies from these surveys and recent advances in the visualization community, focusing on two main categories: (1) visualization for scientific data, and (2) the science of science within the visualization community.

**Visualization for scientific data.** Many visualization techniques have been developed to reveal insights from scientific databases, such as dynamic and heterogeneous networks [25, 76], sequences and time series [70, 75], multi-dimension measures [54], and texts [26, 32]. These studies introduce general visualization techniques for specific data structures and use scientific datasets as illustrative examples. Other research focuses on developing visualization systems to streamline literature queries (e.g., VitalITy [53] and VisualBib [21]), scientific discoveries (e.g., VISStory [24] and Galex [46]), and academic evaluations (e.g., SD<sup>2</sup> [33]). These studies only use data within science (i.e., papers).

Similar to papers, patent data are also of a textual, network, and temporal nature. Studies using patent data cover several themes: patent document classification [8, 39] and information retrieval [11, 41], knowledge discovery using patent citation, semantic, or agent (e.g., inventor and organization) networks [11, 39, 72], and technology evolution and emerging technology exploration [8, 72]. In addition, a few studies provide a comprehensive analysis by visualizing multifaceted patent data. PatViz [41] summarizes patent search results using various factors (e.g., patent categories), but it does not focus on the relationship between

papers and patents. DIVA [51], on the other hand, creates network linkages between papers and patents using keyword similarity. However, it fails to provide deep insights between science and technology due to the absence of paper-patent citations.

**Science of science within the visualization community.** A growing number of studies have explored scientific data within the visualization community to reflect on community development. These studies involve the creation of specific datasets like Vispubdata [36], VIS30K [19], and VisImages [22], and the development of platforms such as VIS Author Profiles [44] and VISPubComPAS [71]. Additionally, many works statistically analyze the community’s development, focusing on aspects such as authors [34], topics [34], genders [59, 62], collaborations [59], peer reviews [73], and so on.

Overall, these studies have focused on papers or patents separately, ignoring the complex interactions between them. Our work combines multiple data sources for papers and patents to systematically study the dual frontiers of science and technology.

### 2.3 Graph and Tree Visualization

Graph [5, 9, 23, 35, 69] and tree [45, 55, 68] visualization have been studied extensively in the visualization community. Here we discuss the most relevant techniques for paper-patent citation networks, including bipartite graphs, compound graphs, and tree visualization.

Patent-paper citations may be represented as a bipartite graph problem. Sun et al. [61] proposed a technique, bicluster-based seriation, to reduce edge crossings in a bipartite graph. Chan et al. [17] used the minimum description length (MDL) principle to aggregate bipartite relations for scalability. However, they disregard complex node attributes, such as temporal and hierarchical structures, rendering them unsuitable for our scenario. The compound graph technique, often used for large-scale networks to group nodes for scalability, is another relevant method [9]. Our work also uses this technique to visualize the large-scale citation graph by grouping paper and patent nodes.

Tree visualization displays hierarchical node connections [55]. Various tree types use the width of the link to represent the flow quantity between the parent and child nodes, such as decision trees (e.g., BaobabView [64] and TreePOD [52]) and flow maps [12, 56]. However, most methods focus on static graphs and neglect the hierarchical aspect of nodes. They are insufficient for our needs, as we require temporal dimensions and detailed exploration across various levels of detail.

We, therefore, propose a scalable node-link representation that summarizes the citation between papers and patents, overcoming various domain-specific constraints.

## 3 SYSTEM DESIGN

In this section, we summarize the analysis goals and design tasks and introduce the system overview.

### 3.1 Analysis Goals

Over the past two years, we have been working closely with leaders from a premier US research university to understand and predict a university’s innovation landscape and potential. We first collected private data from various organizations in the university, including (1) Technology Transfer Office (TTO) on invention disclosures and outcomes, licensing, and startups; (2) HR Office on faculty roster with demographic data (e.g., name, gender, rank, and department); and (3) Sponsored Research Office on grant applications and their outcomes (granted or rejected). We then linked these data with global innovation databases on science and technology, capturing publications, patents, and how these publications are cited in patents as prior art, spanning all scientific fields and patenting domains. After a massive data cleaning and linkage, we used statistical methods for data analysis. We then presented our findings to leaders from seven top US research universities (i.e., two public land-grant, three public non-land-grant, and two private universities), four R&D-based companies and venture firms, and four science funders. During these interactions with leaders in science, industry, and government, we saw great interest in identifying untapped innovation potential in research institutions across a wide range of stakeholders. Given the novelty of the research question and the diverse array of stakeholders it informs, we quickly realized the

© 2023 IEEE. This is the author's version of the article that has been published in IEEE Transactions on Visualization and Computer Graphics. The final version of this record is available at: [xx.xxxx/TVCG.201x.xxxxxxx/](https://doi.org/10.1109/TVCG.2023.3327387)

need for a visual system, which would be crucial for efficient analysis and communication among multiple stakeholders.

To this end, we initiated an interdisciplinary collaboration with domain experts from various fields to design a visual analytics system nine months ago. Two experts,  $E_A$  and  $E_B$ , are from the TTO of our partner university.  $E_A$  is a senior director of invention management who helps faculty convert their scientific output into patents.  $E_B$  is an analyst providing data-driven support. The other two experts,  $E_C$  and  $E_D$ , are SciSci researchers.  $E_C$  is a well-known professor in the field of SciSci.  $E_D$  is a postdoc who focuses on statistical measurements of university innovations. They all aim to understand how scientific research influences technology development and to find promising scientific directions and researchers with high innovation potential. To achieve this objective, they seek to explore the interplay between science and technology comprehensively. Existing methods rely on statistical analysis via ad-hoc analysis procedures, lacking an integrated system to help identify key innovators and research topics. The experts particularly identified three critical analysis goals:

- (I) **Researcher Identification.** Identifying researchers and research ideas with high potential for practical applications and commercialization opportunities will help research institutions to better support researchers with a greater level of equity and efficiency.
- (E) **Interplay Exploration.** Exploring the interplay between science and technology is another goal frequently mentioned by experts. The exploration requires not only the extraction of relational patterns (e.g., paper-patent citations) but also the interpretation of these connections in the context of knowledge transfer.
- (P) **Invention Prediction.** Being able to predict the future likelihood of patentability and the commercialization potential of scientific research and technological inventions is a highly desirable capability by our domain experts.

### 3.2 Visualization Design Tasks

Given the goals above, we devised the design tasks by following the expert-focused design study methodology [60]. To this end, literature reviews guided by our experts, requirement analysis via expert interviews, as well as brainstorming sessions with both visualization and SciSci experts, were conducted iteratively. The design tasks for each analysis goal are outlined below:

#### Tasks for Researcher Identification (I)

- I1: **Provide Researcher Overview.** The visual analysis system should provide an overview of researchers to help analysts select individuals of interest for further investigation, by aligning researchers based on their research profiles.
- I2: **Create Productivity Portrait.** The system should illustrate the detailed characteristics of individual researchers to help the analysts identify talent and potential. Two types of information are of particular relevance: (1) demographic information—such as gender, age, and job rank—for equity and inclusion; (2) productivity and impact measurements for research output.

#### Tasks for Interplay Exploration (E)

- E1: **Inspect the Interplay.** The visualization design should be scalable enough to display large-scale linkages between scientific papers and technical inventions at both individual and field levels intuitively. The design should also clearly reveal the papers that are cited heavily by patents and highlight their characteristics to help analysts understand the key factors in knowledge transfer.
- E2: **Reveal Temporal Trends.** The design should reveal the temporal changes in research topics and the corresponding technical developments to help analysts understand the evolving frontiers of science and technology.
- E3: **Reason with Contextual Information.** Beyond showing the interplay based on explicit linkages between patents and papers, the visualization should capture contextual information to help analysts uncover implicit linkages between research and technology. The contextual information can be captured by

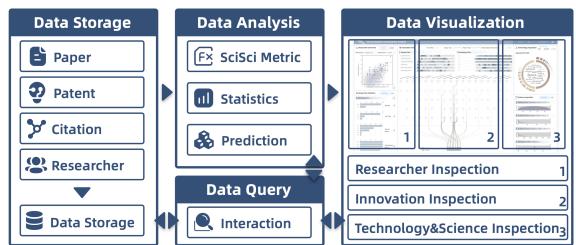


Fig. 1: System overview. *InnovationInsights* consists of a data storage module, a data analysis module, and a data visualization module.

SciSci measures of papers, fields, and the assignee information of patents, which is crucial for identifying important research or estimating the commercialization potential of an invention.

#### Tasks for Invention Prediction (P)

- P1: **Identify Untapped Innovation Potential.** Our system should uncover untapped innovation potential in researchers, research topics, and scientific fields. This key capability building on the prediction model and intuitive visualization will significantly facilitate knowledge transfer for the latest scientific advances.

### 3.3 System Overview

Following the above analysis goals and design tasks, we design *InnovationInsights* as an online system that consists of three modules (Fig. 1): (1) the data storage module, (2) the data analysis module, and (3) the visualization module. The data storage module preprocesses data from multiple sources and stores it in a database. The data analysis module conducts a series of measurements on different entities (e.g., papers, patents, and researchers) and employs a prediction model to recommend papers with high patentability potential. The two modules form the backend of the system. Both historical and prediction data are fed into the data visualization module to display intuitive data insights.

## 4 DATA ANALYSIS

The data analysis module is designed to calculate the contextual information for visual analysis and decision-making. Specifically, we consider two types of information: (1) the data facts about papers, patents, researchers, and assignees that are calculated based on a set of statistical metrics; (2) the potential of scientific research (i.e., a paper) to be transferred, which is estimated by a deep prediction model implemented based on a graph convolutional network (GCN). Before drilling into technique details, we will first introduce the data we use.

### 4.1 Data Preprocessing

Analyzing dual frontiers of science and technology needs to integrate data from various sources (Fig. 2(A)), which are listed as follows:

- **Scientific Research Records.** We leverage the Microsoft Academic Graph (MAG) dataset [67] to retrieve information about scientific research. The dataset consists of 270M research papers and their corresponding meta information, including the title, publication year, topic keywords, doi, author list, author affiliations, and citations.
- **Technical Inventions.** We use the patent records collected in PatentsView [3] to capture the technical inventions and reveal the development of technologies. This dataset contains over 7.9M patents filed through the United States Patent and Trademark Office (USPTO [4]). A subset of the most relevant patent attributes is carefully selected for analysis, including patent ID, title, application year, assignee name (i.e., the owner of the patent), and cooperative patent classification (CPC, i.e., the category of the patent [1]). The CPC category we used consists of three levels: section, subsection, and group (the lowest level). Private data (e.g., invention disclosures and patents collected by research institutions) are also used.
- **Science-Technology Linkages.** To analyze the interplay between scientific research and the development of technology, we use data collected from “Reliance on Science” [49, 50], which include more than 40M citation data that record the details about how technical innovations (i.e., patents) cite research papers.

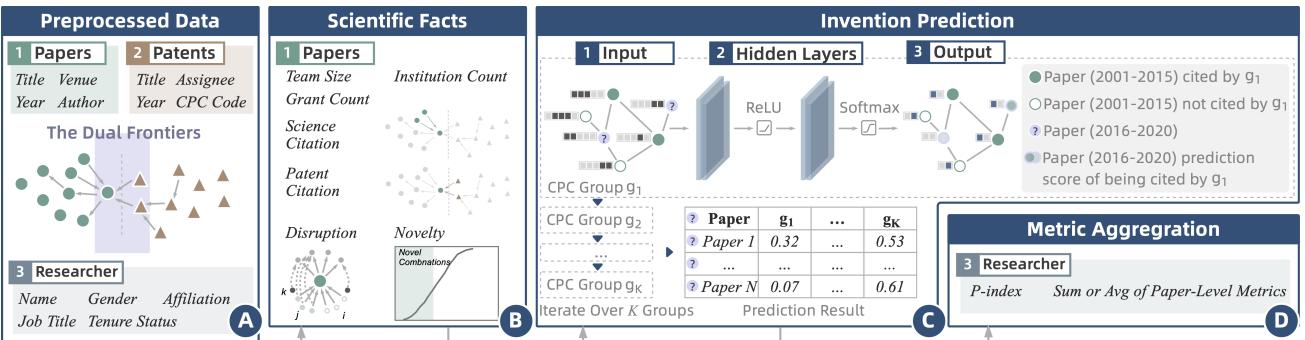


Fig. 2: The analytical process. (A) We preprocess data into the network and multi-dimensional structures. (B)-(C) Next, we construct SciSci metrics using scientific facts and predicted results at the paper level. (D) Finally, we aggregate paper-level metrics to get researcher-level metrics.

- **Researcher Profiles.** This dataset provides demographic information (e.g., gender, rank, and affiliation) for each researcher, collected by research institutions (e.g., gender and rank) or automatically inferred by algorithms (e.g., gender [38]). The data also contain each researcher’s publication records, which are collected from the public (e.g., MAG [67]) and private sources (e.g., university libraries).

We extracted a subset of the data above (supplementary material) to demonstrate our idea of analyzing the dual frontiers of science and technology. This subset includes papers from a 20-year period (2001–2020). A group of researchers and patent assignees were also filtered out for analysis. In particular, the patent assignees were classified into three categories, i.e., university assignees, company assignees, and others, to provide additional context information.

## 4.2 Scientific Facts

We analyze the data to capture the scientific facts for each research paper and each individual researcher based on a set of carefully defined metrics. Specifically, given a research paper  $\mathcal{P}$ , the following metrics are designed to help analysts estimate the quality and impact of  $\mathcal{P}$  in the context of knowledge transfer (Fig. 2(B)), which are calculated based on the entire dataset in Section 4.1:

- *Team Size*: the total number of co-authors of the paper.
- *Institution Count*: the total number of different affiliations regarding the co-authors of the paper.
- *Grant Count*: the total number of grants sponsoring the research for the paper. Due to data availability, we focus primarily on grants from NSF and NIH as a demonstration.
- *Science Citation*: the total number of citations the paper received within 5 years of publication.
- *Disruption*: the degree to which papers citing the focal paper  $\mathcal{P}$  tends not to cite  $\mathcal{P}$ ’s references [74], which is formally defined as:

$$D = \frac{n_i - n_j}{n_i + n_j + n_k} \quad (1)$$

where  $n_i$  is the number of subsequent papers that only cite the focal paper,  $n_j$  represents the number of subsequent papers that cite both the focal paper and its references, and  $n_k$  represents the number of subsequent papers that only cite the references of the focal paper.

- *Novelty*: the extent to which the focal paper’s combination of existing knowledge deviates from the norm among all journal pairs. We first calculate the z-score for each journal pair by comparing its observed frequency to the expected frequency in randomized citation graphs [63]. The focal paper’s novelty score is determined by the 10th percentile z-score of the journal pairs cited in its references.
- *Patent Citation*: the total number of patents that cite  $\mathcal{P}$  within 5 years of its publication. This metric is used to measure  $\mathcal{P}$ ’s impact on technical inventions.

In addition, in order to estimate a researcher  $\mathcal{R}$ ’s performance in scientific research and impact on technical inventions, we define the following metrics (Fig. 2(D)):

- *Paper Count*: the total number of papers that  $\mathcal{R}$  has ever published.
- *Invention Count*: the total number of invention disclosures that  $\mathcal{R}$  has ever disclosed to the university.

- *Scientific Citation*: the number of research papers that cite  $\mathcal{R}$ ’s papers within 5 years of each paper’s publication.
- *Number of Papers Cited by Patents*: the total number of papers that have been cited by at least one patent. This metric measures  $\mathcal{R}$ ’s impact on technical inventions.

## 4.3 Invention Prediction

The ability to estimate the potential of a paper for future inventions helps analysts to identify the next promising research topics as well as potential inventors. Here we introduce an invention prediction model to compute the probability that a research paper will spur future inventions in a given area. The model is designed based on the observation that when the knowledge obtained from a scientific research paper is used in a technical invention, the paper will be cited directly by the corresponding patent. Therefore, we use the citation links from patents to papers as a key feature to train a graph convolutional network (GCN) to help us estimate how likely (i.e., the probability) a paper  $\mathcal{P}$  will be cited directly by patents from a specific technical area, as indicated by class labels (Fig. 2(C)). We briefly review the architecture of GCN, followed by our implementation details below.

**Graph Convolutional Network.** GCN is a crucial technique for deep learning on graph-based data [40]. It has transformed the field by providing a powerful way to analyze and model graph-structured data, which is common in many real-world applications such as social graphs, molecular graphs, and citation graphs [57, 58]. In our case, papers are connected by citations that form a citation graph, and thus GCN is naturally suited to our prediction task. In addition, compared to traditional deep learning models for grid-like data (e.g., images), GCN has several advantages. First, it can capture both local and global structures of the graph and also can handle graphs of varying sizes and structures. Second, it can be trained using both labeled and unlabeled data, i.e., in a semi-supervised learning way. By using unlabeled data, the model can learn more generalized representations of the data, which can improve its performance on the labeled data.

The input for GCN is a graph, in which each node is associated with an  $F$ -dimensional vector arranged as a row of the feature matrix  $X \in \mathbb{R}^{N \times F}$ . The adjacency matrix  $A \in \mathbb{R}^{N \times N}$  represents the relationships between all nodes in the graph where the element  $A_{ij}$  indicates the presence (1) or absence (0) of an edge between node  $i$  and node  $j$ . The label matrix  $Y \in \mathbb{R}^{N_s \times C}$ , where  $N_s$  is the number of nodes with labels in the graph and  $C$  is the number of classes. The element  $Y_{ij}$  is 1 if node  $i$  belongs to class  $j$ , and 0 otherwise.

**Layer-wise Propagation Rule.** The standard GCN proposed by Kipf and Welling [40] takes two propagation layers to perform graph convolution operations on the input data. The first layer is defined as:

$$H = \text{ReLU}(\hat{A}XW^0) \quad (2)$$

where  $W^0 \in \mathbb{R}^{F \times B}$  is the weight matrix connecting the inputs and the first layer of the GCN. The graph is encoded in  $\hat{A} = \tilde{D}^{-1/2}(A + I_N)\tilde{D}^{-1/2}$ , where  $I_N$  is the identity matrix, and  $\tilde{D}$  is a diagonal matrix with  $\tilde{D}_{ii} = 1 + \sum_j A_{ij}$ .  $\text{ReLU}(\cdot) = \max(\cdot, 0)$  is the activation function.  $H \in \mathbb{R}^{N \times B}$  is the matrix of activations of the first layer.

© 2023 IEEE. This is the author's version of the article that has been published in IEEE Transactions on Visualization and Computer Graphics. The final version of this record is available at: [xx.xxxx/TVCG.2023.xxxxxxx](https://doi.org/10.1109/TVCG.2023.3327387)

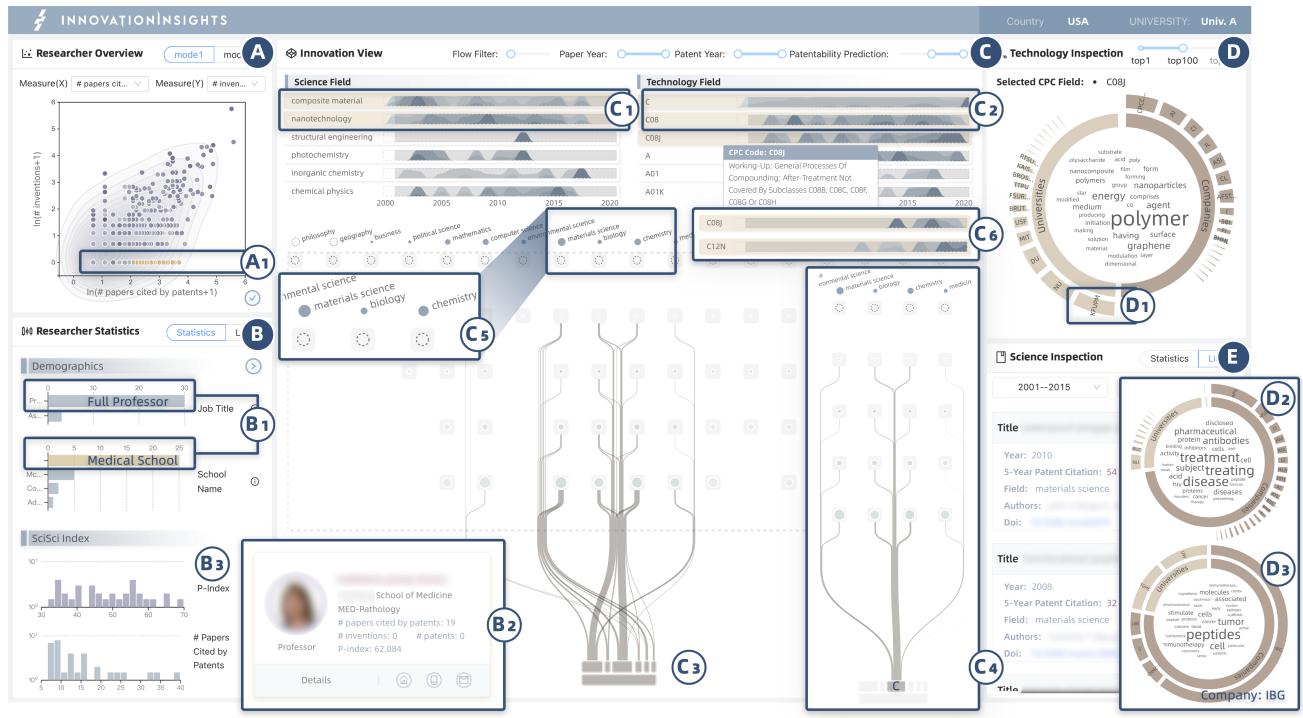


Fig. 3: The system UI of *InnovationInsights*. The *Researcher Overview View* (A) and *Researcher Statistics View* (B) are for individual-level analysis. The *Innovation View* (C) shows the detailed interplay between science and technology. The *Technology Inspection View* (D) and *Science Inspection View* (E) provide additional contextual information about patents and papers.

The output layer is formally defined as follows:

$$Z = \text{softmax}(\widehat{AHW}^1) \quad (3)$$

where  $W^1 \in \mathbb{R}^{B \times C}$  is the weight matrix connecting the first layer and output layer of the GCN.  $\text{softmax}(x)_i = \exp(x_i) / \sum_j \exp(x_j)$  where  $x$  is a vector.  $Z \in \mathbb{R}^{N \times C}$  is the output matrix of GCN where the element  $Z_{ij}$  represents the probability of node  $i$  belonging to class  $j$ .

**Loss Function.** GCN evaluates the cross-entropy error between the predicted class probabilities  $Z$  and the true labels  $Y$  for the nodes in the training set:

$$\mathcal{L} = - \sum_{l \in \mathbb{N}_T} \sum_{c=1}^C Y_{lc} \ln Z_{lc} \quad (4)$$

where  $\mathbb{N}_T$  is the set of nodes in the training set.

**Implementation.** In our work, the input feature matrix  $X$  is composed of paper title embeddings using SPECTER [20], a popular API to generate embeddings for research papers. The input graph  $A$  represents the citation graph between papers. Given a patent CPC category  $g$ , the label matrix  $Y$  provides information on whether a paper is cited by a patent from CPC category  $g$  within 5 years of publication. We selected a 5-year duration for the experiment because our experts prioritized recently published papers and aimed to determine if a paper would be cited by patents soon after its publication. We split the papers published between 2001 and 2014 into a training set (70%) and a validation set (30%), and use papers published in 2015 and 2020 as the test set. For each paper published between 2016 and 2020, we predict its likelihood of being cited by patents in the CPC group  $g$ . We use the PyTorch Geometric implementation of GCN [30] and follow the experimental setup proposed by Kipf and Welling [40]. Our model has 200 epochs of training iterations, a learning rate of 0.01, a dropout rate of 0.5, and 16 hidden units. The weights of the neural network ( $W^0$  and  $W^1$ ) are trained using gradient descent to minimize the loss  $\mathcal{L}$ . To determine the likelihood of a paper  $P$  published between 2016 and 2020 being cited by patents in the CPC group  $g$  within 5 years of publication, we use the predicted probabilities in the softmax output matrix  $Z$  obtained from the final model. In addition, to assess its relative importance within a specific range of papers (e.g., those within a research institution), we further convert its probability to a percentile, denoted as  $\text{Patentability}_P^g$ , which

is a scalar ranging from 0 to 100.

We apply the above prediction pipeline to the top  $K$  patent CPC groups  $g$  (denoted as  $G$ ) based on the number of patents citing our target papers. To assess a paper's overall patentability across different CPC groups, we compute its average likelihood of being cited by patent CPC groups  $g$  in  $G$  (i.e.,  $\text{Patentability}_P^g$ ), denoted as  $\text{Patentability}_P$ .

To evaluate a researcher's overall performance, we calculate the average  $\text{Patentability}_P$  of all their papers published between 2016 and 2020. This aggregated value is called the P-index. To the best of our knowledge, the P-index is the first index proposed in the SciSci literature that measures the extent to which a researcher's recent papers will be cited by patents in the future, acting as an indicator of a researcher's potential for commercial success. The higher the P-index, the higher the commercialization potential of the researcher.

## 5 VISUALIZATION

This section presents the visual design of *InnovationInsights*. We introduce the user interface through a usage scenario followed by detailed descriptions of visualization views and corresponding interactions.

### 5.1 User Interface

Fig. 3 illustrates the user interface of the proposed system, which consists of five coordinated views (Fig. 3(A-E)). A user can start from the *Researcher Overview View* (Fig. 3(A)) to choose a group of researchers (I1), whose scientific facts and profile information are summarized as the context in the *Researcher Statistics View* (I2, Fig. 3(B)). The user can then filter to find interested researchers based on this contextual information. Two sets of horizon graphs are used to illustrate the trend of science and technologies (E2) by displaying the changes in the numbers of the papers (Fig. 3(C1)) and patents (Fig. 3(C2)) in different fields. The user can brush a period of time to filter the papers or patents through these horizon graphs. Once the data (i.e., researchers, papers, and patents) are filtered out, an interplay graph (Fig. 3(C3)) will visualize the citations between the selected papers and patents to reveal the interplay between science and technology (E1). In this view, the user can interactively filter the citations via fields to learn the science-technology connections at different levels of detail. To give the user a better understanding of the connections, additional context information such as keywords for the selected patents (Fig. 3(D)) and the list of the selected papers (Fig. 3(E)) is also displayed (E3).

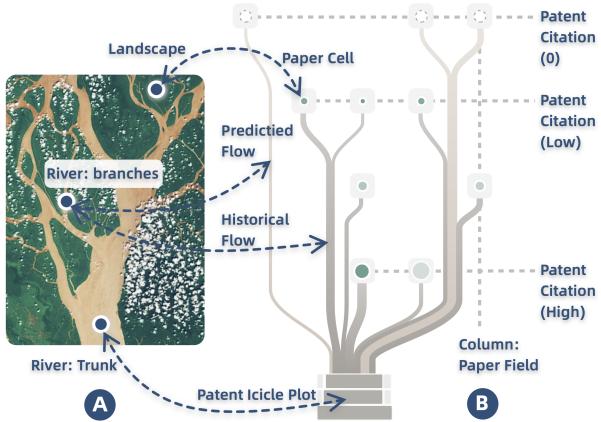


Fig. 4: The design of the *Interplay Graph* is inspired by the river metaphor: (A) the structure of the river; (B) the river-like visual metaphor shows citation linkages between papers and patents with three components: Paper Matrix, Patent Icicle Plot, and Citation Flow.

Despite the exploitative analysis above, the user can also investigate the potential inventors and inventions from multiple views based on the invention prediction results (**P1**). For example, the user can observe the P-index scores of researchers in the *Researcher Overview View* via the opacity of the circles. The user can also examine the researcher distribution (Fig. 3(B3)) and rank the researchers based on P-index in the *Researcher Statistics View*. In the *Innovation View*, the user can filter recent papers with high prediction scores using a slider (Fig. 3(C)) and observe the prediction flow to identify the promising paper fields with high patentability potential. The user can also click a paper cell to check the details in the *Science Inspection View*.

## 5.2 Interplay Graph

The *Interplay Graph* (Fig. 3(C3)) is the primary visualization component that enables users to explore the detailed citations from patents to papers and reveal the interplay between scientific research and technological inventions (**E1, P1**). It consists of three components *Paper Matrix*, *Patent Icicle Plot*, and *Citation Flow*, whose design is inspired by a river metaphor (Fig. 4). The *Paper Matrix* symbolizes the knowledge landscape, and the *Citation Flow* illustrates the diverse branches of knowledge that ultimately merge into the *Patent Icicle Plot*, representing vast technological rivers.

**Paper Matrix.** The *Paper Matrix* summarizes all the selected papers. As the experts sought to explore the interplay at the level of the scientific field, we use each column to represent a research field and each row to show a numerical citation range that indicates the extent to which the papers displayed in the row have been cited by patents. The citation number increases from the top to the bottom, i.e., the papers in the last row at the bottom of the view are those that are most cited by the patents. To deal with large-scale datasets, the *Paper Matrix* supports interactive hierarchical aggregation via both columns and rows. In particular, the rows can be aggregated by directly merging the citation ranges; the columns, i.e., research fields, can be aggregated by following the research field hierarchy introduced in the MAG.

A node in the *Paper Matrix* indicates a collection of aggregated papers, which can be illustrated either by a circle or by a star glyph (Fig. 7(B2)). The size of the circle indicates the number of papers in the collection, and the opacity represents the papers' averaged patent citation number. The star glyph summarizes the papers' statistical features (i.e., scientific facts in Section 4.2).

To reveal how broadly the papers in a research field  $\mathcal{F}$  affect technologies across different areas, we compute a diversity score:

$$\text{diversity} = - \sum_{i=1}^n P(x_i) \log P(x_i)$$

where  $i$  is a patent area. The score indicates the diversity of the area of the patents that cite the papers within  $\mathcal{F}$ . Intuitively, the larger the diversity score is, the larger  $\mathcal{F}$ 's influence will be. We use the size of the blue circle on top of each column (Fig. 3(C5)) to show this score.

**Patent Icicle Plot.** The *Patent Icicle Plot* uses an upturned icicle plot [42] to summarize the patent CPC categories from a three-level hierarchy (Fig. 3(C3)): section, subsection, and group [1]. Each rectangle represents a category, with the length encoding the number of patents in that category. The patent fields are by default in alphabetical order, as required by our experts to query patent categories more efficiently.

**Citation Flow.** The *Citation Flow* visualizes citation linkages from patents to papers. The flows start from a node in the *Paper Matrix*, converging at a field at the bottom of the *Paper Matrix*, and finally merging into *Patent Icicle Plot*, as if the knowledge is flowing from the broad scientific landscape to the technology fields (Fig. 4). The width of the flow represents the number of patent citations. The thicker the flow, the heavier the patent category relies on the knowledge from the collection of the connected papers.

We tend to lay out paper fields with similar patent citations near each other to conveniently explore interdisciplinary patent citations. At the same time, we also keep paper fields with more patent citations near the center of the *Paper Matrix* for a balanced visual appearance with thick flows in the center. To reduce the visual clutter caused by a large number of citation links, we reorder the research fields and route the flows to help reduce line crossing. Formally, the layout procedures described above can be formulated as an optimization problem with the following objective:

$$\alpha \sum_{i < j} w_{ij} \|x_{Q_i} - x_{Q_j}\|^2 + \beta \sum_{i=1}^m \|x_{Q_i} - x_{Q'_i}\|^2 + \gamma \sum_{i=1}^m \sum_{j=1}^n \|x_{Q_i} - x_{P_j}\|^2$$

where  $x$  represents the horizontal position of the paper or patent fields.  $Q$  represents the set of total  $m$  paper fields:  $\{Q_1, \dots, Q_m\}$ .  $Q'$  denotes the optimally ordered list of  $m$  paper fields, sorted by the number of patent citations in the field, which positions fields with the most citations in the center:  $\{Q'_1, \dots, Q'_m\}$ .  $P$  represents the set of total  $n$  patent categories:  $\{P_1, \dots, P_n\}$ .  $w_{ij}$  is the cosine similarity of paper pairs based on patent citation similarity. Intuitively, the first term in the objective function puts paper fields with similar patent citations near each other. The second term balances the flows by centralizing paper fields with more patent citations. The third term minimizes the paper-patent citation flow crossings. We balance the three parts based on the parameters  $\alpha$ ,  $\beta$ , and  $\gamma$ . The flows are rendered using cubic Bézier curves, which are bundled to reduce visual clutter and emphasize important flows.

## 5.3 Context Views

The system also provides a number of coordinated views for users to explore the connection between science and technology and identify untapped potential more systematically with context information.

**Researcher Overview View.** This view summarizes all researchers in a scatter plot (**I1, P1**, Fig. 3(A)) to help experts locate researchers of interest based on research profiles. Each circle represents a researcher, with the opacity encoding the P-index. We use a contour map to show the distributions of researchers. The x-axis and y-axis indicate two researcher metrics (introduced in Section 4), which can be interactively changed based on users' preferences.

**Researcher Statistics View.** This view summarizes characteristics and detailed researcher information for the selected group (**I2, P1**). It supports two visual modes: (1) bar charts and histograms that summarize the demographic information and SciSci metric distribution (Fig. 3(B1)); and (2) researcher cards that show the list of researchers (Fig. 3(B2)), which can be ranked by different SciSci metrics.

**Field Timeline.** The time dimension is also essential to help users identify trending topics in the dual frontiers. Thus, we designed two groups of horizon graphs to reveal the temporal evolution of different fields in science and technology (**E2**, Fig. 3(C1)). Each paper or patent field is represented by a horizon graph which shows the temporal evolution in a space-saving way. The x-axis is the timeline. The saturation of the area encodes the papers or patents published or granted in each field every year. Darker color indicates a higher value.

**Technology Inspection View.** This view shows additional context information about patent categories (**E3**, Fig. 3(D)). In addition to patent categories displayed in the *Interplay Graph*, the experts wanted more details on the assignee distribution and patent topics to aid in

© 2023 IEEE. This is the author's version of the article that has been published in IEEE Transactions on Visualization and Computer Graphics. The final version of this record is available at: [xx.xxxx/TVCG.201x.xxxxxxx/](https://doi.org/10.1109/TVCG.2023.3327387)

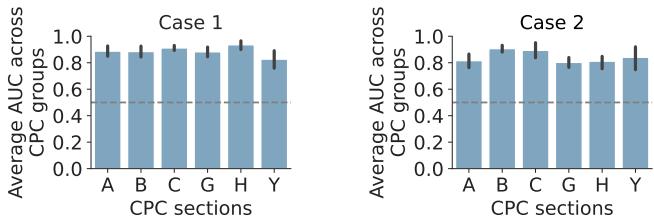


Fig. 5: The prediction performance (AUC) on the test set for two case studies. The average AUC and its 95% confidence interval are reported for each case and CPC section, based on the performance across CPC groups within that section. The dashed lines at 0.5 indicate the baseline performance of random guesses. Overall, the results demonstrate good performance for our prediction tasks.

decision-making (e.g., identifying potential partners and directions for invention commercialization). Thus, this view includes a two-level sunburst showing the proportion of assignees and a word cloud in the center showing the keywords of the selected patents.

**Science Inspection View.** To facilitate paper-level exploration, the *Science Inspection View* shows a paper list (E3, Fig. 3(E)) that can be ranked based on the statistical matrices introduced in Section 4.2 as well as a list of histograms showing distributions of paper-level metrics.

#### 5.4 Iterative Design Process

We went through multiple iterations for the design of the major visual components with our experts. Specifically, for the *Interplay Graph*, we initially used a graph where each node represented a paper or patent. However, it was not scalable for data with large volumes, and experts' feedback indicated that a field-level display was more critical. We thus aggregated the graph to the field level as a bipartite graph. One further feedback was to unfold papers in the same field into different groups based on patent citations, as they wanted to compare the differences among these paper groups. We thus unfolded paper-field nodes into a paper matrix with statistics summarized in each matrix node. Finally, our experts asked us to depict the temporal evolution of the dual frontiers for the purpose of identifying trending topics. We thus presented three choices: a timeline emphasizing temporal evolution [14], and two others emphasizing the citation structure (node-link (Section 5.2) and matrix [13]). The experts ultimately prioritized citation structure with the more intuitive node-link representation, leading us to depict the timeline as a secondary data dimension via the horizon graphs.

## 6 EVALUATION

We evaluated *InnovationInsights* through a quantitative study of the prediction model, two case studies, and interviews with experts.

#### 6.1 Quantitative Evaluation of the Prediction Model

Due to the class imbalance between papers that receive patent citations and those that do not, we evaluate the performance of our model using the AUC (Area Under the Curve) metric. Our final model is chosen from the epoch that produces the highest AUC score on the validation set. We evaluate the model by presenting its prediction performance and scalability in the two datasets used in two case studies (Section 6.2). We apply the prediction pipeline above on the top  $K$  patent CPC groups  $g$  (denoted as  $\mathbb{G}$ ) in the case studies in Section 6.2. We focus on the top 50 patent CPC groups because they cover more than 95% of patents citing our target papers in the case studies.

- AUC.** We present the AUC results on the test set across CPC groups by CPC section in Fig. 5. The overall prediction performance is good and remains robust across different CPC groups.
- Scalability.** The time complexity of GCN has been demonstrated to be linear in the number of graph edges and can scale to millions of edges [40]. In our two cases, the training time per epoch was less than 10 seconds with CPU-only implementations. We also ran the codes in parallel for different CPC groups to accelerate the prediction process. Moreover, the prediction model is pre-executed and does not impact the visualization system in real time.

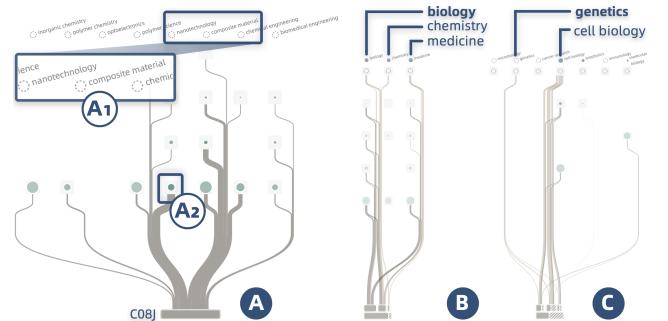


Fig. 6: The paper-patent citation flow in case 1. (A) *Nanotechnology* and *composite material* are the predominant science fields consumed by patent category *C08J*. (B) The historical flow shows most papers published by selected researchers are in basic *biology* journals. (C) The prediction flow shows the papers in *genetics* published by the selected faculty also have a high potential for innovation.

#### 6.2 Case Study

We invited our experts to explore the system. First, each expert explored the system independently. We then summarized their findings and formed two case studies to demonstrate our system.

##### 6.2.1 New Opportunities for University Innovation

We piloted the system with a premier research university and focused on 461 faculty who have at least one paper that has been cited by a patent (37K papers in total). This case study demonstrates how *InnovationInsights* enables our experts ( $E_A$  and  $E_B$ ) to uncover new innovation opportunities and facilitate knowledge transfer for researchers.

**Overview of the innovation landscape.** After loading the data from the university, the experts started from the *Citation Flow* (E1, Fig. 3(C3)), finding *material science*, *biology*, and *chemistry*, as the three predominant disciplines whose knowledge has spurred inventions across many patent categories. Highlighting one category *C* ("Chemistry; Metallurgy", Fig. 3(C4)), the experts noticed that this category draws from papers in not just *chemistry* but also *material science* and *biology*, emphasizing its interdisciplinary orientation. There was also high diversity in the paper-patent citation for *material science* (Fig. 3(C5)). Zooming in, the experts found at least six major patent categories, with *C08J* (*General Processes of Compounding*) as the most rapidly emerging technology area. This area primarily consumed knowledge from *nano**technology* and *composite material* (Fig. 6(A1)), which indicates that these research topics were important in this university. This patent category also relied heavily on papers with large numbers of patent citations (Fig. 6(A2)) in *nano**technology*, showing their importance in technology development. Choosing this cell (E3), the experts found that the top-citation papers were about *polymer nanocomposites*. When the experts went to the *Technology Inspection View* (E3, Fig. 3(D)), they found that unexpectedly, the largest university assignee was KFUPM (a university in Saudi Arabia, Fig. 3(D1)) rather than the university itself, and that this university just started to cite papers in the university in recent years. This finding represented fresh insights for the experts: "in most cases, we expect our own university to be the largest assignee citing our papers. So this finding is rather unexpected. Maybe there are collaboration opportunities with KFUPM, especially in *nano**technology*."

**Uncover hidden talents.** The experts are also interested in uncovering hidden talents and untapped innovation potential. Indeed, when examining the *Researcher Overview View* (I1, Fig. 3(A)), plotting the number of invention disclosures vs. patent-cited papers for each individual, our experts immediately discovered a fascinating insight: at the bottom of Fig. 3(A1) lay an interesting group of researchers. They themselves had no invention disclosures, yet their papers had been cited frequently by other patents. "Who are these people?" our experts asked immediately. Zooming in on the *Researcher Statistics View* (I2, Fig. 3(B1)) revealed that most of them were full professors at the medical school. Most of their papers were published in basic biology journals (Fig. 6(B)), yet surprisingly, they were being cited heavily by companies, finding widespread uses in the private sector (Fig. 3(D2)).

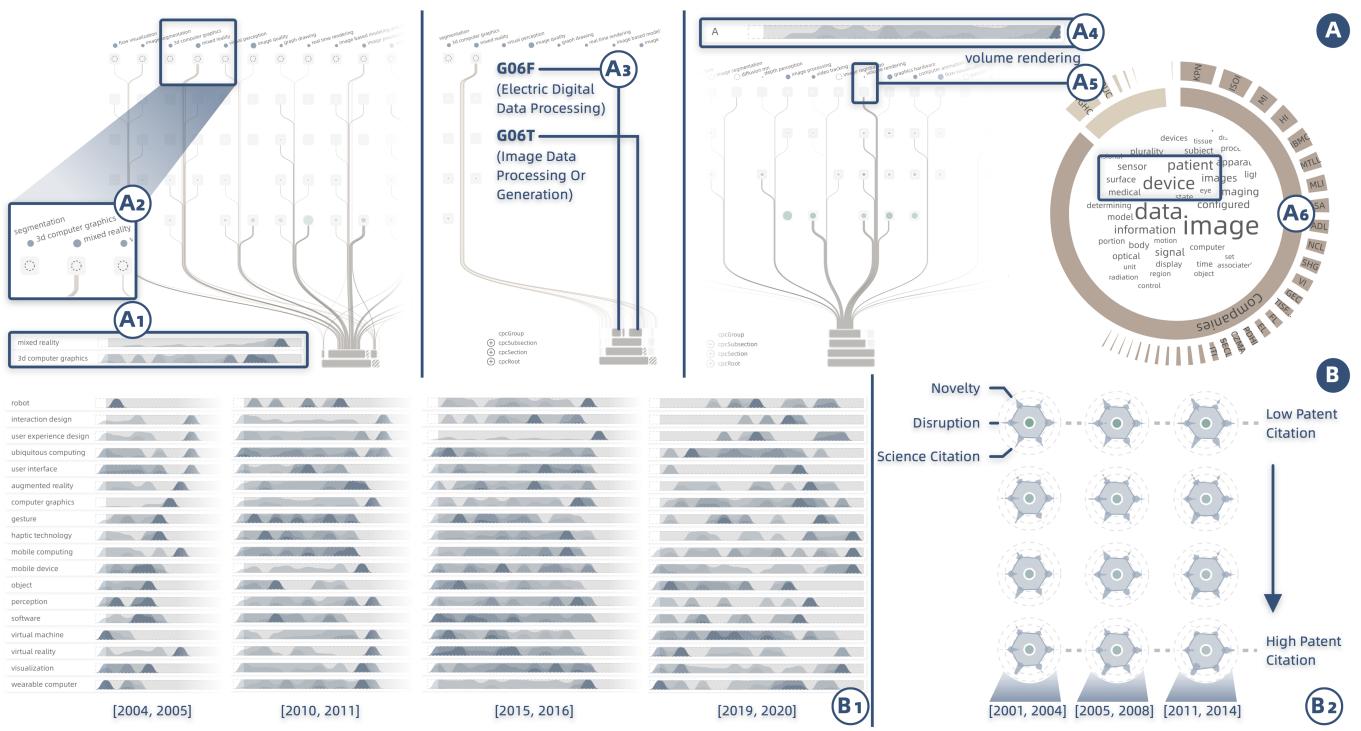


Fig. 7: Innovation insights in case 2. (A1)-(A2) The spotting hot new frontiers in the VIS community include *mixed reality* and *3d computer graphics*. (A3) *mixed reality* is predicted to be used in a variety of patent categories. (A4)-(A6) Papers in *volume rendering* are still being cited by new patents in medicine. (B1) New technologies are increasingly relying on older scientific knowledge. (B2) Relationships between patent citation and other SciSci metrics: science citations and novelty have positive relationships, while disruption has a negative relationship with patent citations.

The experts thus ranked these researchers by P-index to locate those with high commercialization potential. They quickly noticed the second faculty (Fig. 3(B2)) who had a rather high P-index and was the only female faculty among the top P-index researchers. She had no inventions, but many of her papers were cited by patents from private companies. From the *Interplay Graph* (E1, Fig. 6(C)), the experts gathered that most of her papers were in *cell biology* but were increasingly cited by patents in the C12N (*Microorganisms or Enzymes; Mutation or Genetic Engineering*) category (E2, Fig. 3(C6)). And many of these patents were from the same company, namely *Immatics Biotechnologies* in Germany (E3, Fig. 3(D3)). This discovery prompted our experts to hold an immediate follow-up conversation with the faculty member. It turned out that while this faculty had done a sabbatical in Germany, she was completely unaware of this company, or the fact that they were drawing heavily on her research. Using our system, the experts further showed the faculty member other prediction results on which of her other papers in *genetics* (P1, Fig. 6(C)) revealed a high potential for innovation. Two weeks following the conversation, she submitted new invention disclosures to the university, for the first time in her career!

The case shows that *InnovationInsights* is highly effective in uncovering new innovation potential and opportunities in research institutions.

## 6.2.2 Innovation Insights in VIS Communities

Our experts ( $E_C$  and  $E_D$ ) used *visualization* as an exemplary field to explore the dual frontiers of science and technology. Specifically, we identified 2016 top-publishing researchers based on 6 major journals and 13 leading conferences in VIS and analyzed all their publications. **Spotting hot new frontiers.** The experts first focused on the topic of *visualization*. In the paper field timeline (E2, Fig. 7(A1)), two topics quickly stood out: *mixed reality* and *3d computer graphics*. In addition to being highly popular, these two topics were paired together in the *Interplay Graph* (Fig. 7(A2)), suggesting that they were frequently co-cited by similar patent categories. The experts then filtered recent papers with high prediction scores (P1). Interestingly, while other areas (e.g., *image processing* and *real time rendering*) had tended to dominate the field, recent papers in *mixed reality* are characterized by some of the highest prediction scores (Fig. 7(A2)), with prediction flows coming from a variety of patenting domains (e.g., G06F (*Electric Digital Data Processing*) and G06T (*Image Data Processing*)) (Fig. 7(A3)). “*This*

*really speaks to the application potential of ‘mixed reality’*,” as our experts commented. Our experts further noticed that the patent category A (*Human Necessities*) was rising in popularity (Fig. 7(A4)). Zooming in, they found that *volume rendering* was the most applied science topic (E1, Fig. 7(A5)) and was primarily used in A61B (*Diagnosis; Surgery*). Highlighting A61B and in the *Technology Inspection View* (Fig. 7(A6)), the experts found that these patents were related to medical devices. Curious about what papers were highly cited by patents, they clicked the paper cell in the last row (E3). Somewhat unexpectedly, the highly cited papers in this emerging patent category were not new papers; rather, they were canonical papers in the field (e.g., [27]). “*Interesting! Even after ten years, papers in ‘volume rendering’ are still being cited by new patents. These papers are canons in the field. They have a lasting impact on technology development.*”

**New vs. canonical knowledge?** Intrigued by the preceding findings, the experts returned their attention to *human computer interaction* and filtered patent application years based on three ranges in the early, mid, and end of the period between 2001 and 2020 (E2). In recent years (e.g., [2019, 2020] and [2015, 2016], Fig. 7(B1)), the patent citation of old and new papers was diverse. Patents either cited new paper fields or cited fallen fields many years ago. However, when time went back ten more years ago, patents tended to cite more new paper fields at that time (e.g., [2010, 2011] and [2004, 2005], Fig. 7(B1)). “*This finding supports our hypothesis that new technologies are increasingly relying on older scientific knowledge.*”

**What kinds of papers tend to see greater uses in technology?** The experts also explored the characteristics of papers that are heavily cited by patents (E1). To obtain a general view, the experts zoomed out and focused on one of the most prominent fields, *computer vision*, as an example. Inspecting papers published in different periods (Fig. 7(B2)), they found that generally, science citations and novelty had positive relationships with patent citations, whereas the disruption score was negatively correlated with patent citations. “*These results are consistent with our recent findings [77]. Papers that are highly valued within science also see greater practical uses. At the same time, these results also raise new research questions regarding the relationships between novelty, disruption, and patent citations.*”

Overall, these cases demonstrate the effectiveness of our system in navigating multiple dimensions of data and uncovering new insights.

© 2023 IEEE. This is the author's version of the article that has been published in IEEE Transactions on Visualization and Computer Graphics. The final version of this record is available at: [xx.xxxx/TVCG.2023.xxxxxxx](https://doi.org/10.1109/TVCG.2023.3327387)

### 6.3 Expert Interview

We collected feedback from experts in Section 3.1 and interviewed six external experts who had used *InnovationInsights* for the first time.  $P_A$  and  $P_B$  are innovation managers in the TTO of a university.  $P_C$  and  $P_D$  are researchers in SciSci. Although the target users of *InnovationInsights* are the above two groups, the dataset in case 2 may also interest VIS researchers. Thus, we also included two VIS researchers ( $P_E$  and  $P_F$ ) who have three-year visualization expertise. Each interview with an external expert lasted about 90 minutes. We first briefly introduced the project background, including the analytical tasks and data sources. Then we used case 1 in Section 6.2.1 to demonstrate the system workflow and visual encodings. Third, they were asked to explore the system in a think-aloud manner. Finally, we had a semi-structured interview. We took notes on their comments and findings during the process. The feedback from the two groups of experts is summarized below.

**System Workflow.** All experts appreciated the clear workflow. They were able to learn the system logic quickly, from researcher identification to paper-patent citation exploration. We also observed diverse analysis focuses between experts from different fields. Innovation managers from the TTO focused more on researchers and their research fields, while SciSci and VIS experts tended to start directly with scientific fields. As  $P_C$  said, “*we are more interested in the general findings over scientific fields.*” Nevertheless, they agreed that the current workflow was able to meet different needs through filtering schemes.

**SciSci Metrics and Prediction Model.** All of the experts showed interest in these metrics. Those from the TTO were particularly interested in the *P-index* obtained from the prediction model.  $P_B$  was excited about this straightforward approach to locating faculty with high commercialization potential.  $E_A$  suggested making the prediction more transparent to help them understand the mechanism behind the prediction.  $P_E$  and  $P_F$  found metrics such as *team size* and *novelty* interesting, “*these metrics provide us new perspectives to look at papers besides paper citations.*”

**Visualization and Interactions.** The experts noted that the visual components in the system were intuitive and satisfied all the analytical tasks. Many of them appreciated the *Interplay Graph*.  $E_A$  and  $P_F$  reported that it took some time to understand the encoding, but it eventually became very useful and intuitive.  $P_C$  especially appreciated the y-axis in the paper matrix, as she could locate papers with high patent citations more quickly.  $P_B$  and  $P_E$  liked the *Technology Inspection View* and also suggested, “*it would be more interesting if we could check the relationships between these assignees.*”  $P_C$  and  $P_F$  liked the glyph design, “*it makes the comparison between paper groups much easier. But it would be great if labels could be added to show the meaning of dimensions.*”  $E_A$  also mentioned the timeline as important to check recent hot areas as a temporal indicator of innovation potential.

**Suggestions.** Despite universities and research fields,  $E_D$  also wanted to filter researchers at the regional level to compare innovation.  $E_A$  suggested using text analysis to reveal more detail about paper-patent citations, such as the distinction between strong citations (i.e., cite the core knowledge in the paper) and weak citations (i.e., cite a paper in the background).

## 7 DISCUSSION

This section discusses the significance and generalizability, lessons learned, and limitations of our work.

**Significance and Generalizability.** Science provides a foundation for many practical applications in human society, but the pathway through which basic understanding leads to technological development is neither visible nor intuitive. Consider Einstein’s theory of general relativity, deemed as the discovery of the 20th century. Among the myriad innovations it spurred, it proved essential for the Global Positioning System (GPS) through time dilation corrections. The GPS system then provided the technical foundation for applications such as Uber. The ability to effectively trace and visualize the evolving dual frontiers of science and technology is therefore crucial to understanding how science drives practical applications and leads to rising standards of living. Our system not only fulfills our original design purposes, allowing users to better identify the sources of technical inventions and understand the holistic impact of scientific research; it also enables an array of new applications for researchers and research institutions,

ranging from identifying untapped innovation potentials within an institution to forging new partnership opportunities between science and industry. Moreover, the proposed SciSci metrics, prediction model, and visualization system can be adapted for studying other upstream (e.g., funding) and downstream (e.g., policy documents) linkages to science [77]. When adapting the system to other domains, we suggest replacing the patent data with other upstream or downstream data and refining the metrics and prediction model to suit specific scenarios.

**Lessons Learned.** The design study with experts from multiple fields provides valuable insights into conducting interdisciplinary research. First, regularly discussing data insights with experts substantially accelerates progress through timely clarifications of analytical goals. We started with exploratory analysis and used static charts to discuss initial findings. This practice helped us quickly verify insights, facilitating adjustments to our analytical goals. Moreover, the initial findings also inform experts about their practice. During the process, our experts identified researchers with high commercialization potential (Section 6.2.1) and helped them to submit invention disclosures. Second, the visual design of data with a complex structure requires involving multiple design choices. Experts often lack clarity on the level of detail the data should be presented, generally desiring maximum information. It is helpful to offer alternatives at varied granularities and prioritize data dimensions based on their importance for analysis goals.

**Limitations.** Our system is not without limitations. First, the current prediction model relies on citations between papers without considering those between papers and patents. Future work may use other GNN models designed for heterogeneous graphs, which can incorporate different types of nodes (e.g., papers and patents). In addition, the P-index, derived from a GNN model trained on papers published between 2001 and 2014, can be improved by expanding the training dataset with recent papers. We also plan to keep collaborating with TTO experts for model validation and enhancement. Second, due to restrictions on data availability, the scope of our study is limited to a single university or specific field with patent data from the USPTO. Future research may find patent data worldwide and extend the partnership to other universities in order to encompass a wider innovation landscape. Third, the current system exhibits some latency when computing results in real time, particularly with large datasets. We intend to resolve this issue using progressive visual analytics [7, 29]. Lastly, to ensure the system’s long-term utility, we plan to regularly maintain the system and the model with the most recent data (e.g., OpenAlex [2]).

## 8 CONCLUSION AND FUTURE DIRECTIONS

This paper presents *InnovationInsights*, a first-of-its-kind visualization system for researchers and research institutions to explore the complex interactions between science and technology. It supports analyzing multiple entities (e.g., researchers, papers, and patents) through descriptive and predictive analyses. Coordinated views with intuitive interactions are developed to support analysis. Two case studies, expert interviews, and our engagement project with a partner university demonstrate the substantial utility and potential impact of our system. In the future, we plan to integrate more data and release an online system for public use.

This work opens up several fruitful future directions, especially at the intersection of data visualization and the science of science. For example, beyond understanding the impact of science on technological development, visualization approaches would prove fruitful for analyzing the broad uses of science across several crucial downstream applications in society, tracing how science is used in the hallways of governments through science-policy linkages, as well as how science enables life-saving drugs and therapeutics by incorporating clinical trials data. Further, data visualization techniques can also help us better understand the multi-dimensional impacts of funding on scientific progress and individual careers. Developing novel and efficient visual analytics systems to analyze large-scale data, spanning from upstream funding to science to downstream applications, will usefully serve a diverse range of stakeholders, including university leaders, private companies and investors, funding agencies, policymakers, and researchers themselves. Given the crucial role of science in improving the human condition, such systems have the potential to unlock enormous value for science—and for society at large.

## ACKNOWLEDGMENTS

The authors wish to thank Benjamin F. Jones for his helpful comments. They also would like to express their appreciation to anonymous reviewers for their valuable comments. This work is supported by the Air Force Office of Scientific Research under award numbers FA9550-17-1-0089 and FA9550-19-1-0354, the Alfred P. Sloan Foundation G-2019-12485, and the Future Wanxiang Foundation.

## SUPPLEMENTARY MATERIALS

The authors provide the following materials at <https://kellogg-cssi.github.io/InnovationInsights/>: (1) a video introducing the research background and system interface, (2) a demo video presenting a case study of the system, and (3) a document describing two datasets used in case studies.

## REFERENCES

- [1] Cooperative Patent Classification (CPC). <https://www.uspto.gov/patents/search/classification-standards-and-development>. 3, 6
- [2] OpenAlex. <https://openalex.org/>. 9
- [3] PatentsView. <https://patentsview.org/>. 1, 3
- [4] USPTO. <https://www.uspto.gov/>. 3
- [5] M. Abdelaal, N. D. Schiele, K. Angerbauer, K. Kurzhals, M. Sedlmair, and D. Weiskopf. Comparative evaluation of bipartite, node-link, and matrix-based network representations. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):896–906, 2022. doi: [10.1109/TVCG.2022.3209427](https://doi.org/10.1109/TVCG.2022.3209427) 2
- [6] M. Ahmadpoor and B. F. Jones. The dual frontier: Patented inventions and prior scientific advance. *Science*, 357(6351):583–587, 2017. doi: [10.1126/science.aam9527](https://doi.org/10.1126/science.aam9527) 1, 2
- [7] M. Angelini, G. Santucci, H. Schumann, and H.-J. Schulz. A review and characterization of progressive visual analytics. In *Informatics*, vol. 5, p. 31, 2018. doi: [10.3390/informatics5030031](https://doi.org/10.3390/informatics5030031) 9
- [8] E. Ankan, W. Dou, D. Strumsky, D. X. Wang, T. Rabinowitz, and W. Zadrozy. Exploring emerging technologies using patent data and patent classifications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2012. 1, 2
- [9] F. Beck, M. Burch, S. Diehl, and D. Weiskopf. A taxonomy and survey of dynamic graph visualization. In *Computer Graphics Forum*, vol. 36, pp. 133–159. Wiley Online Library, 2017. doi: [10.1111/cgf.12791](https://doi.org/10.1111/cgf.12791) 2
- [10] K. Börner. *Atlas of science: Visualizing what we know*. Mit Press, 2010. doi: [10.1007/s11192-011-0409-7](https://doi.org/10.1007/s11192-011-0409-7) 2
- [11] K. W. Boyack, B. N. Wylie, G. S. Davidson, and D. K. Johnson. Analysis of patent databases using VxInsight. Technical report, Sandia National Lab. (SNL-NM), Albuquerque, NM (United States); Sandia National Lab. (SNL-CA), Livermore, CA (United States), 2000. 2
- [12] K. Buchin, B. Speckmann, and K. Verbeek. Flow map layout via spiral trees. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2536–2544, 2011. doi: [10.1109/TVCG.2011.202](https://doi.org/10.1109/TVCG.2011.202) 2
- [13] M. Burch, B. Schmidt, and D. Weiskopf. A matrix-based visualization for exploring dynamic compound digraphs. In *17th International Conference on Information Visualisation*, pp. 66–73. IEEE, 2013. doi: [10.1109/IV.2013.87](https://doi.org/10.1109/IV.2013.87)
- [14] M. Burch, C. Vehlow, F. Beck, S. Diehl, and D. Weiskopf. Parallel edge splatting for scalable dynamic graph visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2344–2353, 2011. doi: [10.1109/TVCG.2011.226](https://doi.org/10.1109/TVCG.2011.226) 7
- [15] V. Bush. *Science—the Endless Frontier: a Report to the President on a Program for Postwar Scientific Research*, vol. 90. National Science Foundation, 1990. 1
- [16] H. Cao, Y. Lu, Y. Deng, D. A. McFarland, and M. S. Bernstein. Breaking out of the ivory tower: A large-scale analysis of patent citations to HCI research. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2023. doi: [10.1145/3544548.3581108](https://doi.org/10.1145/3544548.3581108) 2
- [17] G. Y.-Y. Chan, P. Xu, Z. Dai, and L. Ren. Vibr: Visualizing bipartite relations at scale with the minimum description length principle. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):321–330, 2018. doi: [10.1109/TVCG.2018.2864826](https://doi.org/10.1109/TVCG.2018.2864826) 2
- [18] C. Chen. *Mapping scientific frontiers: The quest for knowledge visualization*. Springer, 2003. doi: [10.1007/978-1-4471-5128-9\\_2](https://doi.org/10.1007/978-1-4471-5128-9_2)
- [19] J. Chen, M. Ling, R. Li, P. Isenberg, T. Isenberg, M. Sedlmair, T. Möller, R. S. Laramee, H.-W. Shen, K. Wünsche, et al. VIS30K: A collection of figures and tables from IEEE Visualization conference publications. *IEEE Transactions on Visualization and Computer Graphics*, 27(9):3826–3833, 2021. doi: [10.1109/TVCG.2021.3054916](https://doi.org/10.1109/TVCG.2021.3054916) 2
- [20] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *ACL*, 2020. doi: [10.48550/arXiv.2004.07180](https://doi.org/10.48550/arXiv.2004.07180) 5
- [21] A. Dattolo, M. Corbatto, and M. Angelini. Authoring and reviewing bibliographies: Design and development of a visual analytics online platform. *IEEE Access*, 10:21631–21645, 2022. doi: [10.1109/ACCESS.2022.3153027](https://doi.org/10.1109/ACCESS.2022.3153027) 2
- [22] D. Deng, Y. Wu, X. Shu, J. Wu, S. Fu, W. Cui, and Y. Wu. VisImages: A fine-grained expert-annotated visualization dataset. *IEEE Transactions on Visualization & Computer Graphics*, 1(01):1–1, 2022. doi: [10.1109/TVCG.2022.3155440](https://doi.org/10.1109/TVCG.2022.3155440) 2
- [23] Z. Deng, S. Chen, X. Xie, G. Sun, M. Xu, D. Weng, and Y. Wu. Multi-level visual analysis of aggregate geo-networks. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–16, 2022. doi: [10.1109/TVCG.2022.3229953](https://doi.org/10.1109/TVCG.2022.3229953) 2
- [24] A. Dong, W. Zeng, X. Chen, and Z. Cheng. VIStory: Interactive storyboard for exploring visual information in scientific publications. In *Proceedings of the 12th International Symposium on Visual Information Communication and Interaction*, pp. 1–8, 2019. doi: [10.1145/3356422.3356430](https://doi.org/10.1145/3356422.3356430) 2
- [25] M. Dörk, N. H. 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](https://doi.org/10.1109/TVCG.2012.252) 2
- [26] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. HierarchicalTopics: Visually exploring large text collections using topic hierarchies. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2002–2011, 2013. doi: [10.1109/TVCG.2013.162](https://doi.org/10.1109/TVCG.2013.162) 2
- [27] R. Fattal and D. Lischinski. Variational classification for visualization of 3D ultrasound data. In *Proceedings Visualization. VIS'01.*, pp. 403–410. IEEE, 2001. doi: [10.1109/VISUAL.2001.964539](https://doi.org/10.1109/VISUAL.2001.964539) 8
- [28] 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, 2016. doi: [10.1109/TVCG.2016.2610422](https://doi.org/10.1109/TVCG.2016.2610422) 2
- [29] J.-D. Fekete, D. Fisher, A. Nandi, and M. Sedlmair. Progressive data analysis and visualization, 2019. doi: [10.4230/DagRep.8.10.1](https://doi.org/10.4230/DagRep.8.10.1) 9
- [30] M. Fey and J. E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. doi: [10.48550/arXiv.1903.02428](https://doi.org/10.48550/arXiv.1903.02428) 5
- [31] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018. doi: [10.1126/science.aa00185](https://doi.org/10.1126/science.aa00185) 1
- [32] R. González-Márquez, L. Schmidt, B. M. Schmidt, P. Berens, and D. Kobak. The landscape of biomedical research. *bioRxiv*, pp. 2023–04, 2023. doi: [10.1101/2023.04.10.536208](https://doi.org/10.1101/2023.04.10.536208) 2
- [33] Z. Guo, J. Tao, S. Chen, N. Chawla, and C. Wang. SD<sup>2</sup>: Slicing and dicing scholarly data for interactive evaluation of academic performance. *IEEE Transactions on Visualization and Computer Graphics*, 2022. doi: [10.1109/TVCG.2022.3163727](https://doi.org/10.1109/TVCG.2022.3163727) 2
- [34] H. Hao, Y. Cui, Z. Wang, and Y.-S. Kim. Thirty-two years of IEEE VIS: Authors, fields of study and citations. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1016–1025, 2022. doi: [10.1109/TVCG.2022.3209422](https://doi.org/10.1109/TVCG.2022.3209422) 2
- [35] I. Herman, G. Melançon, and M. S. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000. doi: [10.1109/2945.841119](https://doi.org/10.1109/2945.841119) 2
- [36] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. D. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko. Vispubdata.org: A metadata collection about IEEE Visualization (VIS) publications. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2199–2206, 2016. doi: [10.1109/TVCG.2016.2615308](https://doi.org/10.1109/TVCG.2016.2615308) 2
- [37] B. F. Jones. Science and innovation: The under-fueled engine of prosperity. *Rebuilding the Post-Pandemic Economy*, ed. Melissa S. Kearney and Amy Ganz (Washington DC: Aspen Institute Press, 2021), 2021. 1
- [38] F. Karimi, C. Wagner, F. Lemmerich, M. Jadidi, and M. Strohmaier. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the International Conference*

© 2023 IEEE. This is the author's version of the article that has been published in IEEE Transactions on Visualization and Computer Graphics. The final version of this record is available at: [xx.xxxx/TVCG.2023.xxxxxxx](https://doi.org/10.1109/TVCG.2023.3327387)

- Companion on World Wide Web, pp. 53–54, 2016. doi: [10.1145/2872518.2889385](https://doi.org/10.1145/2872518.2889385) 4
- [39] Y. G. Kim, J. H. Suh, and S. C. Park. Visualization of patent analysis for emerging technology. *Expert Systems with Applications*, 34(3):1804–1812, 2008. doi: [10.1016/j.eswa.2007.01.033](https://doi.org/10.1016/j.eswa.2007.01.033) 2
- [40] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017. doi: [10.48550/arXiv.1609.02907](https://doi.org/10.48550/arXiv.1609.02907) 4, 5, 7
- [41] S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative integration of visual insights during scalable patent search and analysis. *IEEE Transactions on Visualization and Computer Graphics*, 17(5):557–569, 2010. doi: [10.1109/TVCG.2010.85](https://doi.org/10.1109/TVCG.2010.85) 1, 2
- [42] J. B. Kruskal and J. M. Landwehr. Icicle plots: Better displays for hierarchical clustering. *The American Statistician*, 37(2):162–168, 1983. doi: [10.2307/2685881](https://doi.org/10.2307/2685881) 6
- [43] D. O. Kutz. Examining the evolution and distribution of patent classifications. In *Proceedings of the International Conference on Information Visualisation*, pp. 983–988. IEEE, 2004. doi: [10.1109/IV.2004.1320261](https://doi.org/10.1109/IV.2004.1320261) 1
- [44] S. Latif and F. Beck. VIS Author Profiles: Interactive descriptions of publication records combining text and visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):152–161, 2018. doi: [10.1109/TVCG.2018.2865022](https://doi.org/10.1109/TVCG.2018.2865022) 2
- [45] G. Li and X. Yuan. GoTreeScape: Navigate and explore the tree visualization design space. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–17, 2022. doi: [10.1109/TVCG.2022.3215070](https://doi.org/10.1109/TVCG.2022.3215070) 2
- [46] Z. Li, C. Zhang, S. Jia, and J. Zhang. Galex: Exploring the evolution and intersection of disciplines. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1182–1192, 2019. doi: [10.1109/TVCG.2019.2934667](https://doi.org/10.1109/TVCG.2019.2934667) 1, 2
- [47] W. Liang, S. Elrod, D. A. McFarland, and J. Zou. Systematic analysis of 50 years of Stanford University technology transfer and commercialization. *Patterns*, 3(9):100584, 2022. doi: [10.1016/j.patter.2022.100584](https://doi.org/10.1016/j.patter.2022.100584) 1
- [48] L. Liu, N. Dehmamy, J. Chown, C. L. Giles, and D. Wang. Understanding the onset of hot streaks across artistic, cultural, and scientific careers. *Nature Communications*, 12(1):5392, 2021. doi: [10.1038/s41467-021-25477-8](https://doi.org/10.1038/s41467-021-25477-8) 2
- [49] M. Marx and A. Fuegi. Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, 41(9):1572–1594, 2020. doi: [10.1002/smj.3145](https://doi.org/10.1002/smj.3145) 1, 2, 3
- [50] M. Marx and A. Fuegi. Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. *Journal of Economics & Management Strategy*, 31(2):369–392, 2022. doi: [10.1111/jems.12455](https://doi.org/10.1111/jems.12455) 2, 3
- [51] S. Morris, C. DeYong, Z. Wu, S. Salman, and D. Yemenu. DIVA: A visualization system for exploring document databases for technology forecasting. *Computers & Industrial Engineering*, 43(4):841–862, 2002. doi: [10.1016/S0360-8352\(02\)00143-2](https://doi.org/10.1016/S0360-8352(02)00143-2)
- [52] T. Mühlbacher, L. Linhardt, T. Möller, and H. Piringer. TreePOD: Sensitivity-aware selection of Pareto-optimal decision trees. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):174–183, 2017. doi: [10.1109/TVCG.2017.2745158](https://doi.org/10.1109/TVCG.2017.2745158) 2
- [53] A. Narechania, A. Karduni, R. Wesslen, and E. Wall. VITALITY: Promoting serendipitous discovery of academic literature with transformers & visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):486–496, 2021. doi: [10.1109/TVCG.2021.3114820](https://doi.org/10.1109/TVCG.2021.3114820) 1, 2
- [54] C. Nobre, M. Streit, and A. Lex. Juniper: A tree+ table approach to multivariate graph visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):544–554, 2018. doi: [10.1109/TVCG.2018.2865149](https://doi.org/10.1109/TVCG.2018.2865149) 2
- [55] A. Pandey, U. H. Syeda, C. Shah, J. A. Guerra-Gomez, and M. A. Borkin. A state-of-the-art survey of tasks for tree design and evaluation with a curated task dataset. *IEEE Transactions on Visualization and Computer Graphics*, 28(10):3563–3584, 2021. doi: [10.1109/TVCG.2021.3064037](https://doi.org/10.1109/TVCG.2021.3064037) 2
- [56] D. Phan, L. Xiao, R. Yeh, and P. Hanrahan. Flow map layout. In *IEEE Symposium on Information Visualization*, pp. 219–224. IEEE, 2005. doi: [10.1109/INFVIS.2005.1532150](https://doi.org/10.1109/INFVIS.2005.1532150) 2
- [57] Y. Qian, P. Expert, P. Panzarasa, and M. Barahona. Geometric graphs from data to aid classification tasks with Graph Convolutional Networks. *Patterns*, 2(4):100237, 2021. doi: [10.1016/j.patter.2021.100237](https://doi.org/10.1016/j.patter.2021.100237) 4
- [58] Y. Qian, P. Expert, T. Rieu, P. Panzarasa, and M. Barahona. Quantifying the alignment of graph and features in deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(4):1663–1672, 2021. doi: [10.1109/TNNLS.2020.3043196](https://doi.org/10.1109/TNNLS.2020.3043196) 4
- [59] A. Sarvghad, R. Franqui-Nadal, R. Reznik-Zellen, R. Chawla, and N. Mahyar. Scientometric analysis of interdisciplinary collaboration and gender trends in 30 years of IEEE VIS publications. *IEEE Transactions on Visualization and Computer Graphics*, 2022. doi: [10.1109/TVCG.2022.3158236](https://doi.org/10.1109/TVCG.2022.3158236) 2
- [60] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, 2012. doi: [10.1109/TVCG.2012.213](https://doi.org/10.1109/TVCG.2012.213) 3
- [61] M. Sun, J. Zhao, H. Wu, K. Luther, C. North, and N. Ramakrishnan. The effect of edge bundling and seriation on sensemaking of biclusters in bipartite graphs. *IEEE Transactions on Visualization and Computer Graphics*, 25(10):2983–2998, 2018. doi: [10.1109/TVCG.2018.2861397](https://doi.org/10.1109/TVCG.2018.2861397) 2
- [62] N. Tovanich, P. Dragicevic, and P. Isenberg. Gender in 30 years of IEEE Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):497–507, 2021. doi: [10.1109/TVCG.2021.3114787](https://doi.org/10.1109/TVCG.2021.3114787) 2
- [63] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013. doi: [10.1126/science.1240474](https://doi.org/10.1126/science.1240474) 4
- [64] S. Van Den Elzen and J. J. Van Wijk. BaobabView: Interactive construction and analysis of decision trees. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 151–160. IEEE, 2011. doi: [10.1109/VAST.2011.6102453](https://doi.org/10.1109/VAST.2011.6102453) 2
- [65] D. Wang and A.-L. Barabási. *The science of science*. Cambridge University Press, 2021. doi: [10.1017/9781108610834](https://doi.org/10.1017/9781108610834) 1
- [66] D. Wang, C. Song, and A.-L. Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013. doi: [10.1126/science.1237825](https://doi.org/10.1126/science.1237825) 2
- [67] K. Wang, Z. Shen, C. Huang, C.-H. Wu, D. Eide, Y. Dong, J. Qian, A. Kanakia, A. Chen, and R. Rogahn. A review of Microsoft Academic Services for science of science studies. *Frontiers in Big Data*, 2:45, 2019. doi: [10.3389/fdata.2019.00045](https://doi.org/10.3389/fdata.2019.00045) 3, 4
- [68] X. Wang, J. He, Z. Jin, M. Yang, Y. Wang, and H. Qu. M2lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):802–812, 2021. doi: [10.1109/TVCG.2021.3114794](https://doi.org/10.1109/TVCG.2021.3114794) 2
- [69] Y. Wang, H. Liang, X. Shu, J. Wang, K. Xu, Z. Deng, C. Campbell, B. Chen, Y. Wu, and H. Qu. Interactive visual exploration of longitudinal historical career mobility data. *IEEE Transactions on Visualization and Computer Graphics*, 28(10):3441–3455, 2021. doi: [10.1109/TVCG.2021.3067200](https://doi.org/10.1109/TVCG.2021.3067200) 2
- [70] Y. Wang, T.-Q. Peng, H. Lu, H. Wang, X. Xie, H. Qu, and Y. Wu. Seek for success: A visualization approach for understanding the dynamics of academic careers. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):475–485, 2021. doi: [10.1109/TVCG.2021.3114790](https://doi.org/10.1109/TVCG.2021.3114790) 2
- [71] Y. Wang, M. Yu, G. Shan, H.-W. Shen, and Z. Lu. VISPubComPAS: a comparative analytical system for visualization publication data. *Journal of Visualization*, 22:941–953, 2019. doi: [10.1007/s12650-019-00585-2](https://doi.org/10.1007/s12650-019-00585-2) 2
- [72] F. Windhager, A. Amor-Amorós, M. Smuc, P. Federico, L. ZenkVxInsight, and S. Miksch. A concept for the exploratory visualization of patent network dynamics. In *IVAPP*, pp. 268–273, 2015. doi: [10.5220/0005360002680273](https://doi.org/10.5220/0005360002680273) 2
- [73] A. Wu, D. Deng, F. Cheng, Y. Wu, S. Liu, and H. Qu. In defence of visual analytics systems: Replies to critics. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1026–1036, 2022. doi: [10.1109/TVCG.2022.3209360](https://doi.org/10.1109/TVCG.2022.3209360) 2
- [74] L. Wu, D. Wang, and J. A. Evans. Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382, 2019. doi: [10.1038/s41586-019-0941-9](https://doi.org/10.1038/s41586-019-0941-9) 2, 4
- [75] 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, 2015. doi: [10.1109/TVCG.2015.2468151](https://doi.org/10.1109/TVCG.2015.2468151) 2
- [76] Y. Ye, R. Huang, and W. Zeng. VISAtlas: An image-based exploration and query system for large visualization collections via neural image embedding. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–15, 2022. doi: [10.1109/TVCG.2022.3229023](https://doi.org/10.1109/TVCG.2022.3229023) 2
- [77] Y. Yin, Y. Dong, K. Wang, D. Wang, and B. F. Jones. Public use and public funding of science. *Nature Human Behaviour*, 6(10):1344–1350, 2022. doi: [10.1038/s41562-022-01397-5](https://doi.org/10.1038/s41562-022-01397-5) 1, 2, 8, 9
- [78] Y. Yin, J. Gao, B. F. Jones, and D. Wang. Coevolution of policy and science during the pandemic. *Science*, 371(6525):128–130, 2021. doi: [10.1126/science.abe3084](https://doi.org/10.1126/science.abe3084) 2