Interactive Visual Exploration of Longitudinal Historical Career Mobility Data

Yifang Wang, Hongye Liang, Xinhuan Shu, Jiachen Wang, Ke Xu, Zikun Deng, Cameron Campbell, Bijia Chen, Yingcai Wu, and Huamin Qu

Abstract—The increased availability of quantitative historical datasets has provided new research opportunities for multiple disciplines in social science. In this paper, we work closely with the constructors of a new dataset, CGED-Q (China Government Employee Database-Qing), that records the career trajectories of over 340,000 government officials in the Qing bureaucracy in China from 1760 to 1912. We use these data to study career mobility from a historical perspective and understand social mobility and inequality. However, existing statistical approaches are inadequate for analyzing career mobility in this historical dataset with its fine-grained attributes and long time span, since they are mostly hypothesis-driven and require substantial effort. We propose *CareerLens*, an interactive visual analytics system for assisting experts in exploring, understanding, and reasoning from historical career data. With *CareerLens*, experts examine mobility patterns in three levels-of-detail, namely, the macro-level providing a summary of overall mobility, the meso-level extracting latent group mobility patterns, and the micro-level revealing social relationships of individuals. We demonstrate the effectiveness and usability of *CareerLens* through two case studies and receive encouraging feedback from follow-up interviews with domain experts.

Index Terms—Digital Humanities; Quantitative History; Career Mobility; Visual Analytics

1 Introduction

Recently, an increasing amount of large-scale population history data has been digitized, including cross-sectional census data and longitudinal data of people's lives [1], [2], [3]. The availability of such data has brought new research opportunities to many fields (e.g., digital humanities, history, and sociology) to study topics such as careers, health, family, and migration from a historical perspective.

One such newly-constructed quantitative historical dataset is the CGED-Q (China Government Employee Database-Qing) [3], [4]. It records the career trajectories of over 340,000 government officials in the bureaucracy of Qing China from 1760 to 1912. This rich dataset is an important new source for the study of career mobility, which refers to the study of career trajectories and the factors influencing them. For sociologists, the long time span and the stable organizational structure enable the study of evolution of mobility patterns over time. It facilitates a better understanding of social mobility in that era [2]. For historians, the voluminous and detailed data provides possibilities for prosopographical studies, which investigate the social relationships and common characteristics of a social group based on biographical data [5]. It can uncover hidden rules and invisible cliques in career mobility [6]. In this paper, we work with the CGED-Q in collaboration with experts in quantitative history who constructed this dataset to study career mobility in a historical context.

However, existing statistical and model-based approaches are inadequate for analyzing career mobility in this new and complex historical dataset. Most of them are hypothesis-driven and ill-suited to exploration. Moreover, describing and interpreting the mobility patterns of the entire population, groups, or individuals presents another challenge. Visual analytics is thus required to analyze career mobility efficiently and intuitively [7]. However, to the best of our knowledge, few studies from the visualization community have explored career mobility. Previous works that studied career data focused mainly on social network extraction [8], [9], similarity analysis [10], [11], [12], and group-level event sequence summarization [13], [14]. They have not targeted career mobility analysis, in which overall mobility trends, along with the career mobility of the groups and individuals of interest need to be examined. Developing a visualization system for supporting such analysis faces several challenges.

First, visualizing a large volume of longitudinal career data with a complex data structure is challenging. Depicting the temporal evolution of mobility patterns in such a longitudinal dataset with multiple attributes will incur severe scalability problems given the long time span. Second, extracting and highlighting social groups and social relationships from this large dataset is non-trivial. Existing sociological methods for extracting social relationships and latent groups based on similarities [15], [16], [17] require specialized expertise and are ad-hoc. Highlighting the groups and social relationships of interest in their overall context without overwhelming the user with detail heightens the challenge. Third, supporting multi-level mobility analysis and reasoning is difficult. Historians and sociologists have

- *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 the State Key Lab of CAD&CG. E-mail: yifang.wang@connect.ust.hk.*
- *Hongye Liang, Jiachen Wang, Zikun Deng, and Yingcai Wu are with the State Key Lab of CAD&CG, Zhejiang University. Yingcai Wu is the corresponding author. E-mail: {21821087, wangjiachen, zikun_rain, ycwu}@zju.edu.cn.*
- *Xinhuan Shu, Cameron Campbell, and Huamin Qu are with the Hong Kong University of Science and Technology. Email: xinhuan.shu@connect.ust.hk, huamin@cse.ust.hk, camcam@ust.hk. Cameron Campbell is also affiliated with Central China Normal University.*
- *Ke Xu is with New York University. Email: kexu@nyu.edu.*
- *Bijia Chen is with Renmin University of China. Email: bjchen@ruc.edu.cn.*

difficulty efficiently exploring overall mobility alongside that of groups and individuals. Summarizing the statistics of groups and individuals to identify targets for further investigation and reasoning requires substantial effort.

To address these challenges, we present *CareerLens*, a visual analytics system which enables experts to generate and verify hypotheses, and explore and reason about insightful patterns of historical career mobility at three levels-of-detail (LODs). At the overall level, the system summarizes the long-term evolution of mobility, allowing experts to explore and compare salient features of groups at different periods. At the group level, *CareerLens* provides recommendations for latent social groups obtained from a group detection algorithm. At the individual level, filtering influential persons is also supported. The groups and individuals of interest with their social relationships can be further highlighted as group sub-flows or career threads using our novel flow design. It utilizes a multi-scale approach to clearly show the position and proportion of people of interest in their overall population context. Moreover, *CareerLens* is a well-coordinated system which enables experts to explore career mobility effectively. We evaluated the effectiveness and usability of the system through two case studies, a longitudinal study, and expert interviews which received positive feedback.

2 Related Work

In this section, we summarize related work in the most relevant fields, including career mobility analysis, career data visualization, and event sequence visualization.

2.1 Career Mobility Analysis

Career mobility refers to the trajectory of individuals from the time they enter the workforce to the time they leave in terms of transitions in job positions, levels, or other job-related attributes [18]. It is an important topic in multiple disciplines. In sociology, it is central in social mobility study to understand social stratification and inequality. In history, it is essential to detect social groups and the effects of social relationships on careers [5]. In data science, it is a typical scenario to study sequences using career data [19], [20].

Social scientists focus on the group-level analysis of transitions between occupation categories [6] or job positions [18] using panel data. Most studies are hypothesis-driven and focus on the empirical analysis of predefined groups over a short period using statistical or model-based approaches [18], [21]. A more recent generation of studies employs sequence analysis to examine the entire career and classifies careers into latent groups according to similarities [16], [17]. These approaches are semi-automatic and *ad hoc*, and have difficulty accommodating large datasets with multiple attributes [15], [16]. Moreover, they assume that careers are independent of one other, while careers are also influenced by social relationships [22]. Experts apply simple visualization for single tasks using tools (e.g., Excel and R), such as heatmap [23] and mosaic diagrams [24] for showing the career inflows and outflows of different occupations. Nevertheless, they are rudimentary and can not support a systematic analysis of career mobility from different LODs.

Alternately, studies in data mining pay more attention to individual-level analysis using Online Professional Networks data (e.g., LinkedIn [25]). Job shift prediction [19], [20], [26] and path similarity analysis [27] are two main categories. However, these works mainly focus on a micro-level analysis and do not target career mobility which emphasizes the characteristics of groups and more complicated social relationships. In this paper, we conduct a multi-level analysis of career mobility utilizing a long-term historical dataset, including inspecting mobility evolution, identifying latent groups, and extracting individuals' social relationships. The ability to explore large volumes of rich career data at different LODs illustrates new opportunities in career mobility studies in both social science and data science.

2.2 Career Data Visualization

Many visualization techniques have been proposed to study career data. Most works have a single focus, such as social network summarization [8], [9], sequence summarization [13], [14], and similarity comparison [10], [11], [12]. For career networks, Fung et al. [8] compared three visual representations (i.e., node-link diagrams, adjacency matrices, and botanical tree) of the academic collaboration networks of individual researchers. PathWay [9] characterizes a researcher's academic career with his collaboration and dispute networks. For multiple career paths, sequence summarization techniques are utilized. Guo et al. proposed EventThread [13] that uses event sequence clustering to visualize academic career events. They further improved the system to find semantically meaningful progression stages [14]. For career path similarity comparison, CV3 [12] compares multiple resumes to suggest suitable jobs to job hunters. EventAction [11] reviews students' historical events and provides suggestions for future planning based on the experiences of similar individuals found by the system. Jänicke et al. [10] developed a visual profiling system to find similar musicians' profiles using biographical information.

Several studies [28], [29] support multi-task analysis, such as exploring groups and social relationships. Khulusi et al. [28] designed a system based on Joseph Priestley's Chart of Biography to show musicologists' relationships and movements among different institutions. However, they do not analyze the effect of relations on musicologists' career development. Zhang and Wang [29] applied text mining to semi-structured resumes to obtain semantic information on career paths. They also incorporate social relationships. However, the visualization (e.g., stepped lines) support a limited number of career comparisons and lose the overall context. They are also limited by the short time range. It is not scalable for large-scale and long-term career data to show mobility evolution and group summary. Our system is a new attempt at visually analyzing complex and long-term historical career mobility data from different LODs.

2.3 Event Sequence Visualization

Many techniques have been proposed for event sequence visualization. A comprehensive survey is available in [30]. Given our focus on visualizing latent social groups based on sequence similarity, we mainly discuss relevant event sequence visualization techniques. Initial works directly reveal original event sequences along a horizontal time axis [31], [32] or in a spiral direction [33], [34] to show the

raw information of each sequence. To show patterns on large event sequence datasets, later studies aggregate multiple event sequences and provide a visual summary. Flow-based visualization is one of the most widely adopted visual representations [35], [36], [37], [38], [39]. EventFlow [35] and CoreFlow [36] utilize a tree-like visual structure to reveal branching patterns in event sequences. Outflow [37] and DecisionFlow [38] strengthen the Sankey diagram with more information such as time duration. Each node represents an event and each edge connects the adjacent nodes according to the event sequences. In addition, some studies further group the event sequences into time-specific clusters based on stage similarity [13], [40]. We also adopt flow-based visualization in our design. To compare multiple event sequences, matrix- and list-based visualization are widely adopted [41], [42].

However, most current approaches to event sequence visualization can not be used directly in our work. First, the sequence distribution along the absolute timeline is important for correlating mobility patterns with special historical periods. However, most works only consider relative time, where the origin time of each sequence is shifted to the same starting point or a specific event for a better alignment and comparison. Second, observing the group or individuals of interest within the whole context should be considered to understand the position and proportion of the group or individual over the entire population. Thus, new visual designs are required to understand career mobility data.

3 BACKGROUND

In this section, we first introduce the background of the historical career mobility data we use. Then we formulate three levels of analytical tasks based on iterative interviews with domain experts to guide the system design. Finally, we give a system overview to summarize the whole pipeline.

3.1 Data Description and Pre-processing

In this study, we use CGED-Q [3], a new resource for the quantitative study of Qing officialdom, to conduct career mobility analysis. It comprises career records for nearly all the regular officials in the Qing bureaucracy (over 340,000), extracted from rosters that were compiled every three months between 1760 and 1912. Such time-depth is quite rare for career data, and multiple attributes make it possible to distill social relationships among officials. Well-defined job levels and a stable organizational structure over a long time period provide unique opportunities for the study of career mobility from a historical perspective. Each record consists of two parts, namely, the officials' background information and the details of their current position. We introduce a subset of attributes we use in our study:

- *Timestamp*: The year and season covered by the record.
- *Name*: The official's real name in the Qing dynasty.
- *Birthplace*: The geographic origin of the official.
- *Family Background*: A identity indicating whether the official was associated with the imperial lineage.
- *Ethnicity*: Three types of officials are identified based on ethnicities: Manchu, Mongol, or Han.
- *Exam Degree*: The examination or purchased degree held by the official. Those with high examination degrees were political elites.
- *Unique ID*: A 12-character unique identifier of each official generated by nominative linkage by our experts.
- *Job Location*: The geographical location of the official's current job.
- *Job Department*: The department in the bureaucracy where the official works. We classified them into fifteen categories according to experts' suggestions.
- *Job Level*: The administrative rank of the job in the bureaucratic hierarchy, represented by a number (ranging from 10 to 1 with 0.5 as a step). It is assigned by our experts based on the rank specified for the job in historical documents.

We further transformed the record-based data into individual-based career paths using the *Unique ID* attribute.

3.2 Design Process and Task Analysis

Our goal is to present a visual analytics system for career mobility analysis from a historical perspective using CGED-Q. In the past year, we have been working closely with four experts in a user-centric method through all stages, including task, design, development, and evaluation. Specifically, two of them are the constructors of CGED-Q and the co-authors of this paper. E_A is a professor who is well-established in historical demography and social mobility. E_B is a postdoctoral scholar studying Qing history. Two other experts include a postgraduate (E_C) knowledgeable in career mobility and a postgraduate (E_D) studying Qing bureaucracy, who had conducted research using the CGED-Q. During the process, we held frequent interviews through video meetings and emails to learn traditional workflows, distill requirements, and collect feedback based on the prototype. We summarize the milestones during the collaboration as follows.

Identifying analytical tasks and learning traditional workflows. In the first month, we discussed the semantics of the data attributes and research problems on which the experts worked. E_A and E_C wanted to study overall trends to understand change and continuity over time. They also wanted to detect latent groups based on the similarities of careers since they may reveal hidden rules within the bureaucracy that favor certain types of officials. E_B and E_D wanted to examine whether social relationships played a role in careers and detect otherwise invisible cliques. We also learned the traditional workflow. Experts would target specific populations and conduct hypothesis-driven analysis via tabulations and regressions. This was time-consuming and not amenable to exploration and discovery. By designing small views and iteratively discussing them with experts, we formed a set of initial tasks.

Developing a prototype. We implemented a prototype fulfilling the initial tasks and analytical workflows. Experts explored the system and provided feedback. They were basically satisfied but proposed several suggestions for improvement (e.g., more intuitive visualization to compare different appointment periods within a group and more flexible ways to choose individuals of interest).

Refining the prototype. We updated the design tasks based on the feedback and refined the system. The experts were more satisfied with this version and found many interesting stories. They also reported bugs during the process which we solved responsively. After the iterative design process, we finally summarized the requirements and formed them into three levels of six analytical tasks.

The **overall-level** tasks give an overview of career mobility and guide users to choose regions of interest quickly.

T1 What are the general characteristics of career mobility? Experts require a quick overview of the data to identify the regions of interest for further analysis. People with vertical movements (i.e., who changed job levels, such as promotion and demotion) or attained high job levels are more likely to attract the experts' attention.

T2 What special features do the groups with vertical movements have at different time periods? Further investigation is required to explain the mobility patterns in different historical periods. Experts want to find out if there are common features among the promotion/demotion groups.

The **social-group-level** tasks emphasize the detection and analysis of latent social groups, including intra-group inspection and inter-group comparison.

T3 What are the characteristics of different social groups? Identifying latent groups based on similarity in patterns of vertical movements is essential to finding hidden rules in career mobility. The system also needs to provide statistics for groups to help find individual similarities.

T4 What is the mobility pattern for each group? After identifying groups, experts want to check the mobility patterns at the job-level and department levels within a group to compare career path similarities (e.g., the promotion speed and department transfer routines).

The **individual-level** tasks require delving into details to inspect the career mobility of individuals of interest.

T5 What are the mobility characteristics for different individuals? Besides macro-level analysis, experts also wish to examine specific individuals at a micro-level. Those more influential with long careers and high positions are interesting targets.

T6 How do the mobility patterns of each individual and his social relationships change over time? To better understand differences in opportunities, experts wish to inspect individuals' different social relationships and how their mobility evolves and interacts with them.

3.3 System Overview

Fig. 1 provides a system overview of *CareerLens*. It is a web-based application with three modules, namely, a data preprocessing module, a data analysis module, and a data visualization module. The data preprocessing model transforms the large-scale record-based personnel raw data into individual-based data and stores them in the database. The data analysis module detects latent groups, computes index statistics, and extracts social relationships. They comprise the back-end and are implemented using Python and MySQL. The data visualization module constructs a front-end application using Vue.js [43] and D3.js [44] with multiple coordinated views to support a comprehensive analysis.

4 Data Analysis

In this section, we describe our analytical methods to process this career mobility data. First, we introduce the algorithm for identifying latent groups. Second, we distill three types of social relationships and a set of indices used in sociology to extract statistical features of different social groups.

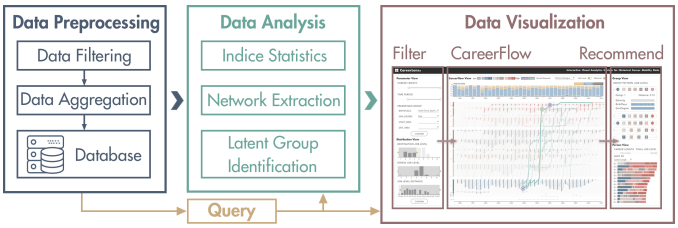


Fig. 1. System overview. *CareerLens* has three parts: a data preprocessing module, a data analysis module, and a data visualization module.

4.1 Latent Social Group Detection Algorithm

After choosing the population of interest, we apply latent group detection to classify officials and distill career patterns. Here we focus on the vertical career mobility analysis. We define a *latent social group* as a cluster of individuals with similar vertical (i.e., job level) movement sequences. As discussed in Section 2.1, latent group identification methods in social science are mostly ad-hoc and insufficient for our experts to understand promotion and demotion rules.

MinDL Algorithm. Considering the diversity of career paths in CGED-Q, we adopt a noise-tolerant method called MinDL [45], [46], [47] for the job-level sequences to better characterize vertical career mobility and cluster similar career paths. It is an optimization-based algorithm to partition all the sequences into different clusters and extract sequential patterns simultaneously at a minimum cost. The method adopts the minimum description length principle [48] to make the extracted clusters and corresponding patterns "best describe and summarize" the original sequence collection. This principle requires to minimize the *description length* to obtain the optimal results. A cluster of original sequences can be summarized by two parts: a sequential pattern (i.e., a subsequence that frequently appears in the cluster) and a set of corrections (e.g., add, delete, and replace) to restore the original sequences from the sequential pattern. Therefore, the total *description length* can be represented by the sum of the pattern lengths and correction lengths:

$$L(C) = \sum_{(P,G)\in C} \|P\| + \left(\alpha \sum_{(P,G)\in C} \sum_{s\in G} \|edits(s,P)\| \right) + \lambda \|C\| \tag{1}$$

, where $P = (e_1, e_2, ..., e_l)$ is a sequential pattern of a cluster, e_i is an event type (i.e., a job level), $s = (e_1, e_2, ...e_m)$ is an individual's original job level sequence, $G = \{s_i, s_j, ...s_k\}$ is a cluster of original sequences belonging to pattern P, $C = \{(P_1, G_1), (P_2, G_2), ..., (P_n, G_n)\}$ is a set of all clusters of the original sequence collection, $\{G_1, G_2, ..., G_n\}$ is a partition of the whole sequence collection, $L(C)$ is the total description length of sequential patterns of all clusters, and $edits(s, P)$ is the minimal set of edits to restore the original sequence s from the pattern P. The equation has two parameters: α controls the importance of information loss in the correction phase and λ controls the total number of sequence clusters. The larger the λ, the less the number of sequence clusters. To minimize the *description length* is to minimize $L(C)$.

Implementation and Improvements. The implementation of MinDL is in a bottom-up manner. Each original

sequence is initially a cluster. Clusters are merged iteratively and greedily by pairs. Considering the merge phase (i.e., $edits$ computation) is time-consuming, we adopt Locality Sensitive Hashing (LSH) [45] using weighted Jaccard similarity to speed up the algorithm. Specifically, group sequential patterns are firstly transformed into multisets. During the iteration, only if the weighted Jaccard similarity of the two multisets is larger than a threshold th, will they enter the merge phase. We also improve the original method to fit our scenario. First, to avoid producing clusters of small sets, we sort the sequential patterns in descending order in every iteration instead of starting the merging in random order. Second, since the origin and destination job levels are important according to experts, we constrain the algorithm to merge sequences only if they have the same origin or destination job level. After a set of experiments, we set two parameters as 0.8 (α) and 0 (λ), and the threshold for Jaccard similarity as 0.5 regarding the reasonable results.

We choose MinDL instead of other sequence mining algorithms (e.g., CM-SPAM, BIDE+, MaxSP, VMSP, and VGEN) because firstly it is highly tolerant of noise and can thus catch sequential patterns of a group while still allowing for slight differences. Such tolerance fits well in our historical data with missing or mistaken records and long career paths that cannot be clustered well using other methods. Second, other methods can only uncover frequent patterns filtered by a minimum support threshold, which is difficult for our experts to understand and define. Moreover, they will ignore unusual patterns if the threshold is set improperly, which experts also wish to check. Thus, MinDL is appropriate for partitioning all the careers into different clusters.

4.2 Data Processing

Besides latent group identification, we also distilled three social relations and computed several mobility indices.

Social Relationships. We have extracted three types of social relations based on experts' suggestions: colleagues (working in the same department at certain periods), townsmen (from the same *Birthplace*), and classmates who passed the exam in the same year. For colleagues, we store the overlapped periods and the departments they served together.

Career Mobility Indices. There are three indices reflecting career mobility at the individual and group levels [49].

- *Direction*: Individual-level index. Promotion and demotion are two types of job-level changing directions. At a time period (t_i, t_j), people are promoted if the job level at t_i is lower than the job level at t_j.
- *Vertical Mobility Rate (VMR)*: Group-level index. It is the proportions of individuals who have changed their job levels within a time period (t_i, t_j). In this paper, we divide VMR into two types, namely, upward VMR for promotion ($UVMR_{(t_i,t_j)} = \frac{\|PromotionGroup\|}{\|All\|}$) and downward VMR for demotion ($DVMR_{(t_i,t_j)} = \frac{\|DemotionGroup\|}{\|All\|}$).
- *Distance*: Individual- and group-level index. The individual-level distance refers to an individual's (referred as x) job level difference within a period (t_i, t_j): $Distance(x, t_i, t_j) = Level(x, t_j) - Level(x, t_i)$. The group-level distance is the average of all the individuals' distances within this group.

5 Visual Design

In this section, we first present a user scenario to give an overview of the system. Then we introduce details of our visual designs and interactions.

5.1 User Scenario

We provide a usage scenario to illustrate how *CareerLens* facilitates experts to explore historical career data. The expert first specifies the population of interest in the *Parameter View* (Fig. 2-A). The statistics on the individuals' job levels are shown in the *Distribution View* (Fig. 2-B). He brushes regions of interest and three views on the right are updated. The *Mobility Rate Timeline* (Fig. 2-C1) reveals the overall evolution of vertical mobility. The expert hovers on one bar unit to compare the statistical features of promotion and steady groups using a glyph (Fig. 2-C3). He then inspects the *Population Flow* below (Fig. 2-C2). After the overview, the expert goes to the *Group View* (Fig. 2-D) and chooses a latent group with special promotions. Highlighted in the CareerFlow View as a subflow (Fig. 4-A), he specifies several individuals in the group as threads (Fig. 4-A1, A2, A3) for comparison. He uses the two switch buttons to change the flow (job-level and department) and time modes (absolute and relative) to check more details. Afterwards, he wonders how the influential individuals' career mobility and their social relationships are revealed. He turns to the *Person View* (Fig. 2-E) and chooses one with a high position. He specifies a relation and corresponding career threads are highlighted (Fig. 2-C2). The expert finds that almost all individuals ended up in high positions. Hovering on the threads for detailed information, he finds they all have strong connections.

5.2 Data Filtering and Initial Inspection

The system provides two views for data query and initial inspection (**T1**). The *Parameter View* (Fig. 2-A) is a filtering dashboard for users to choose a population of interest. From top to bottom, users can select a particular career length, time period, and different official backgrounds. In the *Distribution View* (Fig. 2-B), three bar charts show the vertical movement statistics of the chosen population (**T1**). They include the population distribution of origin and destination job levels, and job level *distance* of the whole career. Users can quickly locate the data of interest for further analysis.

Justification: At first, we provided a list of forms for experts to specify the population of interest. However, they still regarded them as inconvenient. Thus, we improved by providing drop-down menus which were more efficient.

5.3 Group and Individual Recommendation

CareerLens contains two views (Fig. 2-D, E) for summarizing the latent groups (**T3**) and influential officials (**T5**). Users can choose interesting ones for detailed inspection.

Description: The *Group View* (Fig. 2-D) lists latent groups (**T3**) detected by the algorithm in Section 4 given the officials chosen from the *Distribution View*. Each group is represented by an intuitive node-link sequence showing the group sequential pattern. Numbers in rectangles form into group patterns, and two circles at the beginning and end show the actual minimum and maximum job levels within the group.

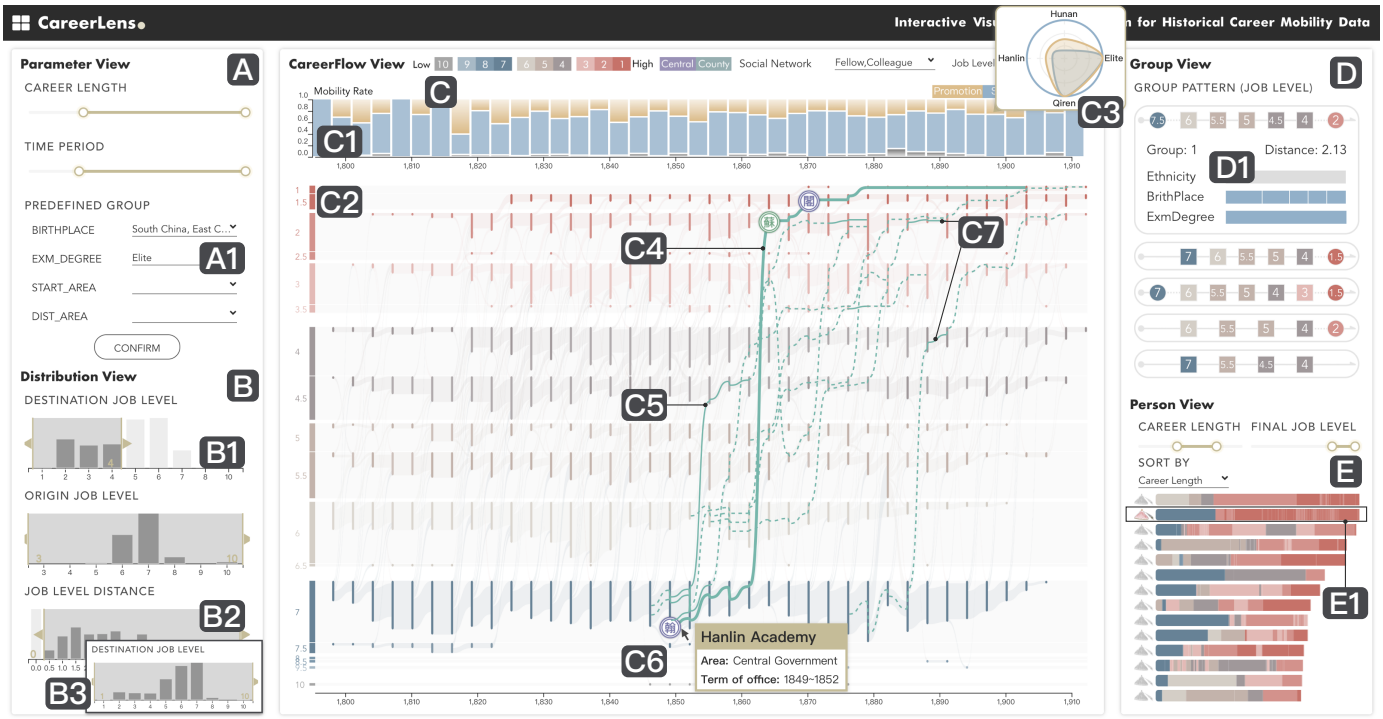


Fig. 2. The system interface of *CareerLens*. The Parameter View (A) provides several constraints for specifying a target population. The Distribution View (B) displays the job level summary for filtering interested individuals. The CareerFlow View (C) presents the mobility patterns in two parts: (1) a percent stacked bar chart showing mobility evolution and a glyph comparing the group features and (2) an enhanced flow showing mobility in job-level and department modes. The Group View (D) and Person View (E) list latent groups and influential individuals respectively for comparison.

Groups are sorted by the numbers of officials in descending order. Experts require a statistical summary of each group to check whether officials have commonalities, so we design a folded information card (Fig. 2-D1). We compute the average job-level *distance* and give four horizontal proportional bars to summarize the demographic profile (i.e., Ethnicity, Birthplace, Family Background, and Exam Degree). Experts can quickly identify whether a value dominates an attribute, thus finding commonalities. They can also hover on each horizontal bar to check detailed statistics of this attribute.

The *Person View* (Fig. 2-E) lists a visual summary of potential influential officials (**T5**). Since domain experts wish to filter influential officials flexibly using career length and final job level, we thus provide these two constraints. We also provide two methods to sort officials (i.e., career starting years or career lengths). Each horizontal bar shows an individual's job level information. The total length encodes the career length. It comprises multiple small bars with length encoding corresponding working periods and color encoding different job levels. According to our experts, we divide job levels into four levels represented by four color categories: red for high-level jobs (from level 3 to 1), brown for middle-level (from level 6 to 4), blue for low-level (from level 9 to 7), and grey for the lowest level (level 10). Inside each level, we use saturation to distinguish job levels. The darker the color, the higher the job level. The color scheme for job levels is consistent in the whole system. Users can have a quick overview of these individuals and find those with long careers and great promotion distance.

Justification: We presented all the information cards at first. However, experts mostly focused on the group patterns and regarded the statistics as overwhelming when shown all at once. Thus, we preserved the group patterns and folded the information cards to be inspected as needed by experts.

5.4 CareerFlow View

The *CareerFlow View* (Fig. 2-C) is the primary visual component of our system, which consists of two parts that share the same time scale: (1) The *Mobility Rate Timeline* (Fig. 2-C1) displays the mobility over time (**T1, T2**) with a glyph comparing features of two groups; (2) The *Population Flow* (Fig. 2-C2) reveals the overall mobility (**T1**) with particular individuals highlighted (**T4, T6**). Different flow modes and interactions are supported to facilitate the analytic process.

5.4.1 Mobility Rate Timeline

The *Mobility Rate Timeline* (Fig. 2-C1) shows the vertical mobility rate (VMR) changing over time, giving an overview of the data (**T1**) filtered in the *Distribution View*. Hovering on a bar unit, a glyph (Fig. 2-C3) is shown to compare statistics of different groups (i.e., promotion, demotion, and steady groups), thereby revealing the salient features influencing vertical mobility at different periods (**T2**).

Description: In Fig. 2-C1, a percent stacked bar chart shows the temporal VMR (**T1**). The x-axis represents the time. We split the timeline into three-year steps based on our experts' suggestions, since officials were reviewed and possibly promoted every three years within the Qing bureaucracy. The y-coordinate of each time step represents the VMR of the current three years. Each stacked bar consists of three bar units in three colors, yellow for UVMR (promotion), black for DVMR (demotion), and blue for the proportion of the unchanged officials (steady). To find salient

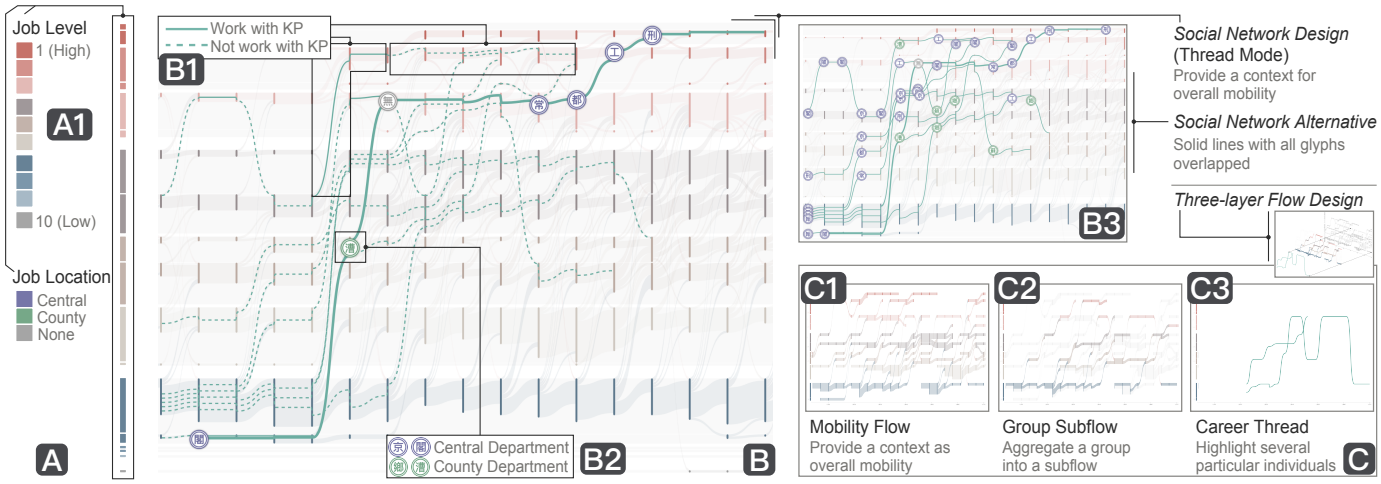


Fig. 3. The visual design of *CareerLens*. (A) The color legend used in two job-level and department flows. (B) The CareerFlow View visualizes the overall career mobility using an enhanced Sankey diagram. (B1) and (B2) are the detailed design components and (B3) is a component alternative in CareerFlow View. (C) The three-layer flow design in *Population Flow*, including individual thread, group subflow, and overall mobility flow.

features in the promotion/demotion groups compared with the steady group (**T2**), we use a glyph (Fig. 2-C3) when hovering on a bar unit. It is a radar chart with four features corresponding to four categorical attributes. The four attributes are *Birthplace*, *Ethnicity*, *Exam Degree*, and *Job Department* at the beginning of the current three years. For the promotion/demotion group, experts want to know whether a value (i.e., feature) dominates an attribute and distinguishes this group from the steady group. We thus distill the value with the largest proportion in an attribute in the promotion or demotion group. We then compute the proportion of the same attribute value in the steady group for comparison. For example, if Anhui-born officials dominate the *Birthplace* attribute by 80% in the promotion group, we would check the proportion of Anhui-born officials in the steady group. If it is only 10% in the steady group, it indicates that being Anhui-born is a special feature for this promotion group. The color scheme is the same as the percent stacked bar chart.

Justification: At first, we tried a step line chart showing the VMR over time with a bar chart tooltip. However, after trying the system, experts wished to see the promotion and demotion rates separately. Moreover, they reported that the bar chart wasted space and was hard to remember with various attributes. Therefore, we changed the bar chart into a radar chart and showed the UVMR and DVMP in a percent stacked bar chart, and finally obtained positive feedback.

5.4.2 Population Flow

The *Population Flow* (Fig. 2-C2) displays a detailed career mobility pattern in two modes (i.e., job-level and department) with two types of time (i.e., absolute and relative), which can be switched by two buttons. We provide a novel flow design adopting a multi-scale approach to form into the superimposition of three layers, namely, overall mobility flow (Fig. 3-C1), group subflow (Fig. 3-C2), and individual career thread (Fig. 3-C3)). The groups of interest (**T4**) or individual social relationships (**T6**) are embedded into the overall population. This allows for the study of the mobility of a specific population within its broader historical context.

Description: The *Population Flow* (Fig. 3-B) is built on a modified Sankey diagram and strengthened with subflows and threads embedded to show interested groups and social relationships. The x-axis is a timeline and consistent with the *Mobility Rate Timeline*. The vertical direction encodes job levels or departments in two modes, respectively.

Mobility Flow. We present two mobility flow modes, namely, job-level mode (Fig. 3-B, Fig. 4-A) and department mode (Fig. 4-B). Job-level mode summarizes vertical career mobility over time. As shown in Fig. 3-A, job levels are double-encoded using both colors and y positions. We use the same color scheme in the *Person View*. Y positions from bottom to top represent job levels from low to high. Each node in the flow represents a job level group at a timestamp. The height of flow encodes the number of individuals. The rectangles with gray backgrounds provide clear alignments. We adopted two layout rules to alleviate the flow crossings and keep the flow trends intuitive. First, we place the nodes representing the same job layer at different timestamps in the same vertical position (aligned to the top). Second, we sort the flows between two adjacent timestamps according to the origin and the destination job levels to reduce the crossings.

In the department mode (Fig. 4-B), we classified departments into fifteen categories sorted by their importance in the bureaucracy based on experts' suggestions. They are encoded by two colors (Fig. 3-A1, purple for the central government and green for the county administration), since experts regard it as essential to distinguish these two places to study personnel transfers among departments. Other encodings are similar to the job-level mode. In both modes, we provide a relative time mode (Fig. 6-A, B) that aligns the beginning timestamps of each individual to support career comparisons within a specific group. Users can also zoom in the flow along the timeline for a detailed inspection.

Group Subflow. The latent groups and individuals' social relationships can be highlighted in the *Population Flow* (Fig. 3-C) for a detailed investigation (**T4, T6**). Two modes are supported to portray the career mobility of the selected individuals. The first mode (Fig. 3-C2) aggregates the career paths within a group into a subflow, where each flow segment is embedded into the corresponding overall flow segment. It is intuitive and scalable, allowing for inspecting group mobility and proportions relative to the whole population.

Career Thread. The second mode (Fig. 3-C3) represents each person as a green thread and embeds it into the mobility flow and group subflow it belongs to. Users can highlight individuals within a latent group or one's social relationships as threads, and switch to the other flow modes to compare.

Social Relationships. To show the social relationships of a selected individual (KP) (**T6**), users can select specific social relationships (as mentioned in Section 4) in a drop-down menu. The career threads that have social connections with KP will be highlighted over the overall mobility flow. In Fig. 3-B1, each thread consists of two line types, namely, solid lines indicating the time period this official worked with KP, and the dashed line meaning they did not work in the same department. Different department glyphs (Fig. 3-B2) in two colors (with the same encodings in department flow) are overlaid onto the KP's career thread. The characters of the glyphs are abbreviations of different departments.

Justification: For the social relationship layout, we first overlaid the department glyphs to all the career threads which are in solid lines. However, our experts found visual clutters distracting (Fig. 3-B3), and the interactions showing co-working periods were not intuitive during the exploration. Therefore, we removed the glyphs in other threads and only left those of KP's. Moreover, we use solid and dashed lines to show the working interactions with KP more clearly.

5.5 Cross View Interactions

CareerLens provides various interactions enabling users to explore multiple coordinated views [7].

Query and Filter. Users can choose individuals of interest in the *Parameter View* and *Distribution View* through a drop-down list and bar charts, respectively.

Scale and Zoom. The *Mobility Rate Timeline* and *Population Flow* support zooming and panning simultaneously to allow a detailed investigation of the region of interest.

Tooltips. We provide tooltips in the *Group View* (i.e., Fig. 5-A1, C1), *Mobility Rate Timeline* (i.e., Fig. 2-C3), and the *CareerFlow View* (i.e., Fig. 2-C6) for detailed information.

Highlights. Users can choose groups of interest or individuals in the *Group View* and *Person View*. They will be highlighted in the *Population Flow*. Users can click threads in a flow mode, which will be highlighted in other modes.

Switching Contexts. Users can inspect four mobility modes in the *CareerFlow View* which are supported using switching buttons. They can also switch between threads and subflows using a context menu.

6 Evaluation

We evaluate the effectiveness and usability of *CareerLens* with two case studies in Section 6.1, an expert interview in Section 6.2, and a longitudinal study in Section 6.3.

6.1 Case Study

We invited experts in Section 3.2 to conduct case studies. Experts from different disciplines have diverse research focuses, so we summarize observations and comments from a free-form exploration and form two cases to fully demonstrate the features of *CareerLens*.

6.1.1 Political Elites from South and East China

Political elites with a high degree in the Imperial Examination have received much attention in the study of Qing history. Three experts aimed to investigate these elites, especially those from South and East China (Fig. 2-A1), where the economy was well developed in the late Qing dynasty. Given different research interests, E_A and E_C focused on the mobility patterns and E_B examined the social connections.

Overall Mobility Inspection. E_A was interested in the overall mobility trend to learn social stability over time (**T1, T2**). He began with the job level summary (**T1**) in the *Distribution View* (Fig. 2-B). From Fig. 2-B1, E_A noticed the proportion of destination job levels higher than level 5 was high compared with that of the whole elite population (Fig. 2-B3). Thus, he chose these officials with vertical movements (Fig. 2-B2). The results were shown on the right (Fig. 2-C, D, E). In the *Mobility Rate Timeline* (Fig. 2-C1), the UVMR was initially unstable (1800–1820) and stabler later (1880–1900) (**T1**). The DVMR was at a low level with a slight rise around 1883. *"I thought the mobility rates would be high in the late Qing, but it seems not."* E_A then checked the *Population Flow* in both job-level and department modes (Fig. 2-C2, Fig. 4-B) (**T1**). The population distributions were consistent with previous findings. However, zooming for the detail, E_A was surprised by various types of flows, *"I never thought so many different job-level transitions happened in a short time period."*

Considering the political unrest in the late 19 century, he wondered how the promoted officials varied (**T2**). Thus, he hovered on corresponding yellow bars to compare the groups. E_A stopped at one bar where the glyph (Fig. 2-C3) showed that officials from Hunan were more likely to be promoted than steady ones during this period. E_A commented that it might relate to a policy changed where many Hunan-born officials were recruited and promoted. The overall mobility provided a holistic view of the political elites based on the mobility rate, and job-level and department transfers.

Group-level Inspection. E_C wanted to identify latent groups to find typical promotion routes in the bureaucracy (**T3, T4**). From Fig. 2-D, she observed a set of groups with different promotion patterns (**T3**) and chose the one with the most officials (Fig. 2-D1) for detailed analysis (**T4**). In the *Population Flow* (job-level mode) (Fig. 4-A), the group was highlighted as a subflow, most of which were distributed in the middle region. E_C wondered whether they also had similar mobility between departments, so she switched to the department mode (Fig. 4-B). Most of them moved from the first two central departments (Fig. 4-B4) to a county administration (Fig. 4-B5). It indicates a strong correlation between job-level and department transfers in this group.

E_C noticed that a few officials moved differently at the later stages at job levels (i.e., got higher levels or suffered great demotions) based on the noise-tolerant group detection method. Given the similar promotion routes at most stages, she wondered about the factors that distinguished them from others (**T4**). Thus, E_C highlighted them as threads (Fig. 4-A1, A2, A3). Switching to the relative time mode (Fig. 4-A4), two officials (yellow and blue threads) obtaining higher positions promoted faster than other group members (subflow), which seem to be successful careers. Curious about the turning points causing the final promotion/demotion, she chose the department mode. The two officials who got higher positions

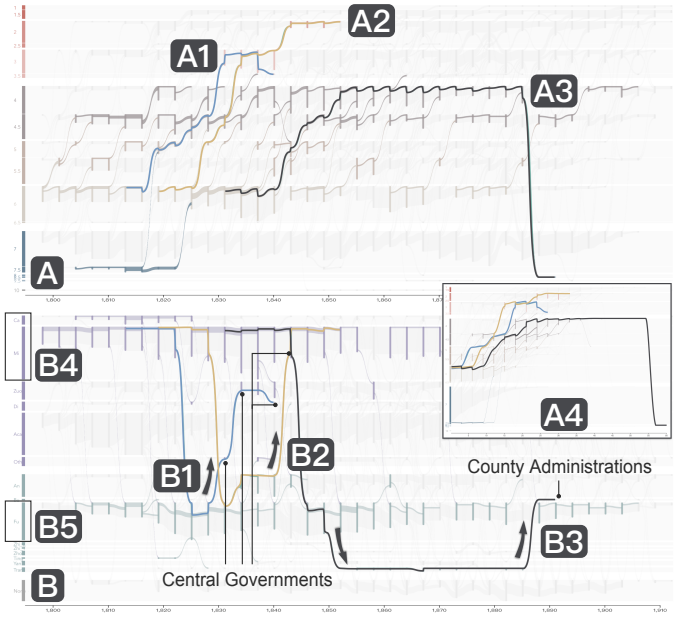


Fig. 4. The career mobility of political elites in South and East China. (A) presents the *Population Flow* in the job-level mode (absolute and relative time). (B) presents the *Population Flow* in the department mode.

were finally called back to the central government (Fig. 4-B1, B2), whereas the demoted one wandered among other county departments (Fig. 4-B3). E_C explained, *"Sending officials to counties was an existing while rare route to have them better trained. They would be called back if they worked excellently. Also, they would be demoted if they failed to manage work well."* She considered this latent group special since officials had similar career features in both job level and department modes, indicating a potential career route in the bureaucracy. The three special officials also revealed the possibility of promotion and demotion based on working performance.

Individual-level Inspection. As a historian focusing on a case-based approach, E_B was concerned with influential officials and their social relationships (**T5, T6**). In the *Person View* (**T5**, Fig. 2-E), based on domain knowledge, she defined those with final job levels higher than 3 and career lengths longer than 20 years as influential officials. Sorting by career length, she quickly noticed that the second official (KP, Fig. 2-E1) jumped directly from level 7 (dark blue) to level 2 (middle red). It indicates that he was a fast-rising star. Such rapid promotion was rare in the Qing bureaucracy. Thus, E_B highlighted KP as a thread (Fig. 2-C4) in the *Population Flow*.

Besides KP's capabilities, E_B supposed that some underlying social relationships might have also played an essential role in his career (**T6**). Thus, she chose those from the same hometown as KP since it was a special social relationship in China. Many officials were highlighted, with a large proportion having no career connection with KP. Thus, she strengthened the relationship by choosing colleagues. The filtered officials' destinations showed that most of them ended up in high positions. *"I like the department glyphs and tooltips, and solid/dashed lines are also clear to know when others worked with this key official."* In Fig. 2-C5, an official had worked with KP in the early stages of KP's career and held a higher job level than his. Other officials were promoted more slowly than KP, but interacted with him in later stages (Fig. 2-C7). E_B hovered on these threads to obtain more information and verified that most officials had strong social connections with KP [1] (e.g., elevated KP, promoted by KP, or nepotism). Therefore, in organizational hiring, social relationships may influence individuals' career mobility (e.g., be promoted via employee referrals [21], [22]).

From this case, we demonstrate the usefulness of *CareerLens* to study career mobility at three LODs. Experts could quickly have an overview of the mobility trends and study groups and individuals of interest with coordinated views.

6.1.2 Career Mobility between Different Origins

The comparison between different populations of interest is another major topic in the study of career mobility. In particular, the mobility of individuals from different geographical origins may vary due to the unbalanced distribution of social resources. Considering the background of the Qing Dynasty, E_D was curious about the differences between coastal and central China. After exploring these regions, she found interesting differences between Hunan- (central) and Zhejiang-born (coastal) officials.

E_D started by loading officials from Hunan. From the *Distribution View* (Fig. 5-A2), the population with the origin job level was abnormally high in level 2 (**T1**). Thus, she selected those with a high job level (i.e., levels 1, 2, and 3) to analyze in the *Population Flow* (Fig. 5-A). The population was uniform over time at the middle and low job levels, while at levels higher than 3, it suddenly increased in the late Qing dynasty (1870-1890) (**T1**). A frequent group from level 2 to 1.5 (Fig. 5-A1, highlighted in Fig. 5-A3) strengthened this result. The *Exam Degrees* of these officials were all military degrees (**T3**). To identify the reason for this population surge, E_D switched to the department mode (**T4**, Fig. 5-B). Most officials in this group worked in a county administration (Fig. 5-B1). It reminded her of a war [2] that had broken out in the late Qing dynasty when many generals from Hunan were directly assigned a high job level because of their military performance. Back to the *Distribution View* (**T1**), she noticed a high distribution in origin job level 7 (Fig. 6-A1) and then chose these officials. The first group in the *Group View* (**T3**, Fig. 6-A) indicated that a large number of officials moved from level 7 to 5.5. After highlighting the group in the *CareerFlow View* at a relative time mode (**T4**, Fig. 6-A), she found almost all officials from level 5.5 came from this group (Fig. 6-A3). It indicated a potential promotion route to level 5.5. Turning to department mode (Fig. 6-A2), the highlighted subflows were distributed in the Hanlin Academy (central) and Fu (county), but hardly had officials transferred between them. E_D supposed there were two types of promotion routes in departments from level 7 to 5.5.

E_D compared the mobility of Zhejiang-born high officials following the same exploration steps. From Fig. 5-C, the mobility flow was stable and no surge occurred (**T1**). Interestingly, she noted the fourth largest group (Fig. 5-C1, C3) showed a frequent switching between job level 7 and 3. Most of them obtained high degrees in the Imperial Examination (**T3**, political elites). Turning to the department mode (**T4**, Fig. 5-D), they all worked in the Hanlin Academy (Fig. 5-D1) and were sent to other unknown administrations (Fig. 5-D2),

1. https://en.wikipedia.org/wiki/Li_Hongzhang
2. https://en.wikipedia.org/wiki/Taiping_Rebellion

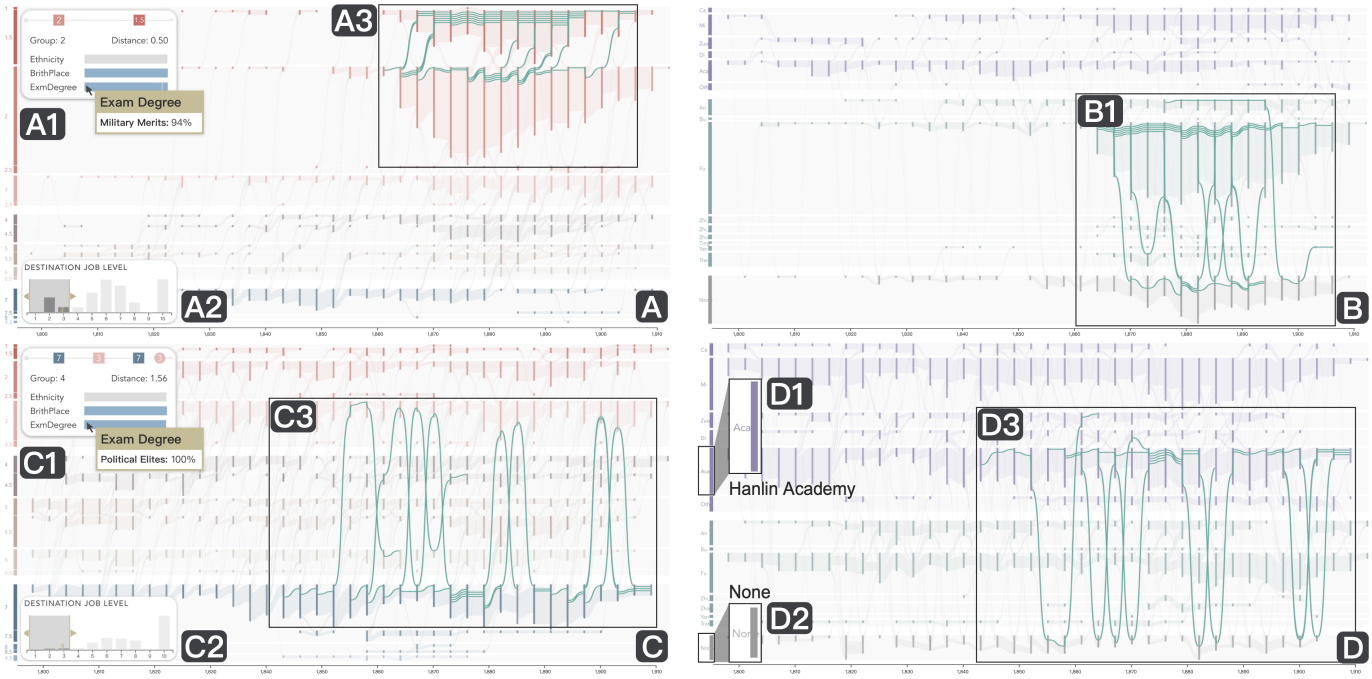


Fig. 5. The career mobility in Hunan and Zhejiang Provinces. For Hunan Province, (A) shows the *Population Flow* in the job-level mode and (B) in department mode. For Zhejiang Province, (C) summarizes the *Population Flow* in the job-level mode and (D) in the department mode.

thereby resulting in a promotion. They then returned to the Hanlin Academy and were restored to level 7. After checking more details, E_D remembered that capable officials from the Hanlin Academy would be sent temporarily to the counties to oversee the Imperial Examinations, which resulted in the brief promotion. After they were done, they went back to the Hanlin Academy to their original job levels. However, she never expected that such trajectories were more common for Zhejiang officials than those from Hunan. This difference required in-depth investigations. Next, choosing officials starting from level 7 (Fig. 6-B1), E_D found a demotion group in the *Group View* (level 6 to 7, Fig. 6-B) (**T3**). Switching to the department flow with a relative time mode (**T4**, Fig. 6-B2), most of them moved from Ministry (central) to Fu (county). Checking more details, she found that they changed from the same position to another position. Although it was a demotion in terms of job level, they gained more real power in the counties. E_D stated that the mobility differences between the two provinces helped her validate the influences of origin differences on career mobility and also find unexpectedly frequent groups.

The case demonstrated that *CareerLens* could help experts effectively conduct a comparative analysis of different populations of interest.

6.2 Expert Interview

To evaluate the usefulness and effectiveness of *CareerLens*, we interviewed the four experts (E_A-E_D) in Section 3 and four new experts (P_A-P_D) who used *CareerLens* for the first time. P_A and P_B are researchers studying social mobility in the early 20th century and contemporary China, respectively. P_C is a postgraduate studying Qing history and geography. P_D is a postgraduate in quantitative social science. P_A and P_B had heard of the CGED-Q but hadn't worked with it before. P_C and P_D had conducted research using the CGED-Q.

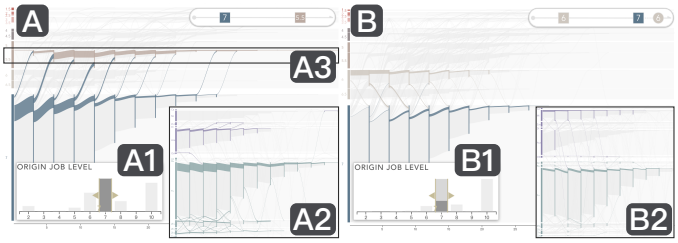


Fig. 6. The career mobility in Hunan and Zhejiang Provinces. (A) presents a group mobility pattern in relative time mode in Hunan, where most of the officials at job level 5.5 come from job level 7. (B) shows the mobility pattern of a demotion group from job level 6 to job level 7 in Zhejiang.

Procedure. Each interview for new experts (P_A-P_D) consisted of five sections and lasted for approximately 90 minutes. First, we spent 10 minutes introducing the project background and analytical tasks. Second, we demonstrated the system workflow and visual encodings with a comprehensive example (20 minutes). We gave another 15 minutes for free exploration and a Q&A to ensure the participants were familiar with the system. Third, experts were asked to conduct a task-driven exploration using Case 1 settings in Section 6.1.1 (20 minutes). We listed six tasks using all the visual components that reflect all the visualization tasks in Section 3.2. Fourth, they explored the system freely to find interesting stories (15 minutes). Finally, we conducted a semi-structured interview by first asking a set of questions and letting them provide open-ended feedback (10 minutes). We also conducted the same semi-structured interview with our domain experts (E_A-E_D) to gain more feedback. Please refer to our supplementary material for more details. The feedback and suggestions are summarized into three categories below.

System. All the experts appreciated the clear workflow of our system, which follows and strengthens their traditional analytical method. *CareerLens* enables them to quickly

locate and analyze the population of interest by showing information at different LODs with a user-friendly interface. *"In the past, we relied on tabulations and narrowed down the range of interests manually. It's impressive that we can explore our data in a new way that is more effective"* (E_A). They also like the techniques (e.g., sequence mining and visualization) adopted, which help them find new insights. *"Previously it was hypothesis-driven and hard to find unknown results directly"* (P_A). *"These techniques help a lot find interesting stories like latent groups and social relations"* (E_B). They mentioned they had not studied social relations much previously due to the lack of advanced methods. But now, they were inspired to dig out underlying political connections among officials.

Visualization and Interaction. E_A-E_D considered *CareerLens* as a comprehensive visual analytics system that fulfilled all the design tasks. The four new experts (P_A-P_D) could understand most of the visual designs after the second phase (i.e., demonstration and Q&A) and finish the task-driven exploration within the allotted time. They were impressed by the aesthetic visualization and intuitive interactions, and also appreciated the labels and legends. Specifically, they regard the *Mobility Rate Timeline* as useful for finding salient features of promotion/demotion groups, *"...the glyph is straightforward for comparing two groups"* (P_A). For the flow design, we observed that new experts could quickly catch the idea of the overall flow and subflow but might not remember all the encodings and interactions at once. For example, in the department mode, P_C took the vertical positions for the job levels and related the flow from the central government to the counties to demotions. The vertical positions actually were not relevant to the vertical movement but only the department movement. P_D forgot how to choose an official of interest using the thread mode, *"a switch button may be more intuitive instead of this context menu"*. However, after exploring the system for a while, they eventually master it. They appreciated many visual components in *Population Flow*. Many of the experts particularly liked the design of the two time modes and the two flow modes. We observed that they would frequently switch among these modes to study the mobility from different perspectives. *"...the relative time mode makes promotion speed comparison much easier"* (P_A). E_B and P_D liked the social relationship visualization, *"...I can clearly see how others interacted with the chosen one through these threads"* (E_B). As for interactions, experts thought most of them were intuitive and useful for exploring career mobility.

Suggestions. Experts also provided valuable suggestions based on their research focuses. P_C and P_D focused on spatial mobility analysis and suggested that a distinction between different county administrations based on provinces would be better for detailed analysis. E_B and P_B preferred a raw data table to show detailed official information, which is more straightforward than the two tooltips. P_C suggested replacing the context menu with a switch button to change between the subflow and the thread modes.

6.3 Longitudinal Investigations

We have deployed our system for the regular use of our experts (E_A-E_D) for six months. During this period, we found that their exploration approaches changed over time, from verifying existing facts to in-depth analysis.

Fact Verification. When the experts first tried *CareerLens*, they started with a population of interest using a single constraint (e.g., people from one place) in the *Parameter View* to verify the facts they already knew and get familiar with the system. While most findings were consistent with previous results, they still found unexpected ones. For example, when E_B studied the group of officials with the Gongsheng examination degree, she was surprised at the population surge in job level 6 in the late Qing dynasty. It was not previously studied in this field. She wrote down the discovery for further investigation.

In-depth Study. Traditional historians tended to focus on a group of interest for a long time. When we went back to them a few months later, we found that though still focusing on one group, they had changed the exploration approaches. First, they became used to conducting comparisons with other groups. For example, E_C was working on officials from one province and preferred to compare it with other provinces recently. In this way, she uncovered the special characteristics of mobility in her target province more quickly. *"Previously, we focused on our own target provinces independently because we do not have time and ability to process so many data in different provinces. This system helps brush data more efficiently"* (E_C). They also gathered interesting findings. For example, officials starting careers from a province near the central government rarely reached high positions, while those starting from the frontier obtained more high positions. *"This is strange and even contradictory with our prior knowledge. We need to find more historical materials to explain it"* (E_C). Second, they would subdivide officials with diverse filtering combinations in the *Parameter View* for a fine-grained study. One example was that they tried to filter officials by different career lengths and narrow them down to a short time segment (e.g., 1830-1850). They could then compare the promotion possibilities with officials from diverse backgrounds and time periods. It had rarely been tried before.

Another interesting observation is the study of social relationships. Experts previously had not worked on it due to their different research focuses and the lack of advanced tools. However, during the exploration of latent groups, many experts came across influential officials with cliques that finally comprised a clique evolution history in the late Qing dynasty (e.g., Mu [3], Zeng [4], and Li [5] cliques). Although experts knew about most of these cliques, they were impressed by these evidence-based methods for revealing these cliques intuitively. *"...this confirmed that officials from the same hometown and working together were more likely to form into implicit connections such as nepotism. This system is potentially very useful for tracing people's political connections"* (E_A).

7 Discussion

In this section, we summarized the significance of our work and lessons learned during our interdisciplinary collaboration with social scientists, and also discussed the limitations and generalizations of our system.

Significance. The availability of new quantitative historical data has provided both opportunities and challenges

3. https://en.wikipedia.org/wiki/Mujangga
4. https://en.wikipedia.org/wiki/Zeng_Guofan
5. https://en.wikipedia.org/wiki/Li_Hongzhang

to the social and visualization communities to better understand human societies. Traditional analytical methods may be limited for analyzing data of large volume and high complexity, especially for traditional historians with little quantitative analytical knowledge. Thus, close collaborations between different communities are required to explore the data and gain new insights. In this work, we study a newly available dataset with social scientists from multiple disciplines to understand career mobility from a historical perspective. We see our work as a first step toward career mobility visual analytics and hope our experience can inspire future collaborations between social science and data science.

Lessons Learned. Collaborations with social scientists during the whole process provide us with valuable experiences in interdisciplinary studies. First, the user-centered process is essential to distill analytical tasks. Such newly available quantitative datasets have attracted experts from multi-disciplines with diverse research focuses. Sociologists and historians are mostly not familiar with visual analytics and cannot clearly describe their requirements. Thus, continuous interviews and a summary of tasks are crucial. Second, choosing better approaches to solve problems requires more experiments since the solution space is larger in interdisciplinary studies. We initially sought methods in sociology to identify latent groups, while these methods (e.g., Poisson regression [16], [17]) are ad-hoc or semi-automatic and cannot be directly adopted. As such, we chose a sequence mining method in data mining and obtained good results.

Limitations. Although the case study and expert interview demonstrated the potential of *CareerLens* to analyze career mobility data, the work still has limitations. First, the primary concern is scalability. Although we have proposed a multi-scale flow design to alleviate the scalability problem when highlighting all the individuals within a group, this issue still exists. For example, although the visual representation of embedding an individual's social relationships into the *Population Flow* is intuitive, visual clutter occurs when the number of officials in a social relationship grows. A possible improvement would be to combine it with a node-link diagram, where users can select individuals in the flow. Second, more straightforward system designs are required to avoid complex interactions, such as using switch buttons uniformly to change flow and thread modes. We also plan to add a user guide for experts to explore the system. Third, our system lacks detailed information to support further reasoning by experts. Some experts aim to investigate specific job positions. Some wish to distinguish county departments at a more fine-grained provincial level. Thus, additional detailed views showing this information should be included in the future. Fourth, the evaluation only involves historians and sociologists, since the primary purpose of *CareerLens* is to help these domain experts to analyze career mobility. However, such a valuable historical dataset may also be of interest to a wider range of users such as the general public to learn more about the history of officials in the Qing dynasty. Additional user studies involving non-experts are needed to collect feedback and requirements to guide the adaptation of *CareerLens* for public use.

Generalizations. Although *CareerLens* is designed for career mobility analysis for a particular historical dataset, it can be applied to other scenarios. First, other longitudinal datasets related to careers and life histories can be studied. The most similar one is career mobility within an organization with a sophisticated and steady organization structure and well-defined job levels, such as the civil service and the military manpower of a country [50], [51]. Other generalized scenarios could be life histories over a whole population with multi-variate statuses (e.g., income, health, and migration) [2], [52]. We suggest practitioners choose a dimension as an analytical focus (e.g., income levels and disease categories) like our application (i.e., vertical mobility). The main effort required is to pre-process the raw data into individual-based sequences along with their attributes. Then the system can be applied with slight changes to analyze this key dimension. The filtering criteria and visual summaries in the four side views should be updated based on the key dimension. In the *Mobility Rate Timeline*, practitioners can define groups based on the key dimension and compare them using user-specified attributes. The *Population Flow* holds almost the same encoding as our application, only with different dimensions that encode the vertical position (i.e., replacing job level and department with other dimensions). Second, the multi-scale design for embedding groups, individuals, and social relationships in the flow provides a new representation for studying individuals of interest over a whole population.

8 Conclusion

In this paper, we have presented *CareerLens*, an interactive visual analytics system for experts to explore career mobility from a historical perspective using a newly available dataset (CGED-Q). It is well-coordinated to show overall mobility evolution, extract latent social groups, and depict the different social relationships of individuals, which provides a new approach to traditional career mobility studies. Our case studies, expert interviews, and longitudinal study demonstrate the effectiveness and usefulness of the system.

Based on *CareerLens*, we envision several promising directions for future researches. First, besides using job level for the latent group detection, we plan to utilize different features (e.g., departments and job titles) and further incorporate multi-features to capture different types of latent groups. Second, we intend to improve the scalability of *CareerLens*, such as using a mixture design of *Population Flow* and node-link diagrams instead of embedding all career threads into the flow. Third, considering that career mobility is always affected by important events, we aim to find the correlations between career mobility and different events (e.g., political events and disasters) for in-depth reasoning.

Acknowledgments

We would like to thank all the reviewers for their constructive comments. We also thank Huan Wei, Jiang Wu, Furui Cheng, Xingbo Wang, Yuzhe Jiang, and Leni Yang for their kind help on this project. The work is partially supported by Hong Kong Research Grants Council (RGC) General Research Fund (GRF) grant 16213317, National Natural Science Foundation of China (62072400), Zhejiang Provincial Natural Science Foundation (LR18F020001), and the 100 Talents Program of Zhejiang University. Construction of the CGED-Q was supported by Hong Kong RGC GRF 16600017.

REFERENCES

[1] S. Manson, J. Schroeder, D. V. Riper, and S. Ruggles, "Ipums national historical geographic information system: Version 14.0." [Online]. Available: https://ipums.org/projects/ipums-nhgis/d050.v14.0

[2] International Institute of Social History, "The historical sample of the netherlands (hsn)." [Online]. Available: https://iisg.amsterdam/en/hsn

[3] R. Yuxue, B. Chen, X. Hao, C. Campbell, and J. Lee, "China Government Employee Dataset-Qing Dynasty Jinshenlu 1900-1912 Public Release User Guide," 2019.

[4] B. Chen, C. Campbell, Y. Ren, and J. Lee, "Big data for the study of qing officialdom: The china government employee database-qing (cged-q)," *Journal of Chinese History*, vol. 4, no. 2, pp. 431–460, 2020.

[5] L. Stone, "Prosopography," *Daedalus*, vol. 100, no. 1, pp. 46–79, 1971.

[6] N. Sicherman and O. Galor, "A theory of career mobility," *Journal of Political Economy*, vol. 98, no. 1, pp. 169–192, 1990.

[7] T. Munzner, *Visualization Analysis and Design*, ser. A.K. Peters visualization series. A K Peters, 2014. [Online]. Available: http://www.cs.ubc.ca/%7Etmm/vadbook/

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

[9] M. Q. Y. Wu, R. Faris, and K. Ma, "Visual exploration of academic career paths," in *Proceedings of Advances in Social Networks Analysis and Mining*. ACM, 2013, pp. 779–786.

[10] S. Jänicke, J. Focht, and G. Scheuermann, "Interactive visual profiling of musicians," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 200–209, 2016.

[11] 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*, 2016, pp. 61–70.

[12] V. A. Filipov, P. Federico, and S. Miksch, "CV3: visual exploration, assessment, and comparison of CVs," in *Proceedings of the Eurographics Conference on Visualization, Posters*, 2018, pp. 1–3.

[13] 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*, vol. 24, no. 1, pp. 56–65, 2018.

[14] 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*, vol. 25, no. 1, pp. 417–426, 2019.

[15] A. V. D'Unger, K. C. Land, P. L. McCall, and D. S. Nagin, "How many latent classes of delinquent/criminal careers? results from mixed poisson regression analyses," *American Journal of Sociology*, vol. 103, no. 6, pp. 1593–1630, 1998.

[16] D. S. Nagin, "Group-based trajectory modeling and criminal career research," *Journal of Research in Crime and Delinquency*, vol. 53, no. 3, pp. 356–371, 2016.

[17] T. F. Liao and R. Y. Gan, "Filipino and indonesian migrant domestic workers in Hong Kong: Their life courses in migration," *American Behavioral Scientist*, vol. 64, no. 6, pp. 740–764, 2020.

[18] J. E. Rosenbaum, "Tournament mobility: Career patterns in a corporation," *Administrative Science Quarterly*, vol. 24, no. 2, pp. 220–241, 1979.

[19] 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 the IEEE International Conference on Data Mining*, 2015, pp. 1057–1062.

[20] H. Qu, Z. Yu, Z. Yu, H. Xu, B. Guo, and X. Xie, "What is my next job: Predicting the company size and position in career changes," in *Proceedings of IEEE Trustcom/BigDataSE/ISPA*, 2016, pp. 1668–1675.

[21] S. M. Lipset and R. Bendix, "Social mobility in industrial society," *Social Forces*, vol. 38, no. 2, pp. 172–172, 1959.

[22] L. Nan, K. Cook, and R. S. Burt, *Social capital: Theory and research*. Aldine de Gruyter New York, 2001.

[23] B. F. Jarvis and X. Song, "Rising intragenerational occupational mobility in the United States, 1969 to 2011," *American Sociological Review*, vol. 82, no. 3, pp. 568–599, 2017.

[24] M. A. Grant, "Methods for exploring and presenting contingency tables: A case study visualizing the 1949 Great Britain occupational mobility table," *Graduate Theses and Dissertations*, 2017.

[25] "Linkedin." [Online]. Available: https://www.linkedin.com/

[26] 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 the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 14–24.

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

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

[29] 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*, vol. 10, no. 1, pp. 8:1–8:25, 2019.

[30] Y. Guo, S. Guo, Z. Jin, S. Kaul, D. Gotz, and N. Cao, "Survey on visual analysis of event sequence data," *CoRR*, vol. abs/2006.14291, 2020.

[31] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman, "Lifelines: using visualization to enhance navigation and analysis of patient records," in *The craft of information visualization*. Elsevier, 2003, pp. 308–312.

[32] M. Krstajic, E. Bertini, and D. A. Keim, "Cloudlines: Compact display of event episodes in multiple time-series," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2432–2439, 2011.

[33] P. Dragicevic and S. Huot, "Spiraclock: a continuous and non-intrusive display for upcoming events," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, L. G. Terveen and D. R. Wixon, Eds., 2002, pp. 604–605.

[34] F. Fischer, J. Fuchs, and F. Mansmann, "Clockmap: Enhancing circular treemaps with temporal glyphs for time-series data," in *Proceedings of Eurographics Conference on Visualization*, 2012.

[35] M. Monroe, "Interactive event sequence query and transformation," Ph.D. dissertation, University of Maryland, College Park, MD, USA, 2014.

[36] Z. Liu, B. Kerr, M. Dontcheva, J. Grover, M. Hoffman, and A. Wilson, "CoreFlow: Extracting and visualizing branching patterns from event sequences," *Computer Graphics Forum*, vol. 36, no. 3, pp. 527–538, 2017.

[37] K. Wongsuphasawat and D. Gotz, "Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2659–2668, 2012.

[38] D. Gotz and H. Stavropoulos, "DecisionFlow: Visual analytics for high-dimensional temporal event sequence data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1783–1792, 2014.

[39] Y. Wu, X. Xie, J. Wang, D. Deng, H. Liang, H. Zhang, S. Cheng, and W. Chen, "Forvizor: Visualizing spatio-temporal team formations in soccer," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 65–75, 2018.

[40] 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*, vol. 25, no. 1, pp. 417–426, 2019.

[41] 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*, vol. 24, no. 1, pp. 709–718, 2017.

[42] 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*, vol. 26, no. 1, pp. 407–417, 2019.

[43] "Vue.js," https://cn.vuejs.org/index.html.

[44] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 data-driven documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

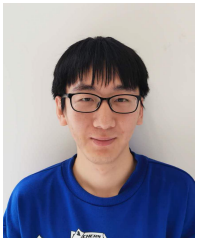
[45] Y. Chen, P. Xu, and L. Ren, "Sequence synopsis: Optimize visual summary of temporal event data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 45–55, 2018.

[46] Z. Deng, D. Weng, J. Chen, R. Liu, Z. Wang, J. Bao, Y. Zheng, and Y. Wu, "Airvis: Visual analytics of air pollution propagation," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 800–810, 2019.

[47] J. Wu, Z. Guo, Z. Wang, Q. Xu, and Y. Wu, "Visual analytics of multivariate event sequence data in racquet sports," in *2020 IEEE*

Conference on Visual Analytics Science and Technology (VAST). IEEE, 2020, pp. 36–47.
[48] P. Grünwald, *The minimum description length principle*, 2007.
[49] P. A. Sorokin, *Social mobility*. Taylor & Francis, 1998, vol. 3.
[50] A. Tziner, "Choice and commitment to a military career," *Social Behavior and Personality: An international journal*, 1983.
[51] H.-J. Voth and G. Xu, "Patronage for Productivity: Selection and Performance in the Age of Sail," *CEPR Discussion Paper No. DP13963*, 2019.
[52] Harvard University, Academia Sinica, and Peking University, "China biographical database (CBDB)," Apr. 2019. [Online]. Available: https://projects.iq.harvard.edu/cbdb

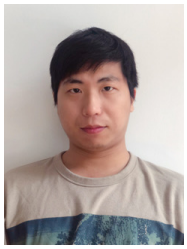
Yifang Wang is currently a Ph.D. candidate in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology (HKUST). She obtained her B.Eng. degree in Software Engineering from Zhejiang University, China in 2018. Her research interests include visual analytics in social science, visual anomaly detection, and immersive visualization. For more information, please visit http://wangyifang.top/about/.



Hongye Liang is currently a master student in the department of Computer Science and Technology at Zhejiang University. He received his B.E. degree in Software Engineering from Zhejiang University. His research interests mainly include sports visualization and information visualization.

Xinhuan Shu is currently a Ph.D. candidate in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology (HKUST). She received her B.E. degree in Computer Science and Technology form Zhejiang University, China in 2017. Her research interests include data-driven storytelling, animated visualization, and visual analytics.

Jiachen Wang received his BS in digital media technology from Zhejiang University, China in 2018. He is currently a Ph.D. student in the State Key Lab of CAD&CG, Zhejiang University in China. His research interests are in sports data visualization and visual analytics.



Ke Xu is currently a postdoctoral research associate in the Visualization and Data Analytics Research Center (VIDA) at New York University (NYU). He obtained his Ph.D. in the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology in 2019, and B.S. in Electronic Science and Engineering from Nanjing University, China in 2015. His research interests include data visualization, human-computer interaction, with focus on visual anomaly detection and explainable AI.

Zikun Deng received the B.S. degree in transportation engineering from Sun Yat-sen University in 2016. He is currently pursuing the Ph.D. degree with the State Key Lab of CAD&CG, Zhejiang University. His research interests mainly include spatiotemporal data mining, visualization, and urban visual analytics.

Cameron Campbell is Professor in the Division of Social Science at the Hong Kong University of Science and Technology. He obtained a BS in Engineering and Applied Science and in History at the California Institute of Technology and a PhD in Sociology and Demography at the University of Pennsylvania. He was named a Changjiang Scholar by the Ministry of Education of the People's Republic of China in 2017 and received a Guggenheim Fellowship in 2004. His research focuses on stratification and inequality. He also studies kinship and demographic behavior in China in comparative perspective using large multi-generational population databases.

Bijia Chen is currently a postdoctoral fellow in the Department of History at Renmin University of China. She obtained her Ph.D. in Social Science at the Hong Kong University of Science and Technology in 2019. She received B.A. degree in History from Lanzhou University in 2013 and M.Phil degree from the Division of Social Science at HKUST. Her research interests include historical demography, social mobility, and career mobility of officials during the Qing.

Yingcai Wu is a ZJU100 Young Professor at the State Key Lab of CAD&CG, Zhejiang University. His main research interests are in visual analytics and information visualization, with focuses on user behavior analysis, urban informatics, social media analysis, and sports analytics. He received his Ph.D. degree in Computer Science from the Hong Kong University of Science and Technology. Prior to his current position, Dr. Wu was a postdoctoral researcher in the University of California, Davis from 2010 to 2012, and a researcher in Microsoft Research Asia from 2012 to 2015. For more information, please visit http://www.ycwu.org.

Huamin Qu is a professor in the Department of Computer Science and Engineering (CSE) at the Hong Kong University of Science and Technology (HKUST) and also the director of the interdisciplinary program office (IPO) of HKUST. He obtained a BS in Mathematics from Xi'an Jiaotong University, China, an MS and a PhD in Computer Science from the Stony Brook University. His main research interests are in visualization and human-computer interaction, with focuses on urban informatics, social network analysis, E-learning, text visualization, and explainable artificial intelligence (XAI).