# Seek for Success: A Visualization Approach for Understanding the Dynamics of Academic Careers

Yifang Wang, Tai-Quan Peng, Huihua Lu, Haoren Wang, Xiao Xie, Huamin Qu, and Yingcai Wu

**Abstract**— How to achieve academic career success has been a long-standing research question in social science research. With the growing availability of large-scale well-documented academic profiles and career trajectories, scholarly interest in career success has been reinvigorated, which has emerged to be an active research domain called the Science of Science (i.e., SciSci). In this study, we adopt an innovative dynamic perspective to examine how individual and social factors will influence career success over time. We propose *ACSeeker*, an interactive visual analytics approach to explore the potential factors of success and how the influence of multiple factors changes at different stages of academic careers. We first applied a Multi-factor Impact Analysis framework to estimate the effect of different factors on academic career success over time. We then developed a visual analytics system to understand the dynamic effects interactively. A novel timeline is designed to reveal and compare the factor impacts based on the whole population. A customized career line showing the individual career development is provided to allow a detailed inspection. To validate the effectiveness and usability of *ACSeeker*, we report two case studies and interviews with a social scientist and general researchers.

**Index Terms**—Career Analysis, Academic Profiles, Science of Science, Publication Data, Citation Data, Sequence Analysis

## 1 Introduction

How to achieve individual career success is a long-standing research question that has been studied in various social science disciplines, such as sociology, organizational behaviors, and information science. With the increased availability of academic profiles such as researchers' careers and scientific outputs, academic careers have become one of the prominent topics in the study of careers that attracts the attention of both social scientists and general researchers. Social scientists want to unravel factors that will positively or negatively contribute to academic career success. Researchers in other disciplines are concerned with how to raise scientific productivity and achieve career success. This line of research has regained its prominence with the emergence of Science of Science [21] in the age of computational social science [36].

Previous studies have focused on career success in terms of an individual's position or promotions within an institution. However, in the current boundaryless career world [8], it is not unusual for researchers to take a more flexible approach to pursue their career success across institutions and even social sectors. This change poses new conceptual and methodological demands for empirical research on academic career success. The first demand is a new perspective to capture the sequential patterns in researchers' career paths. The traditional point-to-point transition perspective fails to capture the long-term impact of historical events on the upcoming career performances. The classical time series modeling assumes only quantitative changes in a career path, making it challenging to capture qualitative changes. The second demand is a more panoramic framework to identify potential factors contributing to academic career success. Previous studies have well documented the impacts of individual factors (e.g., educations). Nevertheless, there is no doubt that researchers' careers are also contingent on their social connections (e.g., collaborations). The third demand lies in a more computationally efficient way to detect driving factors underlying academic career success and an informative way to analyze the intricate and dynamic relationships between factors and academic career success.

By harvesting multiple data sources in the visualization field as a context for illustration, the current study adopts a dynamic perspective to understand how individual and social factors will influence researchers' career success by using an interactive visual analytics approach. However, developing such a system faces three challenges. First, distilling the potential factors on academic career success and analyzing their dynamic impacts over time are difficult. In addition to individual factors, operationalizing social factors (e.g., collaborations) aggravates the complexity of the problem given their network nature, let alone organizing these factors into proper temporal formats. Moreover, capturing the effects of these factors requires a dynamic multivariate analytical framework given the long period of research fields. Second, visually presenting the effects of multiple factors over time is challenging. Such effects are always with complex data structures in social science such as multi-dimensional, temporal, or with pairwise comparisons. Visual data representations require supporting both cross-sectional and longitudinal studies of factors. Third, supporting the exploration of both rich academic profiles and the multi-factor effects on academic career success is non-trivial. Experts desiderate to combine the impacts and patterns with the academic profiles to comprehensively understand the results [51]. Statistical summaries and coordination among diverse data aspects could be challenging given multiple sources and dimensions.

To address the first challenge, we identify a set of individual and social factors based on social theories and our domain expert. We then propose a novel framework to analyze multi-factor effects on academic career success over time. For the second challenge, we design novel visualization to reveal the above time-varying effects. Specifically, the *Impact Timeline* is for comparing the effects of different categories within a factor. A *CareerLine* shows one's academic career development with multiple factors. We solve the third challenge by proposing *ACSeeker*, an interactive visualization system that assists social scientists in exploring academic career success with multiple factors from different levels of detail. Our contributions are listed as follows.

- We characterize the problem domain of visual analytics of time-varying effects of multiple factors on academic career success.
- We propose a novel framework that analyzes the effects of multiple factors on researchers' career success longitudinally.
- We develop *ACSeeker*, an interactive visualization system for social scientists and other scholars to explore academic careers with multiple factors through novel designs.
- We demonstrate the effectiveness and usability of *ACSeeker* with a dataset that involves more than 1,100 visualization researchers.


- *Yifang Wang was with the State Key Lab of CAD&CG, Zhejiang University and the Hong Kong University of Science and Technology. A part of this work was done when she was a visiting student supervised by Yingcai Wu at Zhejiang University. E-mail: yifang.wang@connect.ust.hk.*
- *Tai-Quan Peng is with Michigan State University. E-mail: pengtaiq@msu.edu and winsonpeng@gmail.com.*
- *Huihua Lu, Haoren Wang, and Yingcai Wu are with the State Key Lab of CAD&CG and Xiao Xie is with the Department of Sport Science, Zhejiang University. Yingcai Wu is the corresponding author. E-mail: {huihualu, haorenwang, xxie, ycwu}@zju.edu.cn.*
- *Huamin Qu is with the Hong Kong University of Science and Technology. E-mail: huamin@cse.ust.hk.*


*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*

## 2 Related Work

In this section, we discuss related work in both career data analysis and sequence data analysis.

### 2.1 Career Data Analysis and Visualization

Careers have been widely studied in social science and data science. Most studies focus on the common patterns or similarity comparisons of career paths in terms of job-related attributes, such as title ranks and organizations. Traditional social science studies focus on a macro-level analysis, which uses cross-sectional analysis to learn inflow and outflows across occupations [30, 47]. Such point-to-point transition analysis fails to capture the careers from a long-term perspective to learn the impact of historical events on upcoming career performance. Works in data mining study from a micro-level to compare individual careers [72] and predict future jobs [39, 70]. It lacks a macro summary of careers, which is essential to understand the aggregated behavioral patterns in social science. Visualization studies use career data as a scenario in the event sequence analysis, including similarity analysis [18, 20, 32, 60] and visual sequence summarization [24, 25]. CV3 [20] is a system for comparing multiple resumes that assist recruiters in finding suitable candidates. Du et al. [18] and Jänicke et al. [32] developed two systems respectively to recommend similar career paths of students or musicians for a target individual. Instead of clustering the whole sequences, EventThread [25] and ET2 [24] use a fine-grained clustering to summarize career paths into latent stages.

The increased availability of academic profile data (e.g., Aminer [54] and Vispubdata [31]) has called for a renovated perspective to study academic careers. Existing works analyze these data from multiple perspectives such as publication, citation, and collaboration networks. Egoslider [65] and Egolines [78] distill the evolutionary collaboration networks to learn the academic interactions. Fung et al.designed a tree metaphor to visualize one's publications [22]. VIS Author Profiles utilizes a template-based natural language generation to present a researcher's profiles. Other works use novel designs (e.g., linked matrix and flower metaphor) to show the scientific influence of different research entities (e.g., researchers and publications) [50, 61].

However, most studies focus on career data, lacking the analysis of potential individual and social factors that can contribute to career success. A few visualization systems contain social factors (e.g., social networks) in career analysis. In the Interactive Chart of Biography [34], institutional and denominational ties are distilled to show the relationships of musicologists. ResumeVis [76] contains an ego-network-based graph that summarizes the co-working relations of the target individual. Nevertheless, these social networks have not been correlated with career success. In this work, we use multiple sources of academic profiles and propose a new visual analytics framework to understand how multi-factors affect academic career success from a dynamic perspective.

### 2.2 Sequence Mining and Visualization

Event sequences are continually studied in the past decades in both social science and computer science with many shared research concerns. Since our focuses are analyzing the dynamic impacts of multiple factors on career success, we would discuss the most relevant works in both sequence mining and visualization.

Sequence mining in both disciplines can be summarized into three categories: pattern discovery [66], sequence inference [68], and sequence modeling [59], based on the survey of Guo et al. [26] and Piccarreta [45]. Sequence clustering is a universal technique to extract common sequential patterns using unsupervised learning. Xu et al. [71] summarized different clustering strategies into three categories: proximity-based [45, 46, 71], feature-based [27], and model-based [41, 75]. Specifically, proximity-based methods use the distance matrix to measure the similarity of sequences. Analysts first define and compute the pairwise dissimilarities between sequences and use clustering approaches to obtain sequential patterns. A critical step in proximity-based methods is to construct the distance matrix. Many sequence dissimilarity measures are proposed [53, 71] such as Euclidean distance and Levenshtein distance. Besides clustering on the entire sequences, several studies propose to use stage analysis to represent the sequence progression with more details [24, 25]. In addition to clustering, sequence inference (also noted as event history analysis in sociology) is another enduring topic. It estimates the influence of historical events or trajectories on upcoming events. Graphical models [10] and regression analysis [52] are widely used. We enhance the approach by Rossignon et al. [48] which combines sequence clustering with sequence inference to estimate the time-varying impacts of multiple factors on the upcoming career performance.

A variety of visualization techniques are also developed to analyze sequential data. Guo et al. [26] have proposed a comprehensive survey of event sequence visual analytics approaches. Here we mainly summarize the most relevant visualization techniques in our work. A large number of designs adopt an intuitive horizontal timeline encoding. Flow-based visualization is a typical representation with great scalability to summarize large-scale sequence data, including tree-based [37, 42] and Sankey-based [23, 63] structures. Two strategies are commonly used to reveal sequential patterns with fine granularity. One uses stage analysis to extract latent stages within the whole sequence to show progressions [24, 25]. The other directly visualize original individual sequences as horizontal lines [13, 16, 67, 69]. Recent work by Bartolomeo et al. [17] further uses layered directed acyclic network together with layout optimization to align specific events among sequences. Similar visual representations are adopted in storyline visualization [55, 56], where each line represents an actor, and the vertical position encodes groups of actors based on different events. Another set of studies use matrix- or list-based techniques to visualize event sequences that are scalable [18, 58, 62, 64, 79]. However, most of the state-of-the-art studies cannot be directly used in the study of academic career success. First, they focus on pattern extraction while the effects of multiple factors are of great importance in our scenario. Second, they consider the relative time which aligns sequences to the same starting point, while absolute time is also important to learn different generations of researchers. We have proposed novel visual designs to fulfill the above requirements.

## 3 Background and System Overview

In this section, we introduce the background and concepts used in our study, summarize the analytical tasks, and provide a system overview to demonstrate the whole pipeline.

### 3.1 Background and Concepts

The study of academic career success has been an enduring topic in multiple fields such as SciSci [21] and sequence analysis in social science [46]. Learning the time-varying impacts of multiple factors on academic career success can benefit the understanding of typical career patterns for social scientists and individual career development for general researchers. We have worked closely with a social scientist ($E_A$) to solve this research problem. He has been conducting multi-factor analysis with sequential data in different social domains. By working with $E_A$, we summarized the following concepts to characterize the problem of the multi-factor effect on academic career success.

- *Career Success*, or *Career Performance*, refers to the outcomes (e.g., title ranks and incomes) of one's working experiences [8,49]. In academic careers, *citations* of a scholar's research outputs are commonly used to measure career success [21].
- *Career Factors* refer to potential factors (i.e., variates) that can affect one's career success. It includes both individual (e.g., personal characteristics) and social factors (e.g., social relations) [74]. In academic careers, due to the data accessibility, we collected four factors that may affect career success based on previous empirical studies and our expert's suggestions. Job title ranks, sectors (e.g., academia, industry, and government agencies), and research domains are individual factors. As collaboration is an essential type of occupation network in academia that can have significant effects on careers [8, 21], we regard collaborators' individual factors as social factors [74].
- *Factor Categories* are different groups identified based on the values within a factor according to user-specified definitions. For example, users may define people within the same value range of a factor as one category.

### 3.2 Data Description

The analysis of dynamic multi-factor effects on academic career success is based on data from multiple sources, some of which require complicated and even manual data preparation. Considering the data availability and our familiarity, we focus on researchers from the visualization (i.e., VIS) community as a context to illustrate academic career analysis. We consider those who have published more than two TVCG papers in which the largest time gap is more than five years as potential VIS researchers after discussing with our expert. We then manually check and filter out those from other fields (about 90 researchers) and finally obtain over 1,100 VIS researchers. We use researchers' names as inputs to search different data sources below, which are transformed into multiple sequences by each researcher for in-depth analysis.

- **Career data** of researchers record the job-related attributes such as the institutions and titles. We collected it from LinkedIn [5], researchers' personal websites, and their institutional webpages.
- **Bibliographic data** is directly gathered from Aminer [54], which includes over 21,600 papers in total based on these researchers. It includes all the publication metadata of a researcher (e.g., authors, year, venue, title, and abstract) by year.
- **Citation data** (by year) is crawled from Google Scholar [4] as a measure of career success [21].

### 3.3 Task Analysis

Our goal is to analyze the dynamic effects of multiple factors on academic career success. Guided by the nested model of visualization design [43], we conducted literature reviews and frequent interviews with the expert to iteratively distill and refine the tasks. Finally, we form the analytical tasks into three levels [9] as follows.

The **inter-factor-level** task provides an overview of how different factors contribute to career success.

**T1** **How does a factor influence career success over time?** For longitudinal comparison, the expert wishes to know how the impact of a specific factor on career success develops over time. This could be related to the development of specific academic fields.

**T2** **How do multiple factors differ in their impacts on career success?** The expert wants to know the effects of different factors at a specific time as a cross-sectional comparison to determine the dominant factors influencing career success. Specifically, it is interesting to compare the impacts of individual and social factors.

The **intra-factor-level** tasks drive into one specific factor to compare the change of effects of different categories on career success over time.

**T3** **How does a category within a factor change over time to affect career success?** The impact of a category can change at different periods. It reflects the change of the roles of this category.

**T4** **How do the categories within a factor differ from affecting career success?** For each factor, different categories may have different sizes of impacts. Figuring out those with high impacts can help explain and provide guidelines for academic career development.

The **individual-level** task gives several concrete examples to help understand the factor effects on career success.

**T5** **How does a researcher's career path change over time?** The expert wishes to identify different individuals from the data and study the multi-factor effects on career success from a micro perspective. Revealing the career changes of a researcher over time can help understand the different academic stages he goes over.

**T6** **How do the different factors of a researcher change over time?** Showing the evolution of career factors over time is essential to interpret the career development of a researcher and further validate existing rules or generate new hypotheses.

### 3.4 System Overview

*ACSeeker* is a web-based application with three modules: a data preprocessing module, a data analysis module, and a data visualization module (Fig. 1). The data preprocessing module collects and cleans researchers' career profile, publication, and citation data. Then it organizes these data into multiple sequences and stores them in the database. The data analysis module utilizes a novel framework to analyze the effect of multiple factors on career success. They form into the backend of the system and are implemented using Python and MongoDB. The data visualization module constructs a frontend application using Vue.js [7] and D3.js [12] with three views to support analysis.

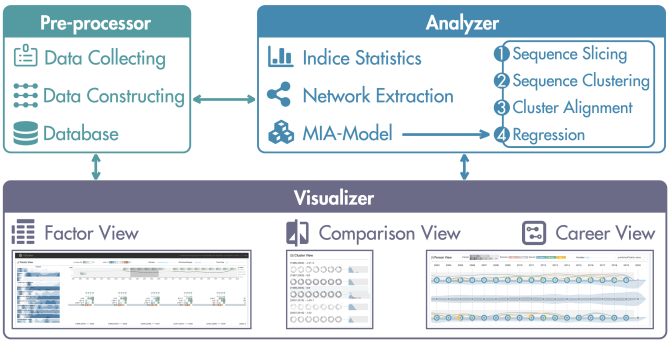


Fig. 1. System overview. *ACSeeker* has three components: a data pre-processor module, an analyzer module, and a visualizer module.

## 4 Data Analysis

In this section, we first introduce a set of data preprocessing measures. Then we introduce the existing Sequence History Analysis (SHA) [48] applied in previous social science research. Finally, we describe our new analytical framework based on SHA to simultaneously extract career patterns and analyze the impacts of multiple types of historical sequences on career success over a long period.

### 4.1 Data Preprocessing

After collecting multiple data sources (Section 3.2), we preprocessed the data in a semi-automatic way. For the career data, we organized each job into an event with a timestamp (by year) and an institution. We manually tagged the job titles and sectors in career data. Job titles were tagged into three ranks (i.e., junior, intermediate, and senior) based on researchers' tenure in academic research (Fig. 2-B1). We also tagged three sectors: academia, industry, and government agency, based on the institutions (Fig. 2-B3). From the bibliographic data, we extracted the paper venues by year and classified them into twelve categories to represent different research domains based on [1] (Fig. 2-B2). We also extracted all the collaborators of a researcher by year to construct his ego-networks. For the citation data, we used Quartile [6] to divide each year's citations into four ranks of equal size (Fig. 2-B1). In addition, we separated the top 3% citation researchers from the top 25% to inspect the pioneers in the visualization field.

Finally, multiple sequences are constructed for each researcher (Fig. 2-C). *Career Sequences* are composed of events defined by both title and citation ranks ordered by year. *Domain Sequences* consist of a list of paper venue categories with corresponding numbers of papers each year. *Sector Sequences* are sequences of sectors where researchers are affiliated over time. *Citation Sequences* are sequences of citation numbers by year. Meanwhile, we record all the collaborators of each researcher by year. We use their career, domain, and sector sequences as social factors. It allows us to quantify how a researcher's collaborators can influence his career success. These sequences, along with the collaboration networks, are fed into our analytical framework.

### 4.2 Sequence History Analysis (SHA)

A body of literature in social science has studied the effect of historical event trajectories on an upcoming target event [38]. Most of them reduce the historical sequences into summary indicators (e.g., event duration) [28], which fail to keep the full information in the original sequences. Sequence History Analysis (SHA) [48] is an innovative approach to preserve more complex sequential information in two steps. First, *Sequence Analysis* [46] is applied to identify representative patterns over the historical sequences. A distance matrix is constructed to document the pairwise distances between raw sequences. Using this matrix, the sequences are clustered into groups (i.e., categories) based

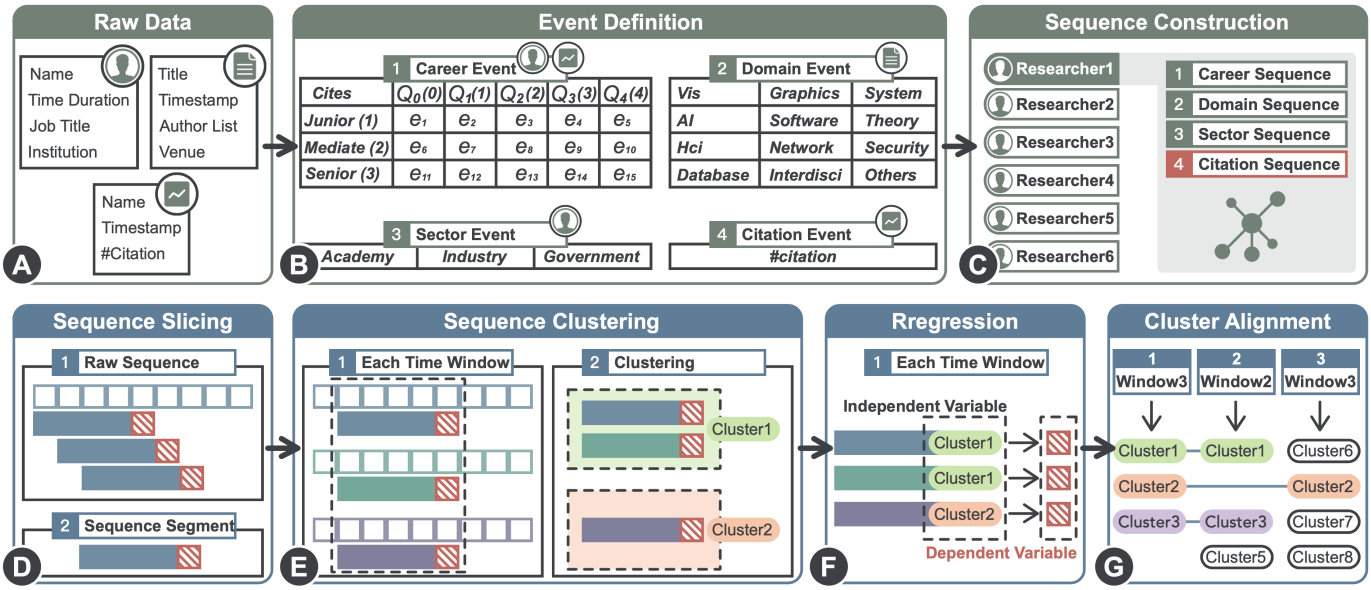


Fig. 2. The pipeline of data preprocessing and the multi-factor impact analysis (MIA) framework. For data preprocessing (A-C), we first define the events based on raw data collected from three sources. Then we construct four sequences for each researcher as the input of the framework. The MIA framework (D-G) consists of four steps: (D) Sequence Slicing, (E) Sequence Clustering, (F) Regression, and (G) Cluster Alignment.

on clustering algorithms (e.g., k-means). It retains the most representative sequential patterns over raw sequences and substantially reduces the computational demand in the subsequent multivariate analysis. Second, *Event History Analysis* [28] is used to analyze how these historical sequential patterns will affect the upcoming event. Different regression models will be applied to obtain the estimation of the effects.

However, the original SHA cannot be directly adopted in our scenario. First, it analyzes the impacts in a static manner by aligning sequences to the same starting point. However, the impacts of historical trajectories of factors could vary over time due to the development of the research field. Second, the SHA is supposed to analyze only one type of sequences, while in our scenario, multiple types of sequences are hypothesized to influence career success.

### 4.3 Multi-factor Impact Analysis (MIA)

We have worked with our domain expert to enhance the SHA approach to support dynamic analysis of the impacts of multiple factors on academic careers over time. The whole framework consists of four components: sequence slicing, sequence clustering, multivariate linear regression, and cluster alignment (Fig. 2-D, E, F, G).

**Sequence Slicing.** We first improve SHA by slicing multiple sequences into different time windows. Given a long period, our expert hoped to inspect the time-varying impacts of different factors. Moreover, he stated the importance of considering the time-sensitivity of the impacts of historical sequences: the farther the historical event, the less relevant it is to the upcoming career performance. We thus apply the sliding window method [15, 73] to meet the two requirements. Each researcher's career-related sequences (Fig. 2-D1) are arranged along the absolute timeline and sliced into different windows of a fixed size (i.e., size $w$, Fig. 2-D2). We aim to analyze the impacts of multiple factors in each time window separately. The window size and the moving step can be adjusted based on domain knowledge in *ACSeeker*.

**Sequence Clustering.** In each time window, we follow the *Sequence Analysis* method in the SHA approach to identify the most representative sequential patterns of each factor. It uses sequence clustering on each type of sequences (i.e., career, domain, and sector) respectively (Fig. 2-E). After comparing different clustering algorithms (e.g., k-means, k-medoids, agglomerative, DBSCAN, optics, and spectral), we finally chose k-means due to its optimal performance and efficiency. We use Euclidean distance to obtain the dissimilarity matrix, which is widely applied in social science with remarkable computational efficiency [53]. As our expert expected to adjust the number of clusters to obtain more meaningful results, the system allows users to customize the range of cluster numbers in k-means.

**Multivariate Linear Regression.** In each time window ($[t, t+w]$) (Fig. 2-F), we then conduct a multi-factor impact analysis on career success. We improve the SHA approach by applying ordinary-least-square (OLS) regression to incorporate multiple factors.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_n x_{in} + \varepsilon_i, (i = 1, ..., n) \quad (1)$$

We have incorporated twelve independent variables (i.e., IVs, including individual and social factors) and a dependent variable (i.e., DV) based on our expert's suggestions as follows:

- $IV_1 - IV_6$. Six categorical IVs are based on the clustering results of three types of historical sequences (i.e., *Career*, *Domain*, and *Sector Sequences*) of researchers ($IV_1 - IV_3$) and their top collaborators ($IV_4 - IV_6$) who have the most co-authored papers in a time window. We choose the top collaborator as a representative that potentially affects one's career significantly (e.g., advisors or long-term collaborators). To include categorical variables as IVs in OLS regression, dummy coding [3, 29] is employed, transforming a categorical IV with $n$ categories into $n-1$ *dummy variables* [2]. Specifically, an arbitrary category of a categorical IV is chosen to serve as the *reference category* and all other categories are set to be the *comparison* or *target categories*. The obtained coefficient of each dummy variable means how a comparison category performs on the career success compared with the *reference category*. The obtained p-value shows the statistical significance of this comparison. As the choice of the *reference category* is arbitrary, we improve SHA by conducting a post-hoc analysis, which enables pairwise comparisons on the career success among all categories of a categorical IV as a panoramic view. Specifically, we traverse to set each of the $n$ categories as a *reference category* in turn in the regression. It results in a matrix-like table recording the coefficients and p-values of pairwise comparisons.
- $IV_7 - IV_{12}$. Besides the top collaborator, our expert suggested having a global summary of one's overall collaborators in the model. We thus include two types of numerical variables. The first type ($IV_7 - IV_9$) is the number of collaborators in each sequence cluster weighted by the collaboration strength. It is the product of the proportion of collaborators of this cluster and the normalized number of papers he collaborated over all the papers he published within the time window. The second type ($IV_{10} - IV_{12}$) counts the number of sequence clusters for each sequence type.
- $DV$. We use the number of citations at the following year of the time window to measure career success [21]. We take the logarithm of the citations as DV since the raw citations each year are with a positively skewed distribution.

In each window, we compute the explanatory power (i.e., impact) of each IV (i.e., factor), which is measured by the difference of $R^2$ for whether adding this IV into the regression.

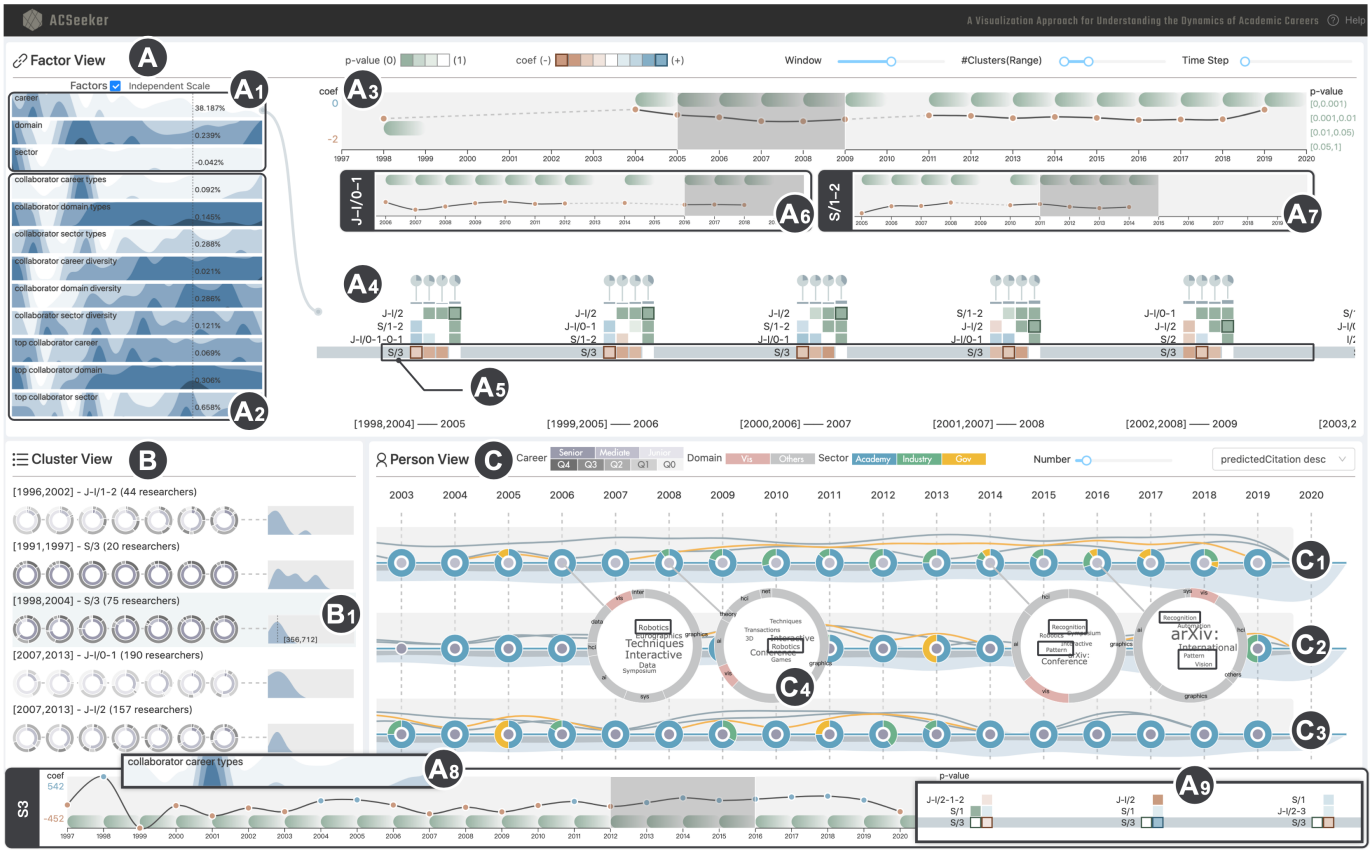


Fig. 3. The system interface of *ACSeeker*. (A) The *Factor View* reveals the regression results. It consists of three parts: *Horizon Chart Group* for inter-factor analysis, *Navigator* and *MatrixLine* for intra-factor analysis. (B) The *Cluster View* summarizes a list of sequence clusters chosen in the *Factor View*. (C) The *Person View* provides detailed information of careers and multi-factors for each researcher using a novel *CareerLine* design.

**Cluster Alignment.** In the last step, our expert wanted to align the same cluster in each sequence type across windows to learn its effect on career success temporally, which is not supported by SHA. We adopt a naive approach to label clusters by the event with the largest proportion in the cluster by year as it is the most representative one. We regard those with the same label as the same cluster and conduct the alignment across windows using these labels (Fig. 2-G).

In addition, for each researcher, we compute three diversity scores (i.e., career, domain, and sector types) of his collaborators by year using entropy ($n$ is the number of total event types):

$$Diversity(k,t) = -\sum_{i=1}^{n} P(x_i,t) log P(x_i,t), k \in \{career, sector, domain\} \quad (2)$$

The regression models of all windows (each includes *p-values*, *coefficients* and the difference of $R^2$ of each IV, and the predicted citations of each researcher) are fed into the *ACSeeker* for further analysis.

## 5 Visual Design

*ACSeeker* consists of three views to facilitate the analytical tasks in Section 3.3: the *Factor View*, the *Cluster View*, and the *Person View*. The workflow is as follows. Users will begin with the *Horizon Chart Group* in the *Factor View* and choose a factor of interest for detailed analysis in the *Impact Timeline*. Through *MatrixLine* and *Navigator*, based on domain knowledge, they may focus on the analysis of a specific category (i.e., sequence cluster) within a factor to learn its time-varying effect on career success. They may also adjust the number of clusters and the window size to obtain better results. During the exploration of *MatrixLine*, users can add interested clusters to the *Cluster View* for sequential comparisons. Finally, they can choose a cluster of interest in the *Cluster View* and use the *Person View* to learn individuals' careers and factor impacts within the cluster through the *CareerLine* design. The rest of this section will follow the order of the workflow to introduce each visual component.

### 5.1 Factor View

The *Factor View* (Fig. 3-A) is the primary visual component to show the regression results. It consists of two parts: (1) the *Horizon Chart Group* supports the inter-factor comparison on career success (**T1**, **T2**); (2) the *Impact Timeline* allows the intra-factor inspection within a factor (**T3**, **T4**). Users can have an overview of the time-varying effects of multiple factors and choose groups of interest for further study.

#### 5.1.1 Inter-Factor-Level Analysis

The *Horizon Chart Group* (Fig. 3-A1, A2) summarizes the trends of factor impacts and supports the comparison among factors (**T1**, **T2**).

*Description*: Each horizon chart represents a factor which is extended from a line chart. The x-axis encodes the time and the y-axis represents the explanatory power of a factor (i.e., $R^2$). The line chart is divided into layered bands with uniform ranges. The y value is encoded by a gradient color scheme in blue. The darker the color, the higher the value. Then the bands are shifted to the center and distributed within a fixed height. Two y scales are provided: a unified scale for impact comparison across factors and an independent scale for temporal inspection within a factor. Users can choose a factor for further analysis.

*Justification*: Initially, we used line charts with two modes. However, a multi-line graph that includes all the factors in one coordinate suffered great visual clutters. Small multiples where each line chart represented a factor were also not appropriate, since the height of each line chart was too narrow to show the temporal trends, let alone the comparison among factors. Thus, we chose the horizon chart to show the impact of multiple factors in a compact way.

#### 5.1.2 Intra-Factor-Level Analysis

After specifying the factor in the *Horizon Chart Group*, users can use the *Impact Timeline* (Fig. 4) to study the detailed impacts of the factor obtained from the regression model. Users can observe the temporal trends of the impact and compare the effects of different categories within the factor through intuitive interactions (**T3**, **T4**).

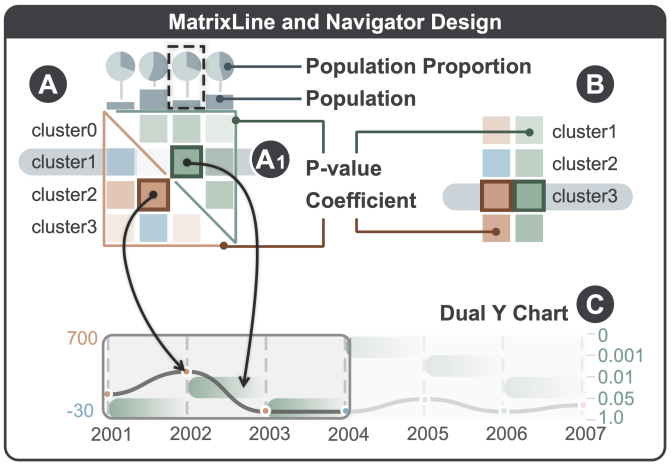


Fig. 4. The visual design of *MatrixLine* and *Navigator*. (A) The $n*n$ matrix design for categorical variables with $n$ clusters. (B) The $n*2$ matrix design for numerical variables. (C) The dual-y chart with coefficients and p-values obtained in the regression.

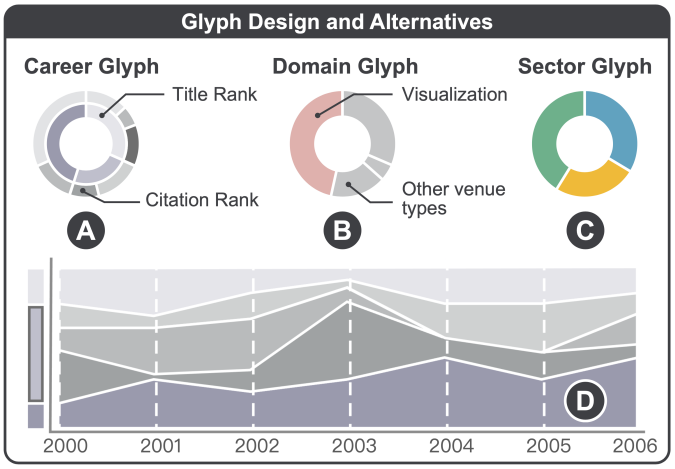


Fig. 5. Glyph designs and an alternative to summarize group information of careers and factors. The career glyph (A) is a sunburst graph showing the title and citation ranks. The domain (B) and sector (C) glyphs are donut charts showing the distribution of different categories of a group.

*Description*: The *Impact Timeline* consists of two parts: a *MatrixLine* showing the detailed impact of a time window (Fig. 4-A, B) and a *Navigator* revealing the impact evolvement over time (Fig. 4-C).

**MatrixLine.** Each matrix shows the MIA model results of a time window (**T4**). The design targets independent variables related to sequence clustering results (i.e., $IV_1 - IV_9$ in Section 4.3). To show the impacts of different categories (i.e., clusters) in individual's ($IV_1 - IV_3$) and his top collaborator's factors ($IV_4 - IV_6$), our MIA model transforms these categories into *dummy variables* and uses post-hoc analysis to produce pairwise category comparisons. The output is an $n*n$ matrix for coefficients and p-values in the regression as mentioned in Section 4.3. Thus, in Fig. 4-A, we divide the matrix into an upper triangle and a lower triangle to show the pairwise p-values and coefficients, respectively. In the upper triangle, each cell represents the pairwise p-value. The p-value is partitioned to four ranges (i.e., [0, 0.001), [0.001, 0.01), [0.01, 0.05), [0.05, 1]) based on the traditional social science approaches [11, 57] and our expert's suggestions to show the statistical significance. We use white for range [0.05, 1] (i.e., not statistically significant) and green in different saturations for the other three ranges to distinguish the statistical significance at different levels. The smaller the p-values, the darker the green. In the lower triangle, each cell shows the coefficient of pairwise categories. Our expert emphasized the importance of studying both the absolute values and the positive and negative of the coefficients. Thus, we use blue and red to encode the positive and negative values, respectively. The saturation encodes the absolute values. The larger the absolute value, the darker the color. The bar charts and pie charts above the matrix summarize the population and proportions of individuals in each category (i.e., sequence cluster) respectively. The dark grey area in the pie chart represents the proportion of this cluster.

To inspect the time-varying impact of a category, users can first specify a *reference category* (Fig. 4-A1) to align all the matrices (Section 4.3) across time. It will skip matrices without the *reference category*. Then they can choose a *target category* by clicking its pie chart. Two cells showing the p-value and coefficient will be highlighted across time. The *Navigator* will also be updated to summarize the temporal trends. The target category will be added to the *Cluster View* for further analysis. Users can customize the number of clusters, the length of the time window and the window moving step to adjust the MIA model.

To show the impact of collaboration strength of each collaborator's category of a social factor (i.e., $IV_7 - IV_9$), we transform the $n*n$ matrix into an $n*2$ one (Fig. 4-B). Each row represents a category of this factor and two columns represent the p-value and coefficient respectively. Users can align a category to study its time-varying impact.

**Navigator.** A *Navigator* provides a temporal overview of two values (i.e., p-value and coefficient) of a selected category mentioned above or the numerical independent variables (i.e., $IV_{10} - IV_{12}$) (**T3**). It is a timeline with a dual y-axis that represents p-value and coefficient, respectively. A stepped line graph in green shows the p-values and the line chart in black reveals the coefficients. The color of the circle on the line chart encodes the positive and negative values with the same color scheme in the matrix. Users can quickly find the period of interest and drag to focus the time window in *MatrixLine*. For the numerical factors, users can directly use *Navigator* to study the time-varying impacts.

*Justification*: We designed two line charts for two values as the navigator while it was space-wasting. We then assembled them into a dual-y chart with an area chart encoding the p-value. However, after trying the system, our expert suggested that the ranges of the p-values were more practical to learn the statistical significance than the raw values. Thus, we adopted a stepped line graph to fulfill the requirement.

### 5.2 Cluster View

The *Cluster View* (Fig. 3-B) presents a list of categories (i.e., sequence clusters) chosen from the *Impact Timeline*. Three glyphs (Fig. 5-A, B, C) are designed to summarize the cluster of different sequences (**T4**).

*Description*: Each row shows a sequence cluster of a time window with a title listing the window, the cluster label, and the number of researchers within the cluster. Each glyph of the row summarizes the distribution of researchers within the cluster at one year. We have designed three glyphs to show the cluster summary of three types of sequences (Fig. 5-A, B, C). The career glyph uses a sunburst structure with two levels of hierarchy to show the career sequence event distribution. The inner ring shows three title ranks (i.e., junior, intermediate, and senior) with purple in three saturation categories (Fig. 3-C). The outer ring encodes five citation ranks in black at five saturation categories. The domain glyph is a doughnut chart that records the distribution of individuals' most frequently published paper venue types in the cluster. We highlight the visualization venues as pink and others as grey to show visualization researchers' domain diversity (Fig. 3-C). The sector glyph is also a doughnut chart showing the social sector (i.e., academia, industry, and government) distribution: green for the industry, blue for the academia, and yellow for the government agency (Fig. 3-C). The histogram on the right shows the citation distribution of the following year (i.e., DVs) of the sequence window. Users can compare the distribution of different clusters using the sequence list and choose a cluster of interest for further analysis.

*Justification*: Before designing three glyphs, we tried the proportional stacked area graph at first. For example, in Fig. 5-D, we used the same colors to represent three title ranks. We used a navigator on the left to show the citation ranks within each title rank. However, it was space-wasting and hard to be integrated with other career information. Thus, we used glyphs to summarize multiple information of a group which could be reused in *Cluster View* and *Person View*.

### 5.3 Person View

After choosing a cluster from the *Cluster View*, we use a *CareerLine* to visualize an individual's careers and the effects of factors with folded and unfolded modes in the *Person View* (Fig. 3-C, **T5**, **T6**).

*Description*: In the folded mode (Fig. 6-A), the color of each circle shows the career title rank and the middle strip of the *CareerLine* repre-

sents the social sectors of the researcher, all with the same encoding in glyphs. Two flat gray areas distributed above and below the sector strip depict the collaborator and domain diversity scores of the researcher, respectively. The outer light blue area shows the predicted citations of the researcher based on the regression model. When hovering on a circle, a tooltip summarizing the researcher's domain diversity is shown (Fig. 3-C4). The outer ring is a doughnut chart summarizing the paper distribution in different domains. The inner word cloud is generated based on the paper venue names from the bibliographic data. We break the venue names into separate words and count the frequency statistics of each word to reflect the research topics. The word size encodes the number of papers. Users can unfold the area of collaborator diversity score to study details about the researcher's collaborators (Fig. 6-B). A multi-line graph shows the collaborators' different diversity scores (i.e., career, domain, and sector). They can click one of the lines (i.e., diversities) to show the summary of the collaborators' population distribution via corresponding glyphs (Fig. 5-A, B, C). We have provided sorting and filtering for users to choose researchers of interest. Users can rank them by average predicted citations or the collaborator diversity scores.

*Justification*: We tried to reveal all the collaborators' career paths as curves around the target researcher, where the distance of the two career circles is proportional to the number of papers they co-authored each year. However, it caused severe visual clutters when the number of coauthors increased. In addition, according to our expert's feedback, showing the raw career paths of collaborators was useless for comparing individuals and finding drivers that may affect the target researcher's career success. Thus, we computed different diversity scores to summarize the information of collaborators at each timestamp.

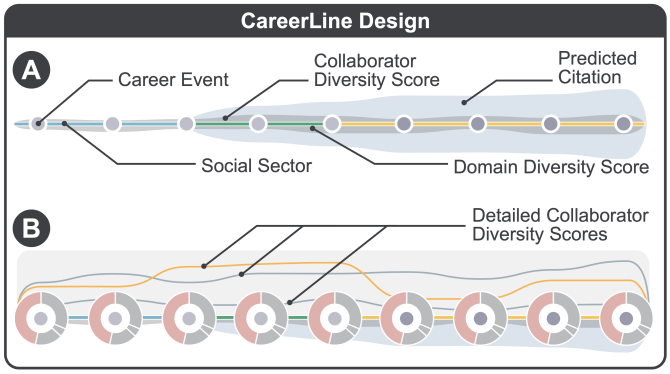


Fig. 6. The visual design of *CareerLine* with folded and unfolded modes. In the unfolded mode, users can have a detailed inspection of three collaborator diversity scores (i.e., career, domain, and sector diversities).

## 6 Evaluation

We evaluate the effectiveness and usability of the system using two case studies in Section 6.1 and interviews in Section 6.2.

### 6.1 Case Study

We invited our expert in Section 3.1 to freely explore *ACSeeker*. We then summarize the observations and comments and form them into two cases to fully demonstrate the system.

#### 6.1.1 Case1: Self-effort or Leverage?

Experts in social science have been engaged in studying multi-factor effects on academic career success for a long time.

**Inter-factor-level Inspection.** $E_A$ began with the *Horizon Chart Group* (Fig. 3-A) to obtain an overview of the impacts of all factors (**T1**, **T2**). Hovering on the horizon charts, he found that in an overall sense individual factors outperformed social factors (**T2**). Nevertheless, the individual factors (Fig. 3-A1) had a great impact in the early era and gradually declined. On the contrary, social factors (Fig. 3-A2), though did not contribute much at first, had become increasingly important (**T1**). *"It has been a long-standing concern in social science whether one's career success should be attributed to human capital (i.e., individual factors) or social capital (i.e., social factors). With the booming of interdisciplinary collaborations, the role of social factors will become increasingly important. Nevertheless, the human capital still plays a pivotal role in their acquisition and accumulation of social capital, which validates that individual factors are still dominant factors."*

**Intra-factor-level analysis.** $E_A$ noticed that the first individual factor (i.e., career) outperformed the other two individual factors (Fig. 3-A1). Wondering how different types of historical careers affected their upcoming career success (**T3**, **T4**), he chose this factor and turned to the *Impact Timeline* (Fig. 3-A3, A4). From the upper triangles, the pairwise comparisons among all categories were mostly statistically significant (green). *"...the differences among career sequence clusters are important in both statistical and substantial sense."* Viewing the *MatrixLine*, he found that senior researchers with a high citation rank (i.e., S/3 in Fig. 3-A5) existed across most of the periods. He clicked to align S/3 across time. From the *MatrixLine*, other clusters all performed worse than it (red cells) to affect career performance. It indicated that being a senior with high historical citations had obvious advantages on their upcoming career success. *"This is related to the accumulated citation impacts of their previous works."* Then, he set S/3 as the *reference category* and chose other clusters (e.g., J-I/2 (Fig. 3-A3), J-I/0-1 (Fig. 3-A6), and S/1-2 (Fig. 3-A7)) as targets to learn the impact evolvement. From the *Navigator*, the temporal trend of each cluster's impact had not changed substantially. *"Researchers' historical careers always have a strong and stable effect on their upcoming career performances. It has supported the view that one has to fight for themselves."*

$E_A$ also wondered how the same sequence type (i.e., careers) in the social factor would affect researchers' careers. He thus set the *collaborators' career types* as the target factor. In the *MatrixLine*, the S/3 group had a strong positive effect on researchers' career performance in the second window (Fig. 3-A9). Thus, he aligned this cluster and went to the *Navigator* (Fig. 3-A8). This group of collaborators negatively affected career performance at first (i.e., before 2010) and began to positively influence the career after around 2011. In addition, three windows (i.e., (1991, 1997), (1997, 2003), and (1998, 2004)) held a rare positive effect before 2010. $E_A$ added these three clusters into the *Cluster View* for further analysis. *"Interesting...Working with well-established researchers may not influence researchers' career performance immediately. They are usually working on innovative ideas, and the research outputs may take time to be seen by the world, which is always regarded as the 'sleeping beauties in science' [33]."*

**Individual-level Analysis.** After having an overview of the whole population, $E_A$ wondered if individuals from a similar starting point would differ in their career success at the later stages (**T5**, **T6**). Turning to the *Cluster View* (Fig. 3-B), he found that the third cluster (i.e., (1998, 2004)) was distinct (Fig. 3-B1) due to the high homogeneity of career distribution and the high citations in the upcoming year (i.e., 2005). Thus, he chose it and went to the *Person View*. He ranked researchers by average predicted citations. The predicted citations of most researchers in the cluster displayed a similar growth trend during (1998, 2004) with a slight rise later. However, the first researcher ($R_1$) acted differently with a citation surge in the last several years (Fig. 3-C1). $R_1$ also had a higher collaboration diversity score compared with others. Thus, $E_A$ unfolded his *CareerLine*. $R_1$ had been collaborating with scholars from different sectors (Fig. 3-C1). Hovering on career dots (Fig. 3-C4), $E_A$ found that $R_1$ changed research topics from robotics to vision. $E_A$ also checked the collaboration and domain diversities of the first other ten researchers as representatives. Most of them had low diversities in several aspects (e.g., mostly published papers within VIS or working with those in academia such as Fig. 3-C2, C3). *"Individuals in the same cluster can differ from one another later, which is known as 'within-group variance' in social science. $R_1$ is a case of the Matthew effect [19, 40]. Different strategies may cause diverse career performances even with the same starting point. For $R_1$, pursuing trendy topics and seeking diverse collaborations were potentially essential to improve his career."*

The case demonstrated that *ACSeeker* could help experts explore the dynamic multi-factor effects on academic career success at different levels of detail effectively.

#### 6.1.2 Case2: Comparative Study

Comparison studies are commonly applied in social science. After exploring multiple factors, $E_A$ found several stories summarized below.

**Social sectors diversity effects.** $E_A$ wondered whether historical

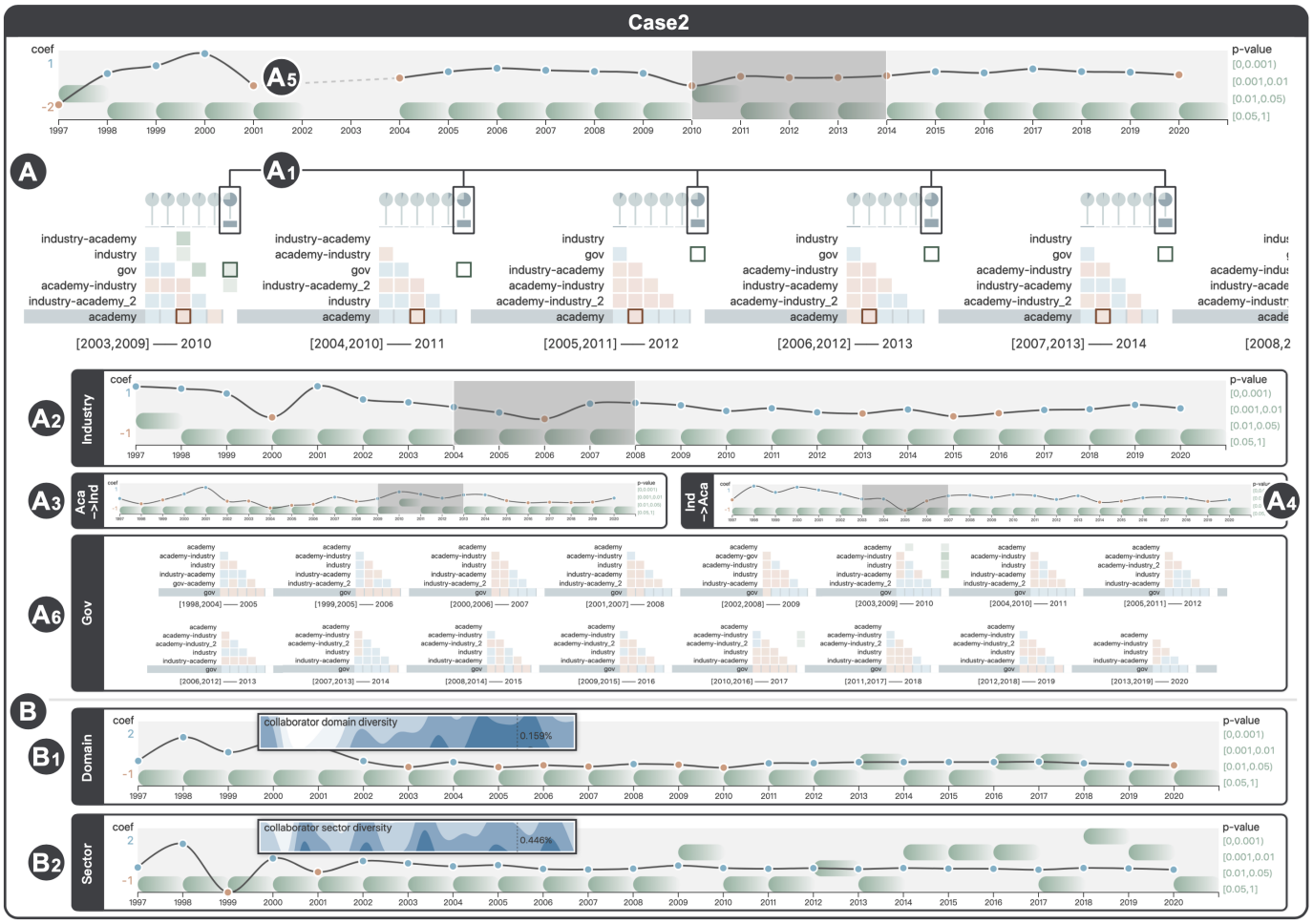


Fig. 7. The comparative study of (A) individual social sector (e.g., academia, industry, and government agency) and (B) collaborator diversity effects.

sectors would influence career success (**T3, T4**). He thus chose the *sector* factor and observed the *MatrixLine* (Fig. 7-A). The sector clusters were stable across different windows and the academia cluster had been in the largest proportion (Fig. 7-A1). He thus specified this cluster as the *reference category* and compared it with industry-related clusters (i.e., industry, academia-industry, and industry-academia, Fig. 7-A2, A3, A4). Generally, the career performance of individuals with industry experience performed better than those in academia alone. $E_A$ inferred that it might be due to the imminent analytical needs in the industry that facilitate career performances. *"Such borderless career movements could be a decent option for researchers' development."* Temporally, he also noticed that three coefficient curves in the *Navigator* all suffered a decline before 2005, then the values all started to increase. After obtaining the history of the VIS field, he learned that these special turning points might be related to the start of the VAST conference, which required cross-sector collaborations.

For comparison, he highlighted those who had been working in the government agencies as the target category. From the coefficient curve (Fig. 7-A5), the career performance presented a rare periodic pattern (i.e., rotating positive and negative effects every four or five years) compared with those staying in academia. $E_A$ further aligned this cluster to have a panoramic view by comparing it with all other clusters. From the *MatrixLine* (Fig 7-A6), this cluster also presented a periodic pattern. It outperformed most clusters to affect the career performance in certain periods (e.g., 2005 to 2006, and 2015 to 2019) and was less important in other periods (e.g., 2010 to 2013). $E_A$ checked different window lengths and was surprised at the similar results. *"I had never expected this before. It is possible that political policies may also affect the career performance of those in government agencies."*

**Collaborator diversity effects.** $E_A$ wanted to learn how the diversities of categories in each social factor could facilitate one's career success (**T1, T2**). Thus, he compared three social factors which describe the number of categories in each factor (Fig. 7-B). Interestingly, the diversity of collaborators' historical domain sequences (Fig. 7-B1) had a negative effect at first and turned to be positive after around 2010. *"Interdisciplinary collaborations always take time,"* $E_A$ explained, *"the collaborations at the early stage are always exploratory. Only after the trials will substantive collaborations appear and have a positive effect."* For the collaborators' historical sector diversity (Fig. 7-B2), it appeared to be almost all positive results over time. *"It has cross-validated our previous findings. Researchers with more cross-border collaborators could outperform those working with collaborators in a single sector."*

The case showed that *ACSeeker* could support comparative studies of multiple factors on academic career performance.

## 6.2 Expert Interview

We first interviewed our domain expert $E_A$. Although the initial purpose of the system is for social scientists to explore the dynamic multi-factor effect on academic career success, the dataset we use may also be of interest to visualization (i.e., VIS) researchers for their career developments. Thus, we also invited four VIS researchers ($P_A$-$P_D$) who had not used *ACSeeker* before. $P_A$ and $P_D$ are third-year PhD students with three-year VIS experience. $P_C$ is a researcher with six-year of expertise. $P_B$ just started her research as a first-year PhD student.

**Procedure.** Each interview for VIS researchers lasted about 60 minutes. We first introduced the background of the project (e.g., the research problem, the data, and the analytical tasks). Second, we used a comprehensive example to demonstrate visual encoding and interactions. Third, we introduced the cases in Section 6.1 to show the insights and the usefulness of the system. Fourth, they could freely explore the system in a think-aloud manner. Finally, we conducted a post-study interview to ask for suggestions on system workflow, analytical framework, and visualization and interactions. We recorded their comments and findings during the process. The feedback of the two groups of users is summarized into three categories.

**System Workflow.** Two groups of users all considered the system workflow clear. $E_A$ commented that *ACSeeker* followed and strengthened their traditional analytical workflow, *"it provides a panoramic*

*overview to compare the impacts of multiple factors, which makes the analytical findings more valid and straightforward."* He would follow the system logic to get familiar with *ACSeeker* from factor comparison to individual inspection. Then he would focus on analyzing the regression results (e.g., *Factor View* and *Cluster View*). The individual level was for case illustration and verification. For VIS researchers, those with statistical background (i.e., $P_C$) reacted similarly to $E_A$. Most insights he found were from the *Factor View*. However, others (i.e., $P_A$, $P_B$, and $P_D$) found it hard to understand the matrix encoding and mostly used the *MatrixLine* to choose clusters and explored from the individual level. They would conclude with the evolving patterns of careers of specific researchers. VIS researchers also had different analytical foci compared with $E_A$. They preferred to filter based on attributes (e.g., choosing collaborators from a specific domain) instead of clustering. They also wanted more information such as the real names of researchers and institutions to find role models. In general, there was a barrier for non-experts (i.e., VIS researchers) to explore *ACSeeker* due to diverse analytical foci and the lack of statistical background. It reflected that the target users of *ACSeeker* were still social scientists.

**Dynamic Multi-factor Impact Analysis.** $E_A$ appreciated the enhancement we added to the traditional SHA model, which made the whole framework more efficient. *"The analytical approaches in social studies always follow a multi-step recipe. However, each step is troublesome that requires a lot of detailed processing semi-automatically. This automated framework indeed eases our burden."* He was particularly impressed by the sequence slicing and cluster alignment. It helped him analyze the impact from a dynamic perspective, which he had never tried before. He also suggested making the lengths of sliding windows more flexible to capture both long-term and short-term impacts.

**Visualization and Interactions.** All the users regarded the system as comprehensive to study the academic careers, which fulfilled all the analytical tasks. Most users liked the horizon charts, which gave a compact summary of multi-factors' dynamic impacts. $E_A$ particularly liked the *Impact Timeline*, *"each matrix is with the same form as the table of coefficients and p-values we obtained from the R program, but in a more intuitive visual representation. The MatrixLine and the animation to align clusters across time are creative."* For the *CareerLine* design, they could not remember all the design components at once. For example, at first, $E_A$ mistook the two gray areas for the same diversity score in a symmetrical manner, which actually represented two types of diversities. $P_A$ and $P_D$ could not fully understand the matrix encoding since they were unfamiliar with the regression model. Nevertheless, after our explanation and exploring for a while, they finally got the point. All the users commented that most of the interactions were intuitive and straightforward. $E_A$ appreciated the well-coordinated system with multiple views displayed, *"I can brush and study the data across views conveniently."* $E_A$ and $P_C$ suggested improving the system readability, such as adding annotations for visual designs and abbreviations. $P_B$ wanted more interpretations related to the regression model.

## 7 Discussion

This section summarizes the significance of our work, the lessons we learned during our collaboration with social scientists, and the limitations and generalizations of *ACSeeker*.

**Significance.** In today's boundaryless career world, researchers' career choices have been increasingly flexible with more potential factors that may affect career success. We believe that the multi-factor impacts on academic career success will benefit both individuals' careers and the emerging literature in Science of Science. Traditional sociological methods are limited to analyzing from a static perspective. Therefore, we work closely with a social scientist to formulate the analytical tasks and perform the analysis longitudinally. Our sequential explication of trajectories, integration of different factors, MIA model, and the visual analytical framework can also provide a new attempt for social scientists to study multi-factor impacts from a dynamic perspective beyond career studies (e.g., mobile news consumption [77]).

**Lessons Learned.** The most important lesson is to take strengths from both domain and visualization fields to solve the problem more efficiently. For the analytical model, we contributed to the sociological framework by adding the sliding window and cluster alignment to transform the traditional static analysis into a dynamic one. For the visual design, *MatrixLine* is also derived from the table-based output from the R program and strengthened with new visualization and interactions (e.g., alignment). In summary, we obtained initial forms of the algorithm and visual design (e.g., SHA and matrix) from classical sociological studies. One crucial strategy to fill the analytical gap that previous methods cannot solve is to work with experts and decompose the problem into a list of limitations step by step. Then we can apply approaches from different subject areas (e.g., computer science or other social science disciplines) to solve each limitation. Moreover, from the interview, although our expert and general researchers all expressed strong interest in the same dataset, they presented different analytical foci. Social scientists aim to find aggregated patterns underlying human behaviors while general researchers prefer detailed 'real' information. In addition to diverse analytical foci, background knowledge (e.g., regression) should also be considered when designing the system.

**Limitations.** Although our evaluation has demonstrated the usefulness of *ACSeeker*, it still has limitations. The first one is the lack of large-scale researcher data. Currently, we have collected around 1100 VIS researchers to demonstrate the system due to the time-consuming data preparation process. However, to support a comprehensive understanding of dynamic multi-factor effects on academic career success, larger datasets with a broader range are necessary. We plan to collect data from computer science or the whole science field to find more stories. Second, the fixed size of sliding windows smooths the variations in different timespans and may ignore short-term influences, which is also of interest to experts. We plan to dynamically set the window size based on the sequential trend to address this limitation.

**Generalizability.** First, the method for operationalizing factors can be generalized, especially for the impact analysis of factor histories. Users can apply MIA to distill typical factor history patterns and analyze the impacts in other domains such as finance and meteorology. Particularly, to deal with the network nature, we have provided initial attempts to measure the social factors in multiple ways (Section 4.3). It can be adapted in other network impact analysis such as the social relation impact on idea innovation [35]. Second, visualization and interactions of showing the post-hoc tests temporally could be used in other pairwise comparisons of impact evolvement [44]. The *CareerLine* (glyph + line chart) also provides new visual representations to show multivariate sequences with state transitions. Third, *ACSeeker* can be applied to other historical trajectory impact analysis using sequential data. The most similar one is to analyze careers beyond academia, such as human resources (HR) management in companies. HR professionals can employ it to provide individualized trainings for employees to enhance their performance. A border usage is to analyze other life-course data (e.g., health and migration [14]) to understand human behaviors. Before applying the system, we suggest practitioners first organize different factors into multiple sequences. Then they can specify a target sequence type (e.g., career success in our study) to analyze.

## 8 Conclusion

This paper has presented *ACSeeker*, an interactive visualization system that enables social scientists to explore the multi-factor impacts on academic career success from a dynamic perspective. It supports the analysis from three levels of detail, including inter-factor level comparison, intra-factor level exploration, and individual level inspection. We have proposed two novel visual designs (i.e., *Impact Timeline* and *CareerLine*) to show the factor impacts. Two case studies and the interviews have shown the usefulness of the system.

In the future, we plan to first enhance the current MIA framework by setting dynamic lengths of sliding windows based on sequential trends. Second, we aim to collect data from a border range (e.g., computer science field) to find more stories. Third, we plan to gather requirements from general researchers to adapt *ACSeeker* for wider use.

## Acknowledgments

This research was supported in part by Hong Kong Theme-based Research Scheme grant T44-707/16-N, GRF 11505119 from HKSAR Research Grants Council, NSFC (62072400), Zhejiang Provincial Natural Science Foundation (LR18F020001), and the Collaborative Innovation Center of Artificial Intelligence by MOE and Zhejiang Provincial Government (ZJU).

## REFERENCES


[1] Catalogue of international academic conferences and journals recommends by china computer federation. https://www.ccf.org.cn/c/2019-04-25/663625.shtml.

[2] Dummy Variable. https://en.wikipedia.org/wiki/Dummy_variable_(statistics).

[3] Dummy Variable in Regression. https://stattrek.com/multiple-regression/dummy-variables.aspx.

[4] Google scholar. https://scholar.google.com/.

[5] LinkedIn. linkedin.com.

[6] Quartile in statistics. https://en.wikipedia.org/wiki/Quartile.

[7] Vue.js. https://cn.vuejs.org/index.html.

[8] M. B. Arthur, S. N. Khapova, and C. P. Wilderom. Career success in a boundaryless career world. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 26(2):177–202, 2005.

[9] J. Bertin. Semiology of graphics; diagrams networks maps. Technical report, 1983.

[10] D. Bhattacharjya, K. Shanmugam, T. Gao, N. Mattei, K. Varshney, and D. Subramanian. Event-driven continuous time bayesian networks. In *Proceedings of AAAI Conference on Artificial Intelligence*, vol. 34, pp. 3259–3266, 2020.

[11] D. D. Boos and L. A. Stefanski. P-value precision and reproducibility. *The American Statistician*, 65(4):213–221, 2011.

[12] M. Bostock, V. Ogievetsky, and J. Heer. $D^3$ data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[13] W. Chen, T. Lao, J. Xia, X. Huang, B. Zhu, W. Hu, and H. Guan. GameFlow: narrative visualization of nba basketball games. *IEEE Transactions on Multimedia*, 18(11):2247–2256, 2016.

[14] N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.

[15] C.-S. J. Chu. Time series segmentation: A sliding window approach. *Information Sciences*, 85(1-3):147–173, 1995.

[16] Z. Deng, D. Weng, Y. Liang, J. Bao, Y. Zheng, T. Schreck, M. Xu, and Y. Wu. Visual cascade analytics of large-scale spatiotemporal data. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[17] S. Di Bartolomeo, Y. Zhang, F. Sheng, and C. Dunne. Sequence Braiding: Visual overviews of temporal event sequences and attributes. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1353–1363, 2020.

[18] F. Du, C. Plaisant, N. Spring, and B. Shneiderman. EventAction: Visual analytics for temporal event sequence recommendation. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology*, pp. 61–70, 2016.

[19] G. Feichtinger, D. Grass, P. M. Kort, and A. Seidl. On the Matthew effect in research careers. *Journal of Economic Dynamics and Control*, 123:104058, 2021.

[20] V. Filipov, A. Arleo, P. Federico, and S. Miksch. CV3: Visual exploration, assessment, and comparison of CVs. In *Computer Graphics Forum*, vol. 38, pp. 107–118, 2019.

[21] 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), 2018.

[22] T.-L. Fung, J.-K. Chou, and K.-L. Ma. A design study of personal bibliographic data visualization. In *Proceedings of IEEE Pacific Visualization Symposium*, pp. 244–248, 2016.

[23] D. Gotz and H. Stavropoulos. DecisionFlow: Visual analytics for high-dimensional temporal event sequence data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1783–1792, 2014.

[24] S. Guo, Z. Jin, D. Gotz, F. Du, H. Zha, and N. Cao. Visual progression analysis of event sequence data. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):417–426, 2018.

[25] S. Guo, K. Xu, R. Zhao, D. Gotz, H. Zha, and N. Cao. EventThread: Visual summarization and stage analysis of event sequence data. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):56–65, 2017.

[26] Y. Guo, S. Guo, Z. Jin, S. Kaul, D. Gotz, and N. Cao. Survey on visual analysis of event sequence data. *arXiv preprint arXiv:2006.14291*, 2020.

[27] V. Guralnik and G. Karypis. A scalable algorithm for clustering sequential data. In *Proceedings of IEEE International Conference on Data Mining*, pp. 179–186, 2001.

[28] B. Hans-Peter and K. Golsch. *Event History Analysis with Stata*. 2007.

[29] M. A. Hardy. *Regression with dummy variables*, vol. 93. 1993.

[30] M. Hout. *Analysing Mobility Tables*. 1983.

[31] 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.

[32] S. Jänicke, J. Focht, and G. Scheuermann. Interactive visual profiling of musicians. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):200–209, 2015.

[33] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, 112(24):7426–7431, 2015.

[34] R. Khulusi, J. Kusnick, J. Focht, and S. Jänicke. An interactive chart of biography. In *Proceedings of IEEE Pacific Visualization Symposium*, pp. 257–266, 2019.

[35] B. Kijkuit and J. van den Ende. With a little help from our colleagues: A longitudinal study of social networks for innovation. *Organization Studies*, 31(4):451–479, 2010.

[36] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Life in the network: the coming age of computational social science. *Science*, 323(5915):721, 2009.

[37] Z. Liu, B. Kerr, M. Dontcheva, J. Grover, M. Hoffman, and A. Wilson. CoreFlow: Extracting and visualizing branching patterns from event sequences. In *Computer Graphics Forum*, vol. 36, pp. 527–538, 2017.

[38] I. Madero-Cabib, J.-A. Gauthier, and J.-M. Le Goff. The influence of interlocked employment–family trajectories on retirement timing. *Work, Aging and Retirement*, pp. 38–53, 2016.

[39] Q. Meng, H. Zhu, K. Xiao, L. Zhang, and H. Xiong. A hierarchical career-path-aware neural network for job mobility prediction. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 14–24, 2019.

[40] R. K. Merton. The matthew effect in science, ii: Cumulative advantage and the symbolism of intellectual property. *isis*, 79(4):606–623, 1988.

[41] Y. Ming, P. Xu, F. Cheng, H. Qu, and L. Ren. ProtoSteer: Steering deep sequence model with prototypes. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):238–248, 2019.

[42] M. Monroe. *Interactive Event Sequence Query and Transformation*. PhD thesis, University of Maryland, 2014.

[43] T. Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009.

[44] T.-Q. Peng, Y. Zhou, and J. J. Zhu. From filled to empty time intervals: Quantifying online behaviors with digital traces. *Communication Methods and Measures*, 14(4):219–238, 2020.

[45] R. Piccarreta and M. Studer. Holistic analysis of the life course: Methodological challenges and new perspectives. *Advances in Life Course Research*, 41:100251, 2019.

[46] G. Ritschard and M. Studer. Sequence analysis: Where are we, where are we going? In *Sequence Analysis and Related Approaches*, pp. 1–11. 2018.

[47] J. E. Rosenbaum. Organizational career mobility: Promotion chances in a corporation during periods of growth and contraction. *American Journal of Sociology*, 85(1):21–48, 1979.

[48] F. Rossignon, M. Studer, J.-A. Gauthier, and J.-M. Le Goff. Sequence history analysis (SHA): Estimating the effect of past trajectories on an upcoming event. In *Sequence Analysis and Related Approaches*, pp. 83–100. 2018.

[49] S. E. Seibert, M. L. Kraimer, and R. C. Liden. A social capital theory of career success. *Academy of Management Journal*, 44(2):219–237, 2001.

[50] M. Shin, A. Soen, B. T. Readshaw, S. M. Blackburn, M. Whitelaw, and L. Xie. Influence flowers of academic entities. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology*, pp. 1–10, 2019.

[51] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *The Craft of Information Visualization*, pp. 364–371. 2003.

[52] L. Stopar, P. Skraba, M. Grobelnik, and D. Mladenic. StreamStory: exploring multivariate time series on multiple scales. *IEEE Transactions on Visualization and Computer Graphics*, 25(4):1788–1802, 2018.

[53] M. Studer and G. Ritschard. What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pp. 481–511, 2016.

[54] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. ArnetMiner:

Extraction and mining of academic social networks. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 990–998, 2008.

[55] T. Tang, R. Li, X. Wu, S. Liu, J. Knittel, S. Koch, L. Yu, P. Ren, T. Ertl, and Y. Wu. PlotThread: Creating expressive storyline visualizations using reinforcement learning. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):294–303, 2020.

[56] T. Tang, S. Rubab, J. Lai, W. Cui, L. Yu, and Y. Wu. iStoryline: Effective convergence to hand-drawn storylines. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):769–778, 2018.

[57] M. Walsh, S. K. Srinathan, D. F. McAuley, M. Mrkobrada, O. Levine, C. Ribic, A. O. Molnar, N. D. Dattani, A. Burke, G. Guyatt, et al. The statistical significance of randomized controlled trial results is frequently fragile: a case for a fragility index. *Journal of Clinical Epidemiology*, 67(6):622–628, 2014.

[58] J. Wang, J. Wu, A. Cao, Z. Zhou, H. Zhang, and Y. Wu. Tac-Miner: Visual tactic mining for multiple table tennis matches. *IEEE Transactions on Visualization and Computer Graphics*, 27(6):2770–2782, 2021.

[59] J. Wang, K. Zhao, D. Deng, A. Cao, X. Xie, Z. Zhou, H. Zhang, and Y. Wu. Tac-Simur: Tactic-based simulative visual analytics of table tennis. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):407–417, 2019.

[60] Y. Wang, H. Liang, X. Shu, J. Wang, K. Xu, Z. Deng, C. D. Campbell, B. Chen, Y. Wu, and H. Qu. Interactive visual exploration of longitudinal historical career mobility data. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[61] Y. Wang, C. Shi, L. Li, H. Tong, and H. Qu. Visualizing research impact through citation data. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(1):1–24, 2018.

[62] J. Wu, Z. Guo, Z. Wang, Q. Xu, and Y. Wu. Visual analytics of multivariate event sequence data in racquet sports. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology*, pp. 36–47, 2020.

[63] J. Wu, D. Liu, Z. Guo, Q. Xu, and Y. Wu. TacticFlow: Visual analytics of Ever-Changing tactics in racket sports. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[64] Y. Wu, J. Lan, X. Shu, C. Ji, K. Zhao, J. Wang, and H. Zhang. iTTVis: Interactive visualization of table tennis data. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):709–718, 2017.

[65] 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.

[66] Y. Wu, D. Weng, Z. Deng, J. Bao, M. Xu, Z. Wang, Y. Zheng, Z. Ding, and W. Chen. Towards better detection and analysis of massive spatiotemporal co-occurrence patterns. *IEEE Transactions on Intelligent Transportation Systems*, 22(6):3387–3402, 2020.

[67] C. Xie, W. Chen, X. Huang, Y. Hu, S. Barlowe, and J. Yang. VAET: A visual analytics approach for e-transactions time-series. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1743–1752, 2014.

[68] X. Xie, F. Du, and Y. Wu. A visual analytics approach for exploratory causal analysis: Exploration, validation, and applications. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1448–1458, 2020.

[69] X. Xie, J. Wang, H. Liang, D. Deng, S. Cheng, H. Zhang, W. Chen, and Y. Wu. PassVizor: Toward better understanding of the dynamics of soccer passes. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1322–1331, 2020.

[70] H. Xu, Z. Yu, H. Xiong, B. Guo, and H. Zhu. Learning career mobility and human activity patterns for job change analysis. In *Proceedings of IEEE International Conference on Data Mining*, pp. 1057–1062, 2015.

[71] R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[72] Y. Xu, Z. Li, A. Gupta, A. Bugdayci, and A. Bhasin. Modeling professional similarity by mining professional career trajectories. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1945–1954, 2014.

[73] Y. Yu, Y. Zhu, S. Li, and D. Wan. Time series outlier detection based on sliding window prediction. *Mathematical Problems in Engineering*, 2014, 2014.

[74] H. Zacher, C. W. Rudolph, T. Todorovic, and D. Ammann. Academic career development: A review and research agenda. *Journal of Vocational Behavior*, 110:357–373, 2019.

[75] H. Zeng, X. Wang, A. Wu, Y. Wang, Q. Li, A. Endert, and H. Qu. EmoCo: Visual analysis of emotion coherence in presentation videos. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):927–937, 2019.

[76] C. Zhang and H. Wang. ResumeVis: A visual analytics system to discover semantic information in semi-structured resume data. *ACM Transactions on Intelligent Systems and Technology*, 10(1):1–25, 2018.

[77] L. Zhang, L. Zheng, and T.-Q. Peng. Structurally embedded news consumption on mobile news applications. *Information Processing & Management*, 53(5):1242–1253, 2017.

[78] J. Zhao, M. Glueck, F. Chevalier, Y. Wu, and A. Khan. Egocentric analysis of dynamic networks with egolines. In *Proceedings of ACM Conference on Human Factors in Computing Systems*, pp. 5003–5014, 2016.

[79] J. Zhao, Z. Liu, M. Dontcheva, A. Hertzmann, and A. Wilson. MatrixWave: Visual comparison of event sequence data. In *Proceedings of ACM Conference on Human Factors in Computing Systems*, pp. 259–268, 2015.